Análisis de regresión aplicado

Teoría y práctica

COLECCIÓN
HÉCTOR OCHOA BACELIS
Textos de enseñanza de ciencias básicas

José Manuel Piña Gutiérrez

Rector

Análisis de regresión aplicado

Teoría y práctica

José Isabel López Naranjo
Rodolfo Osorio Osorio

Universidad Juárez
Autónoma de Tabasco

Analisis de regresión aplicado: Teoria y práctica / José Isabel López Naranjo, Rodolfo Osorio Osorio. – Primera edición. – Villahermosa, Tabasco: Universidad Juárez Autónoma de Tabasco, 2015.

514 páginas: Ilustrado. – (Colección: Héctor Ochoa Bacelis. Textos de enseñanza de ciencias básicas).

Incluye referencias bibliográficas (p. 483-497)

ISBN: 978-607-606-218-0

1. Análisis de regresión. I. López Naranjo, Jose Isabel, Autor

L.C. QA278 P67 2015

Primera edición, 2015

ISBN: 978-607-606-218-0

Apoyo editorial:	Francisco Morales Hoil
Corrección de estilo:	Blanca Quiriarte
Diseño y formación:	Ricardo Cámara Córdova
Ilustración de protada:	Leidy Gabriela Moreno Olán

Hecho en Villahermosa, Tabasco, México

Índice

Agradecimientos

Para la elaboración de este documento fueron de gran importancia las contribuciones de muchos de mis compañeros que comparten esta noble labor de la enseñanza. La práctica docente por muchos años ha hecho posible la definición de este libro en el cual participaron con su apoyo y voluntad de manera decidida profesores y estudiantes de esta división académica, con el ánimo de contribuir al aprendizaje de esta área tan importante de la estadística aplicada. Por lo que expreso mi más sincero agradecimiento al Dios Todopoderoso que me permite la vida y la salud y la posibilidad de presentar ante la comunidad universitaria este material que seguramente será de gran utilidad en el aprendizaje y formación de los futuros profesionistas y para la generación e innovación del conocimiento.

Deseo también agradecer el apoyo irrestricto del rector de la Universidad Juárez Autónoma de Tabasco, el Dr. José Manuel Piña Gutiérrez, en el financiamiento para la edición de esta obra; manifestando así su compromiso con la educación y la formación de los jóvenes tabasqueños en aras de una mejor sociedad.

Asimismo agradezco a la directora de la División Académica de Ciencias Agropecuarias de la Universidad Juárez Autónoma de Tabasco, el M.A.A. Alma Catalina Berumen Alatorre, por su gran interés y apoyo para la publicación de este libro; así como su empeño en los trámites editoriales que

permitieron llevarlo a buen término en correlación con el Plan de Desarrollo de la División y el mejoramiento de la calidad de los programas educativos de licenciatura y maestría. De igual forma agradezco al Dr. Rodolfo Osorio Osorio, coordinador de Investigación y Posgrado de la División su gran entusiasmo y apoyo en la integración del comité editorial para la revisión y autorización del manuscrito.

Guardo también gratitud a todos los profesores investigadores que colaboraron en la revisión del documento, al Dr. Fidel Ulín Montejo, de la División Académica de Ciencias Básicas de la Universidad Juárez Autónoma de Tabasco, al Dr. Pedro García Alamilla, al Dr. Efraín de la Crúz Lázaro, de la División Académica de Ciencias Agropecuarias. A todos ellos les agradezco sus observaciones y aportaciones que sin duda enriquecieron la presentación del material aquí tratado, demostrando así su compromiso con la institución y la calidad profesional que los caracteriza.

Con gratitud al director de Difusión Cultural de la Universidad Juárez Autónoma de Tabasco, el Ing. Miguel Ángel Ruíz Magdonel, por su empeño en la gestión y supervisión de la edición de esta obra y sus valiosas sugerencias que seguramente mejoraron la presentación de este libro.

A mi esposa y mis tres hijos que de alguna manera estuvieron involucrados en el manuscrito de este libro con su apoyo en cuestiones computacionales, por su compañerismo, amistad y paciencia durante su elaboración, que Dios los bendiga de manera abundante.

Prólogo

Una de las metas en mi vida profesional es la elaboración de un libro de un tema apasionante como lo es el "análisis de regresión", con el cual se genera e innova el conocimiento actual de cualquier disciplina de las ciencias. Por ello, cuando ingrese a la Universidad Juárez Autónoma de Tabasco con adscripción a la División Académica de Ciencias Económico Administrativas impartí la asignatura de Econometría a estudiantes del Programa Educativo de Economía, inicié la búsqueda y recolección de información relacionada a esta temática y ahora que me encuentro en la División Académica de Ciencias Agropecuarias impartiendo Estadística Aplicada para Médicos Veterinarios y Estadística Inferencial para Ingenieros Agrónomos, he dedicado tiempo para seleccionar, analizar y estructurar dicho material en un manuscrito que facilite el aprendizaje de estas asignaturas que son muy importante para estudiantes e investigadores de las diferentes áreas de las ciencias.

Por la experiencia del autor y de otros profesores investigadores la impartición de esta asignatura se facilita actualmente con la ayuda de un equipo de cómputo, no obstante que no se cuenta con literatura en castellano suficiente para que estudiantes y profesores cuenten con un fácil acceso a esta temática. Por esto, de las funciones de docencia e investigación surge la necesidad de contar con un libro ágil y de fácil entendimiento que integre los principios teóricos y metodológicos del análisis de regresión aplicado en las

ciencias agropecuarias en un lenguaje comprensible y familiar que permita al estudiante y profesor investigador aplicarlos de manera eficiente.

El análisis de regresión y el ajuste de datos a modelos lineales es una herramienta fundamental en las actividades de investigación para la generación e innovación del conocimiento, enriqueciendo las ciencias agropecuarias con información útil y aplicable en la resolución de problemas y la generación de alternativas en los procesos de producción. Por esto, los investigadores deben de conocer las técnicas modernas del análisis de regresión ya que hoy contamos con equipo de cómputo de alta capacidad en el procesamiento de grandes volúmenes de datos, provenientes de experimentos. El análisis de regresión es una parte fundamental de la estadística inferencial, en la cual las conclusiones se basan en contribuciones que una o más variables independientes tienen sobre una variable respuesta.

Este libro lo pueden usar los estudiantes y profesores investigadores de las ciencias agropecuarias, los biólogos, los químicos, los sociólogos, las ciencias médicas y en el diseño y mejoramiento de productos de manufactura. Para el uso de estas técnicas del análisis de regresión el profesor investigador o el estudiante sólo deben poseer un conocimiento básico de estadística descriptiva, ya que los temas han sido tratados con pocos desarrollos matemáticos y estadísticos; lo que lo hace accesible para todo aquel que esté interesado en este tema.

Por todo lo anterior se considera importante ofrecer a la comunidad científica, académica y estudiantil un documento con la información aquí descrita que contribuya a mejorar el aprendizaje y aplicación de esta

herramienta, a enriquecer el acervo bibliográfico de la División Académica de Ciencias Agropecuarias y de otras instituciones de educación superior y centros de investigación con programas educativos y de investigación afines.

José Isabel López Naranjo

Presentación

La presente obra es el fruto de los esfuerzos personales de los autores, familiares y compañeros de que dedican su tiempo a la enseñanza y la investigación. Se puede mencionar que es una síntesis de la experiencia adquirida durante más de veinticinco años dedicados a la enseñanza y la investigación, participando en cursos especializados sobre estadística aplicada en instituciones de educación superior en nuestro país.

Representa una recopilación y selección estricta de la literatura existente sobre la temática, incluyendo información sobre investigaciones que se han realizado en instituciones de educación superior, centros de investigación sobre diversas disciplinas de las ciencias aplicadas, por lo que se considera una fuente de técnicas y principios actualizados y de utilidad en la solución de problemas reales.

El propósito de apoyar la publicación de este libro se orienta a la gran utilidad de esta herramienta de ajustar modelos de regresión a conjuntos de datos de experimentos reales y que contienen fines de predicción. Asimismo la de fortalecer la formación de los estudiantes de licenciatura para que cuenten con un material accesible y didáctico para su uso en sus tareas de investigación. De igual manera profesores investigadores de las diferentes áreas de las ciencias podrán utilizar esta obra como apoyo en sus investigaciones y construir modelos estadísticos que les permitan una interpretación comprensiva de los fenómenos estudiados. Todo profesional podrá utilizar esta obra ya que solo se requiere de un curso básico de estadística para comprender sus alcances y principios que se trataron de describir de una manera sencilla y familiar sin demostraciones matemáticas complicadas que muchas veces hacen que una obra no sea consultada. Los ejemplos de aplicación de esta herramienta se resolvieron con el software

estadístico SAS, muy conocido entre las instituciones que realizan investigación científica y que forman profesionales competitivos y de calidad.

El análisis de regresión es una herramienta de la inferencia estadística aplicada que permite ajustar conjunto de datos proveniente de trabajos de investigación con propósitos de predicción. En esta obra se trata una diversidad de técnicas útiles en esta tarea y se proporcionan las estrategias que permiten la selección de las variables, el tamaño de muestra y el mejor modelo; aquél que explique de mejor manera la relación entre las variables estudiadas. Esta técnica es aplicable en cualquier área de las ciencias: agropecuarias, biológica, económica, salud, genética, matemáticas, del comportamiento humano, etc., donde los objetivos incluyan predicción de resultados futuros de la variable respuesta (dependiente).

Por lo anterior se considera de gran valor ofrecer a la comunidad científica, académica y estudiantil un libro con la información aquí tratada que contribuye a enriquecer el acervo bibliográfico de nuestra universidad y en especial de la División Académica de Ciencias Agropecuarias y de igual forma a otras instituciones de educación superior y de investigación que ofertan programas educativos afines y/o desarrollen actividades de investigación y desarrollo.

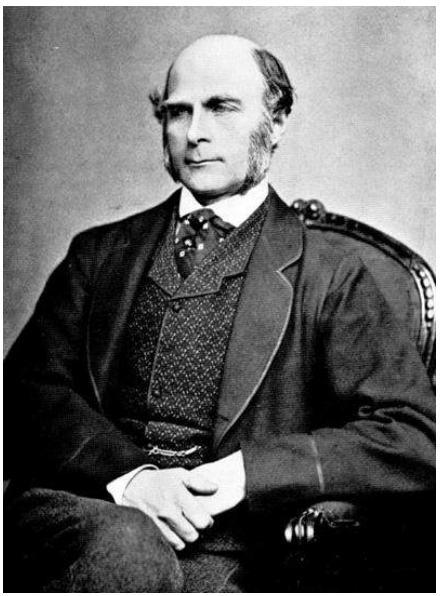
Dr. Roberto Flores Bello

Director de la División Académica de Ciencias Agropecuarias

Universidad Juárez Autónoma de Tabasco

"Estudio en la duda, acción en la fe"

Un visionario: Sir Francis Galton



Antropólogo, geógrafo, explorador, inventor, meteorólogo, estadístico, psicólogo y eugenista británico, Francis Galton (1822-1911) es mundialmente reconocido por su afición a medir y cuantificar prácticamente todo. Cursó estudios en el King's College de la Universidad de Londres y en el Trinity College de Cambridge, aunque es notable que no tuvo cátedras universitarias y realizó la mayoría de sus investigaciones por su cuenta. Sus múltiples contribuciones recibieron reconocimiento formal cuando, a la edad de 87 años, se le concedió el título de Sir o Caballero del Reino.

De intereses muy variados, Galton contribuyó a diferentes áreas de la ciencia. A menudo sus investigaciones fueron continuadas dando lugar a nuevas disciplinas.

Su obsesión por medir todo lo medible (y también lo no medible) le hizo no solamente tener un laboratorio de antropometría que le permitió recoger una inmensa cantidad de estadísticas sobre la estatura, dimensiones,

fuerza y otras características sobre las personas desarrollando nuevas técnicas para sus diversas mediciones.

En sus búsquedas, Galton explicó el fenómeno de la regresión a la media, usó por primera vez la distribución normal, describió las propiedades de la distribución normal bivariada y su relación con el análisis de regresión y también introdujo el concepto de correlación posteriormente desarrollado por Pearson y Spearman. Fue a partir de sus aportaciones que la estadística comenzó a considerarse una ciencia.

Hoy en día prácticamente todo trabajo científico que se precie está respaldado por un buen soporte estadístico siendo uno de los culpables, sin lugar a dudas, Francis Galton, considerado por muchos como uno de los padres de esta ciencia.

Capítulo 1
Análisis de regresión

1.1. Introducción

En 1885 Sir Francis Galton introdujo por primera vez la palabra “regresión” en un estudio; éste demostró que el tamaño entre descendientes (hijos) y padres no presenta una tendencia directa, pero se acerca bastante al promedio general. El método de mínimos cuadrados fue descubierto por Carl Friedrich Gauss, quien usó el procedimiento por primera vez a inicios del siglo XIX. Sin embargo, existe controversia sobre este hecho, ya que se menciona que Adrien Marie Legendre fue la primera en publicar un trabajo sobre su uso en 1805.

A finales de 1960, el análisis de regresión y el método de mínimos cuadrados fueron acoplados a la práctica y considerados elementos compatibles. Es conveniente mencionar que en muchas situaciones no ideales, los mínimos cuadrados ordinarios (MCO) no son lo más apropiados. Sin embargo, para el análisis de muchos datos se acepta el sesgo de la estimación para abatir la multicolinealidad y lograr una regresión robusta.

Actualmente esta técnica de análisis se aplica a datos diversos: la relación probabilística o funcional de procesos químicos o biológicos, la

economía nacional, un grupo de pacientes en un experimento clínico, un grupo de tipos de metal en un estudio de resistencia a la tensión, entre otros. De esta manera, en algunos casos las variables pertinentes son variables aleatorias y están relacionadas en un sentido probabilístico a través de una distribución conjunta. En otros casos, las variables son cantidades matemáticas y los supuestos indican que existe una relación funcional asociativa. En un conjunto de datos que involucra medidas de variables, el análisis de regresión se aplica para descifrar ciertos aspectos del mecanismo que las relaciona. El siguiente ejemplo ilustra qué clase de información se puede obtener con el análisis de regresión de datos.

Considere un estudio en el que se elige al azar un grupo de personas que recorren una distancia específica y en cada individuo se cuantifican las siguientes variables:

X_1: edad
X_2: peso
X_3: frecuencia cardiaca al inicio de la carrera
X_4: frecuencia cardiaca al final de la carrera
X_5: tiempo de recorrido
Y: consumo de oxígeno

Es razonable pensar que la variable Y representa la respuesta con la cual los individuos utilizan el oxígeno. Se puede pensar también que desempeña un papel algo diferente a X_1, X_2, X_3, X_4 o X_5. De hecho, nuestro interés es

visualizar a la variable *X* como cantidades que determinan a *Y*, o mejor aun, que predicen a *Y*. Entonces las *X* son llamadas variables independientes o variables predictoras y la *Y* es la variable dependiente o la variable respuesta. Dado los datos, el analista espera generar información, considerando el papel de cada variable X (edad, peso, frecuencia cardiaca y tiempo de recorrido) en términos de su influencia en la variable *Y* (consumo de oxígeno). Si alguna de estas variables *X* tiene algun efecto sobre el consumo de oxígeno, puede aportar información muy valiosa para describir la variable respuesta *Y*. Finalmente, se puede esperar que los datos permitan la estimación de la relación que existe entre estas variables.

1.2. Modelos de regresión

En las siguientes secciones se discutirá el uso y desarrollo de modelos estadísticos. Simplemente se supone que todos los procedimientos usados y las conclusiones obtenidas del análisis de regresión dependen al menos indirectamente de los supuestos de un modelo de regresión. Un modelo es lo que el analista percibe como el mecanismo que genera los datos sobre los cuales se aplica el análisis de regresión. Los modelos de regresión se encuentran por lo general en una forma algebraica. Por ejemplo en la ilustración del consumo de oxígeno, si el investigador está dispuesto a asumir que la relación es representada por una estructura lineal en las variables predictoras, entonces un modelo adecuado es dado por:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon \quad (1)$$

En donde β_0, β_1, β_2, ..., β_5, son constantes desconocidas llamadas coeficientes de regresión. Los procedimientos comprendidos en el análisis de regresión pueden extraer conclusiones acerca de esos coeficientes. El investigador puede estar involucrado en un intento laborioso para determinar si un incremento en la frecuencia cardiaca al final de la carrera (X_4) verdaderamente disminuye o incrementa la eficiencia del consumo de oxígeno (Y). Es un hecho que el signo y la magnitud de los coeficientes son de gran importancia. El término ε en el modelo es adicionado considerando que el modelo es probabilístico. Esencialmente describe los errores aleatorios del modelo. Cuando se aplica la ecuación (1) a un conjunto de datos se puede determinar el término ε como una ayuda para cuantificar cualquier variación individual, adicional a los parámetros proporcionados por el modelo.

El modelo de la ecuación (1) representa al model lineal (lineal en los parámetros β). Cualquier procedimiento de regresión involucra el ajuste del modelo a un conjunto de datos, que son definidos como lecturas de las variables en las diferentes unidades experimentales muestreadas. El término "ajuste a un conjunto de datos" involucra la estimación de los coeficientes de regresión y la correspondiente formulación de un modelo de regresión ajustado, una estrategia empírica que es la base de cualquier inferencia estadística.

El fundamento de un análisis estadístico de regresión se representa por la calidad del ajuste. Evidentemente, si el modelo de regresión propuesto no

describe los datos satisfactoriamente, la información estadística calculada no es representativas. Este texto no está orientado al tratamiento de modelos no lineales. Los modelos no lineales son comunes en muchas de las ciencias naturales o aplicaciones en ingeniería. Un bioquímico puede postular un modelo de crecimiento del tipo

$$y = \frac{\alpha}{1 + e^{\beta t}} + \varepsilon \tag{2}$$

para representar el crecimiento *y* de un organismo particular como una función del tiempo t. Aquí los parámetros estimados con los datos son α y β. Muchos de los mismos problemas que se intentan resolver por regresión lineal pueden ser manejados con regresión no lineal o viceversa. Sin embargo, los aspectos de cálculo para construir modelos no lineales son más complejos y requieren un tratamiento especial. El modelo de regresión ajustado es producto del modelo propuesto como un estimador de la relación funcional que describe los datos. El tipo de modelo postulado a menudo depende del rango de las variables regresoras definidas en los datos. Por ejemplo, un ingeniero químico puede tener conocimiento de que su sistema necesita el uso de términos que indiquen curvatura en el modelo. Suponer que x_1 y x_2 representan temperatura y concentración de reactivo, y la respuesta *y* es el rendimiento de una reacción química.

Para reflejar la curvatura en el modelo se construye una estructura del tipo:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon \quad (3)$$

Entonces los rangos de las variables regresoras pueden definir el tipo de modelo.

Con rangos estrechos en x_1 y x_2 el ingeniero puede usar el modelo que no incluye términos cuadráticos. Aún cuando (3) incluye potencias y productos de orden dos en las variables regresoras, se considera otro ejemplo de un modelo lineal; ya que los coeficientes son enteramente lineales. Anteriormente todas las aplicaciones de regresión que describen un conjunto de datos, la formulación del modelo es una simplificación de lo que ocurre en los datos de procesos observacionales. Áreas de las ciencias sociales y del comportamiento representan ejemplos donde los sistemas de los cuales se toman los datos son muy complicados para modelos con una estructura correcta. Los modelos lineales usados son aproximaciones que se espera funcionen bien en el rango de los datos usados en su construcción. Cuando lo sofisticado del campo sujeto a estudio no es suficiente para proporcionar un trabajo teórico, entonces una aproximación con un modelo lineal empírico y el sentido común puede ser muy informativa; particularmente cuando se usa en conjunción con un grupo de datos de buena calidad.

1.3 Usos formales del análisis de regresión

Este documento trata de cubrir varias categorías que representan clases específicas de inferencias y distinciones que son extraídas de ellas. Esto es crucial, puesto que todos los procedimientos de estimación, o sólo el modelo que es adoptado, pueden depender del objetivo del estudio. Esto a menudo parece contrario a la ideología de ciertos usuarios de la metodología. Para el analista inexperto puede parecer que el modelo que aparentemente describe mejor a los datos será adoptado para todo propósito. Sin embargo, un modelo que da una solución satisfactoria a un problema no necesariamente sirve para resolver otro. El uso del análisis de regresión está en tres o quizás cuatro categorías, aunque exista algún traslape. Estas categorías son:

1. Predicción
2. Variables de desperdicio
3. Especificación del modelo
4. Estimación de parámetros

Se advierte al analista que debe conocer y llevar en mente cuál es el objetivo principal en su esfuerzo, o esfuerzos, para hacer análisis de regresión. Inicialmente consideremos la tarea (1). Aquí los parámetros estimados no son buscados para su propio fin. No se investiga para la especificación del modelo verdadero, sino de cómo el tipo de modelo tiene influencia en la predicción. No es importante que se defina el papel de cada regresor en el modelo con

estricta precisión. Ciertamente, el ejemplo del proceso químico es un problema de predicción. El rendimiento de la reacción es importante y el ingeniero necesita predecirla de forma adecuada.

La tarea (2) es relevante en un gran número de aplicaciones de la vida real. La formulación del modelo es secundaria, ya que se usa simplemente como un instrumento para detectar el grado de importancia de cada variable en la explicación de la variación en la respuesta. Las variables que explican una cantidad razonable de la variación en la respuesta quizá se pueden utilizar para otros estudios; por otro lado, aquellas que sean regresoras parecen jugar un menor papel y son eliminadas. Esta práctica precede a menudo al estudio extensivo de procesos de construcción de modelos. La especificación del modelo se explica por sí misma. El analista debe tomar con mucho cuidado la postulación del modelo; cualquier analista sabe directa o indirectamente que hay gran cantidad de modelos candidatos en competencia, en diferentes formas funcionales. Cada forma funcional define un papel diferente de las regresoras. Cuando el modelo es lineal, este tipo de ejercicio puede ser limitado; a menos que la complejidad de cada variable regresora esté bien definida en los datos. La estimación de parámetros es por lo general el único propósito al aplicar un análisis de regresión en ciertos campos científicos. En un grupo de datos se muestra que una función de producción agrícola es ajustada a un grupo de variables regresoras que representan gastos. Seis tipos de regresoras de gasto y una variable de precipitación son usadas. La unidad de muestreo es un año y los datos se colectaron en el estado de Chiapas para un periodo de 25 años. Estos 25 datos son ajustados a un modelo lineal y tanto

la predicción como la variable de desperdicio, no son importantes. Sin embargo, los rangos específicos de los coeficientes de regresión justifican una teoría económica particular. Los signos y magnitudes de los coeficientes son cruciales.

Si los datos no reflejan una tendencia que involucre las variables, no hay éxito en el desarrollo del modelo o en la extracción de inferencias del sistema. Si algún tipo de asociación existe, no implica que los datos lo revelarán en una figura claramente detectable; los datos pueden provenir de un experimento, una investigación planificada, una colección y tabulación sobre el tiempo, una simulación, o de otras muchas fuentes. Es claro que el tamaño de muestra es importante; cuando éste es muy pequeño, el analista no puede calcular medidas adecuadas de error en los resultados de regresión y no tiene bases suficientes para verificar los supuestos del modelo. Sin embargo el tamaño del conjunto de datos es de gran consideración. Por ejemplo; no podemos desarrollar un modelo para encontrar una amplia relación si la muestra de unidades experimentales no representa a la población que se está atendiendo con el modelo. No se pueden realizar amplias generalizaciones en ninguna de las tareas del análisis de regresión si los datos son muy específicos. Para el ejemplo del consumo de oxígeno, se podría suponer que todos los individuos usados fueron atletas bien entrenados. Ambos, el rango de las variables y otras características de la inferencia, se aplican sólo a esta población.

En muchas situaciones las dificultades con el análisis de regresión son el resultado del fracaso en uno o más de los supuestos. En particular, el

modelo de regresión lineal múltiple de (1) se analiza bajo el supuesto de que las variables regresoras son medidas sin error. Si existe medida del error excesivo en la regresión, los estimadores de los coeficientes de regresión pueden ser afectados severamente; así como otras inferencias (tales como la predicción, las variables de desperdicios, etc.) pueden ser confundidas con incertidumbre. Quizás las limitaciones más serias en una regresión de un conjunto de datos es el fracaso para recolectar datos en todas las variables regresoras importantes. Esta insuficiencia puede ocurrir debido a que el analista no fue cuidadoso al seleccionar el total de las variables regresoras; sólo si todas o muchas de estas cantidades son identificadas, las limitaciones en el proceso de recolección de datos pueden prevenirse. Cuando esto ocurre el modelo puede ser deficientemente sobreespecificado, resultando en pobres estimadores de los coeficientes de regresión y, en consecuencia, una pobre predicción.

De manera básica, la regresión tiene dos significados: uno surge de la distribución conjunta de probabilidad de dos variables aleatorias, el otro es empírico y nace de la necesidad de ajustar alguna función a un conjunto de datos.

Capítulo 2
Clasificación de variables y selección del análisis

2.1. Introducción

Las variables pueden ser clasificadas en varias formas para determinar qué método de análisis de datos se puede usar. Aquí describiremos tres métodos de clasificación: a) espaciadas, b) orientación descriptiva y c) escala de medida. Asimismo se establecerán diferencias entre los modelos determinísticos y los modelos estadísticos, enfatizando su utilidad en el estudio de fenómenos aleatorios.

2.2 Espaciadas

En este esquema de clasificación llamaremos *variables espaciadas* a aquellas en las cuales se puede determinar si hay espacios entre valores sucesivos. Si entre un valor y otro hay espacios, la variable se denomina *discreta*; si no existen, la variable se denomina *continua*. Para ser más preciso, una variable es discreta si entre dos valores observables no existe otro valor posible; por otro lado, una variable es continua si entre dos valores observables, existe

otro valor potencialmente observable. Ejemplos de variables continuas son: edad, presión sanguínea, niveles de colesterol, altura y peso. Algunos ejemplos de variables discretas son: sexo, número de muertos, número de partos de una vaca, número de sementales en una explotación ganadera, la identificación de grupos y estado de enfermedad, etc.

Cuando se procede con datos, la frecuencia de distribución de muestreo para variables continuas es representada de forma diferente que para variables discretas. Los datos de una variable continua son agrupados por lo general en intervalos de clases y se determina una distribución de frecuencia contando las observaciones en cada intervalo. Tal distribución por lo general es representada por un histograma de frecuencias. A su vez, los datos de una variable discreta generalmente no son agrupados pero son representados por gráficos en segmentos de línea.

Es importante notar que las variables discretas algunas veces pueden ser tratadas como variables continuas. Esto es útil cuando los valores posibles de cada variable, aunque discreta, no cubren un amplio rango de valores. En tal caso, los valores posibles, aunque técnicamente espaciados, muestran pequeños espacios entre valores que una representación visual aproximada a un intervalo. Cuando los datos de estas variables son agrupados en clases (11-15, 16-20, etc.) el histograma de frecuencia resultante es una gráfica más clara de las características de la variable, en relación a una gráfica de líneas. Entonces en este caso, tratar las clases sociales como una variable continua es más útil que hacerlo como si fueran una variable discreta.

Por lo general, es útil tratar una variable discreta como continua. Esto es porque algunas variables que son continuas pueden ser agrupadas en categorías y tratadas como variables discretas en un análisis dado. Por ejemplo, la variable "edad" puede ser discreta agrupando sus valores en dos categorías, "jóvenes" y "viejos". De igual manera, "presión sanguínea" es una variable discreta si es categorizada como "baja", "media" y "alta", o en deciles.

2.3. Descriptivas (orientación descriptiva)

Un segundo esquema para clasificar variables está basado en la cualidad que puede o no tener una variable para describir o ser descrita por otras variables. Tal clasificación depende del objetivo de estudio más que de la estructura matemática inherente de la misma variable. Si la variable investigada es descrita en términos de otras variables, la llamaremos respuesta o variable dependiente. Si podemos usar la variable en conjunción con otras variables para describir una variable respuesta dada, la llamaremos predictor o variable independiente. Otras variables pueden afectar las relaciones pero no ser de interés en un estudio. Tales variables pueden ser referidas como variables de *control* o de *ruido*; o, en algún contexto, también pueden ser nombradas como *covariables o confundidas.* Usualmente la distinción entre variables dependientes e independientes es clara. No obstante, una variable considerada

como dependiente para evaluar cierto objetivo dentro de un estudio, puede ser considerada como independiente para evaluar un objetivo diferente.

2.4. Escala de medida

Un tercer esquema de clasificación se refiere a la precisión de medida de la variable. Se conocen tres: nominal, ordinal y de intervalos. El nivel más débil de medida es el *nominal*. En este nivel los valores asignados a la variable simplemente indican diferentes categorías. La variable sexo, por ejemplo, es nominal puesto que asignamos los números 1 y 0 para denotar macho y hembra, respectivamente; de manera que podemos distinguir las dos categorías de sexo. Una variable que describe grupos de tratamientos también es nominal, dado que el tratamiento involucrado no puede presentar rangos de acuerdo a algún criterio (nivel de dosis).

Un nivel un poco más alto de medida, que permite no solamente agrupar en categorías separadas sino también ordenarlas, se denomina *ordinal*. Los grupos de tratamientos se pueden considerar ordinales, si diferentes tratamientos difieren por la dosis. En este caso no sólo se dice qué individuo cae en el grupo de tratamiento, sino a quién se le administró una dosis alta del mismo. También la clase social es una variable ordinal, ya que se le da un orden entre las diferentes categorías. Por ejemplo; todos los miembros de la clase media alta se ubican, en algún sentido, más arriba en comparación con los de la clase media baja. Una limitación, quizás debatible,

en la precisión de una medida tal como clase social, es la cantidad de información proporcionada dada la magnitud de las diferencias entre categorías. Entonces, aunque la clase media alta es más alta, que la clase media baja, qué tan alta puede ser es debatible. Una variable que puede dar no solamente un orden sino también una medida significativa de la distancia entre categorías es llamada *variable intervalo*. Para ser intervalo una variable debe tener algún tipo de unidad de medida estándar o física aceptada. Altura, peso, presión sanguínea y número de muertes satisfacen este requisito; además medidas subjetivas tales como percepción de preñez, tipo de personalidad, prestigio o estrés social, no cumplen.

Una variable intervalo que tiene una escala con un cero verdadero ocasionalmente se designa como una *variable de razón o escala de razón*. Un ejemplo de una variable de escala de razón es la altura de una persona. La temperatura es comúnmente medida en grados Celsius, una escala de intervalo; por otro lado, las medidas de temperatura en grados Kelvin son referidas a un cero absoluto y representan una variable de razón. Un ejemplo común de variable de razón en estudios de salud es la concentración de colesterol en la sangre. Las variables de escala de razón a menudo involucran errores de medida que tienen una distribución no normal y son proporcionales al tamaño de las medidas. Tales errores proporcionales violan un importante supuesto de regresión lineal denominado igualdad de varianza del error para todas las observaciones.

De manera similar, para variables de otros esquemas de clasificación, la misma variable puede ser considerada al nivel de medida en un análisis y a

diferentes niveles en otros estudios. Entonces, edad puede ser considerada como de intervalo en un análisis de regresión o, si es agrupada en categorías, nominal en un análisis de varianza. Cabe puntualizar que los diferentes niveles de precisión matemática son acumulativos. Una escala ordinal posee todas las propiedades de una escala nominal más ordinalidad; a su vez, una escala de intervalo es también nominal y ordinal. Lo acumulativo de estos niveles permite al investigador caer en uno o más niveles de medida en el análisis de datos. Así, una variable de intervalo puede ser tratada como nominal u ordinal en un análisis particular; y una variable ordinal puede ser analizada como nominal.

2.5. Selección del análisis

Cualquier investigador se enfrenta con la necesidad de analizar datos de acuerdo a la selección de un método específico de análisis. En dicha selección se involucran varias consideraciones, mismas que incluyen: (1) el propósito de la investigación, (2) las características matemáticas de las variables involucradas, (3) los supuestos estadísticos hechos sobre estas variables y (4) cómo se recolectarán los datos (el procedimiento de muestreo). Las primeras dos consideraciones son generalmente suficientes para determinar un análisis apropiado. Sin embargo, el investigador debe considerar los últimos dos puntos antes de finalizar las recomendaciones iniciales.

2.6. Asociación versus causalidad

Es importante ser cuidadoso de los resultados obtenidos en un análisis de regresión o, en términos generales, de cualquier forma de análisis aplicado para cuantificar la asociación entre dos o más variables. Aunque los cálculos estadísticos usados para generar una medida estimada de la asociación sean correctos, la misma estimación puede ser sesgada. Tal sesgo puede ser el resultado del método usado para seleccionar sujetos del estudio, los errores de la información usada en el análisis estadístico o como consecuencia de la inclusión de otras variables en la asociación observada y que no han sido medidas o apropiadamente consideradas en el análisis.

Por ejemplo, si la presión diastólica sanguínea y el nivel de actividad física fueren medidos sobre una muestra de individuos en cualquier tiempo, entonces un análisis de regresión puede sugerir, que en promedio, la presión sanguínea disminuye con el incremento de la actividad física. Así, tal análisis puede dar evidencias que esta asociación es de consistencia moderada y es estadísticamente significativa. Sin embargo, si el estudio involucró solamente adultos saludables y si el nivel de actividad física fue medido inapropiadamente, o si otros factores tales como; edad, raza y sexo no fueron correctamente tomados en cuenta, entonces las conclusiones anteriores pueden ser consideradas inválidas o al menos cuestionables.

Continuando con el ejemplo anterior, si los investigadores están satisfechos de que los hallazgos son completamente válidos; entonces ¿concluirán que el nivel bajo de actividad física es la causa de la alta presión

sanguínea? La respuesta es un inequívoco no. Los hallazgos de una asociación "estadísticamente significativa" en un estudio particular no establece una relación causal. Para evaluar la presencia de causalidad, el investigador debe considerar criterios que son externos a las características específicas y resultados de cualquier estudio.

Muchas de las definiciones estrictas de causalidad, requieren que un cambio en una de las variables (X) produzca siempre un cambio en otra variable (Y). Esto sugiere que para demostrar una relación causa-efecto entre X y Y, la demostración experimental requiere que un cambio en Y se traduzca en un cambio en X. Aunque es necesaria tal evidencia experimental a menudo es impráctico y no ético obtenerla, especialmente cuando se consideran factores de riesgo que son potencialmente nocivos para humanos. Consecuentemente, criterios alternativos basados en la información que no involucran directamente evidencia experimental son típicamente empleados cuando se intenta hacer una inferencia causal; considerándolos como variables relacionadas con poblaciones humanas. Una escuela de pensamiento considera que la inferencia causal ha producido una colección de procedimientos denominados "análisis de senderos" (Blalock, 1971). Datos para tales procedimientos se han aplicado primeramente a estudios sociológicos y ciencias sociales en general. Esencialmente dichos métodos intentan calcular causalidad indirectamente por eliminación, realizando una explicación causal vía el análisis de datos y finalmente llegando a un modelo causal aceptable que obviamente no es contradictorio con los datos. Entonces,

estos métodos más que intentar establecer una teoría causal directa, llega a un modelo final a través de un proceso de eliminación.

Una aproximación más ampliamente usada para hacer conjeturas causales, particularmente en ciencias de la salud y sociales, emplea un juicio de evaluación de los resultados combinados de varios estudios que usan un grupo de criterios operacionales generalmente necesarios para justificar una teoría causal. Esfuerzos para definir un grupo de criterios fueron realizados en 1950 y al inicio de 1960 por investigadores que revisaban estudios aleatorios sobre la salud de fumadores.

Una lista de criterios generales para evaluar la gran cantidad de evidencias disponibles, reportó una relación causal que fue formalizada por Hill (1971) y esta lista fue adoptada por muchos investigadores epidemiólogos. La lista contiene siete criterios, que son:

1. La fuerza de la asociación. La fortaleza de una asociación observada se manifiesta en una serie de estudios diferentes, lo menos probable es que esta asociación es espuria debido al sesgo.
2. Efecto dosis-respuesta. El valor de la variable dependiente cambia en un patrón significativo (incrementa) con la dosis (o nivel) del agente causal sospechado bajo estudio.
3. Falta de ambigüedad temporal. La causa hipotética precede la ocurrencia del efecto. Note que la habilidad para establecer este patrón de tiempo dependerá del diseño usado en el estudio.

4. Consistencia de los hallazgos. Muchos o todos, los estudios relacionados con una hipótesis causal generan resultados similares. De esta forma, estudios que proceden con una cuestión dada, pueden tener serios problemas de sesgo; mismos que pueden disminuir la importancia de la asociación observada.
5. Plausibilidad biológica y teórica de las hipótesis. La relación causal hipotética es consistente con el conocimiento biológico y teórico. Note sin embargo que el estado actual del conocimiento puede ser insuficiente para explicar ciertos hallazgos.
6. Coherencia de la evidencia. Los hallazgos no generan conflictos con los hechos aceptados acerca de los éxitos de las variables estudiadas (conocimiento acerca de la historia natural de alguna enfermedad).
7. Especificidad de la asociación. El factor de estudio (la causa sospechada) es asociado con un efecto (una enfermedad específica). Note, sin embargo, que muchos factores de estudio tienen múltiples efectos y que muchas enfermedades tienen causas múltiples.

Claramente, aplicando los criterios anteriores para una hipótesis causal dada, es un objetivo directo. Sólo si estos criterios son satisfechos, una relación causal no se puede considerar con certidumbre completa. No obstante sin evidencia experimental sólida, el uso de tales criterios puede ser una manera

lógica y práctica para establecer el tipo de causalidad, especialmente cuando se atienden estudios sobre poblaciones humanas.

2.7. Modelos estadísticos versus determinísticos

Aunque la causalidad no se establece por el análisis estadístico, las asociaciones entre variables pueden ser cuantificadas en un sentido estadístico. Con un diseño estadístico propio y el análisis pertinente, un investigador puede explicar, mediante un modelo, los cambios en las variables independientes que están relacionados con cambios en la variable dependiente. Sin embargo, los modelos estadísticos desarrollados usando regresión u otros métodos multivariables necesitan ser distinguidos de los modelos determinísticos. La ley de la caída de los cuerpos, en física, por ejemplo, es un modelo determinístico que asume una condición ideal. La variable dependiente varía en una manera prescrita completamente de acuerdo a la función matemática perfecta de la variable independiente.

Los modelos estadísticos, por el otro lado, permiten la posibilidad de error en la descripción de una relación. Por ejemplo, relacionando la presión sanguínea con la edad, personas de la misma edad no tienen la misma presión sanguínea. No obstante, con métodos estadísticos propios somos capaces de concluir que sobre el promedio, la presión sanguínea se incrementa con la edad. Asimismo, la modelación estadística apropiada puede permitirnos predecir la presión sanguínea esperada para una edad dada y asociar una

medida de variabilidad con la predicción. A través del uso de la teoría de probabilidad y estadística, tales declaraciones toman en cuenta la incertidumbre del mundo real por medio de las medidas del error y la variabilidad individual. De esta forma, aunque tales declaraciones son necesariamente no determinísticas, requieren una interpretación cuidadosa. Desafortunadamente, tales interpretaciones son muy difíciles de hacer.

Capítulo 3
El modelo de regresión simple

3.1. Introducción

La forma más simple (pero no por eso trivial) del problema general de regresión inicia con una variable dependiente Y y una variable independiente X. Este es el caso especial cuando k=1 para establecer una notación simple y comprensible. Para esclarecer los conceptos básicos y los supuestos del análisis de regresión, es útil empezar con una sola variable independiente. En este capítulo pretendemos plantear una discusión detallada sobre una clase especial de modelos lineales, en la cual la expresión para los valores esperados de la respuesta depende solamente de una variable medida en una escala continua. Los conceptos básicos para el ajuste de modelos lineales son introducidos e ilustrados con ejemplos numéricos. Se mencionan muchas ideas aquí que serán útiles en los siguientes capítulos.

3.2. Descripción del modelo

El modelo aplicado en la estructura más simple de regresión es el modelo de regresión lineal simple. Aquí el término *simple* implica una sola variable regresora, X, y el término *lineal* implica lineal en X. Empezamos con la descripción:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (1)$$

donde *y* es la medida de la variable respuesta, β_0 y β_1 son el intercepto y la pendiente, respectivamente y ε es el error del modelo. Podemos ajustar (1) a un conjunto de datos, donde los pares de observaciones (x_1, y_1), (x_2, y_2), (x_3, y_3) ... (x_n, y_n) son tomados de unidades experimentales y se obtienen los estimadores de β_0 y β_1. Entonces el modelo se escribe:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, 2, 3, ..., n \qquad (2)$$

3.3. Supuestos e interpretación de los parámetros del modelo

El propósito de la formulación de un modelo en el análisis de varianza es permitir al investigador conceptualizar cómo se generan las observaciones. Esta formulación de teoría estadística permite estudiar las propiedades de los estimadores de los parámetros. Los supuestos fundamentales del

procedimiento de mínimos cuadrados son importantes y serán detectados por el lector.

Primero, se asume que las X_i no son aleatorias, observadas con un error despreciable, mientras que las ε_i son variables aleatorias con media cero y varianza constante σ^2 (supuesto de homogeneidad de varianzas). En el futuro, se usará el operador esperanza (E) para denotar una *media poblacional*. Entonces se puede establecer que la $E(\varepsilon_i) = 0$ y $E(\varepsilon_i^2) = \sigma^2$. Además, se asume que los ε_i no están correlacionados de observación a observación; de esta manera, en muchas situaciones prácticas las X_i presentan alguna variación aleatoria. Aquí se asume que cualquier variación aleatoria en *X* es despreciable comparada con el rango en el cual es medida. Además cualquier error en las medidas de las X_i se considera muy pequeño comparado con el rango. El caso donde *X* y *Y* son variables aleatorias será tratado más adelante. El modelo de regresión lineal simple de la ecuación (2) describe una situación en la cual a un valor especifico de *X*, por decir X_i, la media de la distribución de las Y_i, está dado por:

$$E(y_i) = \beta_0 + \beta_1 x_i \qquad (3)$$

La varianza de la distribución σ^2 es independiente del nivel de *X*. Entonces, existe una relación lineal fundamental implicada por (2) que relaciona la respuesta media a *X*. Por esta razón podemos ocasionalmente usar la notación:

$$E(y/x) = \beta_0 + \beta_1 x \qquad (4)$$

Con E (*Y/X*) que se refiere a la media de *Y* condicional a un valor específico de *X*.

Una tarea importante del analista es estimar el parámetro β_0, el *intercepto* en el modelo de regresión, y β_1, la *pendiente* de la línea de regresión. Estos parámetros son llamados *coeficientes de regresión* y sus estimadores juegan un papel importante en el procedimiento total de análisis. El parámetro β_1 es el cambio en E (*Y*) por unidad de incremento en *X*. Los valores de β_0 y β_1 dependen de las unidades usadas para *Y* y para *X*. Entonces, el analista debe tomar en cuenta esto en cualquier interpretación dada a los estimadores.

A menudo el investigador se puede confundir acerca de la interpretación de β_0 o su estimador de datos experimentales. Por lo general, las características del modelo que se calcula con los datos son independientes del rango de *X* en que los datos fueron tomados. En otras palabras, existe una presunción de que el modelo dado por la ecuación (2) está limitado a una región confinada de *X*. Si esta región cubre a $X = 0$, entonces el estimador de β_0 puede ser interpretado como la media de *Y* cuando $X = 0$. Pero si la cobertura de los datos está lejos del origen, entonces β_0 es simplemente un término de regresión que significa poco en la interpretación. Por lo general, el estimador de β_0 es considerado como un valor sin una interpretación lógica o no razonable, e imposible en el contexto del problema. El analista debe tener en mente que la interpretación de β_0 es equivalente a extrapolar el modelo fuera del rango en que fue usado. Dos condiciones para una interpretación lógica de β_0 son las siguientes:

- Debe de ser posible que la variable independiente *X* tome el valor de cero en el conjunto de datos de interés.
- Debe de haber suficientes valores de la variable *X* alrededor de cero.

Considere la figura 3.1 y suponga que el conjunto de datos representa medidas del rendimiento de una reacción química (*Y*) a diferentes niveles de temperatura (*X*). Se tomaron 15 puntos mostrados en la gráfica; obviamente en el rango de temperatura cubierto por los datos, un modelo lineal es completamente apropiado para suponer que la curva punteada representa el *verdadero valor esperado* del rendimiento que tiene una relación curvilínea. Una extensión de la porción de la línea recta (línea remarcada) para un valor de $X = 0$ de temperatura produce una *y* cerca del 17% de rendimiento. Esto sería un valor razonable de un estimador de β_0 para estos datos. Sin embargo, su interpretación no se hace en sentido exacto, en este caso, sólo si no se está enterado de la verdadera relación curvilínea involucrada, él o ella puede saber que debajo de los 20°C no hay reacción. Entonces una proyección de rendimiento a 0°C no es necesario considerarlo. El papel del intercepto, será simplemente parte del mecanismo que describe que sólo sucede entre los 80° y los 220°C.

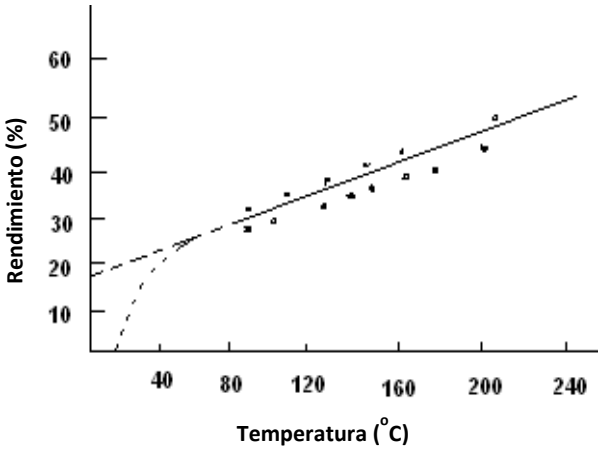


Figura 3.1. Rendimiento de una reacción química en función de la temperatura.

3.4. Supuesto de teoría normal

Claramente nos enfocamos en el supuesto fundamental que se aplica en cualquier procedimiento de análisis estadístico. Para entender totalmente y apreciar los supuestos de regresión, se debe reflexionar sobre qué propiedad o propiedades de los estimadores dependen de cuáles supuestos. Por ejemplo, el sesgo, la varianza y covarianza de los estimadores serán discutidos más adelante y cada una de esas propiedades depende de uno o más supuestos mencionados aquí. En una siguiente sección, se discuten las pruebas de hipótesis y estos procedimientos inferenciales dependen de un supuesto adicional denominado en el modelo los $\varepsilon_{i,}$ que se asume tienen una distribución Gaussiana. Este supuesto se denomina como *supuesto de teoría normal*.

3.5. Formulación de los mínimos cuadrados

El método de mínimos cuadrados se usa de manera extensiva en otros procedimientos de estimación para construir modelos de regresión; antes de los años 70 fue usado casi exclusivamente. Este método está diseñado para generar los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de β_0 y β_1, respectivamente, y el valor ajustado.

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i \qquad (5)$$

de la respuesta; de manera tal que la suma de cuadrados de residuales se minimiza. Esto es:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (6)$$

En la ecuación anterior se sustituye el valor de $\hat{y}_i$ y se obtienen las derivadas parciales; y como un resultado de β_0 y β_1 debe satisfacer

$$\frac{\partial y}{\partial \beta_0}\left[\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right] = 0$$

y

$$\frac{\partial y}{\partial \beta_1}\left[\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right] = 0$$

La minimización de $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ tiene buena exposición y en muchos ejemplos los valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ poseen buenas propiedades. La motivación es que los datos están unidos en la línea de regresión ajustada

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad (7)$$

si los errores en el ajuste, esto es los residuales, están poco separados. La figura 3.2. ilustra el procedimiento con las desviaciones verticales que representan a los residuales.

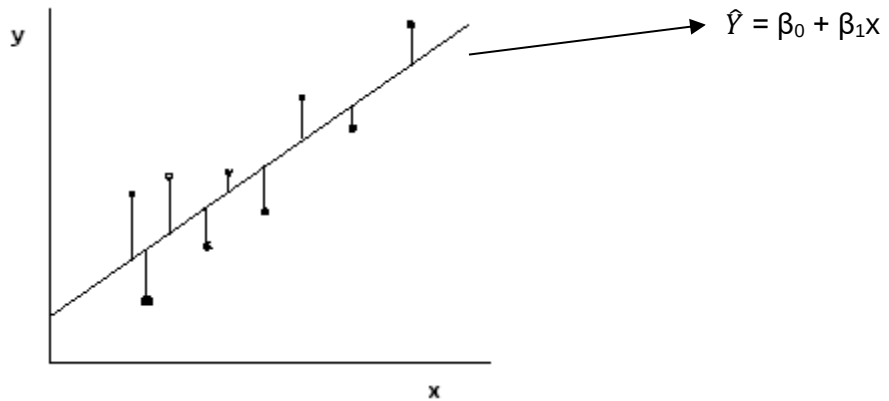


Figura 3.2. Ilustración de los residuales mínimos cuadrados

El desarrollo de los estimadores de mínimos cuadrados se obtiene de las derivadas parciales con respecto a los dos parámetros, esto es

$$2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \qquad (8)$$

$$-2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i) = 0 \quad (9)$$

Simplificando la ecuación (8) tenemos

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0 \quad (10)$$

Simplificando la ecuación (9) tenemos

$$(-)[\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 \sum_{i=1}^{n} x_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2] = 0 \quad (11)$$

Estos dos resultados (ecuaciones 10 y 11) reciben el nombre de ecuaciones normales; de las cuales se obtienen los estimadores de los parámetros, esto es, las sumas de cuadrados (S_{xx}) de las X y la suma de cuadrados de los productos cruzados (S_{xy}).

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n \quad (12)$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n \quad (13)$$

por lo que

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \quad (14)$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{15}$$

Y tenemos los dos estimadores de los parámetros del modelo. Entonces la línea ajustada por mínimos cuadrados es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X \tag{16}$$

En muchas aplicaciones las conclusiones son extraídas directamente del intercepto y la pendiente.

3.5.1. Propiedades de los estimadores

No es difícil mostrar que bajo el supuesto de que las X_i no son aleatorias y la $E(\varepsilon_i) = 0$, los estimadores son insesgados. La esperanza de $\hat{\beta}_1$ está dada por

$$E(\hat{\beta}_1) = E\left(\frac{S_{xy}}{S_{xx}}\right) = \sum_{i=1}^{n}(x_i - \bar{x})\, E(y_i)/ S_{xx}$$

$$= \frac{1}{S_{xx}} \sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i)) = \hat{\beta}_1 \frac{S_{xy}}{S_{xx}} = \beta_1 \tag{17}$$

Para el intercepto, tenemos

$$E(\hat{\beta}_0) = E(\bar{y} - \beta_1 \bar{x}) = \frac{1}{n} E\left(\sum_{i=1}^{n} y_i\right) - \beta_1 \bar{x} = \frac{1}{n}(n\, \hat{\beta}_0) = \beta_0 \tag{18}$$

Las propiedades de varianza en los estimadores de mínimos cuadrados denotan que Var (y_i) = Var (ε_i) = σ^2, enfatizando que la varianza del error del modelo es constante para los valores fijos de la variable regresora. Entonces, podemos tomar la varianza de β_1 y se obtiene

$$\text{Var}(\hat{\beta}_1) = \frac{1}{S_{xx}^2} \sum_{i=1}^{n} \sigma^2 (x_i - \bar{x})^2 = \sigma^2/S_{xx} \quad (19)$$

$$\text{Var}(\hat{\beta}_0) = \text{Var}\left(\sum_{i=1}^{n} y_i/n\right) + \bar{x}^2 \text{ Var}(\hat{\beta}_1) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \quad (20)$$

Por lo anterior es claro que en el desarrollo de la varianza de β_0 y β_1 el supuesto de varianza homogénea fue usado así como el supuesto de que las ε_i no están correlacionadas. Además, los resultados dependen de la condición de que las X_i no son aleatorias. La insesgabilidad y las propiedades de varianza de los estimadores de mínimos cuadrados pueden ser desarrollados usando una aproximación diferente.

3.5.2. Estimación de la varianza del error (grados de libertad de residuales)

En situaciones prácticas es necesario obtener un estimador de la varianza del error σ^2. El estimador se usa en el cálculo de los errores estándar estimados de los coeficientes para la prueba de hipótesis que, en muchas situaciones,

juega un papel importante para asignar calidad al ajuste y la capacidad de predicción del modelo de regresión $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. No es difícil visualizar intuitivamente que el estimador de los residuales es $y_i - \hat{y}_i$, que son los errores de ajuste observados y claramente su contraparte empírica son los ε_i, errores del modelo que no son observados. Entonces, es razonable que la varianza muestral de los residuales proveerá un estimador de σ^2. Si dividimos la suma de cuadrados de los residuales $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ por un denominador apropiado, n – 2 entonces se produce un estimador insesgado. Esto es, definimos

$$s^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / n - 2 \quad \text{como el estimador de } \sigma^2 \qquad (21)$$

La cantidad s^2 por lo general se denomina *cuadrado medio del error.* Se enfatiza que s^2 es insesgado bajo el supuesto importante de que el modelo es correcto. El denominador por lo general se denomina como los *grados de libertad del error o residual*. Se puede visualizar a los grados de libertad del residual en un contexto de regresión como el número de puntos (datos), n, o pedazos de información reducidos por el número de parámetros estimados (dos parámetros, intercepto y pendiente). Hemos usado aquí la explicación de que los grados de libertad de residuales son n - 2, puesto que se pierden dos grados de libertad debido a los requerimientos de estimación de dos parámetros. Una explicación similar puntualiza a s^2 como una varianza muestral de los residuales. Los grados de libertad de los residuales pueden ser vistos como el tamaño de muestra reducido por el número de restricciones de los residuales. Estas restricciones son inducidas por los requerimientos de la

estimación de los parámetros. Sabemos que cuando estamos muestreando variables aleatorias denotadas por y_1, y_2, y_3 ... y_n, los grados de libertad para la estimación de σ^2 son n – 1. La pérdida de un grado de libertad se debe a la restricción$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$ de los residuales que en este caso son y_i - $\hat{y}_i$. Estas restricciones son dadas por:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \qquad (22)$$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = 0 \qquad (23)$$

Las cuales son fáciles de comprobar al estimar los dos parámetros del modelo.

Ejemplo 1. En un estudio clínico se desea definir la relación entre la presión sistólica (y) y la edad de 30 pacientes en un hospital de Villahermosa, Tabasco. Los datos son los siguientes:

Paciente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
X	39	47	45	47	65	46	47	42	67	56	64	56	59	34	42
Y	144	220	138	145	162	142	170	124	158	154	162	150	140	110	128
Paciente	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
X	48	45	17	20	19	36	50	39	21	44	53	63	29	25	69
Y	130	135	114	116	124	136	142	120	120	160	158	144	130	125	175

Ajustamos un modelo de regresión lineal simple, n = 30.

Cálculos:

$\bar{y}$ = 142.53 $\bar{x}$ = 45.13

$\sum_{i=1}^{30} x_i$ = 1354

$\sum_{i=1}^{30} y_i$ = 4276

$\sum_{i=1}^{30} x_i\, y_i$ = 199576

$\sum_{l=1}^{30} x_i^2$ = 67894

Empleando estos resultados tenemos que:

$\hat{\beta}_0$ = 142.53 – (0.97) (45.13) = 98.71

$\hat{\beta}_1$= 6585.87/6783.47 = 0.97

Por lo que la ecuación de la línea recta ajustada es

$$\hat{y} = 98.71 + 0.97x \qquad (24)$$

Esta línea quiere decir que la presión sistólica se incrementa en la medida que se incrementa la edad de las personas. Note que el punto (47, 220) se observa completamente fuera de lugar de los otros datos y tal observación por lo general se denomina *outlier (punto lejano).* Este tipo de datos puede afectar la estimación de mínimos cuadrados y es importante decidir si se eliminan y quedan fuera del conjunto de datos. Definitivamente esta decisión puede ser

hecha después de una evaluación de las condiciones experimentales, el proceso de recolección de datos y los datos mismos. Si la decisión es difícil, se puede determinar el efecto que causa en el ajuste al remover los datos. Entonces se realiza el ajuste de los datos a un nuevo modelo de regresión lineal y el resultado es la línea dada por

$$\hat{y} = 97.08 + 0.95x \tag{25}$$

La cual, si observamos la gráfica, parece que ajusta de mejor forma al conjunto de datos en la figura 3.3.

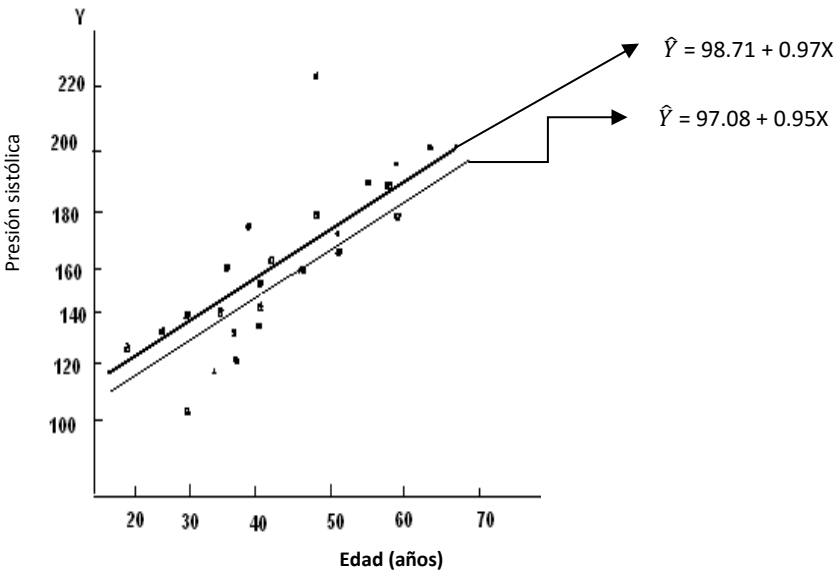


Figura 3.3. Las mejores líneas ajustadas a los datos de edad versus presión sistólica.

3.5.3. Medida de la calidad de la línea ajustada y el estimador de σ^2

Una vez que la línea de mínimos cuadrados ha sido determinada, nos gustaría evaluar si la línea ajustada puede ser utilizada para predecir valores de Y y su extensión. Una medida que ayuda para responder esta cuestión es dado por

$$SC_{error} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (26)$$

Donde $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Claramente si $SC_{error} = 0$, la línea recta ajusta perfectamente los datos; esto es, $Y_i = \hat{Y}_i$ para cada i y todos los puntos observados caen sobre la línea ajustada. Además cuando el ajuste es bajo, la SC_{error} es grande, ya que las desviaciones de los puntos con la línea de regresión son grandes.

Dos posibles factores contribuyen a la inflación de la SC_{error}. El primero es que existe una pérdida de variación en los datos; esto es σ^2 puede ser grande. El segundo se refiere a que el supuesto de un modelo de línea recta puede ser no apropiado. Es importante por lo tanto determinar los efectos separados de cada uno de esos componentes, puesto que ellos dirigen decididamente diferentes resultados con respecto al ajuste del modelo. En este caso asumiremos que el segundo factor no es un resultado. Entonces, asumiendo que el modelo de línea recta es apropiado, podemos obtener un estimador de σ^2 usando SC_{error}. Tal estimador es necesario para hacer inferencias estadísticas de la relación lineal verdadera entre *X* y *Y*. El estimador de σ^2 está dado por

$$S^2_{Y/X} = \frac{1}{n-2} \text{SC}_{\text{error}}$$

Para facilitar su cálculo se puede usar la siguiente ecuación:

$$S^2_{Y/X} = \frac{n-1}{n-2} (S^2_Y - \hat{\beta}_1{}^2 S^2_x) \quad (27)$$

donde

$$S^2_Y = \frac{1}{n-1}[\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/ \text{n}] \quad \text{varianza muestral de } y$$

$$S^2_X = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/\text{n}] \quad \text{varianza muestral de } x$$

En nuestro ejemplo tenemos los valores

$$S^2_Y = 509.91$$

$$S^2_X = 233.91$$

$$S^2_{Y/X} = \frac{29}{28}[509.91 - (0.97)^2(233.91)] = 299.77$$

Si el modelo de la línea recta es apropiado, la respuesta media poblacional $\mu_{y/x}$ cambia con *x*.

Por ejemplo, usando la línea ajustada a los datos como una aproximación a la línea poblacional para los datos de presión sistólica versus

edad, la media estimada de *y* cuando *x* = 40 es aproximadamente 138; mientras que la *y* estimada cuando *x* = 70 es 167.

3.5.4. Inferencia respecto a la pendiente y el intercepto

Para evaluar si la línea ajustada contribuye para predecir *y* de *x* y tomando en cuenta la incertidumbre de usar una muestra, es una práctica estándar calcular intervalos de confianza y/o prueba de hipótesis estadísticas acerca del parámetro desconocido en el modelo de línea recta aceptado. Tales intervalos de confianza y pruebas de hipótesis requieren el supuesto de que la variable aleatoria *y* tiene una distribución normal en cada valor fijo de x. Bajo este supuesto se deduce que los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ son distribuidos normalmente, con medias respectivas β_0 y β_1 cuando se cumple el supuesto de linealidad y con varianza fácilmente derivable. Estos estimadores, junto con los estimadores de sus varianzas, pueden ser usados para formar intervalos de confianza y pruebas estadísticas basadas en la distribución de t de student.

En este punto se advierte al lector que haga una reflexión formal para que la información en su campo de trabajo se extraiga de la línea de regresión ajustada. Las posibilidades pueden ser enumeradas como:

1. ¿Tiene *x*, la variable de regresión, una verdadera influencia sobre *Y* la respuesta?
2. ¿Existe un ajuste adecuado de los datos por el modelo?

3. ¿Predice adecuadamente la respuesta el modelo (a través de interpolación o extrapolación)?

En el primer caso, por lo general se puede responder la pregunta a través de una prueba de hipótesis sobre la pendiente β_1. Como hemos indicado, una hipótesis que siempre se plantea es

$$H_0: \beta_1 = 0 \quad \textit{versus} \quad H_1: \beta_1 \neq 0 \tag{28}$$

Si H_0 es verdadera, la implicación es que el modelo se reduce a $E(y) = \beta_0$, sugiriendo que verdaderamente la variable regresora no tiene influencia sobre la respuesta (al menos no a través de una relación lineal). El rechazo de H_0 a favor de H_1 nos permite inferir que *x* tiene influencia significativamente sobre la respuesta en un contexto lineal.

Existe un peligro de lectura presente demasiado grande para el rechazo de H_0 en las hipótesis anteriores. Por esta razón los resultados de la prueba no poseen la importancia atribuida a ellos. Es completamente posible que se encuentre que *x* afecta significativamente a la respuesta, pero el modelo no provee una solución exitosa al problema que motivó el análisis de regresión. El rechazo de H_0 simplemente significa que se detectó una tendencia; nada está relacionado con la calidad de ajuste de la línea de regresión, con respecto a un estándar preconcebido. Además, y quizás de mayor importancia, nada está implicado con respecto a la capacidad del modelo para predecir de acuerdo a algún grupo estándar. Como se puede esperar, la prueba de

hipótesis hace uso de las propiedades distribucionales de las Sumas de Cuadrados de regresión (SC_{regres}) y las Sumas de Cuadrados de residuales ($SC_{residual}$); y, como se puede esperar, las conclusiones dependen de las magnitudes relativas. Las pruebas de hipótesis más importantes proceden del parámetro del modelo de línea recta; de acuerdo con disyuntiva que nos indicaría si la pendiente de la línea de regresión es significativamente diferente a cero, o equivalentemente, si *x* ayuda a predecir *Y* usando un modelo de línea recta. La hipótesis nula apropiada para este caso es H_0: $\beta_1 = 0$. Se recomienda tener mucho cuidado al interpretar el resultado de la prueba de esta hipótesis.

Si ignoramos la posibilidad siempre presente de cometer el Error Tipo I (rechazar una H_0 verdadera) o un Error Tipo II (no rechazar una H_0 falsa) podemos hacer las siguientes interpretaciones.

1. Si H_0: $\beta_1 = 0$ no es rechazada (es aceptada) esto tiene uno de los dos significados a continuación:
    a. Para un modelo de línea recta verdadero, *x* permite poco o no ayuda en la predicción de *y*; esto es, $\bar{y}$ es esencialmente tan buena como $\bar{y} - \hat{\beta}_1 (x - \bar{x})$ para predecir *y* (figura 3.4.a).
    b. La verdadera relación entre *x* y *y* no es lineal; esto es, el verdadero modelo puede involucrar términos cuadráticos, cúbicos u otras funciones más complejas de *x* (figura 3.4.b).

Combinando (a) y (b), podemos decir que no rechazar H_0: $\beta_1 = 0$ implica que un modelo de línea recta en *X* no es el mejor modelo y no ayuda mucho para predecir a *Y*.

2. Si H_0: $\beta_1 = 0$ es rechazada, esto significa que:
    a. *x* provee información significativa para predecir *y*; esto es, el modelo, definido por $\bar{y} + \hat{\beta}_1(x - \bar{x})$ es mejor que el modelo sencillo $\bar{y}$ para predecir *y* (Figura 3.4.c).
    b. Un mejor modelo puede tener, por ejemplo, un término curvilíneo, aunque existe un componente lineal definido (figura 3.4.d).

Combinando (a) y (b) podemos decir que rechazar H_0: $\beta_1 = 0$ implica que un modelo de línea recta en *x* es mejor que un modelo que no incluye *X*; aunque puede muy bien representar solamente una aproximación lineal para una relación verdaderamente no lineal.

Un punto importante de las interpretaciones anteriores es que si la hipótesis nula es o no rechazada, un modelo de línea recta no es apropiado. En su lugar, alguna otra curva puede describir la relación entre *x* y y de mejor manera.

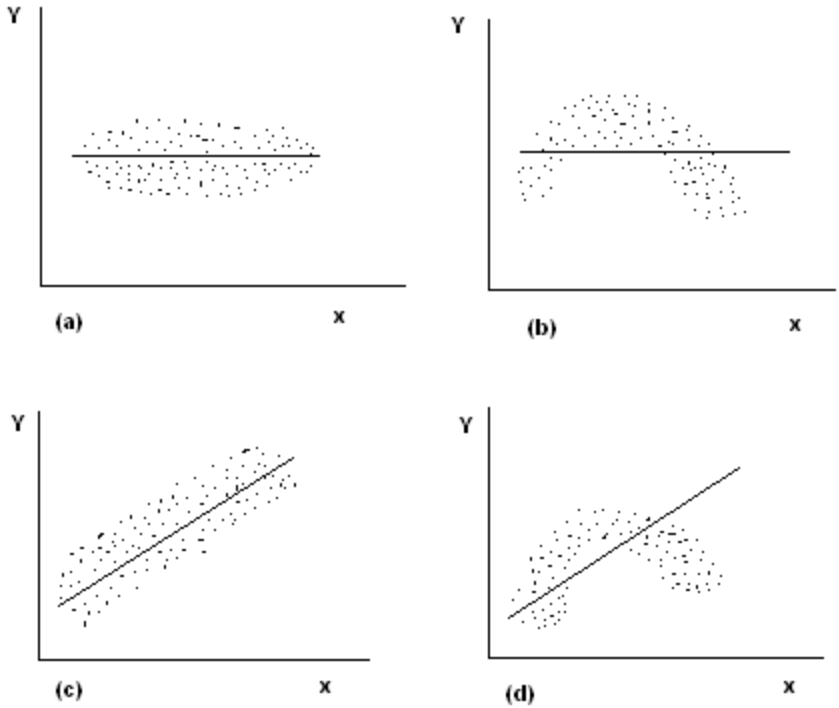


Figura 3. 4. Interpretación de la prueba de pendiente cero.

3.6. Prueba para el intercepto cero

Otra hipótesis que algunas veces es probada se relaciona con el intercepto, si la línea recta ajustada pasa por el origen; esto es, si la intersección entre *y* y β_0 es cero. La hipótesis usual aquí es H_0: $\beta_0 = 0$. Si esta hipótesis nula no es rechazada, puede ser apropiado quitar la constante del modelo de acuerdo a la experiencia anterior; esto sugiere que la línea pasa a través del origen y que se tienen observaciones con valores cercanos al origen para mejorar el

estimador de β_0. En muchos casos la hipótesis es de poco interés, ya que los datos no son agrupados alrededor del origen. Por ejemplo, cuando observamos la edad (x) y presión sanguínea (y) no estamos interesados en conocer qué pasa cuando $x = 0$ y raramente se eligen valores de x cercanos a cero.

3.7. La Tabla de Análisis de Varianza (ANVA)

Un resumen total de los resultados de cualquier análisis de regresión es la Tabla de Análisis de Varianza. Este nombre se deriva principalmente del hecho de que la información básica en una tabla ANVA contiene varios estimadores de varianza. Estos estimadores pueden ser usados para responder la cuestión de la inferencia principal en el análisis de regresión. En el caso de la línea recta estas cuestiones son: a) ¿la pendiente β_1 igual a cero es verdadera?, b) ¿cuál es la fuerza de la relación de la línea recta?, c) ¿es la línea recta un modelo apropiado?. Precisamente los problemas del análisis de varianza pueden ser expresados en un contexto de regresión, ya que resumen todos los resultados obtenidos de los métodos relacionados.

Las propiedades distribucionales de la teoría normal permiten la prueba fácil de hipótesis (20). Una simple prueba de F se puede realizar con los cálculos de la Tabla ANVA. Esto es, la relación entre las dos sumas de cuadrados divididas por sus grados de libertad generan la Suma de Cuadrados Medios de regresión ($SCM_{regresión}$) dividida entre la varianza muestral (s^2).

$$\frac{\frac{SC_{regresión}}{1}}{\frac{SC_{error}}{n-2}} = \text{SCM}_{\text{regresión}} \qquad (29)$$

Que tiene una distribución χ_1^2/ (1)/ χ_{n-2}^2/(n – 2) bajo H_0, entonces $SCM_{regresión}/s^2$ tiene una distribución $F_{1,\, n-2}$ bajo H_0 y es un candidato a un buen estadístico de prueba para probar la hipótesis (28).

Por lo anterior tenemos el siguiente modelo matemático:

Variación total no explicada = variación debido a la regresión + variación residual no explicada

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (30)$$

La ecuación (22) por lo general se denomina *ecuación fundamental del análisis de regres*ión y se ajusta a cualquier situación general de regresión, figura 3.5.

$\hat{Y}_i - Y_i$ = a la cantidad de X_i no explicada por la regresión

$\hat{Y}_i - \bar{Y}$ = a la cantidad de X_i explicada por la regresión

$Y_i - \bar{Y}$ = a la cantidad total no explicada de X_i

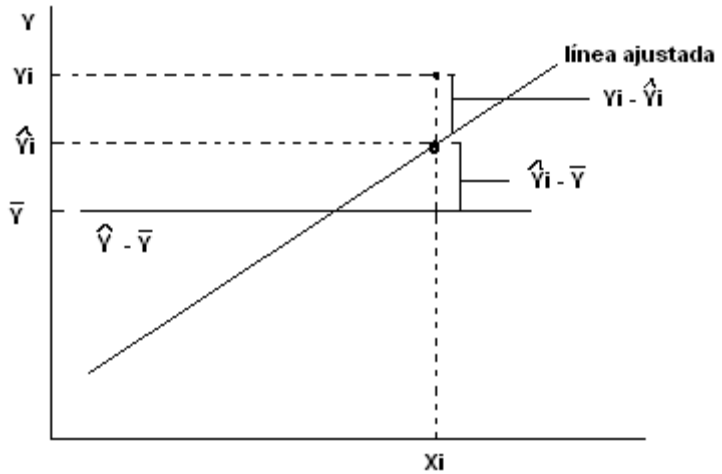


Figura 3.5. La variación explicada y no explicada por la línea recta de regresión.

Es importante observar que el término cuadrado medio de residual es el estimador de $S^2_{Y/X}$. Si el modelo verdadero de regresión es una línea recta, entonces $S^2_{Y/X}$ es un estimador de σ^2. Por otro lado, el término cuadrado medio de regresión ($SC_Y - SC_{error}$) genera un estimador de σ^2 solamente si la variable *x* no ayuda a predecir la variable dependiente *y*; esto es, si la hipótesis Ho: $\beta_1 = 0$ es verdadera. Si en realidad $\beta_1 \neq 0$, el término cuadrado medio de regresión es alterado en proporción a la magnitud de β_1 y por lo tanto sobreestimará a σ^2.

Se puede demostrar que los cuadrados medios de residual y los cuadrados medios de regresión son estadísticamente independientes. Entonces, si H_0: $\beta_1 = 0$ es verdadera, la razón de esos términos representan la razón de dos estimadores independientes de la misma varianza σ^2. Bajo el supuesto de normalidad de las y´s, tal razón tiene una distribución F y este estadístico F puede ser usado para probar la hipótesis H_0 (de que no existe

relación lineal entre *x* y *y*). Afortunadamente, esta manera de probar H_0 es equivalente a usar la prueba de dos colas de la prueba de t, debido principalmente al resultado matemático para *v* grados de libertad.

$$F_{1,v} = T^2_v \quad \text{tal que} \quad F_{1,v,\,1-\alpha} = t^2_{v,\,1-\alpha/2} \tag{31}$$

Estas expresiones en (31) establecen que el punto de la distribución F del 100 (1 – α) % con 1 y *v* grados de libertad es exactamente el mismo que el cuadrado del punto de la distribución t del 100(1 – α/2) con *v* grados de libertad.

Entonces la Tabla del Análisis de Varianza está dada por:

Fuente de variación	Grados de libertad	Sumas de cuadrados	Cuadrados medios	F_0
Regresión	1	$SC_{regresión}$	$SC_{regresión}/1$	$CM_{regresión}/CM_{error}$
Residual	n - 2	SC_{error}	$SC_{error}/n-2$	
Total	n - 1	SC_{total}		

Donde:

$$SC_{total} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i}^{n} y_i^2 - n\,\overline{y}^2 \tag{32}$$

$$SC_{regresión} = \hat{\beta}_1(\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i / n) \tag{33}$$

$$SC_{error} = SC_{total} - SC_{regresión} \tag{34}$$

3.8. Estimación de intervalos de confianza

Las pruebas de hipótesis separadas para la pendiente y el intercepto pueden ser acompañadas utilizando pruebas de t. Además podemos construir intervalos de confianza para los dos parámetros del modelo estimado, esto es:

Un intervalo de confianza del 100(1- α) % para la pendiente es

$$\beta_1 = \hat{\beta}_1 \pm t_{n-2,\,1-\alpha/2}\sqrt{\frac{CME}{(n-2)S_x^2}} \quad \text{donde} \quad S_x^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2] \qquad 35)$$

Donde CME es el cuadrado medio del error en la tabla ANVA. Un intervalo de confianza del 100(1 – α) % para el intercepto es

$$\beta_0 = \hat{\beta}_0 \pm t_{n-2,1-\alpha/2}\left[CME\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}}\right] \qquad (36)$$

3.9. Predicción de nuevos valores de Y a valores de X_0

En la práctica podemos estimar la variable respuesta *Y* sólo en valores basados en la línea de regresión ajustada; esto es, podemos predecir un valor de *Y* dado un $X = X_0$. Note que el estimador puntual que se usa en este caso es $\hat{Y}_{X_0} = \hat{\beta}_0 + \hat{\beta}_1 X_0$. Entonces, $\hat{Y}_{X_0}$ es usada para estimar ambos μ_{Y/X_0} y un valor de la respuesta Y en X_0. De esta manera es necesario poner algunos límites a los

estimadores y conocer su variabilidad. Aquí, sin embargo, no se puede asumir que se está construyendo un intervalo de confianza para *Y*, puesto que *Y* no es un parámetro, no se aplica una prueba de hipótesis por la misma razón. El término usado para describir los "límites híbridos" que requerimos es el *intervalo de predicción*, y está dado por:

$$\bar{y} + \hat{\beta}_1(x_0 - \bar{x}) \pm t_{n-2,\, 1-\alpha/2}\ \text{CME}\ \sqrt{1 + \frac{1}{n} + \ (x_0 - \overline{x})^2 / S_x^2} \qquad 37)$$

Primero se observa que un estimador con una respuesta individual tiene más variabilidad que un estimador de un grupo de valores respuestas. Esto es reflejado por el término extra 1 bajo el signo de raíz cuadrada en la ecuación anterior, que no se encuentra en la parte de la raíz cuadrada de la fórmula de intervalos de confianza para $\mu_{Y/X}$. Para ser más especifico, en la predicción de una *Y* observada para un individuo dado, se tienen dos fuentes de operación de error: error individual medido por σ^2 y el error en la estimación μ_{Y/X_0}.

3.10. Evaluando la calidad del modelo de línea recta

Se ha puntualizado anteriormente que la estrategia usual para la regresión con una sola variable independiente es asumir que el modelo de línea recta es apropiado, este supuesto es rechazado si los datos indican que se ajusta un modelo más complejo. Se pueden usar muchos métodos para evaluar si el

supuesto de línea recta es razonable. La técnica básica incluye pruebas para la falta de ajuste y se entiende mejor en términos de modelos de regresión polinomial. Muchas regresiones diagnósticas ayudan a evaluar los supuestos de la línea recta explícita o implícitamente. Con el modelo lineal, los supuestos de linealidad, homocedasticidad (la igualdad de varianzas) y normalidad están entrejidos y por lo general están en el modelo o son violados como un conjunto.

3.11. Calidad del modelo

Se puede recordar de secciones anteriores, la noción de pruebas de hipótesis de la pendiente, motivados por la necesidad de determinar si existe una relación lineal significativa. Sin embargo, también se menciona otra cuestión que se responde por la regresión ajustada. En muchos problemas de regresión, la prueba de hipótesis no es la forma de inferencia que resolverá los problemas de análisis. Por lo general, el éxito del ejercicio analítico depende de la propia selección de criterios cuantitativos, que determinan la calidad del modelo ajustado. Dos cuestiones aparecen completamente de forma natural.

1. ¿Se ajusta el modelo de forma adecuada a los datos?
2. ¿Predice el modelo bastante bien la respuesta?

Aquí trataremos específicamente con estos dos problemas para el caso simple de regresión lineal. Aunque los conceptos se pueden extender para la regresión lineal múltiple.

3.11.1. Coeficiente de determinación (R^2)

El coeficiente de determinación, referido simbólicamente como R^2, es usado algunas veces de forma incorrecta y mide el ajuste de la línea de regresión. Si se considera el modelo dado en ecuaciones anteriores, la definición es simplemente

$$R^2 = \frac{SC_{regresión}}{SC_{total}} = 1 - \frac{SC_{error}}{SC_{total}} \tag{38}$$

De la ecuación anterior se nota que representa la proporción de la variación en los datos de la variable, respuesta que se explica con el modelo. Claramente $0 \leq R^2 \leq 1$, y el límite superior es alcanzado cuando el ajuste del modelo en los datos es perfecto; esto es, cuando todos los residuales son cero. ¿Cuál es un valor de R^2 aceptable? Ésta es una cuestión difícil para responder y, en verdad, que sea aceptable depende del campo científico del que los datos son tomados. Un químico puede encontrar un R^2 que quizás exceda a 0.99, mientras que un investigador del comportamiento humano sólo espera encontrar un R^2 tan alto como de 0.70. Claramente algunos fenómenos científicos se ajustan mejor a modelos lineales que otros.

Aunque el coeficiente de determinación es fácil para interpretar y puede ser comprendido por muchos investigadores, existen algunas trampas en su uso que son muy notorias. Por ejemplo: es un criterio peligroso, para comparar modelos candidatos, simplemente porque al adicionar cualquier término (cuadrático, cúbico, etc) se incrementarán los Cuadrados Medios del Error y por lo tanto se incrementa R^2. Esto implica que R^2 puede ser hecho artificialmente alto, por una práctica poco prudente del ajuste. Un incremento en R^2 no implica que el término adicional en el modelo sea necesario. Verdaderamente, se observa, que si la predicción es el propósito del modelo, entonces una estructura de un modelo más complicado con un R^2 más alto que el de un modelo simple, no necesariamente indica que es el mejor. Entonces no se realizará una selección de modelos de un proceso que solamente involucre las consideraciones de R^2. La noción del sobreajuste y su papel en la evaluación de las capacidades de predicción se discutirán en los siguientes capítulos.

El coeficiente de determinación también puede ser artificialmente alto, debido a que la pendiente de la regresión es grande y por la dispersión de los datos de las regresoras x_1, x_2, x_3... x_n, que es grande. Se puede ilustrar esto, sin mucha dificultad, observando las estructuras del denominador y el numerador separadamente. Se ha demostrado que

$$E\,(SC_{regresión}) = \sigma^2 + \beta_1^2 S_x^2 \qquad (39)$$

y que

$$E\,(SC_{error}) = \sigma^2(n-2)$$

y como un resultado, tenemos que

$$E\,(SC_{total}) = (SC_{regresion} + SC_{error}) = \sigma^2\,(n-1) + \beta_1^2 S_x^2 \qquad (40)$$

Ahora podemos establecer que $E(R^2)$ es simplemente la razón de los valores esperados.

Entonces comprendemos que R^2 = variación explicada por la regresión / variación total. Donde la palabra "variación" está en la variable respuesta.

En el caso de la regresión a través del origen, tenemos que el coeficiente de variación es dado por

$$R^2_{(0)} = \frac{SC_{regresión}}{SC_{total}} \qquad (41)$$

Por lo general existe una fuerte tendencia de que $R^2_{(0)}$ es más grande que R^2. Pero no significa que la calidad del ajuste sea mejor. Es el resultado del uso de sumas de cuadrados no corregidos por la media.

3.11.2. Coeficiente de Variación

El Coeficiente de Variación (C.V.) es un criterio razonable para representar la calidad del ajuste y medida de la dispersión alrededor de la línea de regresión. El C.V. se define como,

$$\text{C.V.} = \frac{s}{\bar{y}} 100 \tag{42}$$

El C.V. es el estimador residual de las desviaciones del error estándar, medidas como un porcentaje del valor de la respuesta promedio. Su uso se debe a que es el estimador de las desviaciones del error estándar (s), por lo general no es satisfactorio como una medida de calidad de ajuste, ya que no tiene una escala libre. Por ejemplo, un investigador que tiene un valor de *s* = 14 partes por millón como una desviación del error estándar y si encuentra un C.V. = 5%, entonces el investigador sabe que la dispersión natural alrededor de la línea es del 5% de la respuesta promedio; lo cual es más claro que el sólo valor de *s*.

3.11.3. Factores que alteran la capacidad de predicción del modelo

La intuición ciertamente sugiere cuáles factores influyen en la capacidad de predicción de un modelo de regresión. Por ejemplo,un incremento en el tamaño de la muestra obviamente disminuye la varianza de la predicción,

asumiendo que las otras cosas son iguales. Además, la predicción mejora por un incremento en la cantidad S_x^2 implicando que la dispersión más grande del rango de los datos de la variable regresora es la mejor capacidad predictora del modelo. Esto es una importante consideración en situaciones en dónde el investigador tiene el control en los valores de X. Una nota de cuidado es: *la varianza de predicción es una medida de la calidad de la predicción*, asumiendo que el modelo es apropiado. Si el modelo no es apropiado y si se incluyen pocos términos en el modelo, el sesgo en la respuesta predicha es un factor que se debe tomar en cuenta. Claramente, si el investigador usa una dispersión excesiva en los niveles de *X* para obtener un s_x^2 mayor, existe el riesgo de violar los supuestos del modelo. Por ejemplo, un modelo lineal en *X* en una región estrecha puede ser agradable, mientras que en un rango más grande en *X* puede requerir un modelo que contenga términos cuadráticos.

Finalmente, otro resultado que es consistente con la intuición relacionada a la localización de X_0, es el punto en el cual se desea predecir. Se establece que la predicción será mejor cuando X_0 se acerque a $\bar{X}$, el promedio de la variable regresora. Por esto se pueden observar resultados pobres cuando se usa la extrapolación.

3.11.4. Una mirada a los residuales

En este caso se ilustra una presentación preliminar del tipo de información que puede ser obtenida de los residuales ordinarios, esto es $\varepsilon_i = y_i - \hat{y}_i$, denominado *los errores del ajuste*. No se presentarán muchos detalles ya que

serán tratados en puntos donde se describa la regresión múltiple como una herramienta del estudio de las propiedades de los residuales. Por eso en regresión lineal simple se mostrará el análisis y gráficos de residuales. El lector podrá observar los residuales como cantidades a través de los cuales se determina si las condiciones son o no ideales. Por ideal se refiere a los supuestos hechos sobre los ε_i.

Las severas violaciones de los supuestos E (ε_i) = 0 y E (ε_i^2) = σ^2 para i = 1, 2, 3 ... n, se pueden por lo general confirmar a través del análisis de residuales. El investigador usa los residuales como las cantidades que más se aproximan a los errores del modelo, los ε_i. La presentación de una gráfica de residuales versus X, la variable regresora, describe valores aleatorios centrados alrededor de cero (figura 3.6.).

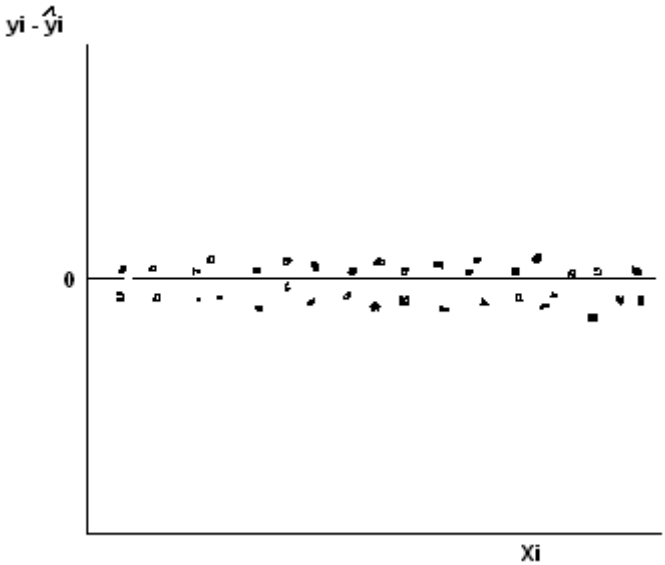


Figura 3.6. Gráfica ideal de (Y_i - $\hat{Y}_i$) *versus* X_i de una regresión lineal simple.

3.12. Examen de los datos y el modelo

La discusión en este punto está basada en la premisa de que el modelo está correctamente especificado. El supuesto de que las observaciones son independientemente distribuidas como N ($\beta_0 + \beta_1 x_i$, σ^2) tiene varias implicaciones que se pueden verificar. En particular, se pueden examinar las siguientes cuestiones:

a) ¿Es éste un supuesto razonable para la función media, al menos sobre el rango de los datos disponibles?
b) ¿Es la varianza la misma sobre el rango de los valores de entrada?
c) ¿Fueron todos los datos colectados del mismo modelo?
d) ¿Es aceptable el supuesto de normalidad?

Se consideran varias gráficas y procedimientos numéricos para verificar estos supuestos. Estas técnicas están comprendidas en lo que se denomina *regresión diagnóstica*.

Para el modelo de regresión lineal simple, una forma obvia de verificar la correspondencia de la función media es la gráfica de la línea ajustada sobre el diagrama de dispersión de los datos observados. Esta gráfica puede revelar puntos o grupos de puntos que no son explicados por el modelo ajustado. El gráfico puede sugerir una función de respuesta alternativa, o que la varianza depende de los valores de entrada. En la figura 3.7. se muestran dos gráficos de dos tipos de datos. En ambos casos se observa que el modelo se ajusta bien

a los datos y parece ser poco razonable dudar de los incisos (a) y (b) citados anteriormente.

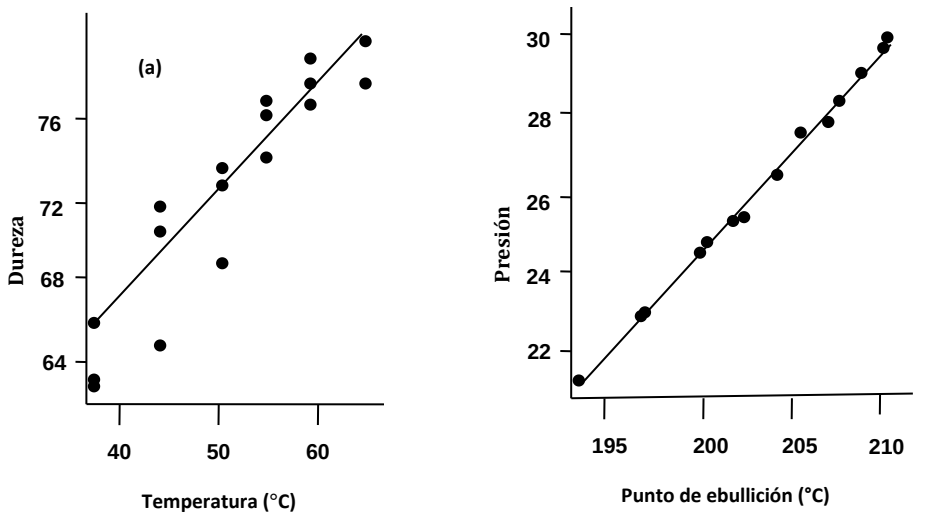


Figura 3.7. Ecuaciones ajustadas. (a) Partículas de alimento y (b) datos Forbes.

Los modelos ajustados para ambos conjuntos de datos son:

Modelo para las partículas de alimento			**Modelo para los datos Forbes**		
Variable	**Parámetro**	**Error estándar**	**Variable**	**Parámetro**	**Error estándar**
Intercepto	45.457	2.845	Intercepto	-81.064	2.052
Pendiente	0.517	0.053	Pendiente	0.522	0.010
R^2	0.854		R^2	0.9944	

Un problema potencial en estas gráficas es que la magnitud de la respuesta puede predominar con algunos problemas en desviaciones de los supuestos.

Para enfatizar algún problema que esté presente podemos considerar una gráfica de los residuales en función de la entrada. Estos residuales jugarán un papel importante en el análisis y en una discusión detallada de sus propiedades.

3.12.1. Residuales

Si todos los parámetros fueran conocidos, el análisis consistiría en la verificación de que las diferencias $\varepsilon_i = y_i - \beta_0 - \beta_1 X_i$ sean una muestra N (0, σ^2). Hay muchas pruebas para la normalidad basadas en cada una de las muestras o en la función de densidad de la distribución cumulativa muestral. Una de tales muestras se discute más adelante. En la práctica, los parámetros no son conocidos y es natural considerar los residuales observados ε_i. El vector de residuales es distribuido N [0, σ^2 (I – H)]. Las varianzas y covarianzas de los residuales son denotados por Var (ε_i)= σ^2 (I- h_{ii}) y la covarianza Cov (ε_i, ε_j) = $-\sigma^2 h_{ij}$.

Donde

$$H = \frac{1}{n} JJ^T + \frac{1}{S_{xx}} x^* x^{*T}, \text{ en particular } h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \qquad (43)$$

Entonces los residuales por lo general están correlacionados con diferentes varianzas y no se pueden ver como una muestra aleatoria N(0, σ^2). A pesar de

estas observaciones, los residuales son intuitivamente atrayentes; ya que revelan cómo nuestro modelo ajusta los datos. Una gráfica de los ε_i contra los X_i puede revelar aparentes desviaciones de los supuestos. Un problema común es que la variabilidad se incrementa con la magnitud de la respuesta, ya que si β_1 es positivo, la variabilidad incrementará con los incrementos de las entradas. Este comportamiento se puede observar en la figura 3.8 para el caso de pendiente positiva (a) y pendiente negativa (b).

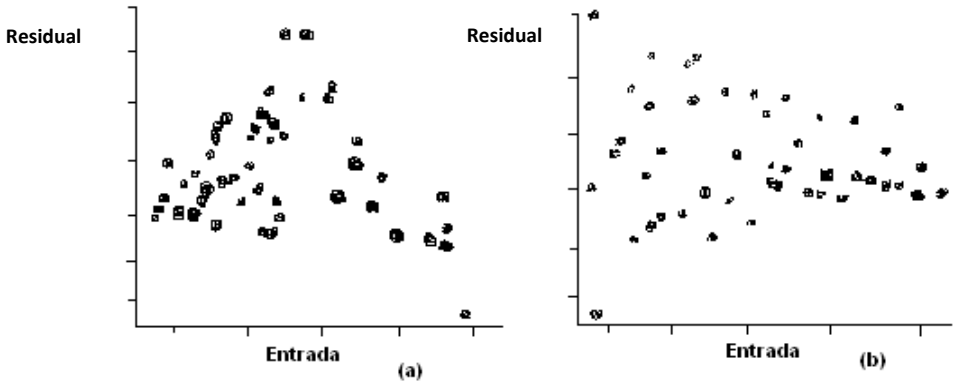


Figura 3.8. Gráficas de residuales con incrementos de la varianza y la respuesta
(a) Pendiente positiva (b) Pendiente negativa

Para el modelo de regresión lineal simple, se grafican los residuales ε_i contra las entradas X_i lo que equivale a graficar los residuales contra los valores predichos $\hat{y}_i$; ya que los valores predichos son una función lineal de las entradas. El comportamiento de las gráficas de los pares $(\hat{y}_i, \varepsilon_i)$ está dado por el hecho de que los coeficientes de correlación muestral calculados para estos

pares es cero. Entonces las gráficas de ($\hat{y}_i$, ε_i) exhiben una tendencia no lineal; si el modelo es correcto, no muestran un patrón distinto.

En la figura 3.9 se muestran las gráficas de los residuales contra las entradas para los datos de los dos modelos anteriores. La gráfica de los datos de las partículas de alimento no muestran ningún patrón pero la gráfica de los datos de Forbes tiene dos rasgos distintos. Esto es, muchos de los residuales indican un patrón fuerte en forma de U, pero una observación se desvía de este patrón. Esta gráfica sugiere que (1) puede existir un problema con el dato lejano (punto 12), que (2) no se puede tener una forma funcional correcta, o que (3) puede ser mejor modelar alguna función de la respuesta. Este patrón no se vio en la figura 3.7, ya que la magnitud de los residuales es más pequeña relativamente a la magnitud de la respuesta. El modelo lineal da un buen ajuste de los datos, lo que es completamente aceptable para la predicción.

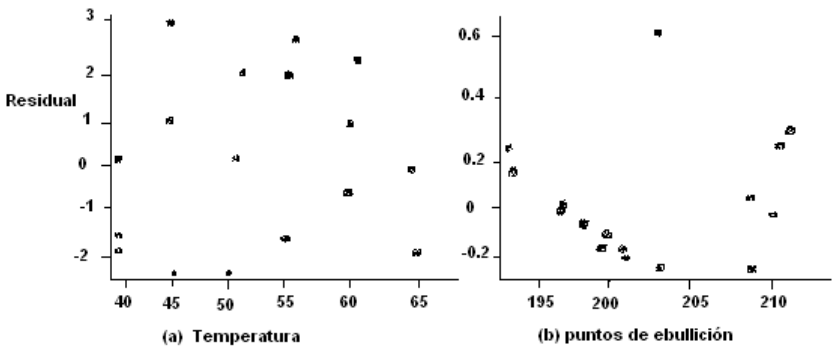


Figura 3.9. Gráficos de los residuales contra las entradas.
(a) Datos de partículas de alimentos (B) Datos Forbes

Un error muy común en esta etapa del análisis es graficar los residuales contra los valores observados. Lo cual muestra un patrón distinto que sugiere un modelo incorrecto. Para saber por qué sucede esto debemos ver que cor (ε, y) = $\sqrt{1 - R^2}$. Con esto se puede observar que la gráfica de (y_i, ε_iε) muestra una tendencia positiva si $R^2 < 1$, sólo cuando los supuestos del modelo se cumplen. El uso de residuales para indicar patrones o casos inusuales es muy común. Se pueden examinar algunas funciones alternativas de los residuales que son más adecuadas para este propósito.

1. Residuales estandarizados y estudentizados. Como se ha dicho anteriormente, los gráficos de residuales pueden revelar un problema con el supuesto de varianza homogénea, como se observa en la figura 3.8. Mientras estas gráficas pueden ser alternativas, se recomienda tener mucho cuidado. Recordamos que la varianza de un residual en regresión lineal simple está dada por

$$\text{Var}(r_i) = \sigma^2[\ 1 - \frac{1}{n} - (x_i - \bar{x})^2\ ]/\ S_{xx} \qquad (44)$$

Esto significa que los residuales para entradas lejanas de $\bar{x}$ tendrán una varianza más pequeña que las que están cercanas a la media. Entonces, la gráfica de residuales puede sugerir que los datos son más variables para valores intermedios y menos variables para valores extremos de las entradas; sólo cuando los supuestos del modelo se cumplen. Para hacer comparables

los residuales, se balancean por sus errores estándares y esto se refiere a los *residuales estandarizados*, definidos como

$$(\varepsilon_{is})_i = \frac{\varepsilon_i}{\sqrt{S^2(1-h_{ii})}} \qquad (45)$$

Donde S^2 es el cuadrado medio del residual del análisis de varianza. Las gráficas de estos residuales contra los valores predichos generalmente son semejantes a las gráficas usuales de residuales, excepto para casos extremos cuando se presentan diferencias dramáticas en los elementos de la diagonal de la matriz H. Estos residuales son usados como una indicación del hecho de que una observación es pobremente ajustada por el modelo. Mientras esta estandarización es razonable, porque no es completamente satisfecha, si un caso no está bien ajustado, la suma de cuadrados de residual es inflada por los grandes residuales. Con estos se reduce la magnitud de los residuales estandarizados. Un procedimiento alternativo es reajustar el modelo excluyendo el *i-ésimo* caso y se usan los cuadrados medios de residual resultantes denotados por $S^2_{(i)}$ con [(N – 1) – 2] grados de libertad como la medida de variabilidad. La omisión de ajustar el *i-ésimo* caso es la diferencia entre y_i y el valor predicho obtenido de la regresión con N – 1 observaciones, no incluyendo el caso i. Denotamos por $\hat{y}_{i(i)}$ el valor predicho para el *i-ésimo* caso de la regresión calculada sin el caso i y usamos $y_i - \hat{y}_{i(i)}$ como la medida del ajuste del caso i. Estandarizando estas diferencias se tienen los residuales estudentizados, definidos por

$$t_i = y_i - \hat{y}_{i(i)} / \sqrt{var\,[y_i - \hat{y}_{i(i)}]} \quad (46)$$

Un primer vistazo de los cálculos de estos residuales que representan un esfuerzo importante es el que debemos evaluar N regresiones adicionales, cada una con n – 1 observaciones, correspondientes a la eliminación del caso i para i = 1, 2, 3... n.

Se puede mostrar que

$$t_i = \frac{\varepsilon_i}{\sqrt{S_i^2(1-h_{ii})}} \quad (47)$$

Además, $S^2_{(i)}$ se calcula de información disponible como:

$$S^2_{(i)} = \frac{S^2}{n-1-p}\,[n - p - (\varepsilon_s)_i^2] \quad (48)$$

Donde p = 2 en el modelo de regresión lineal simple. Si un caso es pobremente ajustado, vemos que $t_i > (\varepsilon_s)_i$ y da una indicación más objetiva para el pobre ajuste de ese caso.

El concepto “estudentizado” fue introducido por Hartley (1959) y ha sido discutido en muchos trabajos. Por lo general, nos referimos a (46) como los residuales estandarizados y a (47) como los residuales estudentizados.

3.12.2. Outliers, puntos extremos e influencia

No es anormal en el análisis de regresión encontrar uno o más casos que se ven inusuales, en el sentido de que las observaciones difieren del cuerpo principal del conjunto de datos. Se tienen dos situaciones: primero; el valor de la variable de entrada puede ser aceptable, pero el valor de la respuesta es muy grande o muy pequeño. Tales casos son llamados *outliers* (puntos lejanos). Segundo, el valor de la variable de entrada es completamente diferente del resto de los datos (en regresión lineal simple, podemos tener $|X_i - \bar{X}|$). Tales observaciones serán llamadas *puntos extremos*. El término caso influyente se refiere a cualquiera de los dos casos, un outlier o punto extremo, cuya inclusión altera sustancialmente los estimadores, las sumas de cuadrados de los residuales o los valores predichos. Nuestra tarea es identificar tales casos y valorar la influencia que tienen sobre el análisis. Aquí se discuten estos conceptos en el contexto de la regresión lineal simple, ya que son fáciles de visualizar. Esto ayuda a la intuición cuando consideremos la extensión al caso de la regresión lineal múltiple. Se utiliza el término de *diagnóstico de casos* para referirse a las técnicas, gráficas o numéricas que serán usadas para identificar las observaciones inusuales.

Outliers. Se considera primero el caso de los outlier (puntos lejanos) en una situación típica como en la figura 3.10, se muestra una gráfica de residuales versus la variable de entrada en un análisis de regresión lineal simple. Muchos de los datos fueron bien ajustados, pero una observación tiene un gran

residual y la respuesta está arriba de la línea recta. Seguramente la línea recta se ha movido hacia arriba, generando también residuales negativos en un intento de acomodar el caso (la observación). Esto ilustra un problema que es inherente al método de mínimos cuadrados. Esto es, el método intenta ajustar casos inusuales a costa de no ajustar bien el resto de los datos.

Tales observaciones pueden ser obvias en un diagrama de dispersión de los datos en una regresión lineal simple. La gráfica de residuales puntualiza la situación. Con más de una variable predictora puede ser difícil detectar outliers en una gráfica simple de los datos, pero las gráficas de residuales ayudan a hacerlo. En cualquier caso, es útil cuantificar la magnitud del outlier. Los residuales ordinarios ε_i hacen esto, como se ha visto, pero no se pueden comparar ya que sus varianzas pueden diferir. Además tiene la misma dimensión que las respuestas, en consecuencia aparece la cuestión de significancia. Los residuales estandarizados definidos anteriormente son candidatos naturales y podemos observar su distribución. Usando los resultados de la distribución conjunta de formas lineales y cuadráticas se puede mostrar que los residuales ε_i no son independientes de s^2, en consecuencia estos residuales no tienen la distribución de t. Puesto que $(\varepsilon_s)_i$ es N(0, 1) si s^2 fuera reemplazada por σ^2, es común usar esta aproximación para proveer un indicador de la magnitud de estos residuales. Entonces, una observación puede ser identificada como un outlier, si $| (\varepsilon_s)_i | > 3.0$, un valor que esperaríamos exceda el 0.27% de las veces. Si aplicamos este criterio a todos los residuales, debemos cuidar los efectos simultáneos.

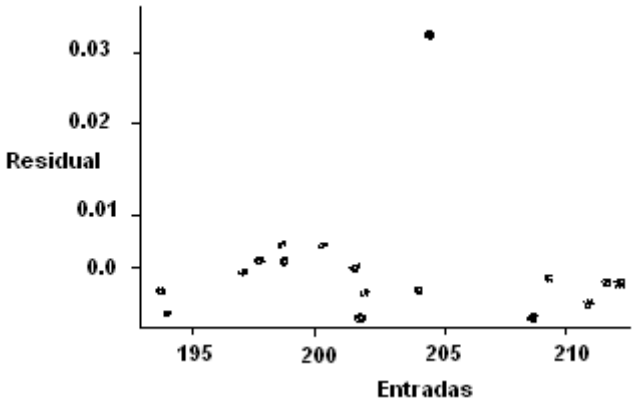


Figura 3.10. Gráfica de residuales indicando una observación outlier.

Los residuales estudentizados dan una mejor indicación de la incapacidad del modelo para ajustar el caso i y su distribución se puede obtener fácilmente. También se puede mostrar que $t_i \sim t_{(N-1-p)}$, donde en el modelo p = 2. Para una muestra de tamaño razonable , el criterio $| t_i | > 3$ indica un residual inusualmente grande.

Puntos extremos. El concepto de un punto extremo en regresión lineal simple es obvio. Esto es, X_i es extremo si está muy lejos del resto de las observaciones. Las consecuencias de un punto extremo dependerán del valor asociado a la respuesta y en consecuencia el término *influencia potencial* es aplicado por lo general. El término *punto de apoyo* también se usa para reflejar esta influencia potencial en el análisis. Esta idea se ilustra en la figura 3.11. Donde se puede ver que el valor de una de las observaciones es mucho más grande que las restantes. La figura 3.11(a) ilustra la situación en la cual

este caso es influyente ya que la pendiente de la línea ajustada es mucho menor si este caso está incluido y más si es excluido. La figura 3.11(b) muestra un punto extremo que no tendrá mucho efecto en la línea ajustada pero extiende el rango de aplicación del análisis. En cualquier caso se intenta checar los valores de la entrada y la salida de esta situación.

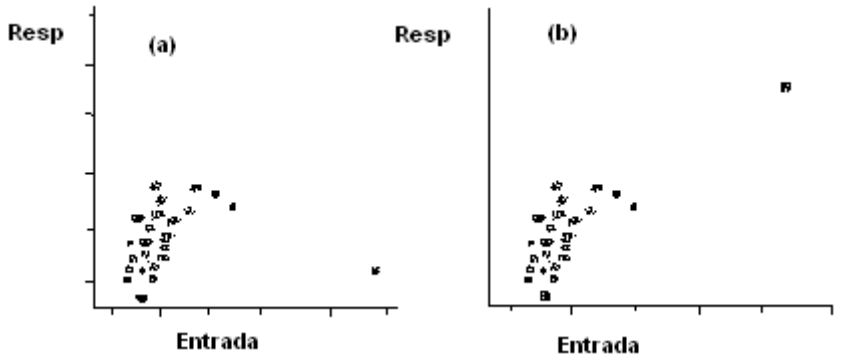


Figura 3.11. Diagramas de puntos que indican un punto extremo.
(a) un caso influyente (b) un caso no influyente.

La situación de la figura 3.11 (a) presenta un dilema. Se puede descartar esta observación y hacer una buena descripción del resto de los datos. Esto se justifica si se identifica un error, al registrar los valores de la entrada o de la respuesta. Por el otro lado, si es válida esta observación, el rango de aplicación del modelo es mucho mayor, pero para el modelo de regresión lineal simple no se considera apropiado. Un modelo que inicialmente se incrementa y después decrece, es una función cuadrática que puede ser más apropiada. Desafortunadamente, no se tiene más información para justificar

este modelo. El punto que se enfatiza es el que las observaciones inusuales no deben ser tratadas ligeramente. Su inclusión puede tener un efecto substancial en el análisis, por eso se sugiere eliminarlas. El peligro con esto es que la observación puede ser correcta y que contiene información valiosa, que puede contribuir el entendimiento del sistema bajo estudio. La situación de la Figura 3.11 (b), es menos que un problema ya que la línea ajustada no cambia substancialmente por la eliminación de estos puntos extremos. Aquí permanece la cuestión de si se aplica el modelo para intervenir las entradas, ya que no se tienen datos para justificar esta afirmación.

Mientras que la identificación de casos extremos es fácil cuando tenemos una sola variable de entrada, se convierte en un problema cuando tenemos muchas variables de predicción. En esta situación estamos forzados a confiar en cualquier indicador numérico o en métodos gráficos más complejos en nuestras investigaciones. Para motivar esta discusión podemos notar que en regresión lineal simple, una medida natural de la distancia de X_i del resto de los datos es $(x_i - \bar{x}_{(i)})$, donde $\bar{x}_{(i)}$, es la media obtenida de los n – 1 observaciones restantes. En la práctica usamos $(x_i - \bar{x})$, reconociendo que esta medida está influenciada por el hecho de que $\bar{X}$ se calcula usando el caso i. Si elevamos al cuadrado esta cantidad y la escala, y la dividimos por la suma de cuadrados S_{xx} en base a todos los datos, obtenemos la medida de la distancia escalada,

$$h_{cii} = \frac{(x_i - \bar{x})^2}{S_{xx}} \qquad (49)$$

De esta manera Belsley, Kuh and Welsch (1980) sugirieron que valores de $h_{ii} > \frac{2p}{N}$ es un indicativo de la presencia de valores outlier en el vector de entrada. Como otra indicación del papel de la matriz Hat en la valoración de la influencia potencial, recordamos que los valores predichos son dados por $\hat{y}$ = Hy. Esto es,

$$\hat{y}_i = h_{ii}\, y_i + \sum_{j \neq i}^{n} h_{ij}\, y_j \qquad (50)$$

Se puede demostrar que si un elemento de la diagonal de H está cercano a uno, los elementos restantes en esta hilera son pequeños. La influencia del caso i, es aparente en la ecuación anterior, ya que $h_{ii} = 1$, entonces $\hat{y}_i = y_i$. Esto es, este caso es influyente que fuerza a la línea de regresión a pasar más cerca del punto en cuestión. También es interesante valorar el efecto de un caso extremo sobre su residual. Como se ha observado, un caso influyente tendrá $\hat{y}_i = y_i$ o equivalentemente, $\varepsilon_i = 0$. Recordando que el *i-ésimo* residual es $\sigma^2 (1 - h_{ii})$, se observa que los casos influyentes tendrán una varianza pequeña. Para entender este resultado, hay que recordar que esta varianza mide la variabilidad en muestreo repetido con la misma matriz de diseño. Un punto extremo siempre se podrá ajustar ya que la varianza de los residuales será pequeña. Los residuales estandarizados y los estudentizados parcialmente se calculan dividiendo los residuales por su error estándar pero el escalado de los residuales por lo general no es suficiente para señalar las observaciones que son inusuales.

Influencia. La discusión de outliers y puntos extremos ha sugerido su potencial para influenciar su análisis, pero no podemos ser tan específicos de la manera en que la influencia será medida. Hay varias cuestiones obvias:

1. ¿Cómo están afectados los estimadores de los parámetros β_0 y β_1 por la inclusión de los casos sospechosos?
2. ¿Cuál es el efecto sobre el estimador de σ^2?
3. ¿Cuál es el efecto sobre los valores predichos?

Sin muchas preocupaciones de la aparente cantidad de cálculos, se sugiere que se calcule para cada caso, i = 1, 2, 3, ..., N, $\hat{\beta}_{0(i)}$, $\hat{\beta}_{1(i)}$, $S^2_{(i)}$ y $\hat{y}_{(i)}$, los estimadores de los parámetros y el vector de los valores predichos basados en el análisis sin considerar el *i-ésimo* caso.

Ejemplo 2. Los datos presentados en la siguiente tabla corresponden a 50 refinerías de petróleo y muestra los galones totales de agua usados mensualmente (H_2O) y el número de barriles de petróleo procesado (PROD), también muestra la capacidad de las torres refrigerantes (CAP). Ajustar el modelo lineal de regresión del H_2O sobre PROD. Ajustar el modelo H_2O/PROD y CAP/PROD.

H_2O	PROD	CAP	H_2O	PROD	CAP	H_2O	PROD	CAP
19.01	13.5	1.77	35.75	20.0	2.40	14.98	16.5	1.34
8.92	11.0	0.79	31.25	21.0	2.89	18.88	17.0	1.73
14.21	12.5	1.28	26.78	23.0	2.37	33.13	19.0	3.11
21.15	12.0	1.98	38.02	22.5	3.58	24.73	17.5	2.33
12.99	11.5	1.49	17.09	27.0	1.43	15.12	18.5	1.32
25.37	14.5	2.36	33.65	26.0	2.94	20.97	19.5	1.91
15.71	10.0	1.5	36.87	28.0	3.52	45.31	24.5	4.46
21.43	16.0	2.04	53.81	28.5	5.22	18.26	21.0	1.59
11.7	15.5	1.05	35.65	29.5	3.21	28.06	23.0	2.67
22.54	18.5	2.22	46.74	26.0	4.34	26.32	24.5	2.36
22.74	16.5	2.04	12.35	10.5	1.10	15.83	24.0	1.36
19.95	15.0	1.75	25.52	13.0	2.37	16.84	22.0	1.48
25.65	15.5	2.45	12.68	14.0	1.16	40.17	23.5	3.97
35.34	23.0	3.27	17.25	12.5	1.58	23.38	27.5	1.93
14.91	20.0	1.40	11.5	10.0	1.00	45.23	28.5	4.21
25.63	24.5	2.36	17.76	12.0	1.68	24.51	26.5	2.24
21.81	21.5	1.89	38.28	19.0	3.51			

Usando software SAS, tenemos

data regres;

input ho prod cap;

cards;

19.01 13.5 1.77
8.92 11.0 0.79
14.21 12.5 1.28
21.15 12.0 1.98
12.99 11.5 1.49
25.37 14.5 2.36
15.71 10.0 1.5
21.43 16.0 2.04
11.7 15.5 1.05

22.54	18.5	2.22
22.74	16.5	2.04
19.95	15.0	1.75
25.65	15.5	2.45
35.34	23.0	3.27
14.91	20.0	1.40
25.63	24.5	2.36
21.81	21.5	1.89
35.75	20.0	2.40
31.25	21.0	2.89
26.78	23.0	2.37
38.02	22.5	3.58
17.09	27.0	1.43
33.65	26.0	2.94
36.87	28.0	3.52
53.81	28.5	5.22
35.65	29.5	3.21
46.74	26.0	4.34
12.35	10.5	1.10
25.52	13.0	2.37
12.68	14.0	1.16
17.25	12.5	1.58
11.5	10.0	1.00
17.76	12.0	1.68
38.28	19.0	3.51
14.98	16.5	1.34
18.88	17.0	1.73
33.13	19.0	3.11
24.73	17.5	2.33
15.12	18.5	1.32
20.97	19.5	1.91
45.31	24.5	4.46
18.26	21.0	1.59
28.06	23.0	2.67
26.32	24.5	2.36

15.83 24.0 1.36
16.84 22.0 1.48
40.17 23.5 3.97
23.38 27.5 1.93
45.23 28.5 4.21
24.51 26.5 2.24

proc glm;

model prod = cap;

proc plot;

plot prod*cap = "x";

title "Gráfica de la producción sobre la capacidad";

run;

Resultados

El análisis de prod. sobre cap.

The GLM Procedure

Number of observations 50

Dependent Variable: prod

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	645.439609	645.439609	32.60	<.0001
Error	48	950.365391	19.799279		
Corrected Total	49	1595.805000			

R-Square	Coeff Var	Root MSE	prod Mean
0.404460	23.01934	4.449638	19.33000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Cap	1	645.4396094	645.4396094	32.60	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Cap	1	645.4396094	645.4396094	32.60	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	11.17168257	1.56131349	7.16	<.0001
Cap	3.57977948	0.62697923	5.71	<.0001

Los resultados del análisis de prod *versus* H_2o

The GLM Procedure

Number of observations 50

Dependent Variable: prod

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	714.345584	714.345584	38.90	<.0001
Error	48	881.459416	18.363738		
Corrected Total	49	1595.805000			

R-Square	Coeff Var	Root MSE	prod Mean
0.447640	22.16913	4.285293	19.33000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Ho	1	714.3455837	714.3455837	38.90	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Ho	1	714.3455837	714.3455837	38.90	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
0.36818521	1.55946024	6.65	<.0001	
ho	0.36086004	0.05785823	6.24	<.0001

	R-Square	Coeff Var	Root MSE	prod Mean
Modelo 1 (CAP)	0.404460	23.01934	4.449638	19.33000
Modelo 2 (H_2O)	0.447640	22.16913	4.285293	19.33000

Como podemos ver en el resumen los valores de R^2 son bajos y el coeficiente de variación similar y alto para este tipo de datos, por lo que se puede considerar la transformación de las variables, para lograr un mejor ajuste y capacidad de predicción del modelo.

Transformando las variables de la siguiente manera: H_2Opro = H_2O/prod y capro = cap/prod y ajustamos el modelo lineal los resultados son:

data regres;

input ho prod cap;

hopro = ho/prod;

capro = cap/prod;

cards;

19.01 13.5 1.77
8.92 11.0 0.79
14.21 12.5 1.28
21.15 12.0 1.98
12.99 11.5 1.49
25.37 14.5 2.36
15.71 10.0 1.5
21.43 16.0 2.04
11.7 15.5 1.05
22.54 18.5 2.22
22.74 16.5 2.04
19.95 15.0 1.75
25.65 15.5 2.45
35.34 23.0 3.27
14.91 20.0 1.40
25.63 24.5 2.36
21.81 21.5 1.89
35.75 20.0 2.40
31.25 21.0 2.89
26.78 23.0 2.37
38.02 22.5 3.58
17.09 27.0 1.43
33.65 26.0 2.94
36.87 28.0 3.52
53.81 28.5 5.22
35.65 29.5 3.21

```
46.74   26.0    4.34
12.35   10.5    1.10
25.52   13.0    2.37
12.68   14.0    1.16
17.25   12.5    1.58
11.5    10.0    1.00
17.76   12.0    1.68
38.28   19.0    3.51
14.98   16.5    1.34
18.88   17.0    1.73
33.13   19.0    3.11
24.73   17.5    2.33
15.12   18.5    1.32
20.97   19.5    1.91
45.31   24.5    4.46
18.26   21.0    1.59
28.06   23.0    2.67
26.32   24.5    2.36
15.83   24.0    1.36
16.84   22.0    1.48
40.17   23.5    3.97
23.38   27.5    1.93
45.23   28.5    4.21
24.51   26.5    2.24
```

proc glm;

model hopro= capro;

proc plot;

plot hopro*capro = "x";

title "Grafica de la producción sobre el Capacidad";

run;

Los resultados del análisis son:

The GLM Procedure

Number of observations 50

Dependent Variable: $h_2$opro

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6.56223546	6.56223546	860.83	<.0001
Error	48	0.36591054	0.00762314		
Corrected Total	49	6.92814601			

R-Square	Coeff Var	Root MSE	hopro Mean
0.947185	6.768043	0.087311	1.290042

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Capro	1	6.56223546	6.56223546	860.83	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Capro	1	6.56223546	6.56223546	860.83	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.114228313	0.04193461	2.72	0.0090
Capro	9.916227678	0.33797722	29.34	<.0001

R-Square	Coeff Var	Root MSE	hopro Mean
0.947185	6.768043	0.087311	1.290042

Este modelo con las variables transformadas explica de mejor manera la relación entre ellas el R^2 se incrementa considerablemente y el coeficiente de variación disminuye.

3.13. Ejercicios

3.1 Los siguientes datos son de un estudio del peso seco (Y) de embriones de pollos y la edad de 6 a 16 días (X). También se proporcionan los datos transformados de los pesos a logaritmos naturales (Z).

Edad (X)	6	7	8	9	10	11	12	13	14	15	16
Peso seco (Y)	0.029	0.052	0.079	0.125	0.181	0.261	0.425	0.738	1.130	1.882	2.812
Log (Y)	-1.538	-1.284	-1.102	-0.903	-0.742	-0.583	-0.372	-0.132	0.053	0.275	0.449

a) Construya gráficas con la variable *Y* y otra con la variable *Z*.

b) Dibuje líneas rectas sobre las dos gráficas anteriores y explique cuál de las dos líneas ajusta mejor a los datos.

c) Calcule los estimadores mínimos cuadrados de los parámetros de la línea de regresión para cada diagrama y dibuje las líneas de predicción sobre los datos.

d) Para cada una de las líneas ajustadas calcular un intervalo de confianza del 95% para la pendiente verdadera de ambas líneas e interpretar los intervalos en cada caso, con relación a la hipótesis nula de que la pendiente verdadera es cero.

e) Para cada una de las líneas estimadas calcular y grafique las bandas de confianza y de predicción al 90%.

f) En tu opinión; ¿Cuál de las dos líneas de regresión tiene el mejor ajuste? Explique.

3.2 Un sociólogo de una institución correccional realiza un estudio para determinar la relación entre inteligencia y delincuencia. Un índice de delincuencia en el rango (0 a 50) fue establecido para registrar la severidad y frecuencia de crímenes cometidos y la inteligencia fue medida por el IQ. Los datos muestran el índice de delincuencia (ID) y el IQ de 18 menores convictos.

ID (Y)	26.2	33.0	17.5	25.25	20.3	31.9	21.1	22.7	10.7
IQ (X)	110	89	102	98	110	98	122	119	120
ID (Y)	22.1	18.6	35.5	38.0	30.0	19.7	41.1	39.6	25.15
IQ (X)	92	116	85	73	90	104	82	134	114

a) Construya una gráfica con las dos variables.

b) Dado que $\hat{\beta}_1$ = -0.249 y $\hat{\beta}_0$ = 52.273, definir la línea de regresión estimada y graficarla en el diagrama anterior.

c) ¿Cómo interpretas el hecho cuando el IQ = 0, $\hat{Y}$ = 52.273, cuando el índice de delincuencia no es mayor que 50?

d) Construya un intervalo de confianza del 95% para la verdadera pendiente considerando que $S_{Y/X}$ = 7.704 y S_x = 16.192.

e) Interprete este intervalo de confianza con relación a la prueba de hipótesis nula de que la pendiente es cero con un nivel de $\alpha = 0.05$.

f) Pruebe la hipótesis nula de pendiente cero dado que $S_{Y/X} = 4.933$, $S_X = 14.693$ y $n = 17$.

g) Con estos datos ¿podrías concluir que el índice de delincuencia disminuye cuando el IQ aumenta?

3.3 Un grupo de 14 niños y adolescentes participaron en un estudio para analizar la relación entre la edad y el tiempo promedio de sueño. Para obtener una medida del tiempo promedio de sueño se registraron tres noches consecutivas los tiempos y se obtuvo un promedio. Los resultados fueron:

TS(min/24 hr)	586	461.75	491.1	565	462	532.1	477.6
Edad	4.4	14	10.1	6.7	11.5	9.6	12.4
TS(min/24 hr)	515.2	493	528.3	575.9	532.5	530.5	478.6
Edad	8.9	11.1	7.75	5.5	8.6	7.2	9.7

a) Construya una gráfica de los pares de datos para las dos variables.

b) Calcule los estimadores mínimos cuadrados de la pendiente y el intercepto para la línea recta de regresión y verificar tus resultados con los dados por la computadora.

c) Grafique la línea de regresión sobre el diagrama del punto (a).

d)¿Son violados algunos de los supuestos de la regresión de línea recta?

e) Construya un intervalo de confianza del 95% para la pendiente.

Medias	Desviación estándar
$\overline{Y}$ = 519.30385	40.95056
$\overline{X}$ = 9.05769	2.77518
Corr. Pearson = 0. 9515	Corr. Spearman = 0. 9285
Regresión lineal	
Pendiente = -14.04105	1.3681
Intercepto = 646. 4833	12.91773
Estadístico $F_{(1,11)}$ = 105.33021	

f) ¿Rechazaría la hipótesis nula en base al intervalo de confianza del inciso anterior?

g) Calcule y grafique la banda de confianza del 95% en la gráfica de la línea estimada.

3.4 La prueba de aptitud física es un aspecto importante en un entrenamiento atlético. Es una medida común de la magnitud de la aptitud cardiovascular que indica el volumen máximo de ingestión de oxigeno durante un ejercicio extenuante. Se realizó un estudio con 24 hombres jóvenes para determinar la influencia del tiempo que toma para correr dos millas. La medida de ingestión de oxigeno fue realizada por métodos estándares de laboratorio, obteniéndose los siguientes resultados.

Sujetos	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Vol. De O_2 (y)	42.33	53.10	42.08	50.06	42.45	42.46	47.82	49.92	36.23	49.66	41.49	46.17
Tiempo (x)	918	805	892	962	968	907	770	743	1045	810	927	813
Sujetos	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
Vol. De O_2 (y)	48.18	43.21	51.81	53.28	53.29	47.18	56.91	47.80	48.65	53.69	60.60	56.73
Tiempo (x)	858	860	760	747	743	803	683	844	755	700	748	775

a) Estime los parámetros de una regresión lineal simple: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (i = 0 1 2, ..., 24)

b) Pruebe la hipótesis $H_0: \beta_1 = 0$ *vs* $H_1: \beta_1 \neq 0$. ¿Tiene el tiempo una influencia significativa sobre la ingestión máxima de oxígeno para correr dos millas? Use un $\alpha = 0.01$.

3.5 Se realizó un experimento para determinar la influencia de ciertas medidas físicas sobre el rendimiento de pateadores en futbol americano. 14 pateadores fueron utilizados y se midió la distancia promedio alcanzada en cada tiro. Además se tomaron medidas de la fuerza de las piernas derecha e izquierda (en lb alzadas) vía una prueba de pesos levantados con cada una de las piernas. Los datos son:

Sujetos	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Pierna izquierda	170	130	170	160	150	150	180
Pierna derecha	170	140	180	160	170	150	170
Distancia	162.6	144.0	147.6	163.6	192.0	171.9	162.0
Sujetos	**8**	**9**	**10**	**11**	**12**	**13**	**14**
Pierna izquierda	110	110	120	140	130	150	145
Pierna derecha	110	120	130	120	140	160	152
Distancia	104.1	105.8	117.7	140.3	150.2	165.2	147.6

a) Ajuste una regresión lineal simple con la distancia como la variable respuesta y la pierna derecha e izquierda como las variables regresoras.

b) Ajuste una regresión lineal simple con la fuerza de la pierna izquierda como la variable regresora.

3.6 Se realizó un estudio para determinar la relación entre las horas-hombre para realizar una actividad y una variable regresora que representa una medida del esfuerzo para hacer la actividad (en unidades producidas). Los datos son.

Área	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Unidades	15	25	57	67	197	166	162	131	158	241	399
Horas - hombre	85	125	203	293	763	639	673	499	657	939	1546
Área	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**
Unidades	527	533	563	563	932	986	1021	1643	1985	1640	2143
Horas - hombre	2158	2182	2302	2202	3678	3894	4034	6622	7890	6610	8522

a) Ajuste un modelo de regresión que contenga el intercepto.

b) Calcule el coeficiente de determinación (R^2) del modelo anterior, R^2_0 y $R^2_{(0)*}$ para el modelo con intercepto cero.

c) Calcule S^2 para los modelos con intercepto y con intercepto cero.

d) Calcule los límites de confianza de E (y/x_i) para las 22 áreas. Aplique los cálculos para ambos modelos (con intercepto y con intercepto cero).

e) Use la información de los incisos anteriores para seleccionar el mejor modelo de los dos.

3.7 Discuta los cálculos de un intervalo de confianza del 100 (1 – α)% de la pendiente para el caso del modelo con intercepto cero.3.8 Considere los datos del ejercicio 3.4.

a) Construya intervalos de confianza simultáneos de la media del volumen máximo de O_2 para los siguientes valores de *X* el tiempo para correr dos millas. (i) X = 750 (ii) X = 775 (iii) X = 800 (iv) X = 825 (v) X = 850.
b) Construya intervalos de confianza conjuntos del 95% para la pendiente y el intercepto. Use una gráfica para ilustrar tus resultados.

3.8 Se desarrolló un estudio en el instituto politécnico de la Universidad de Virginia para determinar si cierta medida de resistencia del brazo estático (X) tiene influencia sobre las características de "levantamiento dinámico" (Y) de un individuo. Se usó una muestra de 27 individuos que fueron sometidos a pruebas de resistencia y se les aplicó una prueba de levantamiento de peso. Los datos son,

Sujeto	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
X	17.3	19.3	19.5	19.7	22.9	23.1	26.4	26.8	27.6
Y	71.4	48.3	88.3	75.0	91.7	100	73.3	65.0	75.0
Sujeto	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
X	28.1	28.2	28.7	29.0	29.6	29.9	29.9	30.3	31.3
Y	88.3	68.3	96.7	76.7	78.3	60.0	71.7	85.0	85.0
Sujeto	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**
X	36.0	39.5	40.4	44.3	44.6	50.4	55.9	37.9	43.8
Y	88.3	100	100	100	91.7	100	71.7	96.8	104.7

a) Estime, con mínimos cuadrados la ecuación de regresión lineal.

b) Es considerado que $\beta_0 = 0$ y que $\beta_1 = 2.2$. Pruebe la hipótesis apropiada conjunta. Use una región de confianza conjunta del 95% para β_0 y para β_1 para ilustrar tus conclusiones.

c) Grafique los residuales estudentizados contra X y comentar como se pueden utilizar para aplicar otros análisis a los datos.

3.9 Se tomaron las observaciones de una reacción química sometida a varias temperaturas en un experimento para estudiar su relación. Los datos son,

X(°C)	150	150	150	200	200	200	250	250	250	300	300	300
Y(%)	77.4	76.7	78.2	84.1	84.5	83.7	88.9	89.2	89.7	94.8	94.7	95.9

a) Ajuste una regresión lineal simple, estimando β_0 y β_1.

b) Construya intervalos de confianza del 95% sobre E (y/x) en los cuatro niveles de temperatura en los datos. Graficar los límites superiores e inferiores del intervalo de confianza alrededor de la línea de regresión ajustada.

c) Grafique una región de confianza del 95% en la línea de regresión. Graficarla en la misma figura del inciso (b). Comentar.

3.10 Se realizó un estudio demográfico con parejas casadas con un descendiente para determinar el efecto del ingreso anual del matrimonio sobre el tiempo (en meses) del nacimiento del primer hijo después de casados. Los datos de 20 parejas son los siguientes.

Ingresos	5775	9800	13795	4120	25015	12200	7400	9340	20170	22400
Meses	16.2	35.0	37.2	9.0	24.4	36.75	31.75	30.0	36.0	30.8
Ingresos	4608	24210	19625	18000	13000	5400	6440	9000	18180	15385
Meses	9.7	20.0	38.2	41.25	44.0	9.2	20.0	40.2	332.0	39.2

a) Construya un diagrama de los puntos de tiempo (Y) y los ingresos (X).

b) Intente construir a ojo una línea que ajuste los datos de la mejor manera.

c) Qué te dice esta línea acerca de la relación descrita?

d) Use los siguientes cálculos para determinar los estimadores mínimos cuadrados de la pendiente (β_1) y el intercepto (β_0) para una regresión de línea recta del tiempo sobre los ingresos.

Medias	**Desviación estándar**
$\bar{X}$ = 13193.150	6824.8780
$\bar{Y}$ = 29.04250	11.31785
Correlaciones	
Pearson	0.4304
Spearman	0.42873
Kendall	0.3113
Estimadores	
Pendiente	7.1376×10^{-4}
Intercepto	19.62575
S(Y/X)	10.4958
Prueba de $F_{1,18}$	4.09277

e) Defina la línea de regresión sobre el diagrama del inciso anterior.

f) ¿Algunos de los supuestos de la línea recta no son satisfechos claramente en este ejemplo?

g) Pruebe la hipótesis nula de que la pendiente verdadera $\beta_1 = 0$, al nivel de $\alpha = 0.01$. Interprete los resultados de esta prueba.

h) ¿Puedes sugerir otro modelo de línea recta que describa mejor la relación entre el tiempo y los ingresos?

3.11 Siguiendo las elecciones de Octubre de 2009, un investigador político intenta determinar la relación entre los gastos de campaña sobre los anuncios televisivos y subsecuentes votantes. La siguiente tabla presenta el porcentaje de los gastos totales de campaña relacionados a los anuncios de televisión (X) y el porcentaje de votantes registrados (Y) para una muestra de 20 distritos en Mèxico.

Y	**X**	**Y**	**X**
35.4	28.5	40.8	31.3
58.2	48.3	61.9	50.1
46.1	40.2	36.5	31.3
45.5	34.8	32.7	24.8
64.8	50.1	53.8	42.2
52.0	44.0	24.6	23.0
37.6	27.2	31.2	30.1
48.2	37.8	42.6	36.5
41.8	27.2	49.6	40.2
54.0	46.1	56.6	46.1

En un análisis preliminar se obtuvieron los siguientes resultados.

Medias	Desviación estándar
$\bar{X}$ = 36.990	8.7676
$\bar{Y}$ = 45.710	10.8167
Estimadores	
Pendiente = 1.1770	8.718 x10^{-2}
Intercepto = 2.174	3.309
Prueba de hipótesis de $F_{1,18}$ = 182.2589	
Correlaciones	
Pearson = 0.9540	
Spearman = 0.9551	
Kendall = 0.8587	

a) Construya un diagrama de puntos de Y y X.

b) Use la tabla de los cálculos anteriores para determinar la línea de regresión de Y sobre X.

c) Dibuje la línea estimada sobre el diagrama del inciso (a).

d) ¿Algunos de los supuestos de la línea recta no son satisfechos claramente en este ejemplo?

e) Pruebe la hipótesis de que la pendiente $\beta_1 = 0$ usando un $\alpha = 0.01$ e interpreta tus resultados.

f) Pruebe la hipótesis H_0: $\mu_{Y/X_0} = 45$ para $X_0 = \bar{X} = 36.99$ con un nivel de $\alpha = 0.05$.

g) Calcule un intervalo de confianza del 95% para μ_{Y/X_0}=36.99.

3.12 La tabla presenta el primer salario anual (Y) de un grupo de 30 graduados de licenciatura que recientemente se han integrado al mercado de trabajo con sus calificaciones promedios obtenidas (X) durante sus estudios.

Salario	Calif.	Salario	Calif.	Salario	Calif.
10455	8.58	12500	9.55	13255	9.55
9680	8.31	13310	9.64	13004	9.55
7300	8.47	12105	9.72	8000	8.47
9388	8.52	6200	8.24	8224	8.47
12496	9.22	11522	8.70	10750	8.78
11812	9.37	8000	8.30	11669	8.78
9224	8.43	12548	8.83	12322	8.98
11725	9.08	7700	8.37	11002	8.58
11320	8.78	10028	8.52	10666	8.58
12000	8.98	13176	8.22	10839	8.58

a) Grafique un diagrama de dispersión de los datos para las dos variables.

b) Calcule los estimadores mínimos cuadrados para la pendiente y el intercepto de la línea de regresión de las dos variables. Verificar los resultados con los de la tabla de abajo.

c) Graficar la línea de regresión ajustada sobre el diagrama de dispersión del inciso (a).

d) Construya un intervalo de confianza del 95% para la pendiente.

e) ¿Rechazaría la hipótesis nula H_0: $\beta_1 = 4000$ con un nivel de $\alpha = 0.01$?

Medias	Desviación estándar
$\bar{X}$ = 2.83833	0.4484
$\bar{Y}$ = 10740.6667	1967.6524
Estimadores	
Pendiente = 3630.56	465.7687
Intercepto = 435.923	1337.857
Prueba de hipótesis de $F_{1,28}$ = 60.7585	
Correlaciones	
Pearson = 0.8264	
Spearman = 0.933515	
Kendall = 0.79921	

f) Calcule una banda de confianza y de predicción del 95%.

g) Grafique las respectivas bandas de confianza y de predicción del inciso anterior.

h) ¿Rechazaría la hipótesis H_0: μ_{Y/X_0} = 11500 cuando X_0 = 2.75? Use un α = 0.05.

3.13 Se realizó un estudio para determinar la curva de respuesta de dosis de la vitamina K, en ratas individuales que fueron reducidas en sus reservas de esta vitamina y alimentadas durante cuatro días con diferentes dosis de vitamina K. La respuesta de cada rata fue medida como la concentración de un agente coagulante necesario para coagular una muestra de su sangre en tres minutos. Los resultados del experimento con 12 ratas se dan en la tabla y se expresan como el logaritmo para ambas variables.

Rata	**Log_{10} (Y)**	**Log_{10}(X)**	**Rata**	**Log_{10}(Y)**	**Log_{10}(X)**
1	2.65	0.18	7	1.55	0.83
2	2.25	0.33	8	1.32	0.92
3	2.26	0.42	9	1.13	1.01
4	1.95	0.54	10	1.07	1.04
5	1.72	0.65	11	0.95	1.09
6	1.60	0.75	12	0.88	1.15

a) Construya un diagrama de dispersión de los datos.

b) Determine los estimadores mínimos cuadrados para la pendiente y el intercepto para la línea recta de regresión de Y sobre X.

c) Grafique la línea ajustada sobre el diagrama de dispersión del inciso (a).

d) Determine y bosqueje la banda de confianza del 99% basada en la línea de regresión ajustada.

e) Convierta la línea ajustada en una ecuación con las unidades originales de $Y' = 10^{Y}$ y $X' = 10^{X}$

Medias	**Desviación estándar**
$\bar{X}$ = 0.7425	0.3200
$\bar{Y}$ = 1.6108	0.5736
Estimadores	
Pendiente = -1.7850	0.05267
Intercepto = 2.9362	0.04230
S(Y/X) = 0.05589	
Prueba de hipótesis de $F_{1,20}$ = 1148.762	
Correlaciones	
Pearson = -0.9968	

f) Para la ecuación del inciso (e) determinar intervalos de confianza del 99% para la respuesta media verdadera de las dosis máximas y mínimas usadas en el experimento ($Y_{máxima}$ = 0.8834, $Y_{mínima}$ = 2.6149).

g) Si cada uno de los valores de X y Y fueran convertidos en sus unidades originales, el modelo ajustado a estos datos es: $\hat{Y}$ = 237.16095 – 21.32117X'. ¿Cómo evaluarías cuál es el mejor modelo, si se usan los datos en las unidades originales o los datos transformados?

3.14 Se condujo un experimento para determinar cómo sería afectado el impacto en la tasa de crecimiento de cierto hongo por tubos de pruebas que contienen el mismo medio, la misma temperatura con diferentes gases inertes. Se condujeron tres experimentos en cada uno de los 6 gases y su tasa de

crecimiento promedio sobre estas tres pruebas fue usada como las respuestas. La tabla presenta el peso molecular (X) de cada gas usado y la tasa de crecimiento promedio (Y) en mililitros por hora para las tres pruebas.

Gas	Y	X
A	3.85	4.0
B	3.48	20.2
C	3.27	28.2
D	3.08	39.9
E	2.56	83.8
F	2.21	131.3

a) Calcule los estimadores mínimos cuadrados de la pendiente, el intercepto para la ecuación lineal de Y sobre X y dibujar la línea ajustada en un diagrama de dispersión de los datos.

b) Pruebe la significancia de la pendiente de la línea ajustada.

c) ¿Qué información no se ha usado que puede mejorar la sensibilidad del análisis?

d) ¿Cuál es el intervalo de confianza del 90% para el verdadero promedio de la tasa de crecimiento cuando el gas usado tiene un peso molecular de 100?

e) ¿Por qué sería inapropiado usar la línea ajustada para estimar la tasa de crecimiento con un gas que tenga un peso molecular de 200?

f) Basado en la selección de los valores de X usados en el estudio; ¿cómo criticaría lo adecuado de la predicción obtenida en este experimento usando la línea recta ajustada?

Capítulo 4
El coeficiente de correlación y el análisis de regresión lineal

4.1. Introducción

En el análisis de datos por lo general se encuentran variables que están relacionadas entre sí y dicha relación es de interés para el investigador por lo que se emplea el análisis de correlación para determinar la magnitud y la dirección de dicha relación. Es importante entender que la relación puede mostrar una tendencia positiva y/o negativa, por lo que el coeficiente toma valores entre $-1 < r < 1$. Aunado a esta técnica se puede calcular la magnitud de la relación de las dos variables una llamada dependiente y la otra independiente a través del análisis de regresión lineal el cual calcula la pendiente de la relación como una tasa de cambio de una variable dependiente al cambiar en una unidad la variable independiente.

4.2. Definición de *r*

El coeficiente de correlación es un estadístico muy usado, que no solamente provee una medida de cómo dos variables aleatorias están asociadas en una

muestra, sino que tiene propiedades que se relacionan con el análisis de regresión de la línea recta. Definimos el coeficiente de correlación muestral r para dos variables *X* y *Y* por la expresión,

$$r=\frac{\sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)/n}{\sqrt{\left[\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_1)^2}{n}\right]\left[\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}\right]}} \quad (1)$$

Una expresión equivalente para *r* que ilustra su relación matemática con los estimadores mínimos cuadrados de la pendiente de una línea de regresión ajustada es,

$$r = \frac{S_x}{S_y}\hat{\beta}_1 \quad (2)$$

Tres importantes propiedades matemáticas de *r* son:

1. Los posibles valores de *r* están en el rango de -1 a 1
2. El coeficiente de correlación *r* es una dimensión cuantitativa, es independiente de las unidades de medidas de *X* y de *Y*.
3. Los valores de *r* pueden ser, positivos, negativos o cero, si $\hat{\beta}_1$, es positivo, negativo o cero y viceversa.

4.3. *r* como una medida de asociación

En los supuestos estadísticos para el análisis de regresión lineal, se consideró anteriormente que la variable *X* no es aleatoria. No obstante, por lo general se considera un problema cuando las dos variables *X* y *Y* son variables aleatorias. La medida de *r* se puede interpretar como un índice de asociación entre *X* y *Y* en el siguiente sentido:

1. Lo más positivo de *r* es la asociación más positiva. Esto significa que cuando *r* está cercano a 1, un individuo con un valor alto para una variable, tendrá un valor alto para la otra y un individuo con un valor bajo para una variable, tendrá un valor bajo para la otra (Figura 4.1.a).
2. Lo más negativo de *r* es la asociación más negativa; un individuo con un valor alto de una variable, tendrá un valor bajo para la otra, cuando *r* está cercano a -1 e inversamente (Figura 4.1b).
3. Si *r* está cercano a cero, existe poca o baja asociación lineal entre *X* y *Y* (Figura 4.1c).

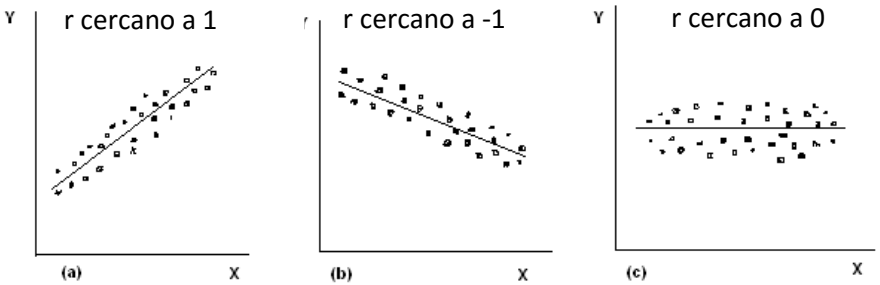


Figura 4.1. Coeficientes de correlación como una medida de asociación.

Puesto que *r* es un índice obtenido de una muestra de n observaciones, puede ser considerado como un estimador de un parámetro poblacional desconocido. Este parámetro desconocido se denomina coeficiente de correlación poblacional y por lo general se denota por ρ_{XY} o simplemente por ρ y es claramente entendido que dos variables están involucradas. Por lo que aquí solamente usaremos ρ para evitar confusión.

> *"El parámetro ρ_{XY} se define como; $\rho_{XY} = \sigma_{xy}/\sigma_X\sigma_Y$, donde σ_X y σ_Y denotan las desviaciones estándares poblacionales de las variables aleatorias X y Y y donde σ_{XY} se denomina covarianza de X y Y. La covarianza σ_{XY} es un parámetro poblacional que describe la cantidad promedio que dos variables "covarían".*

4.4. *r* y la fuerza de la relación con la línea recta

Para iniciar y definir qué estamos pensando por la *fuerza* de la relación lineal entre *X* y *Y*, primero consideramos que el predictor de *Y* será *X*. El mejor predictor en este caso simplemente será $\bar{Y}$ la media muestral de la variable *Y*. La suma de cuadrados de las desviaciones asociadas con el predictor $\overline{Y}$ está dado por la expresión,

$$SC_Y = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad (3)$$

Ahora si la variable *X*, toma cualquier valor para predecir la variable *Y*, la suma de cuadrados de residuales dado por,

$$SC_{error} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (4)$$

Será considerablemente menor que SC_Y. Tal que el modelo de mínimos cuadrados, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ ajusta mejor los datos que la línea horizontal $\hat{Y} = \bar{Y}$ (Figura 4.2). Una medida cuantitativa del mejoramiento del ajuste obtenido usando *X* es dado por el cuadrado del coeficiente de correlación muestral r, que es obtenido por,

$$R^2 = \frac{SC_y - SC_E}{SC_y} \quad (5)$$

La cantidad naturalmente varía entre cero y uno ya que *r* varía entre -1 y 1.

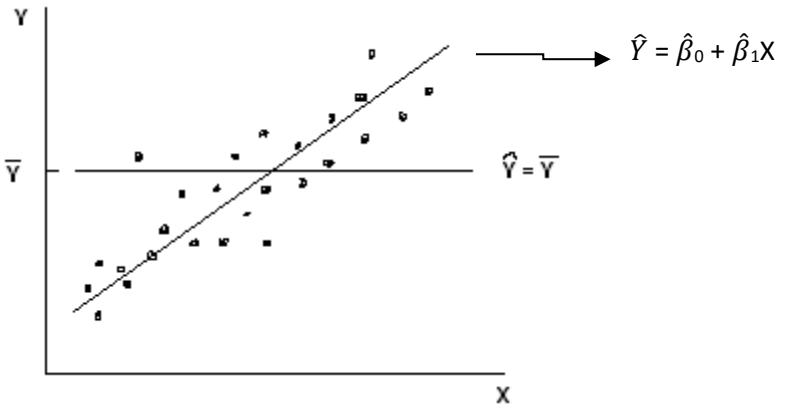


Figura 4.2. Predicción de Y usando y no usando X.

El valor más grande que R^2 puede tomar es 1, que ocurre cuando $\hat{\beta}_1$, no es cero y cuando la $SC_{error} = 0$ (cuando existe una línea perfecta positiva o negativa entre la relación de *X* y *Y*). Por *perfecta*, entendemos que todos los puntos caen sobre la línea recta ajustada. En otras palabras, cuando $Y_i = \hat{Y}_i$ para toda i, entonces tenemos (Figura 4.3).

$$SC_{error} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Tal que

$$R^2 = \frac{SCy - SCerror}{SC_y} = \frac{SC_y}{SC_y} = 1 \qquad (6)$$

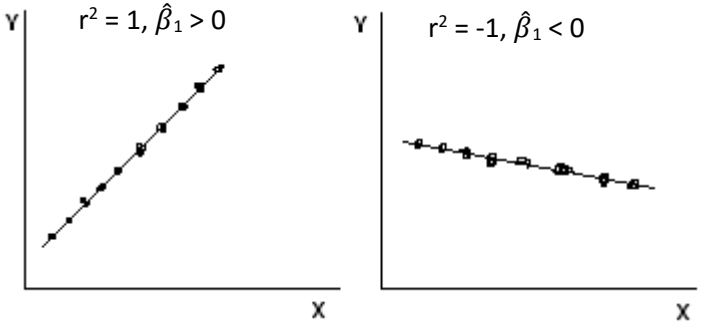


Figura 4.3. Ejemplos de asociaciones lineales perfectas.

El valor más pequeño que R^2 puede tomar es cero. Este valor indica que no existe un poder predictivo apropiado al usar *X;* esto es, $SC_{error} = SC_Y$. Además, recordando en regresión lineal, vemos que un coeficiente de correlación de cero, implica una pendiente de cero y consecuentemente, la ausencia de cualquier relación lineal.

4.5. Lo que no mide *r*

Hay dos malas concepciones comunes de *r* (o equivalentemente, acerca de R^2) que ocasionalmente permiten al investigador hacer interpretaciones espurias acerca de la relación entre *X* y *Y*:

A).- R^2 no es una medida de la magnitud de la pendiente de la línea de regresión. Esto es, si el valor de R^2 es alto, no necesariamente significa que

la magnitud de la pendiente $\hat{\beta}_1$ es grande. Esto se ilustra en la Figura 4.3. Note que R^2 es igual a 1 en ambas partes, no obstante el hecho de que las pendientes son diferentes. Otra manera de mirar esto viene del hecho de que,

$$\hat{\beta}_1^2 = \frac{S_Y^2}{S_X^2}, \qquad \text{cuando } R^2 = 1 \tag{7}$$

Si dos conjuntos de datos diferentes tienen la misma cantidad de variación *X*, pero el primer grupo tiene menos variación de *Y* que el segundo grupo, la magnitud de la pendiente para el primer grupo será más pequeño que para el segundo.

B).- R^2 no es una medida de lo apropiado del modelo de línea recta. Note que $R^2 = 0$ en los incisos a y b de la Figura 4.4. No existe evidencia de ninguna asociación entre *X* y *Y* en (a) y fuerte evidencia de una asociación no lineal en (b). También note que R^2 es alta en los incisos (c) y (d) solo en (c) el modelo de línea recta es apropiado, pero en (d) es completamente inapropiado.

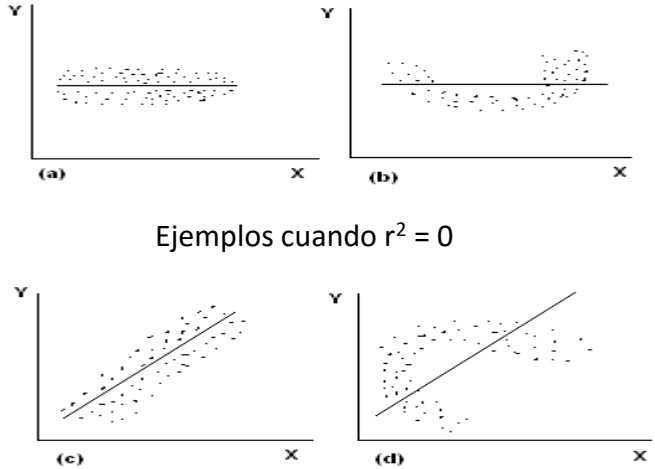


Figura 4.4. Ejemplos mostrando valores de R^2 altos.

4.6. Prueba de hipótesis e intervalos de confianza para el coeficiente de correlación

Investigadores interesados en la asociación entre dos variables X y Y a menudo, desean una prueba de la hipótesis nula H_0: $\rho = 0$.

4.6.1. Prueba de H_0: $\rho = 0$

Una prueba de H_0: $\rho = 0$ es equivalente matemáticamente a probar la hipótesis H_0: $\beta_1 = 0$ para la pendiente de la línea recta. Esta equivalencia es sugerida por la expresión $\beta_1 = \rho\frac{\sigma_y}{\sigma_x}$ y $\hat{\beta}_1 = \frac{rS_y}{S_x}$, que indica que β_1 es positivo, negativo o cero, así como ρ, es positivo, negativo o cero por lo que existe una relación

análoga entre $\hat{\beta}_1$ y r. El estadístico de prueba para la hipótesis H_0: $\rho = 0$, puede ser escrito en términos de r y n, ya que se puede aplicar la prueba sin tener la línea ajustada. El estadístico de prueba es dado por la expresión, $T = (r\sqrt{n-2})/\sqrt{1-r^2}$ el cual tiene una distribución t-student con n – 2 grados de libertad, cuando la hipótesis nula H_0: $\rho = 0$ es verdadera. También se puede utilizar la siguiente expresión. $T = \hat{\beta}_1 S_x \sqrt{n - \frac{1}{S_{y/x}}}$, cualquiera de las dos expresiones da el mismo resultado.

4.6.2. Intervalos de confianza para ρ

Un intervalo de confianza del 100(1 - α) % para ρ, puede ser obtenido utilizando la transformación Z de Fisher. Primero, se calcula un intervalo de confianza de 100(1 – α) % para el parámetro $\frac{1}{2}\ln[(1+\rho)/(1-\rho)]$ usando la siguiente expresión, $\frac{1}{2}\ln\frac{1+r}{1-r} \pm Z_{1-\alpha/2}/\sqrt{n-3}$, donde $Z_{1-\alpha/2}$ es asignado por el investigador.

Por lo anterior denotamos por L_z el límite inferior del intervalo de confianza y por U_z, el límite superior. Entonces para determinar los valores de L_z y U_z.

$$L_z = \frac{1}{2}\ln(1+L_\rho)/(1-L_\rho) \quad y \quad U_z = \frac{1}{2}\ln(1+U_\rho)/(1-U_\rho) \qquad (8)$$

El intervalo de confianza del 100 (1 – α) % es de la forma,

$$L_\rho < \rho < U_\rho$$

Utilizando los datos de la presión y la edad de las personas (Ejemplo 1 del Capítulo 1) tenemos, que $r = 0.66$ y $n = 30$, entonces,

$$\frac{1}{2} \ln (1 + 0.66) /(1 - 0.66) \pm 1.96 / \sqrt{30 - 3}$$

$$0.793 \pm 0.377$$

Entonces el límite inferior $L_z = 0.416$ y el límite superior de $U_z = 1.170$, esto es,

$$0.416 < \frac{1}{2} \ln (1 + \rho)/ (1 - \rho) < 1.170$$

Para transformar esto a un intervalo de confianza para ρ, determinamos los valores de L_ρ y U_ρ, que satisfacen,

$$0.416 = \frac{1}{2} \ln (1 + L_\rho) / (1 - L_\rho)$$

$$1.170 = \frac{1}{2} \ln (1 + U_\rho) / (1 - U_\rho)$$

Usando la Tabla G del anexo, vemos que el valor de 0. 416 corresponde a un r cercano a 0.394, tal que $L_\rho = 0.394$. De igual manera vemos que el valor de

1.170 corresponde a un *r* cercano a 0.824, tal que $U_\rho = 0.824$. Por lo que el intervalo de confianza del 95% para ρ está dado por.

$$0.394 < \rho < 0.824$$

En donde se considera se encuentra el valor verdadero del parámetro ρ.

4.7. Prueba de hipótesis para dos Correlaciones

Se tienen dos muestras aleatorias independientes de tamaños n_1 y n_2 que son seleccionadas de dos poblaciones. Además se puede suponer que es de interés probar H_0: $\rho_1 = \rho_2$ *versus* H_1: $\rho_1 \neq \rho_2$. Una prueba estadística apropiada puede ser desarrollada basada en la sección anterior. También se considera la situación donde las correlaciones muestrales que son comparadas son calculadas utilizando el mismo grupo de datos, en cada caso, estas correlaciones muestrales están correlacionadas entre ellas mismas. En cada población se cumplen los supuestos del análisis de regresión, incluyendo el de normalidad.

Una prueba aproximada de H_0: $\rho_1 = \rho_2$, se puede basar en las transformaciones de Z de Fisher. Sea r_1 la correlación calculada usando n_1 observaciones de la primera población y sea r_2 la correlación muestral de la segunda población. Entonces sea

$$Z_1 = \frac{1}{2} \ln (1 + r_1) / (1 - r_1) \quad (9)$$

$$Z_2 = \frac{1}{2} \ln (1 + r_2) / (1 - r_2) \quad (10)$$

Para probar H_0: $\rho_1 = \rho_2$, podemos calcular el estadístico de prueba

$$Z = (Z_1 - Z_2) / \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \quad (11)$$

Para tamaños de muestras n_1 y n_2 grandes, este estadístico de prueba tiene aproximadamente una distribución normal estándar, cuando H_0 es verdadera. Las siguientes regiones críticas para la significancia del nivel α, son usadas.

$Z \geq z_{1-\alpha}$ (prueba de la cola superior para la alternativa H_1: $\rho_1 > \rho_2$)
$Z \leq -z_{1-\alpha}$ (prueba de la cola inferior para la alternativa H_0: $\rho_1 < \rho_2$)
$|Z| \geq z_{1-\alpha/2}$ (prueba de dos colas para la alternativa H_1: $\rho_1 \neq \rho_2$)

Un ejemplo para aplicar los procedimientos anteriores.

Deseamos probar la hipótesis H_0: $\rho_1 = \rho_2$ *versus* H_1: $\rho_1 \neq \rho_2$

Los datos de las dos poblaciones son:

$r_1 = 0.\ 22$, $n_1 = 30$ y $r_2 = 0.342$ y $n_2 = 30$.

Entonces.

$$Z_1 = \frac{1}{2} \ln (1 + 0.22) / (1 - 0.22) = 0.\ 2337$$

$$Z_2 = \frac{1}{2} \ln (1 + 0.342) / (1 - 0.342) = 0.\ 3564$$

Por lo anterior el estadístico de prueba es,

$$Z = (0.2237 - 0.3564) / \sqrt{\frac{1}{30-3} + \frac{1}{30-3}} = -\ 0.488$$

Para un $\alpha = 0.01$, tenemos que el $Z_{0.005} = 2.576$

Por lo que no se rechaza la hipótesis nula, las dos correlaciones son aproximadamente iguales.

4.8. Ejercicios

4.1 Examine los cinco pares de datos en la siguiente tabla.

i	**1**	**2**	**3**	**4**	**5**
X_i	-2	-1	0	1	2
Y_i	4	1	0	1	4

a) ¿Cuál es la relación matemática entre X y Y?

b) Demuestre que para la línea recta de regresión de Y sobre X, $\hat{\beta}_1 = 0$.

c) Demuestre que r = 0.

d) ¿Porqué no existe relación aparentemente entre X y Y, como se demostró con los estimadores de β_1 y ρ?

4.2 Considere los datos de la tabla.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
X_i	1	1	1	2	2	2	3	3	3	20
Y_i	1	2	3	1	2	3	1	2	3	20

a) Calcule el coeficiente de correlación muestral, r.

b) Use el estadístico T' = r $\sqrt{r-2}$ / $\sqrt{1-r^2}$, para probar H_0: $\rho = 0$ *versus* H_1: $\rho \neq 0$.

c) No obstante, las conclusiones obtenidas en el inciso (b); ¿porqué es incómodo que concluyas que las dos variables están linealmente relacionadas?

4.3 Suponga que en un estudio de la variación geográfica de una cierta especie de leopardo, se tomó la media de la longitud de la tibia (U) y la longitud media del tarso (V) en una muestra de 50 ejemplares de 10 diferentes regiones del sureste mexicano, obteniéndose los siguientes resultados.

i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
U_i	7.500	7.164	7.512	8.544	7.380	7.860	7.836	8.100	7.584	7.344
V_iV	1.680	1.596	1.680	1.908	1.632	1.752	1.776	1.860	1.692	1.680

a) Calcule el coeficiente de correlación entre la longitud del tarso y la tibia.
b) Construya un intervalo de confianza del 95% para ρ.

4.4 En una muestra de 23 adultos hombres la correlación entre la medida total de hemoglobina medida de una punción en uno de los dedos de la mano, fue de 0.82. Para una muestra de 32 mujeres de edades similares la correlación fue de 0.74. Las dos muestras de cada persona fueron recolectadas con una hora de diferencia. Asumir que los supuestos de regresión lineal se cumplen.

a) Pruebe la hipótesis de que las dos correlaciones poblacionales son iguales. Use una prueba de dos colas. ¿Qué puedes concluir?
b) Repita el inciso (a) con una prueba de dos colas para probar si la correlación de las mujeres es más baja que la de los hombres. ¿Qué concluye?
c) Asuma que el investigador planeó conducir una prueba de una cola de la hipótesis de que la correlación para las mujeres es más alta que la de los hombres. ¿Qué prueba haría? ¿Qué puedes concluir?

4.5 El secretario académico de una universidad administró una prueba a todos los estudiantes de nuevo ingreso. Una nueva versión de la prueba es aplicada por la empresa evaluadora. Para aplicar la nueva prueba el secretario académico seleccionó 121 estudiantes de nuevo ingreso y aplica las dos pruebas. Al final del primer año escolar el secretario correlaciona los dos puntajes y con el promedio obtenido en el primer periodo, denotando con 1 la versión vieja, con 2 la nueva versión y con G el promedio del primer periodo escolar. Esto es,

$$r_{12} = 0.\ 6969 \qquad r_{1G} = 0.\ 5514 \qquad r_{2G} = 0.\ 4188$$

Pruebe la hipótesis de que las dos versiones de la prueba están igualmente correlacionadas con el promedio final del primer año. Use una prueba de dos colas con un $\alpha = 0.01$. Asumir que se cumplen todos los supuestos de la regresión lineal.

4.6 Calcule e interprete el coeficiente de correlación para las calificaciones de 8 estudiantes seleccionados al azar de un grupo de ingenieros agrónomos.

Calificación de matemáticas	92	72	80	74	65	83	90	88
Calificación. de Ecología	90	74	63	87	78	74	82	68

4.7 Pruebe la hipótesis de que $\rho = 0$ para el ejercicio anterior. Utilice un $\alpha = 0.05$.

4.8 Los siguientes datos se obtienen de un estudio de la relación entre el peso y tamaño del tórax de recién nacidos.

Peso(kg)	2.75	2.15	4.41	5.52	3.21	4.32	2.31	4.30	3.71
Tamaño de tórax	29.5	26.3	32.2	36.5	27.2	27.7	28.3	30.3	28.7

a) Calcule r.

b) Pruebe la hipótesis nula de que $\rho = 0$ con un $\alpha = 0.01$.

c) ¿Qué porcentaje de la variación en los tamaños del tórax de los recién nacidos se explica por la diferencia de pesos?

4.9. Las cantidades de sólidos eliminados de un material químico, cuando se expone a periodos diferentes de secado son los siguientes.

Horas(x)	4.4	4.5	4.8	5.5	5.7	5.9	6.3	6.9	7.5	7.8
Gr(y)	13.1	9.0	10.4	13.8	12.7	9.9	13.8	16.4	17.6	18.3

a) Calcule el coeficiente de correlación muestral.

b) Pruebe la hipótesis de que $\rho = 0.5$ contra la alternativa de que $\rho > 0.5$. Utilice el valor de p para la conclusión.

4.10 El departamento de medicina veterinaria de la universidad de Virginia analizó datos de marmotas normales. Las variables de interés fueron el peso del cuerpo en gr y el peso del corazón (en gr). Se desea conocer la relación entre las dos variables. Los datos son:

P. del cuerpo	4050	2465	3120	5700	2595	3640	2050	4235	2935	4975	3690	2800	2775
P. del corazón	11.2	12.4	10.5	13.2	9.8	11.0	10.8	10.4	12.2	11.2	10.8	14.2	12.2

a) Calcule e interprete el coeficiente de correlación muestral.

b) Pruebe la hipótesis de que $\rho = 0.75$ contra la alternativa de que $\rho > 0.75$. Use un valor de $\alpha = 0.01$.

CAPÍTULO 5
El modelo de regresión múltiple

5.1. Introducción

El análisis de regresión múltiple puede ser observado como una extensión del análisis de regresión de línea recta, para las situaciones en dónde se involucran más de una variable independiente. Aquí se describe el método de regresión múltiple, se establecen los supuestos requeridos y se describen los procedimientos para la estimación de los parámetros importantes, explicando cómo hacer e interpretar la inferencia de estos parámetros desarrollando ejemplos que ilustran el uso de la técnica del análisis de regresión múltiple. Sin embargo, antes de proceder, es importante mencionar que el modo de trabajar con varias variables independientes simultáneamente en un análisis de regresión es considerablemente más difícil que procediendo con una sola variable independiente, por las siguientes razones.

1. Es más difícil seleccionar el mejor modelo, ya que algunas veces hay varios candidatos razonables,
2. Es más difícil visualizar lo que se observa en el modelo ajustado (si hay más de dos variables independientes), puesto que no es

posible graficarlo directamente en más de tres dimensiones cada uno de los datos o el modelo ajustado.

3. En algunas veces es más difícil interpretar el significado del mejor modelo ajustado en términos de veracidad.
4. Los cálculos son virtualmente imposibles sin acceso a una computadora de alta capacidad de procesamiento de datos y un software confiable.

5.2. Modelos de Regresión Múltiple

Un ejemplo de un modelo de regresión múltiple es dado por cualquier polinomio de segundo orden o de mayor orden. La adición de términos de alto orden al modelo puede ser considerado como equivalente a la adición de nuevas variables independientes. Entonces si denotamos a X como X_1 y a X^2 como X_2, el modelo de segundo orden es

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \quad (1)$$

Puede ser reescrito como

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad (2)$$

De esta forma, en regresión polinomial tenemos una sola variable independiente base, los otros son simples funciones matemáticas de esta variable básica. En muchos problemas de regresión múltiple, el número de variables independientes básicas puede ser mayor a una. La forma general de un modelo de regresión para k variables independientes es dado por

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon \quad (3)$$

Donde β_0, β_1, β_2, β_3, ... β_k, son los coeficientes de regresión que deben ser estimados. Las variables independientes X_1, X_2, X_3, ..., X_k, pueden ser todas variables básicas, o algunas pueden ser funciones de unas pocas variables básicas y ε, el término de error.

5.3. Descripción del Modelo y supuestos

Considerar un experimento en que los datos generados son del tipo:

y	x_1	x_2	...	x_k
y_1	x_{11}	x_{21}	...	x_{k1}
y_2	x_{12}	x_{22}	...	x_{k2}
...	...	...	...	...
y_n	x_{1n}	x_{2n}	...	x_{kn}

Cada hilera de este arreglo representa un dato-punto. Si el investigador está dispuesto a asumir que la región de las x está definida por los datos, y_i está

relacionada aproximadamente de forma lineal con las variables regresoras, entonces la formulación del modelo está dado por,

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_k x_{ki} + \varepsilon_i \quad (4)$$

Como en el caso de una regresión simple, ε_i es un error del modelo, que se asume que no está correlacionado de observación a observación, con media cero y varianza constante σ^2. Además, las variables x no son aleatorias y son medidas con un error muy pequeño (despreciable). El modelo de regresión lineal múltiple representa una desviación del modelo de regresión lineal simple discutido anteriormente. No podemos mostrar los datos fácilmente en una gráfica. El significado de los β_j en el modelo anterior (4) requieren de mayor discusión y una extensión gráfica del procedimiento de mínimos cuadrados es más complicado y requiere más desarrollo. Existe por lo general confusión en algunos científicos que realizan modelación estadística sobre el significado del término lineal.

5.4. ¿Qué es un modelo lineal?

El investigador necesita entender las diferencias entre un modelo lineal y un modelo que no es lineal.

Un modelo lineal se define como un modelo que es lineal en los parámetros; esto es, lineal en los coeficientes, los β's, en el modelo anterior.

Aquí trataremos con modelos lineales. Consideramos la situación en la cual el investigador desea construir un modelo de la relación entre *Y* y *X* pero la relación es por ejemplo polinomial. Este es un ejemplo de un modelo lineal. Por ejemplo, un modelo cuadrático en X (pero lineal en las β's) está dado por,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \quad (5)$$

Otro ejemplo de un modelo lineal es el que contiene interacciones entre un par de variables regresoras. Esto es,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon \quad (6)$$

En algunas aplicaciones existe la necesidad de aplicar transformaciones a las variables regresoras. Por ejemplo, considerar el caso de tres variables regresoras X_1, X_2, X_3. El siguiente es un modelo lineal en el que se hace la transformación logarítmica en cada una de las variables.

$$Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \beta_3 \ln X_3 + \varepsilon \quad (7)$$

En cada uno de los tres ejemplos se puede hacer transformaciones en la variable regresora pero el modelo continúa siendo lineal en los parámetros. La variable respuesta *Y*, también es susceptible de transformaciones. Por ejemplo, el investigador puede estar interesado en usar una transformación logarítmica para *Y*, así como una transformación del recíproco de X_1 y X_2. Como resultado se tiene el modelo lineal escrito como,

$$ln\ y = \beta_0 + \beta_1(\frac{1}{X_1}) + \beta_2(\frac{1}{X_2}) + \varepsilon \quad (8)$$

Todos los modelos descritos aquí, son modelos lineales por definición. Es importante notar que el analista de datos ajusta modelos a datos transformados. Además esperamos dejar claro que el procedimiento de regresión lineal múltiple, se aplica donde las variables involucradas no están en su forma natural.

En este punto el lector, puede investigar, qué clase de ejemplos se pueden considerar en la categoría de modelos no lineales. Dado una respuesta *Y* con dos variables regresoras, X_1, X_2, se pueden construir ejemplos de modelos no lineales, como,

$$Y = \beta_0 + \beta_1 X_1^{Y_1} + \beta_2 X_2^{Y_2} + \varepsilon \quad (9)$$

$$Y = \frac{\beta_0}{1+e^{-(\beta_1 X_1 + \beta_2 X_2)}} + \varepsilon \quad (10)$$

El primer modelo (9) contiene cinco parámetros y el segundo (10) contiene tres parámetros. En ambos casos, los parámetros están en el modelo de una manera no lineal.

5.5. Interpretación del modelo y sus parámetros

Una tarea importante del análisis de regresión es el estimar los β's del modelo. Los β's son por lo general los *coeficientes de regresión parcial*. Con esto queremos decir que un modelo de regresión múltiple con parámetros, β_1, β_2, y β_3 se puede interpretar que, β_1 es el cambio esperado en la respuesta (positiva o negativa) por unidad de cambio en X_1, manteniendo constantes las otras X's. Interpretaciones similares son dadas a los otros β's. Es completamente simple tratar con la interpretación del modelo de regresión lineal simple. En el caso de la regresión lineal múltiple, la interpretación del modelo es una extensión interesante.

Considere la Figura 5.1, que muestra datos de una regresión verdadera, en un plano de regresión. Los ε's son distancias verticales al plano y los puntos en el plano representan la respuesta esperada. Los coeficientes de regresión β_1 y β_2 son las pendientes del plano de regresión en la dirección de X_1 y X_2 respectivamente. El intercepto β_0, se muestra en la Figura 5.1. Obviamente el propósito del análisis de regresión es estimar los tres parámetros y por lo tanto estimar el *plano de regresión poblacional*.

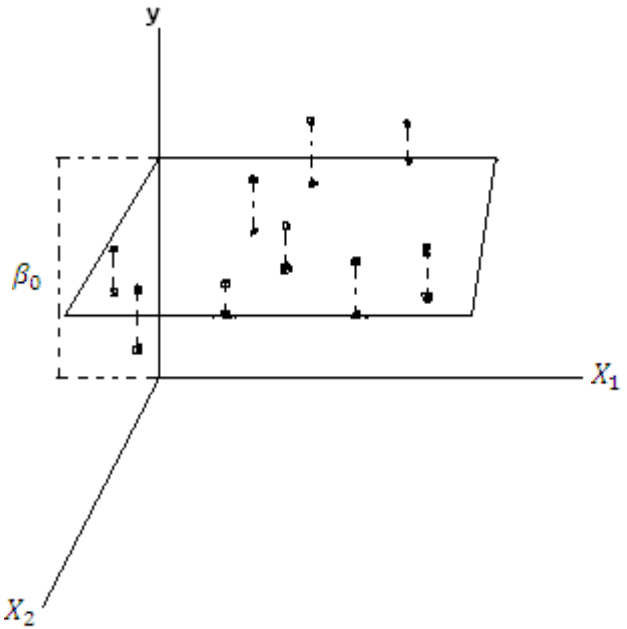


Figura 5.1. Regresión lineal múltiple con dos variables independientes.

5.6 El modelo lineal general y el procedimiento de mínimos cuadrados

Una vez más consideramos el procedimiento de mínimos cuadrados para la estimación de los parámetros del modelo. Es conveniente introducir el modelo general en su notación matricial. El modelo puede ser expresado como,

$$y = X\beta + \varepsilon \qquad (11)$$

donde

$$y = [\, y_1, y_2, y_3, ..., y_n]^T \quad \beta = [\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_k]^T \quad \varepsilon = [\varepsilon_1, \varepsilon_2, \varepsilon_3, ..., \varepsilon_n)^T$$

$$X = \begin{pmatrix} 1 & x_{11} & x_{21} \ldots, & x_{k1} \\ 1 & x_{12} & x_{22} \ldots, & x_{k2} \\ 1 & x_{13} & x_{23} ..., & x_{k3} \\ 1 & x_{1n} & x_{2n} \ldots, & x_{kn} \end{pmatrix}$$

La matriz *X* tiene una dimensión de n x p, donde p = k +1 que denota el número total de parámetros en el modelo. La *j-ésima* columna de *X* contiene lecturas de la variable regresora X_j medida con un error despreciable. Esta matriz es la matriz de los datos, que algunas veces suele llamarse *matriz diseño o matriz del modelo.* La matriz *X* muestra los datos de las regresoras y en los casos del diseño experimental se considera que los niveles de las variables regresoras son planeados. Sin embargo *X* es una función del modelo. Para el caso de lectores que no están familiarizados con el uso de la notación matricial en regresión es recomendable una pequeña capacitación al respecto.

5.6.1. Desarrollo del procedimiento de mínimos cuadrados

Los estimadores $\hat{\beta}$ para los coeficientes de regresión en β, es el vector que satisface,

$$\frac{\partial L}{\partial \beta} = [(y - X\beta)^T (y - X\beta)] = 0 \quad (12)$$

Tenemos el modelo general $Y = X\beta + \varepsilon$, que de acuerdo al procedimiento de mínimos cuadrados se despeja el vector de los errores se suman y se eleva al cuadrado, que en notación matricial se denota por: $L = \varepsilon^T \varepsilon$

$$L = [(y - X\beta)^T (y - X\beta)] \quad (13)$$

$$\frac{\partial L}{\partial \beta} = [(y - X\beta)^T (y - X\beta)] = 0$$

Aquí, la expresión $(y - X\beta)^T (y - X\beta)$, representa la suma de cuadrados de residual. Donde el superíndice T representa a la matriz traspuesta en la suma de cuadrados. Obteniéndose la derivada parcial anterior, tenemos,

$$(-2X^T y + 2X^T X\hat{\beta}) = 0$$

las ecuaciones mínimos cuadrados normales son

$$(X^T X)\hat{\beta} = X^T y$$

asumiendo que X es de rango completo se tiene la solución al modelo

$$\hat{\beta} = (X^T X)^{-1} X^T y \qquad (14)$$

La matriz $(X^T X)$ de la ecuación anterior es simétrica (k + 1) x (k + 1) cuyos elementos de la diagonal son las sumas de cuadrados de los elementos en la columna de la matriz *X* y cuyos elementos de la diagonal superior son las sumas de productos cruzados de elementos en la misma columna. La naturaleza de $(X^T X)$ juega un papel importante en las propiedades de los estimadores de β y por lo general son grandes factores en el éxito (o fracaso) de los cuadrados mínimos ordinarios, como un procedimiento de estimación.

5.6.2. Estimación de σ^2

Es necesario obtener un buen estimador de σ^2 en la regresión múltiple. Este estimador se usa en la prueba de hipótesis o en la evaluación de la calidad del modelo.

Discutimos el estimador y consideramos la identidad de la partición de los grados de libertad total en la regresión lineal múltiple. La suma de cuadrados de regresión $SC_{regresión}$ explica la variación para el modelo de k términos. Entonces la partición de los grados de libertad totales es,

$$n - 1 = k + (n - k - 1) \qquad (15)$$

El estimador insesgado s^2, expresa la variación en los residuales, es decir, variación de la regresión $\hat{y} = X\beta$ con el denominador, ahora será n – p, donde

p es el número de parámetros estimados. En la notación del modelo general, p = k + 1. Como un resultado tenemos,

$$s^2 = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (16)$$

Donde $\hat{Y}_i$ es la respuesta predicha o ajustada al *i-ésimo* dato. Como en la regresión lineal simple, este estimador del cuadrado medio residual expresa la variación natural o la varianza del error experimental y es un estimador insesgado, asumiendo que el modelo postulado y ajustado es el correcto. Podemos notar que un cambio en el modelo vía eliminación de términos del modelo incrementa la $SC_{residuales}$ y también incrementa, n – p, los grados de libertad de los residuales. Entonces el estimador muestral, s^2, puede o no aumentar. Adicionando variables regresoras puede reducir o incrementar s^2. Por lo general la suma de cuadrados del residual sirve más que un simple estimador de la varianza del error. Muchos analistas de datos lo usan (en un sentido comparativo) como un criterio de selección del modelo, para discriminar entre modelos en competencia, es decir, se favorece el modelo con un cuadrado medio residual más pequeño.

5.6.3. Geometría de los mínimos cuadrados

Para los que tienen un amplio campo en geometría vectorial, es prácticamente simple entender la geometría del procedimiento de mínimos cuadrados.

Además, la identidad de la suma de cuadrados para regresión lineal múltiple se puede ilustrar. Considere la Figura 5.2, que muestra una situación de regresión con n = 3 observaciones y p = 2 parámetros. El sistema de ejes tridimensionales mostrado está en un espacio *y*-observaciones. El vector *y* representa el vector de observaciones en el espacio. El plano dimensional de la figura es el espacio de estimación. Por el espacio de estimación entendemos, el espacio que contiene los puntos de la forma $x\hat{\beta}$, dónde $\hat{\beta}$, es un estimador del vector β. Consideramos el punto y*, como un punto candidato arbitrario en el espacio de estimación. Ahora, ¿cuál punto en el espacio de estimación produce una $\hat{y}$ para la cual la suma de residuales es mínima?

La distancia cuadrada de y* a *y* es seguramente $(y - y^*)^T (y - y^*)$ o sea, $y^* = x\hat{\beta}$ la distancia cuadrada será, entonces, $(y - x\hat{\beta})^T (y - x\hat{\beta})$. Entonces el procedimiento de mínimos cuadrados se aplica cuando, seleccionamos el punto en el espacio de estimación que minimiza esta distancia cuadrada.

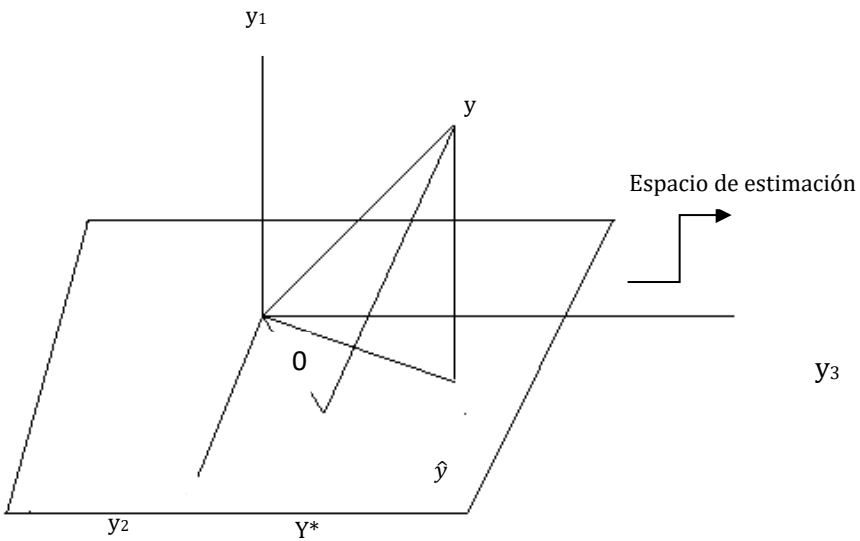


Figura 5.2. La geometría del procedimiento de mínimos cuadrados.

Es obvio que esto acompaña al punto $\hat{y}$ cuando trazamos una perpendicular de *y* al espacio de estimación. Ahora, sabemos que cuando hacemos esto, este vector más corto y - $\hat{y}$ debe ser tal que, $x^T(y - \hat{y}) = 0$ usando la notación $\hat{y} = x\beta$, para este punto, es,

$$x^T(y - x\hat{\beta}) = 0 \qquad (17)$$

Ahora es muy simple visualizar que este resultado implica que $\hat{\beta}$ es dado por

$$(x^Tx)\hat{\beta} = x^Ty \qquad (18)$$

Que representa las ecuaciones normales con las soluciones, con estimadores mínimos cuadrados.

5.7. Propiedades de los estimadores mínimos cuadrados bajo condiciones ideales

Aquí es importante recordar que las condiciones ideales del modelo general son:

a. Los ε_i tienen media cero (la forma del modelo funcional es correcta),
b. Los ε_i no están correlacionados y tienen varianza común σ^2 (varianza homogénea)

Sería anticipado realizar pruebas de hipótesis de los β's, ya que se tiene que considerar el supuesto de normalidad de los ε_i. Revisamos este supuesto para considerarlo en la discusión de las propiedades del vector $\hat{\beta}$.

5.7.1. Sesgo y propiedades de varianza de los parámetros estimados

Bajo la condición de que la E (ε_i) = 0, entonces $\hat{\beta}$ es un estimador insesgado de β. Esto se puede verificar fácilmente, ya que E (y) = xβ y *x* no es aleatoria,

$$E\,(x^Tx)^{-1}x^Ty = (x^Tx)^{-1}\,x^T\,E\,(y) = (x^Tx)^{-1}x^Tx\hat{\beta} = \beta \qquad (19)$$

Es conveniente expresar el vector de estimadores $\hat{\beta}$, en términos de los coeficientes verdaderos, los β's. Podemos escribir,

$$\hat{\beta} = \beta + R\varepsilon \qquad (20)$$

donde

$$R = (x^Tx)^{-1}\,x$$

Para la varianza o las propiedades de dispersión de $\hat{\beta}$, podemos ver la matriz de varianza-covarianzas $E\,(\hat{\beta} - \beta)\,(\hat{\beta} - \beta)^T$ y considerar el supuesto (b). Por lo que la matriz de varianza-covarianza de $\hat{\beta}$ es dada por,

$$\text{Var}\,(\hat{\beta}) = \sigma^2\,(x^Tx)^{-1} \qquad (21)$$

Podemos notar que la varianza de los coeficientes aparece en los elementos de la diagonal de $(x^Tx)^{-1}$ aparte de la varianza del error σ^2. Similarmente la covarianza entre los $\hat{\beta}$'s aparecen como los elementos de la diagonal superior de la misma matriz. Para los casos de las condiciones a y b, pero sin asumir normalidad de los ε_i, establecemos una importante propiedad óptima para los

estimadores de mínimos cuadrados de β. El resultado es el teorema de Gauss–Markoff.

> *En el caso del modelo general $y = x\beta + \varepsilon$, si la $E(\varepsilon) = 0$ y la $Var(\varepsilon) = \sigma^2 I$ (matriz de varianza-covarianza) entonces los estimadores obtienen varianza mínima de la clase de estimadores lineales insesgados.*

A menudo se establece que los estimadores mínimos cuadrados son MELI (mejor estimador linealmente insesgado). El término "mejor" aquí es usado en el sentido de mínima varianza.

5.8. Supuestos en Regresión Múltiple

En secciones anteriores describimos el problema de regresión múltiple con alguna generalidad y también se mencionan algunos supuestos que están involucrados. Ahora se pueden establecer estos supuestos más formalmente.

- Supuesto 1 *Existencia*. Para cada combinación de valores de las variables independientes X_1, X_2, X_3, ..., X_k, Y es una variable aleatoria con una cierta distribución de probabilidad con una media y varianza finita.

- Supuesto 2 *Independencia*. Las observaciones *Y* son estadísticamente independientes una de otra.
- Supuesto 3 *Linearidad*. El valor medio de *Y* para cada combinación específica de X_1, X_2, X_3, ..., X_k es una función lineal de X_1, X_2, X_3, ..., X_k; esto es

$$\mu_{Y/X_1X_2....X_k} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + ..., + \beta_kX_k \quad (22)$$

o también

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + ..., + \beta_kX_k + \varepsilon \quad (23)$$

Donde ε es el componente de error que refleja la diferencia entre una respuesta observada individualmente *Y* y la respuesta promedio verdadera $\mu_{Y/X_1X_2,X_3....X_k}$. podemos hacer algunos comentarios adicionales.

a. La superficie descrita en la primera ecuación es denominada la ecuación de regresión (superficie de regresión).
b. Si alguna de las variables independientes son funciones de alto orden de unas pocas variables independientes básicas (es decir; $X_3 = X_{12}$, $X_5 = X_1X_2$), la expresión $\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + ... + \beta_kX_k$ es realmente no lineal en las variables básicas (de aquí el término superficie más que un plano).

c. Así como en la regresión de línea recta, ε es la cantidad por la cual cualquier respuesta observada individualmente se desvía de la superficie de respuesta. Entonces, ε es el componente del error en el modelo.

- Supuesto 4 *Homocedasticidad*. La varianza de *Y* es la misma para cualquier combinación fija de X_1, X_2, X_3 ..., X_k; esto es,

$$\sigma^2_{Y/X_1X_2X_3....X_k}\sigma^2{}_{Y/}X_{1,}\,x_{2,}\,x_{3,\,\ldots}\,x_k = \text{Var}\,(Y/\,X_1,\,X_2,\,X_3,\,...,\,X_k) = \sigma^2$$

a esto se le denomina el supuesto de Homocedasticidad.

Este supuesto se puede ver muy restrictivo. Sin embargo, como mencionamos antes, en la discusión del supuesto en el análisis de regresión de línea recta, la heterocedasticidad de varianza debe ser considerada, solamente cuando los datos muestran desviaciones muy obvias y significativas de la homogeneidad. En general desviaciones suaves o pequeñas no tendrán un efecto adverso sobre los resultados.

Supuesto 5 *Normalidad*. Para cualquier combinación fija de X_1, X_2, X_3, ..., X_k, la variable *Y* está normalmente distribuida, en otras palabras,

$$Y \sim \text{N}\,(\mu_{Y/X_1X_2X_3,X_k}) \qquad (24)$$

Este supuesto no es necesario para el ajuste de mínimos cuadrados del modelo de regresión, pero se requiere, para hacer la inferencia. En este sentido, las pruebas paramétricas de hipótesis y de intervalos de confianza del análisis de regresión son robustos en el sentido de que solamente desviaciones extremas de la distribución de normalidad de *Y* pueden dar resultados engañosos. Este supuesto está basado en la evidencia teórica y experimental.

5.9. La Tabla del ANVA en regresión múltiple

Como se estableció con la regresión lineal simple, una tabla de ANVA, puede ser usada para definir un resumen total de un análisis de regresión múltiple. La forma particular de una tabla de ANVA, puede variar, dependiendo de cómo las contribuciones de las variables independientes son consideradas (es decir, individualmente o colectivamente en alguna forma). Una forma simple refleja la contribución que todas las variables independientes consideradas colectivamente hacen para la predicción.

Como podemos observar en la Tabla 5.1, la suma de cuadrados totales $SC_{totales} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$, representa la variabilidad total en las observaciones de *Y* antes de contar con el efecto conjunto de todas las variables independientes consideradas. El término $SC_{error} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ es la suma de cuadrados del residual (o la suma de cuadrados debido al error) y representa la cantidad de variación de *Y* que no es explicada después de que la variable independiente ha sido usada en la ecuación de regresión para predecir *Y*.

Finalmente $SC_{total} - SC_{error} = SC_{regresión}$, se llama la suma de cuadrados de regresión y representa la variación explicada (mide la reducción de la variación) debido a las variables independientes en la ecuación de regresión. Entonces tenemos la partición,

> *Suma de Cuadrados Totales = Suma de Cuadrados de regresión + Suma de Cuadrado de residual*

La columna de grados de libertad, da los correspondientes grados de libertad. Los grados de libertad para la regresión es k (el número de variables independientes en el modelo), los grados de libertad del residual es n – k - 1 y los grados de libertad para la suma de cuadrados totales es n - 1. La columna de cuadrados medios contiene, el término de cuadrados medios, se obtiene, dividiendo la suma de cuadrados por sus correspondientes valores de grados de libertad. La razón F es obtenida dividiendo los cuadrados medios de regresión por los cuadrados medios de residual, la interpretación de esta razón de F se discute más adelante. El R^2 en la tabla proporciona una medida cuantitativa de qué tan bien el modelo ajustado que contiene las variables independientes, predice la variable dependiente *Y*; la expresión para su cálculo es

$$R^2 = SC_{total} - \frac{SC_{residual}}{SC_{total}} \qquad (25)$$

La cantidad R^2 está entre los valores de 0 y 1. Si el valor es 1, decimos que el ajuste del modelo es perfecto. Cuando el R^2 toma el valor de cero, significa que el modelo ajustado es incapaz de explicar la variabilidad de *Y*. Siempre ha sido verdad que R^2 se incrementa, en la medida que más variables independientes se adicionen al modelo. Note sin embargo, que un incremento muy pequeño en R^2 puede ser estadísticamente no importante.

Tabla 5.1
La Tabla del análisis de varianza en regresión múltiple

Fuente de variación	**Grados de libertad**	**Sumas de cuadrados**	**Cuadrados medios**	**F_0**
Regresión	k	$SC_{regresión}$	$SC_{regresión}/k$	$CM_{regresión}/CM_{error}$
Residual	n – k - 1	SC_{error}	$SC_{error}/n – k - 1$	
Total	n - 1	SC_{total}		

5.10. Prueba de hipótesis en regresión múltiple

En muchas áreas científicas tradicionales, la función principal de los modelos, es determinar cuál de las variables regresoras, tiene una influencia verdadera sobre la variable respuesta *Y*. De igual manera se tiene que hacer una investigación preliminar para saber cuales factores, son verdaderamente relevantes. Mientras que nuestro arsenal estadístico total es abundante para manejar tal situación, el análisis de regresión es la técnica seleccionada. Por esto, es importante determinar cuales variables regresoras son responsables de una variación significativa en la respuesta *Y*.

Una vez que se tiene un modelo ajustado y los estimadores de los parámetros de interés, se pueden responder preguntas acerca de la contribución de varias variables regresoras para la predicción de *Y*. Consideramos tres tipos básicos de preguntas, que son:

1. *Una Prueba total*. Tomando colectivamente a todo el conjunto de variables independientes (o equivalentemente, al mismo modelo ajustado) contribuye significativamente a la predicción de *Y*?
2. *Prueba para la adición de una sola variable*. ¿La adición de una variable independiente particular de interés adiciona significancia a la predicción de *Y*, que el obtenido, por las otras variables independientes presentes en el modelo?
3. *Prueba para la adición de un grupo de variables*. ¿La adición de algún grupo de variables independientes de interés, incrementa significancia a la predicción de *Y*, obtenida con las otras variables independientes presentes en el modelo?

Estas cuestiones son típicamente respondidas por la aplicación de pruebas de hipótesis estadísticas. Las hipótesis nulas para estas pruebas pueden ser establecidas en términos de los parámetros desconocidos (los coeficientes de regresión) en el modelo. Las formas de estas hipótesis difieren dependiendo de la cuestión. En las siguientes secciones, describiremos las pruebas estadísticas apropiadas, para cada una de las tres cuestiones anteriores. Cada una de estas pruebas puede ser expresada como una prueba F, esto es, la

prueba estadística tendrá una distribución F, cuando la hipótesis nula establecida sea verdadera. En algunos casos, la prueba puede ser expresada como una prueba t.

Un rasgo clave de la prueba F usado en análisis de regresión, es que todas involucran una razón de dos estimadores de varianzas independientes, esto es, $F = \frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}$. Bajo el supuesto para el análisis de regresión lineal múltiple estándar dado anteriormente, el término $\hat{\sigma}_0^2$ estima a σ^2, si H_0 es verdadera; el término $\hat{\sigma}^2$ estima a σ^2, si H_0 es verdadera o no. Las formas especificas de estos estimadores de varianza serán tratados más adelante. En general, cada uno será un término de suma de cuadrados que se puede encontrar en una tabla de análisis de varianza apropiado. Si H_0 no es verdadera, entonces, $\hat{\sigma}_0^2$, estima alguna cantidad más grande que σ^2. Por lo que esperamos un valor de F cercano a 1, si H_0 es verdadera, pero más grande que 1 si H_0 no es verdadera. El valor más grande de F, significa que H_0 no es verdadera. Otra característica general de estas pruebas, *es que cada prueba puede ser interpretada como una comparación de dos modelos.* Uno de estos modelos es el modelo completo o total, el otro se denomina modelo reducido (esto es, el modelo en el cuál el modelo completo se reduce bajo la hipótesis nula).

Como un simple ejemplo considere los dos modelos,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad (26)$$

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad (27)$$

Bajo H_0: $\beta_2 = 0$, el modelo completo se reduce al modelo más pequeño. Una prueba de H_0: $\beta_2 = 0$, es esencialmente equivalente a determinar cuál de los dos modelos es más apropiado.

El lector puede notar, del ejemplo anterior, que el grupo de variables independientes, en el modelo reducido (llamado, X_1) es un subconjunto de las variables independientes en el modelo completo (llamado, X_1 y X_2). Esta es una característica común para todos los tipos básicos de prueba, descritas en esta sección. Suponga por ejemplo, que tenemos H_0: $\beta_1 = \beta_2$. Entonces el modelo reducido se puede escribir como; $Y = \beta_0 + \beta X + \varepsilon$ con $\beta = \beta_1 = \beta_2$ y $X = X_1 = X_2$.

5.10.1. Prueba de significancia para la regresión total

Consideremos la primera cuestión, establecida anteriormente, una prueba total para un modelo que contiene k variables independientes, es decir,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \varepsilon \qquad (28)$$

La hipótesis nula para esta prueba puede ser establecida generalmente como H_0: "todas las k variables independientes consideradas juntas, no explican una cantidad significativa de la variación en *Y*". De forma equivalente, podemos establecer la hipótesis nula como H_0: "no existe significancia en la regresión total, usando las k variables independientes en el modelo", o como, H_0: $\beta_1 =$

$\beta_2 = \beta_3 = ... = \beta_k = 0$. Bajo esta última versión de H_0, el modelo completo es reducido a un modelo que contiene solamente el término intercepto β_0.

Para aplicar la prueba, hacemos uso de las cantidades de los cuadrados medios, calculados en la tabla del análisis de varianza. Entonces, podemos calcular el estadístico F como,

$$F = \frac{CM_{regresiòn}}{CM_{residual}} = \frac{(SC_{total} - SC_{residual})/k}{SC_{residual}/(n-k-1)} \quad (29)$$

Dónde $SC_{total} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ y $SC_{residual} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ son las sumas de cuadrados totales y las sumas de cuadrados de residuales, respectivamente. El valor calculado de F, puede ser comparado con el punto crítico, $F_{k,n-k-1,\ 1-\alpha}$ con un valor de α seleccionado como nivel de significancia. Rechazamos la H_0, si el valor calculado de F es mayor que el punto crítico de la prueba. Asimismo podemos calcular el valor-p, para esta prueba, como el área bajo la curva de la distribución $F_{k,\ n-k-1}$, al lado derecho del estadístico F- calculado. También se puede mostrar una equivalencia de cálculo para el estadístico de prueba F, en términos de R^2, esto es,

$$F = (R^2/k) / (1 - R^2)/ (n-k-1) \quad (30)$$

5.10.2. Prueba Parcial de F

Alguna información adicional importante, relacionada al modelo de regresión ajustado puede ser obtenida presentando la tabla del análisis de varianza como se muestra en la Tabla 5.2.

Tabla 5.2
Análisis de varianza para el peso (Y) con la altura (X_1), edad (X_2) y (edad)^2 = X_3
Componentes de Sumas de Cuadrados de regresión

Fuente		Grados de libertad	Sumas de Cuadrados	Cuadrados Medios	F_0	R^2
Regresión	X_1	1	588.92	588.92	19.67 **	0.7802
	X_2/X_1	1	103.90	103.90	4.78	
	$X_3/X_1, X_2$	1	0.24	0.24	0.01	
Residual		8	195.19	24.40		
Total		11	888.25			

** Altamente significativo

Lo que se ha hecho en esta presentación es la partición de la suma de cuadrados de regresión en tres componentes, con un grado de libertad cada uno:

1. SC (X_1) es la suma de cuadrados explicada, usando solamente X_1 para predecir *Y*,
2. SC (X_2/X_1): la suma de cuadrados extra explicada, usando a X_2 en adición a X_1 para predecir *Y*.

3. SC ($X_3/X_1, X_2$): la suma de cuadrados extra explicada, usando a X_3 en adición a X_1 y X_2 para predecir *Y*.

Podemos usar la información extra, en la tabla para responder a las siguientes cuestiones.

a. ¿Ayuda X_1 significativamente para predecir *Y*?
b. ¿La adición de X_2 contribuye significativamente a la predicción de *Y* después de tomar en cuenta la contribución de X_1?
c. ¿La adición de X_3, contribuye significativamente en la predicción de *Y*, después de tomar en cuenta las contribuciones de X_1 y X_2?

Ya sabemos cómo responder la cuestión (a). Esto simplemente involucra el ajuste de un modelo de regresión de línea recta, usando a X_1, cómo la única variable independiente. El valor de 588.92, es la suma de cuadrados de regresión para este modelo de regresión de línea recta. La suma de cuadrados de residual para este modelo puede ser obtenida sumando: 195.19 + 103.90 + 0.24 = 299.33 con 10 grados de libertad. El estadístico F para probar si hay significancia en la regresión de línea recta, cuando usamos solamente X_1, se calcula por F = (588.92/1)/(299.33/10) = 19.67, el cual se compara con el punto crítico y tiene un valor de $p < 0.01$, esto es, X_1, contribuye significativamente en la predicción lineal de *Y*.

Para responder las cuestiones b y c, podemos usar lo que denominamos *prueba de F –parcial*. Esta prueba evalúa si la suma de

cualquier variable independiente específica, estando otras ya en el modelo, contribuye significativamente a la predicción de *Y*. Esta prueba, permite la eliminación de variables que no son de ayuda para la predicción de *Y*, permitiendo la reducción del conjunto de variables independientes a un conjunto económico de predictores importantes.

5.10.3. La Hipótesis Nula

Asumimos que deseamos probar si la suma de una variable X* mejora significativamente la predicción de *Y* una vez que las variables X_1, X_2, X_3, ..., X_p, ya están en el modelo. La hipótesis nula puede ser establecida como H_0: "X* no se suma significativamente a la predicción de *Y*, dado que X_1, X_2, X_3, ..., X_p ya están en el modelo" o equivalentemente, como H_0: $\beta^* = 0$ en el modelo $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + ...+ \beta_pX_p + \beta^*X^* + \varepsilon$.

Se puede inferir de la segunda declaración (b), que el procedimiento de prueba esencialmente compara dos modelos. El modelo completo contiene X_1, X_2, X_3, ..., X_p y X*, como variables independientes; el modelo reducido contiene, X_1, X_2, X_3, ..., X_p, pero no X*(puesto que $\beta^* = 0$ bajo la hipótesis nula). La tarea es determinar cual modelo es más apropiado basado sobre cuánta información adicional provee X* acerca de *Y* sobre la que ya está dada por X_1, X_2, X_3 ..., X_p.

5.10.4. El procedimiento

Para aplicar una prueba F-parcial, a una variable X*, dado que las variables X_1, X_2, X_3 ..., X_p ya están en el modelo, se debe primero calcular la suma de cuadrados extra de sumar X*, dado que X_1, X_2, X_3, ..., X_p se pueden colocar en la tabla del análisis de varianza, bajo la fuente, el encabezado "regresión X*/X_1, X_2, X_3, ..., X_p". Esta suma de cuadrados es calculada por la expresión

Suma de cuadrados extra de adicionar X* dadas X_1, X_2, X_3, ..., X_p	=	Sumas de cuadrados de regresión cuando X_1, X_2, X_3, ..., X_p y X* están en el modelo	=	Suma de cuadrado de regresión cuando X_1, X_2, X_3 ..., X_p están en el modelo

O en forma más compacta,

$$SC(x^*/ x_1, x_2, x_3, ..., x_p) = SC_{regresión}(x_1, x_2, x_3, ..., x_p, x^*) - SC_{regresión}(x_1, x_2, x_3, ..., x_p)$$

Para el ejemplo anterior tenemos

$$SC(x_2/x_1) = SC_{regresión}(x_1, x_2) - SC_{regresión}(x_1) = 692.82 - 588.92 = 103.90$$

$$SC(x_3/x_1, x_2) = SC_{regresión}(x_1, x_2, x_3) - SC_{regresión}(x_1, x_2) = 693.06 - 692.82 = 0.24$$

Para probar la hipótesis nula H_0: "la adición de X* al modelo que ya contiene X_1, X_2, ..., X_p no mejora la significancia para predecir *Y*", calculamos,

$$F(X^*/ X_1, X_2, X_3, ..., X_p) = \frac{\text{sumas de cuadrados extra de adicionar } X*, \text{dado } X_1X_2X_3.....X_P}{\text{cuadrados medios de residual para el modelo que contiene todas las variables } X_1X_2...X_P\ X*}$$

O de forma más compacta,

$$F(X^*/ X_1, X_2, X_3, ..., X_p) = \frac{SC(X*/X_1X_2....X_P)}{Cuadrados\ Medios\ de\ Residual(X_1X_2.....X_P,X*)} \quad (31)$$

Este estadístico de F tiene una distribución F con 1 y n-p-1 grados de libertad, bajo H_0, tal que rechazamos H_0 si el F-calculado es mayor a $F_{1, n-p-2, 1-\alpha}$. Para nuestro ejemplo los estadísticos de F- parciales son,

$F(X_2/X_1)$ = SC (X_2/X_1) / [Cuadrados medios de residual (X_1, X_2)] = 103.90/ (195.19 + 0.24) = 4.78

$F(X_3/X_1, X_2)$ = SC $(X_3/ X_1, X_2)$/Cuadrado medio de residual (X_1, X_2, X_3) = 0.24/24.40 = 0.01

En ninguna de las dos pruebas se rechaza la H_0 con un nivel de $\alpha = 0.05$, sin embargo para la primera prueba se rechaza H_0 con un $\alpha = 0.10$. Por lo que se puede decir que X_1 y X_2 están en el modelo y que la adición de X_3 es superflua.

5.10. 5. La prueba alternativade t

Una manera alternativa para aplicar la prueba F – parcial para las últimas variables adicionales, es usar una prueba de t. La prueba de t alternativa, enfoca la prueba de la hipótesis nula H_0: $\beta^* = 0$ donde β^* es el coeficiente de X^* en la ecuación de regresión $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + ... + \beta_pX_p + \beta^*X^* + \varepsilon$. El estadístico equivalente para probar esta hipótesis nula es,

$$T = \hat{\beta}^*/S_{\beta^*} \qquad (32)$$

Donde $\hat{\beta}^*$ es el coeficiente estimado correspondiente y S_{β^*} es el estimador del error estándar de $\hat{\beta}^*$ ambos son calculados por programas de regresión estándar.

En la aplicación de esta prueba, rechazamos H_0: $\beta^* = 0$ si,

$|T| > t_{n-p-2,\ 1-\alpha/2}$ (prueba de dos colas; H_A: $\beta^* \neq 0$)

$T > t_{n-p-2,\ 1-\alpha}$ (prueba de la cola superior H_1: $\beta^* > 0$)

$T < -t_{n-p-2,\ 1-\alpha}$ (prueba de la cola inferior; H_1: $\beta^* < 0$)

Se puede demostrar que una prueba de dos colas es equivalente a una prueba F – parcial descrita anteriormente. Por ejemplo, para probar H_0: $\beta_3 = 0$ en el modelo $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon$, ajustado a los datos anteriores, se calcula.

$$T = \hat{\beta}^*/S_{\beta^*} = - 0.0417/0.4224 = - 0.10$$

Elevando al cuadrado este resultado, obtenemos

$$T^2 = 0.01 = \text{a la F parcial } F(X_3/X_1, X_2)$$

de la Tabla 5.2.

5.11. Ejercicios

5.1 Se realizó un estudio en el cual se desea conocer la relación entre la variable dependiente Y y tres variables independientes X_1, X_2, y X_3. Se ajustaron tres modelos de regresión y los estimadores y las tablas del análisis de varianza fueron;

Modelos	variables	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$S_{\beta 1}$	$S_{\beta 2}$	$S_{\beta 3}$
1	X_1	59.092	1.605			0.2387		
2	X_1 X_2	48.050	1.709	10.294		0.2018	2.7681	
3	X_1 X_2 X_3	45.103	1.213	9.946	8.592	0.3238	2.6561	4.4987

Las tablas de análisis de varianza son.

Modelo 1			Modelo 2			Modelo 3		
Fuente	**G. L.**	**SC**	**Fuente**	**G. L.**	**SC**	**Fuente**	**G. L.**	**SC**
Regres	1	3861.630	Regres	2	4689.684	Regres	3	4889.826
Residual	30	2564.338	Residual	29	1736.285	Residual	28	1536.143

a) Use el modelo 3 ¿Cuál es el valor predicho para Y si X_1= 50 y X_2 = 3.5?

b) ¿Cuál es el valor para Y cuando X_1 = 50 y X_3 = 3.5?

c) Use las tablas de ANVA, calcular y comparar los valores de R^2 de los tres modelos.

d) Construir las pruebas de F totales para la significancia de las regresiones y establecer las hipótesis nulas en términos de los coeficientes de regresión de los tres modelos.

5.2 Se realizó un estudio para evaluar la relación de los niveles de colesterol en 25 pacientes antes de su terapia, tomando en cuenta el peso corporal (X_1) y la edad (X_2). Se ajustaron los siguientes modelos.

Modelo 1			**Modelo 2**		
Fuente	**G.L.**	**SC**	**Fuente**	**G.L.**	**SC**
Regresión(X_1)	1	10231.7	Regresión(X_2)	1	101932.7
Residual	23	135144.3	Residual	23	43444.3

a) Dado los dos análisis de varianza mencionar cuál de las dos variables es la predictor más importante de *Y*?
b) Los modelos de regresión estimados resultan del ajuste por separado de Y para X_1 y X_2 así cómo de X_1 y para X_2, esto es;

$$\hat{Y} = 77.983 + 0.417X_1 + 5.217X_2$$
$$\hat{Y} = 199.2975 + 1.622X_1$$
$$\hat{Y} = 102.5751 + 5.321X_2$$

Para cada uno de estos modelos, determinar el valor del nivel (Y) de colesterol predicho para el paciente 4 (con Y = 263, X_1 = 70, X_2 = 30) y comparar los valores con el valor observado. Comente su hallazgo.

c) Dada la siguiente tabla de ANVA, de la regresión usando las dos variables independientes, definir la prueba de F para el modelo de las dos variables y la prueba de F parcial, para la adición de X_1 al modelo, dado que X_2 ya forma parte del modelo.

Fuente	G.L.	SC
Regresión(X_1,X_2)	2	102570.80
Residual	22	42806.20

d) Calcule y compare los valores de R^2 para cada uno de los tres modelos considerados en el inciso (b).

e) Basado en los resultados obtenidos de los incisos (a – d) ¿cuál considera que es el mejor modelo de predicción considerando una o las dos variables independientes?

5.3 Un panel de educación de una comunidad tabasqueña está interesada en evaluar el efecto de los recursos educativos en el aprovechamiento de sus estudiantes. Evaluaron la relación entre los registros promedios de un examen (Y) y las siguientes variables independientes, en una muestra de 25 estudiantes de nivel secundaria: X_1 = gastos por alumnos, X_2 = porcentaje de profesores con maestrías o doctorado y X_3 = proporción alumnos/profesor. La tabla del ANVA de los datos proporciona el ajuste del modelo de regresión de Y sobre X_1, X_2 y X_3.

Fuente	G.L.	SC
Regresión (X_1, X_2, X_3)	3	25974.00
Residual	21	2248.23
Total	24	28222.23

a) Realice la prueba de F total para el modelo

b) Calcule el valor de R^2.

5.4 Se midió la supervivencia porcentual de los espermatozoides en semen de cerdos después de almacenarlo en varias combinaciones de concentraciones de tres materiales usados para aumentar la posibilidad de supervivencia. Los datos son:

Y (% de supervivencia)	X_1 (peso %)	X_2 (peso %)	X_3 (peso %)
25.5	1.74	5.30	10.80
31.2	6.32	5.42	9.40
25.9	6.22	8.41	7.20
38.4	10.52	4.63	8.50
18.4	1.19	11.60	9.40
26.7	1.22	5.85	9.90
26.4	4.10	6.62	8.00
25.9	6.32	8.72	9.10
32.0	4.08	4.42	8.70
25.2	4.15	7.60	9.20
39.7	10.15	4.83	9.40
35.7	1.72	3.12	7.60
26.5	1.70	5.30	8.20

a) Calcule el modelo de regresión completo que explique la relación. Explicar.

b) Calcule el coeficiente de determinación e interprete.

5.5 Se realizó un estudio para determinar la posibilidad de pronosticar el peso de un animal después de un periodo de tiempo determinado sobre la base de su peso inicial y la cantidad de alimento que recibe. Se obtuvieron los siguientes datos:

Peso final (y)	95	77	80	100	97	70	50	80	92	84
Peso inicial (x_1)	42	33	33	45	39	36	32	41	40	38
Alimento consumido (x_2)	272	226	259	292	311	183	173	236	230	235

a) Ajuste una ecuación de regresión múltiple de la forma $\mu_{Y/X_1X_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

b) Calcule el coeficiente de determinación e interprete.

c) Pronostique el peso final de un animal que tiene un peso inicial de 35 kg y que recibe 250 g de alimento.

5.6 Los siguientes datos resultaron de quince eventos experimentales realizados acerca de cuatro variables independientes y una dependiente (y).

y	X_1	X_2	X_3	X_4
14.8	7.8	4.3	11.5	6.3
12.1	6.9	3.9	14.3	7.4
19.0	9.3	8.4	9.4	5.9
14.5	6.8	10.3	15.2	8.7
16.6	11.7	6.4	8.8	9.1
17.2	8.5	5.7	9.8	5.6
17.5	12.6	6.8	11.2	6.8
14.1	7.5	4.2	10.9	7.4
13.8	8.4	7.3	14.7	8.2
14.7	11.3	8.8	15.1	9.2
17.7	10.7	3.6	8.7	4.7
17.0	7.3	4.9	8.6	5.5
17.6	8.4	7.3	9.3	6.6
16.3	6.7	9.7	10.8	8.7
18.2	9.6	8.4	11.9	5.4

a) Ajuste el modelo de regresión de *y* con las cuatro variables.

b) Ajuste el modelo de regresión de *y* con x_1, x_1^2, x_2, x_3, x_2^2, x_4.

c) Calcule el coeficiente de determinación de los dos modelos y compare.

5.7 Un estudio sociológico mostró que recientemente se ha incrementado la incidencia de homicidios en México, ya que la tasa de homicidios por cada 100 mil habitantes (Y) está relacionado con el tamaño de la población (X_1), el porcentaje de familias con ingresos anuales menores a \$10,000 (X_2), y la tasa de desempleo (X_3). Los datos de 20 estados del país se muestran en la siguiente tabla.

Estado	Y	X_1	X_2	X_3	Estado	Y	X_1	X_2	X_3
1	11.2	587	16.5	6.2	11	14.5	7895	18.1	6.0
2	13.4	643	20.5	6.4	12	26.9	762	23.1	7.4
3	40.7	635	26.3	9.3	13	15.7	2793	19.1	5.8
4	5.3	692	16.5	5.3	14	36.2	741	24.7	8.6
5	24.8	1248	19.2	7.3	15	18.1	625	18.6	6.5
6	12.7	643	16.5	5.9	16	28.9	854	24.9	8.3
7	20.9	1964	20.2	6.4	17	14.9	716	17.9	6.7
8	35.7	1531	21.3	7.6	18	25.8	921	22.4	8.6
9	8.7	713	17.2	4.9	19	21.7	595	20.2	8.4
10	9.6	749	14.3	6.4	20	25.7	3353	16.9	6.7

a) Dadas las siguientes tablas ANVA, de las regresiones calculadas, calcule las pruebas de F para cada modelo de regresión de dos variables.

Fuente	G.L	SC
Regresión (X_1, X_2)	2	1317.80
Residual	17	537.40

Fuente	G.L.	SC
Regresión (X_1, X_3)	2	1395.49
Residual	17	459.71

Fuente	G.L	SC
Regresión (X_2, X_3)	2	1477.37
Residual	17	377.82

b) Basándose en los resultados del inciso (a), responda ¿cuál modelo con dos variables recomendaría?

c) Calcule los valores de R^2 para cada modelo de dos variables anteriores y establezca los resultados de acuerdo al inciso (b).

d) Dada la siguiente tabla de la regresión con tres variables, realice la prueba de F para la regresión total.

Fuente	G.L	SC
Regresión (X_1, X_2, X_3)	3	1507.18
Residual	67	348.03

e) Determine y comente sobre el incremento de R^2, tomando en cuenta el modelo que incluye X_2 y X_3 y el modelo que incluye las tres variables independientes.

f) La tabla del ANVA que resulta de la regresión con las variables independientes X_2, X_3 y $X_4 = X_2X_3$ es la siguiente. Use esta tabla para hacer la prueba de la regresión total.

Fuente	G.L	SC
Regresión (X_2, X_3, X_4)	3	1480.46
Residual	16	377.73

5.8 Un equipo de ambientalistas usó datos de 23 ciudades para investigar la relación entre la tasa de mortalidad por cáncer (Y) para el 2005 y las siguientes variables independientes: X_1 = índice de polución del aire en la ciudad, X_2 = edad media (más de 21) en la ciudad, X_3 = porcentaje de la fuerza de trabajo en la ciudad empleada en cierta industria. La tabla del ANVA se muestra resumiendo los resultados de la regresión.

Fuente	G.L	SC
Regresión (X_1, X_2, X_3)	3	1835.93
Residual	19	551.723
Total	22	2387.653

a) Realice la prueba de F para el modelo de regresión de las tres variables.

b) Determine R^2 para el modelo e interprete.

5.9 Se realizó un experimento para evaluar el efecto de ciertas variables sobre la erosión del suelo y se aplicó en parcelas de 10 ft^2 con pendiente, sujetas a 2 pulgadas de lluvia artificial aplicada por periodos de 20 minutos. Los datos y los análisis son:

Parcela	1	2	3	4	5	6	7	8	9	10	11
Y	27.1	35.6	31.4	37.8	40.2	39.8	55.5	43.6	52.1	43.8	35.7
X_1	0.43	0.47	0.44	0.48	0.48	0.49	0.53	0.50	0.55	0.51	0.48
X_2	1.95	5.13	3.98	6.25	7.12	6.50	10.67	7.08	9.88	8.72	4.96
X_3	0.34	0.32	0.29	0.30	0.25	0.26	0.10	0.16	0.19	0.18	0.28

Fuente	G.L	SC
Regresión	3	680.4913
Residual	7	16.0943
Total		696.5856

a) El modelo ajustado con las tres variables independientes está dado por

$$\hat{Y} = -1.879 + 77.326X_1 + 1.559X_2 - 23.904X_3$$

Calcule y compare los valores observados y estimados de Y para las parcelas 1, 5 y 7.

b) Haga la prueba de significancia de la regresión para el modelo con las tres variables independientes.

5.10 En un estudio de Yoshida (1961) se midió el consumo de oxígeno de un grupo de larvas del óxido del alambre a cinco temperaturas. La tasa del consumo de oxígeno por el grupo de larvas en mililitros por hora, variable dependiente, fue transformada a 0.5 menos el logaritmo común. Otra variable independiente de importancia fue el peso del grupo de larvas transformado a logaritmo común. Las tablas del ANVA son las siguientes:

Fuente	G.L	SC		Fuente	G.L	SC
Regresión (X_1)	1	0.0661		Regresión (X_2)	1	2.7742
Residual	45	3.3399		Residual	45	0.6318

Fuente	G.L	SC
Regresión (X_1, X_2)	2	3.2112
Residual	4	0.1948

a) El modelo de regresión múltiple con las dos variables independientes es

$$\hat{Y} = -0.6835 + 0.5917X_1 + 0.0393X_2$$

Usando este modelo ajustado, ¿cuánto cambiaría el consumo de oxígeno predicho para un grupo de larvas con el peso X_1 fijo, si la temperatura cambia de $X_2 = 20$ a $X_2 = 25$?

b) Para una temperatura de 20°C, calcular y comparar los valores predichos de $\hat{Y}$ para pesos de 0.250 y 0.500.

c) ¿Cuál es el R^2 de cada uno de los tres modelos anteriores?

5.11 Un psiquiatra examinó la relación entre niveles de ansiedad (Y) medida en una escala de 1 a 50 como el promedio de tres puntos en un periodo de 2 semanas y las siguientes variables independientes; X_1 = presión sistólica, X_2 = IQ y X_3 = satisfacción en el trabajo (medido en una escala de 1 a 25). La tabla del ANVA resume los resultados obtenidos en un análisis de regresión de una variable adicionada en orden de 22 pacientes bajo terapia en una clínica.

Fuente	G.L	SC
Regresión X_1	1	981.326
X_2/X_1	1	190.232
$X_3/X_1, X_2$	1	129.431
Residual	18	442.292

a) Pruebe la significancia de cada variable independiente de acuerdo a su entrada al modelo. Establezca las hipótesis nulas para cada prueba en términos de los coeficientes de regresión.

b) Pruebe la significancia de la adición de ambas variables X_2 y X_3 al modelo cuando ya contiene a X_1. Establezca la hipótesis nula en términos de los coeficientes de regresión y en términos de coeficientes de correlación (parcial-múltiples).

c) En términos de sumas de cuadrados, cuál prueba corresponde para comparar los dos modelos; $Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon$ y el modelo $Y = \beta_0 + \beta_3X_3 + \varepsilon$. ¿Porqué esta prueba debe ser hecha usando la tabla del ANVA? Describa el procedimiento apropiado.

d) Basados en las pruebas anteriores, ¿cuál recomienda como el modelo estadístico más apropiado? Use un $\alpha = 0.01$.

5.12 Las siguientes tablas proveen información adicional de los datos del ejercicio 5.7.

Fuente	G.L	SC		Fuente	G:L.	SC
X_2	1	1308.34		X_2	1	1360.14
X_1/X_2	1	9.46		X_1/X_3	1	35.35
Residual	17	537.40		Residual	17	459.71

Fuente	G.L	SC		Fuente	G:L.	SC
X_3	1	1360.14		Regresión (X_1, X_2, X_3)	3	1507.18
X_2/X_3	1	117.23				
Residual	17	377.82		Residual	16	348.03

Responda las siguientes cuestiones usado las tablas de ANVA anteriores y las del ejercicio 5.7.

a) Haga la prueba de la variable añadida en orden para el orden X_2, X_1 y X_3.
b) Haga las pruebas de las variables añadidas en orden para el orden X_3, X_1 y X_2.
c) Enliste todos los órdenes que pueden ser probados usando la tabla del ANVA anteriores y las del ejercicio 5.7. Enliste todos los órdenes que no puedan ser calculados.
d) Haga las pruebas de las últimas variables añadidas para X_1, X_2 y X_3.

e) La tabla de ANVA que resulta del ajuste de un modelo con las variables X_2, X_3 y $X_4 = X_2X_3$, es dada en la tabla.

Fuente	G.L	SC	SCM
Regresión X_3	1	1360.14	1360.14
X_2/X_3	1	117.23	117.23
$X_4/X_2, X_3$	1	0.09	0.09
Residual	16	377.73	23.61

Use esta tabla para probar si X_4 mejora significativamente la predicción de Y dado que X_2 y X_3 ya están en el modelo.

5.13 La siguiente tabla de ANVA está basada en los datos del ejercicio 5.3.

Fuente	G.L	SC
Regresión X_1	1	48953.04
X_3/X_1	1	7010.03
$X_2/X_1, X_3$	1	10.93
Residual	21	2248.23
Total	24	28222.23

a) Haga una prueba para comparar los siguientes dos modelos.

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon$$
$$Y = \beta_0 + \beta_1X_1 + \varepsilon$$

b) Haga una prueba para comparar los siguientes dos modelos

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

$$Y = \beta_0 + \varepsilon$$

c) Establezca cuál de los dos modelos se compara cuando se calcula

$$F = \frac{(18953.04+7010.03+10.93)/3}{2248.23/21}$$

5.12 Ejemplo usando SAS

Estos datos provienen de un estudio realizado sobre el tiempo de vida de organismos y cinco variables independientes. Los datos provienen de una muestra de 54 pacientes que esperan una cirugía de hígado. Es un estudio observacional en el cual la variable respuesta es el tiempo de sobrevivencia del paciente. El objetivo del estudio fue obtener una ecuación usando la información obtenida en el diagnóstico pre-operatorio. Antes de la operación, los datos fueron obtenidos de cuatro variables que son posibles predictoras. Estas variables son:

CLOT = Registro del índice de coagulación de la sangre.
PROG = Índice de pronóstico que incluye la edad del paciente.
ENZ = Prueba de la función enzimática
LIV = Registro de la prueba funcional del hígado.
TIEM = Tiempo de supervivencia del paciente.
VITIEM = Logaritmo de la variable TIEM.

El modelo puede ser: $TIEM = \xi_0 + \beta_1 CLOT + \beta_2 PROG + \beta_3 ENZ + \beta_4 LIV + \varepsilon_i$
La variable respuesta es el tiempo (TIEM), que equivale al número de días que vivirá después de la operación.

CLOT	PROG	ENZ	LIV	TIEM	VITIEM	CLOT	PROG	ENZ	LIV	TIEM	VITIEM
6.7	62	81	2.59	200	2.3010	11.2	76	90	5.59	574	2.7589
51.	59	66	1.70	101	2.0043	5.2	54	56	2.71	72	1.8573
7.4	57	83	2.16	204	2.3096	5.8	76	59	2.58	178	2.2504
6.5	73	41	2.01	101	2.0043	3.2	64	65	0.78	71	1.8513
7.8	65	115	4.30	509	2.7067	8.7	45	23	2.52	58	1.7634
5.8	38	72	1.42	80	1.9031	5.0	59	73	3.50	116	2.0645
5.7	46	63	1.91	80	1.9031	5.8	72	93	3.30	295	2.4698
3.7	68	81	2.57	127	2.1038	5.4	58	70	2.64	115	2.0607
6.0	67	93	2.50	202	2.3054	5.3	51	99	2.60	184	2.2648
3.7	76	94	2.40	203	2.3075	2.6	74	86	2.05	118	2.0719
6.3	84	83	4.13	329	2.5172	4.3	8	119	2.85	120	2.0792
6.7	51	43	1.86	65	1.8129	4.8	61	76	2.45	151	2.1790
5.8	96	114	3.95	830	2.9191	5.4	52	88	1.81	148	2.1703
5.8	83	88	3.95	330	2.5185	5.2	49	72	1.84	95	1.9777
7.7	62	67	3.40	168	2.2253	3.6	28	99	1.30	75	1.8751
7.4	74	68	2.40	217	2.3365	8.8	86	88	6.40	483	2.6840
6.0	85	28	2.98	87	1.9395	6.5	56	77	2.85	153	2.1847
3.7	51	41	1.55	34	1.5315	3.4	77	93	1.48	191	2.2810
7.3	68	74	3.56	215	2.3324	6.5	40	84	3.00	123	2.0899
5.6	57	87	3.02	172	2.2355	4.5	73	106	3.05	311	2.4928
5.2	52	76	2.85	109	2.0374	4.8	86	101	4.10	398	2.5999
3.4	83	53	1.12	136	2.1335	5.1	67	77	2.86	158	2.1987
6.7	26	68	2.10	70	1.8451	3.9	82	103	4.55	310	2.4914
5.8	67	86	3.40	220	2.3424	6.6	77	46	1.95	124	2.0934
6.3	59	100	2.95	276	2.4409	6.4	85	40	1.21	125	2.0969
5.8	61	73	3.50	144	2.1584	6.4	59	85	2.33	198	2.2967
5.2	52	86	2.45	181	2.2577	8.8	78	72	3.20	313	2.4955

El programa de análisis de SAS.

data regres;

input clot prog enz liv tiem vitiem;

cards;

6.7	62	81	2.59	200	2.3010
51.	59	66	1.70	101	2.0043
7.4	57	83	2.16	204	2.3096
6.5	73	41	2.01	101	2.0043
7.8	65	115	4.30	509	2.7067
5.8	38	72	1.42	80	1.9031
5.7	46	63	1.91	80	1.9031
3.7	68	81	2.57	127	2.1038
6.0	67	93	2.50	202	2.3054
3.7	76	94	2.40	203	2.3075
6.3	84	83	4.13	329	2.5172
6.7	51	43	1.86	65	1.8129
5.8	96	114	3.95	830	2.9191
5.8	83	88	3.95	330	2.5185
7.7	62	67	3.40	168	2.2253
7.4	74	68	2.40	217	2.3365
6.0	85	28	2.98	87	1.9395
3.7	51	41	1.55	34	1.5315
7.3	68	74	3.56	215	2.3324
5.6	57	87	3.02	172	2.2355
5.2	52	76	2.85	109	2.0374
3.4	83	53	1.12	136	2.1335
6.7	26	68	2.10	70	1.8451
5.8	67	86	3.40	220	2.3424
6.3	59	100	2.95	276	2.4409
5.8	61	73	3.50	144	2.1584
5.2	52	86	2.45	181	2.2577
11.2	76	90	5.59	574	2.7589
5.2	54	56	2.71	72	1.8573
5.8	76	59	2.58	178	2.2504
3.2	64	65	0.78	71	1.8513
8.7	45	23	2.52	58	1.7634
5.0	59	73	3.50	116	2.0645
5.8	72	93	3.30	295	2.4698
5.4	58	70	2.64	115	2.0607

5.3	51	99	2.60	184	2.2648
2.6	74	86	2.05	118	2.0719
4.3	8	119	2.85	120	2.0792
4.8	61	76	2.45	151	2.1790
5.4	52	88	1.81	148	2.1703
5.2	49	72	1.84	95	1.9777
3.6	28	99	1.30	75	1.8751
8.8	86	88	6.40	483	2.6840
6.5	56	77	2.85	153	2.1847
3.4	77	93	1.48	191	2.2810
6.5	40	84	3.00	123	2.0899
4.5	73	106	3.05	311	2.4928
4.8	86	101	4.10	398	2.5999
5.1	67	77	2.86	158	2.1987
3.9	82	103	4.55	310	2.4914
6.6	77	46	1.95	124	2.0934
6.4	85	40	1.21	125	2.0969
6.4	59	85	2.33	198	2.2967
8.8	78	72	3.20	313	2.4955

proc glm;

model tiem vitiem= clot prog enz liv;

run;

Los resultados

The GLM Procedure

Dependent Variable: tiem

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	856844.878	214211.219	40.05	<.0001
Error	49	262086.622	5348.707		
Corrected Total	53	1118931.500			

R-Square	Coeff Var	Root MSE	tiem Mean
0.765771	37.09291	73.13485	197.1667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Clot	1	15.9145	15.9145	0.00	0.9567
Prog	1	343486.8479	343486.8479	64.22	<.0001
Enz	1	400367.1510	400367.1510	74.85	<.0001
Liv	1	112974.9644	112974.9644	21.12	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Clot	1	3796.0734	3796.0734	0.71	0.4036
Prog	1	164715.0695	164715.0695	30.80	<.0001
Enz	1	165662.9490	165662.9490	30.97	<.0001
Liv	1	112974.9644	112974.9644	21.12	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-415.1200629	56.26556738	-7.38	<.0001
Clot	1.3415201	1.59240774	0.84	0.4036
Prog	3.6328013	0.65463522	5.55	<.0001
Enz	2.9817386	0.53577354	5.57	<.0001
Liv	52.3576536	11.39235359	4.60	<.0001

Ahora, usando la variable dependiente logaritmo del tiempo tenemos:

Number of observations 54

The GLM Procedure

Dependent Variable: vitiem

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	3.51850708	0.87962677	94.88	<.0001
Error	49	0.45426517	0.00927072		
Corrected Total	53	3.97277225			

R-Square	Coeff Var	Root MSE	vitiem Mean
0.885655	4.364382	0.096285	2.206144

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Clot	1	0.00068034	0.00068034	0.07	0.7876
Prog	1	1.39597474	1.39597474	150.58	<.0001
Enz	1	1.85166411	1.85166411	199.73	<.0001
Liv	1	0.27018790	0.27018790	29.14	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Clot	1	0.01168738	0.01168738	1.26	0.2670
Prog	1	0.78526474	0.78526474	84.70	<.0001
Enz	1	0.93852717	0.93852717	101.24	<.0001
Liv	1	0.27018790	0.27018790	29.14	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.9193782589	0.07407557	12.41	<.0001
Clot	0.0023539030	0.00209646	1.12	0.2670
Prog	0.0079320058	0.00086185	9.20	<.0001
Enz	0.0070970897	0.00070536	10.06	<.0001
Liv	0.0809696037	0.01499843	5.40	<.0001

	Variables	**R-Square**	**Coeff Var**	**Root MSE**	**Vitiem Mean**
Modelo 1	Tiempo	0.765771	37.09291	73.13485	197.1667
Modelo 2	Log. tiempo	0.885655	4.364382	0.096285	2.206144

Como podemos observar, la variación de los R^2 con sólo transformar la variable tiempo en el logaritmo del tiempo es bastante importante pues permite obtener un modelo con mayor capacidad de predicción. Asimismo, la gran variación del coeficiente de variación en ambos modelos es de 37.0929% del modelo 1 hasta de 4.3633% con el modelo 2. En los modelos 1 y 2 la variable CLOT no es significativa. Se puede considerar que el R^2 del modelo 2 se incrementa si adicionamos los cuadrados de las últimas tres variables predictoras.

Capítulo 6
Correlaciones: Múltiple, Parcial y Parcial múltiple

6.1. Introducción

Los rasgos esenciales de la regresión de línea recta, excepto para la predicción cuantitativa proporcionada por la ecuación de regresión ajustada, pueden ser descritos en términos de los coeficientes de correlación (r). Estos rasgos se expresan como:

1. Los coeficientes de correlación al cuadrado (r^2) miden la fuerza de la relación lineal entre la variable dependiente *Y* y la variable independiente *X*. La cercanía de r^2 a 1, es la relación lineal más fuerte; lo más cercano de r^2 a cero, es la relación lineal más débil.
2. La $r^2 = (SC_{totales} - SC_{residual})/SC_{totales}$ es la reducción proporcional en la suma de cuadrados totales obtenidos usando un modelo de línea recta en *X* para predecir *Y*.
3. La $r = \hat{\beta}_1 \frac{S_x}{S_y}$, dónde $\hat{\beta}_1$ es la pendiente estimada de la línea de regresión.

4. La *r* (o r_{xy}) es un estimador del parámetro poblacional ρ (o ρ_{XY}) que describe la correlación entre *X* y *Y*, considerando ambas variables aleatorias.
5. Asumiendo que *X* y *Y* tienen una distribución bivariada, con parámetros μ_x, μ_Y, σ_x^2 y σ_y^2 y ρ_{XY}, la distribución condicional de *Y* dado *X* es N ($\mu_{Y/X}$, $\sigma_{Y/X}^2$), donde

$$\mu_{Y/X} = \mu_Y + \rho(\sigma_Y/\sigma_X)(X - \mu_X) \text{ y } \sigma_{Y/X}^2 = \sigma^2_Y(1 - \rho^2) \quad (1)$$

Aquí r^2 estima a ρ^2, que puede ser expresado como

$$\rho^2 = (\sigma_Y^2 - \sigma_{Y/X}^2) / \sigma_Y^2 \quad (2)$$

6. La *r* puede ser usada como un índice general de la asociación lineal entre dos variables aleatorias en el siguiente sentido:
   a. El valor positivo más alto de *r* es la asociación lineal más "positiva"; esto es, un individuo con un valor alto de una variable parece tener un valor alto del otro y un individuo con un valor bajo de una variable probablemente tendrá un valor bajo de la otra.
   b. El valor negativo más alto de *r* es la asociación lineal más "negativa"; esto es, un individuo con un valor alto de una variable, tiene un valor bajo de la otra y viceversa.

c. Si *r* está cercana a cero, existe una pequeña evidencia de una asociación lineal, que indica que existe una asociación no lineal o que no hay asociación.

Esta conexión entre regresión y correlación puede ser generalizada como el caso de regresión múltiple. Sin embargo, cuando varias variables independientes están involucradas, los rasgos esenciales de regresión son descritos no por un solo coeficiente de correlación como en el caso de la línea recta, sino por varios. Estos incluyen un conjunto de correlaciones tal como *r*, más un grupo total de índices adicionales (de alto orden) llamados correlaciones múltiples, correlaciones parciales y correlaciones parcial-múltiples. Estos coeficientes de correlación de alto orden nos permiten responder a muchas de las mismas cuestiones que pueden ser respondidas ajustando un modelo de regresión múltiple. Además, la correlación análoga es encontrada y usada particularmente en relaciones espurias no protegidas entre variables, identificando variables de intervención y haciendo ciertos tipos de inferencias causales.

6.2. Matriz de correlación

Cuando se procede con más de una variable independiente, la colección de todos los coeficientes de correlación de orden cero puede ser representada de manera compacta en una forma de matriz de correlación. Por ejemplo, cuando

k = 3 variables independientes X_1, X_2 y X_3 y una variable dependiente Y, hay $\binom{4}{2}$= 6 correlaciones de orden cero y la matriz de correlación tiene la forma general

$$\begin{array}{c} \\ Y \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{cccc} Y & X_1 & X_2 & X_3 \end{array} \\ \begin{pmatrix} 1 & r_{Y1} & r_{Y2} & r_{Y1} \\ & 1 & r_{Y2} & r_{Y13} \\ & & 1 & r_{Y23} \\ & & & 1 \end{pmatrix} \end{array} \tag{3}$$

Aquí r_{Y_j}(j = 1, 2, 3) es la correlación entre Y y X_j y r_{ij} (i,j = 1, 2, 3) es la correlación entre X_i y X_j. De un ejemplo se toman los resultados de la matriz de correlación,

$$\begin{array}{c} \\ Y \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{array}{c} \begin{array}{cccc} Y & X_1 & X_2 & X_3 \end{array} \\ \begin{pmatrix} 1 & .814 & .770 & .767 \\ & 1 & .614 & .615 \\ & & 1 & .994 \\ & & & 1 \end{pmatrix} \end{array} \tag{4}$$

Cada una de estas correlaciones tomadas de forma separada describen la fuerza de la relación lineal entre las dos variables involucradas. En particular, las correlaciones r_{Y_1} = 0.814, r_{Y_2} = 0.770 y r_{Y_3}= 0.767 miden la fuerza de la

asociación lineal con la variable dependiente *Y* para cada una de las variables independientes tomadas separadamente. Como podemos ver, X_1 es la variable independiente con la relación lineal más fuerte con *Y*, seguida de X_2 y X_3. Sin embargo, estas correlaciones de orden cero tampoco describen:

1. la relación total de la variable dependiente *Y* con las variables independientes X_1, X_2 y X_3, consideradas en conjunto;
2. no describen la relación entre *Y* y X_2 después de controlar X_1; y
3. no describen la relación entre *Y* y los efectos combinados de X_2 y X_3, después de controlar a X_1.

La medida que describe (1) es llamada coeficiente de correlación múltiple de *Y* sobre X_1, X_2 y X_3. La medida que describe (2) es llamada coeficiente de correlación parcial entre *Y* y X_2 controlando X_1. Finalmente, la medida que describe (3) es llamada coeficiente de correlación parcial múltiple entre *Y* y los efectos combinados de X_2 y X_3 controlando X_1.

Puede ser observado que sólo X_2 y X_3 están altamente correlacionados en el ejemplo, esto es posible si la relación general de X_2 y *Y* no es lineal, puesto que X_3 está significativamente correlacionada con *Y* después de que X_2 ha sido controlada. En realidad, esto es lo que pasa en general, cuando la adición de un término de segundo orden en regresión polinomial mejora significativamente la predicción de la variable dependiente.

6.3. Coeficiente de correlación múltiple

El coeficiente de correlación múltiple, denotado por $R_{Y/X_1,x_2,\ X_k}$ es una medida de la asociación lineal total de una variable dependiente *Y* con varias variables independientes, X_1, X_2, X_3, ..., X_k.

> "Por asociación lineal entendemos que $R_{Y/X_1,x_2,\ X_k}$mide la fuerza de la asociación entre *Y* y la combinación lineal mejor ajustada de las *X*'s, que es la solución mínimos cuadrados $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + ... + \hat{\beta}_k X_k$. En realidad ninguna otra combinación lineal de las *X*'s tendrá una correlación tan grande con *Y*. También se puede mostrar que $R_{Y/X_1,x_2,\ X_k}$ siempre es no-negativo".

Entonces, el coeficiente de correlación múltiple es una generalización directa del coeficiente de correlación simple (r) para el caso de varias variables independientes. Hemos mencionado esta medida con el nombre de R^2, que es el cuadrado del coeficiente de correlación múltiple. Dos expresiones computacionales proveen las interpretaciones útiles del coeficiente de correlación múltiple $R_{Y/X_1,x_2,.\ X_k}$ y su cuadrado. Éstas son:

$$R_{Y/,x_1,x_2,x_3\ \ldots,X_k} = \{ \sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\} /\{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \ \ \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2\}^{½} \quad (5)$$

$$R_{Y/,x_1,x_2,x_3\ \ldots,X_k} =\{ \sum_{i=1}^{n} Y_i\hat{Y} - n\bar{Y}^2\}/\{ (\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2)(\sum_{i=1}^{n} \hat{Y}_i^2 - n\bar{Y}^2)\}^{½}$$

Y la expresión

$$R^2{}_{Y/,x_1,x_2,x_3\ \ldots,X_k}=\sum_{i=1}^{n}(Y_i-\bar{Y})^2-\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2/\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \quad (6)$$

La expresión (6) es más usada para evaluar el ajuste del modelo de regresión. También puede observarse que de la expresión (5) se deriva $R_{Y/X_1,x_2,\ X_k}$ = r(Y,$\hat{Y}$), la correlación lineal simple entre el valor observado Y y los valores predichos $\hat{Y}$.

Ejemplo: sean X_1 y X_2 dos variables independientes y Y la variable dependiente. El modelo de regresión para estas variables es $\hat{Y}$ = 6.553 + 0.722X_1 + 2.050X_2, entonces se tienen los datos predichos:

Niños	Y_i	$\hat{Y}_i$
1	64	64.11
2	71	69.65
3	53	54.23
4	67	73.87
5	55	59.78
6	58	57.01
7	77	55.77
8	57	59.66
9	56	57.38
10	51	49.18
11	76	75.20
12	68	66.16

Se tiene que $\bar{Y}$ = 62.75 utilizando la expresión (5) resulta

R_{Y/X_1X_2}= (47,943.60 - 12(62.75)^2 / {[48,139 – 12(62.75)^2][47,943.544 – 12(62.72)^2]}^{½} = 0.8832

Otros dos coeficientes de correlación calculados con este conjunto de datos, que corresponden a otros modelos de regresión, son:

$R_{Y/X_1X_2X_3}$ = 0.8833

Donde $X_3 = (X_2)^2$

R_{Y/X_1X_3} = 0.8811

Se puede observar que si se aumenta una variable al modelo de regresión el coeficiente de correlación no muestra un incremento importante ($R_{Y/X_1X_2X_3}$ = 0.8833), asimismo cuando se elimina una variable X_2 y se adiciona X_3 la correlación tiende a disminuir (R_{Y/X_1X_3} = 0.8811). Por lo que el mejor modelo de acuerdo al coeficiente de correlación múltiple es el que tiene un R = 0.8832.

6.4. Coeficiente de correlación parcial

El coeficiente de correlación parcial es una medida de la fuerza de la relación lineal entre dos variables después de controlar los efectos de otras variables en el modelo. Si las dos variables de interés son *Y* y X, y las variables control son Z_1, Z_2, Z_3 ... Z_p, entonces se denotan los correspondientes coeficientes de correlación parcial por $r_{YX/Z_1Z_2Z_3\ \ Z_P}$. El orden de las correlaciones parciales depende del número de variables que se están controlando. Entonces, las parciales de primer orden tienen la forma de $r_{YX/Z}$, las parciales de segundo orden r_{YX/Z_1Z_3}; y en general las parciales de orden p tienen la forma $r_{YX/Z_1Z_2Z_3\ \ Z_P}$.

En el ejemplo anterior de *Y* con X_1, X_2, X_3, el parcial de mayor orden es el de segundo orden. Los valores de las demás correlaciones parciales se pueden calcular de los datos y se muestran en la tabla 6.1.

Si observamos la matriz de correlaciones anterior (4), se puede ver que la variable altamente correlacionada es *Y* con X_1 (r_{Y_1} = 0.814). Entonces, para las tres variables independientes se considera a X_1 como la más importante de acuerdo a la fuerza de su relación lineal con *Y*.

Tabla 6.1
Correlaciones parciales para Y con X_1, X_2, X_3.

Orden	Variables controladas	Forma de correlación	Valor calculado
1	X_1	$r_{Y,X_2/X_1}$	0.589
1	X_1	$r_{Y,X_3/X_1}$	0.580
1	X_1	$r_{X2,X3/X1}$	0.988
1	X_2	$r_{Y,X_1/X_2}$	0.678
1	X_2	$r_{Y,X_3/X_2}$	0.015
1	X_2	$r_{X1,X3/X2}$	0.060
1	X_3	$r_{Y,X_1/X_3}$	0.677
1	X_3	$r_{Y,X_2/X_3}$	0.111
1	X_3	$r_{X1,X2/X3}$	0.022
2	X_1, X_2	$r_{Y,X3/X1,X2}$	-0.015
2	X_1, X_3	$r_{Y,X2/X1,X3}$	0.131
2	X_2, X_3	$r_{Y,X1/X2,X3}$	0.679

De lo anterior surge la pregunta: después de X_1, ¿cuál es la siguiente variable más importante para la predicción lineal de Y? Puesto que el parcial de primer orden r_{YX_2/X_1} = 0.589 es más grande que el de $r_{Y,X_3/X_1}$ = 0.580, se determina que X_2 es la siguiente más importante después de que tomamos en cuenta a X_1. Si solamente se considera a la variable X_3 a la izquierda, se puede preguntar: una vez que se ha tomado en cuenta a X_1 y X_2, entonces, ¿X_3 no contribuye en nada para conocer a Y? Para responder esto se pueden observar las correlaciones parciales de segundo orden, $r_{Y,X_3/X_1X_2}$ = - 0.015. Note que la magnitud de esta correlación parcial es muy pequeña. Entonces nos inclinamos a concluir que X_3 no provee información esencial adicional acerca de Y, una vez que X_1 y X_2 han sido usadas como predictores.

El procedimiento para seleccionar variables se inicia con la variable más importante y se continúa adicionando variables en orden de importancia;

mientras se controlan las variables que ya han sido seleccionadas. A este procedimiento se le ha denominado *selección forward* o *selección hacia adelante.* Alternativamente, se tiene controlado el problema de selección de las variables trabajando hacia atrás. Por esto, iniciamos con todas las variables y se eliminan (paso a paso) variables que no contribuyen a la descripción de la variable dependiente.

6.4.1. Pruebas de significancia para correlaciones parciales

Independientemente del procedimiento usado para seleccionar variables, es necesario decidir en cada paso si un coeficiente de correlación parcial particular es significativamente diferente de cero o no. Se ha descrito cómo probar tal significancia en un contexto ligeramente diferente cuando se consideran los distintos usos de la tabla de análisis de varianza en un análisis de regresión. Cuando se desea probar si el adicionar una variable al modelo de regresión fue valioso, dado que otras variables ya están en el modelo, se utiliza una prueba parcial de F. Se puede mostrar que la prueba parcial de F es exactamente equivalente a una prueba de significancia para los correspondientes coeficientes de correlación parcial. Entonces, para probar si $r_{YX/Z_1Z_2Z_3\ \ Z_P,}$ es significativamente diferente de cero, se calcula la correspondiente F parcial $F_{(X/Z_1Z_2\ldots Z_P)}$ y se puede rechazar la hipótesis nula si el estadístico de F es mayor a un valor crítico apropiado de la distribución $F_{1,\ n\text{-}p\text{-}2}$.

La hipótesis nula para esta prueba puede ser establecida con mayor formalidad considerando la población análoga del coeficiente de correlación parcial muestral $r_{YX/Z_1Z_2Z_3\ z_P,}$. Este parámetro poblacional, usualmente denotado por $\rho_{YX/Z_1Z_2Z_3\ z_P,}$, se denomina *el coeficiente de correlación parcial poblacional*. Entonces, la hipótesis nula puede ser establecida como H_0: $\rho_{YX/Z_1Z_2Z_3\ z_P,}= 0$ y la hipótesis alternativa asociada como H_1: $\rho_{YX/Z_1Z_2Z_3\ z_P,} \neq 0$.

6.4.2. Relacionando la prueba para la correlación parcial y la prueba parcial de F

La estructura de la correlación parcial poblacional contribuye a relacionar las correlaciones de alto orden y la regresión. Por simplicidad se considera esta relación para el caso especial de dos variables independientes. La expresión para el cuadrado de ρ_{YX_1/X_2} se puede escribir

$$\rho^2_{YX_1/X_2} = (\sigma^2_{Y/X_2} - \sigma^2_{Y/X_1X_2})/\sigma^2_{y/X_1} \quad (7)$$

Entonces, el cuadrado de la correlación parcial muestral r_{YX_1/X_2} es un estimador de la reducción proporcional en la varianza condicional de *Y* dado X_2 debido a condiciones de ambos X_1 y X_2.

> "La correlación parcial ρ_{YX_1/X_2} también se puede describir como una correlación de orden cero para una distribución condicional bivariada. Si la distribución conjunta de Y y X_1 dado X_2 es una distribución bivariada. La correlación de orden cero entre Y y X_1 para esta distribución condicional es la que llamamos, ρ_{YX_1/X_2}; esto es exactamente la correlación parcial entre X_1 y Y controlando a X_2".

Entonces podemos definir una expresión para calcular el coeficiente de correlación parcial cuadrado muestral como la

$$r^2_{YX_1/X_2} = \frac{SC_{residual}(usando\ solo\ a\ X_1\ y\ X_2\ en\ el\ modelo) - SC_{residual}\ (usando\ solo\ a\ X1\ en\ el\ modelo)}{SC_{residual}(usando\ solo\ a\ X_2 en\ el\ modelo)}$$

$$= \frac{SC_{extra}\ debido\ a\ la\ adición\ de\ X_1\ al\ modelo,\ \ dado\ que\ X_2\ ya\ está\ en\ el\ modelo}{SC_{residual}\ (usando\ solo\ a\ x_2\ en\ el\ modelo)} \quad (8)$$

Dada la estructura de la expresión anterior y la discusión del estadístico de F parcial, es claro el por qué la prueba de la H_0: $\rho_{YX_1/X_2} = 0$ se aplica usando $F(X_1/X_2)$ como el estadístico de prueba.

6.4.3. Otra manera de describir las correlaciones parciales

Para calcular correlaciones parciales de primer orden se usa la fórmula:

$$r_{YX/Z} = r_{YX} - r_{YZ}\, r_{XZ}) / \sqrt{(1 - r_{YZ}^2)(1 - r_{XZ}^2)} \quad (9)$$

Por ejemplo, para calcular $r_{Y,X2/X1}$, de acuerdo a los valores de la matriz de correlaciones (4) se tiene

$$r_{YX_2/X_1} = (r_{YX_2} - r_{Y,X_1}\, r_{X_{2,X_1}}) / \sqrt{(1 - r_{YX_1}^2)(1 - r_{X_2X_1}^2)}$$

$$= [0.770 - 0.814(0.614)] / \sqrt{(1 - 0.81^2)(1 - 0.614^2)} = 0.589$$

Note que la primera correlación en el numerador es la correlación de orden cero entre *Y* y X_2. La variable controlada X_1 aparece en la segunda expresión en el numerador (donde es correlacionado separadamente con cada una de las variables *Y* y X_2) y en ambos términos en el denominador. Por el uso de la expresión anterior, se pueden interpretar los coeficientes de correlación parciales como un ajuste de los coeficientes de correlación simples para tomar en cuenta el efecto de las variables controladas. En particular, si r_{YZ} y r_{XZ} tienen el mismo signo y si se controla Z se reduce (esto es, hace menos positivo o más negativo, según sea el caso) la correlación de orden cero r_{YX} entre *Y* y *X*. Por el otro lado, si r_{YZ} y r_{XZ} tienen signos opuestos, controlando Z, se incrementa r_{YX}. Para calcular correlaciones de alto orden, simplemente repetimos esta expresión usando las parciales apropiadamente siguiendo el orden más bajo. Por ejemplo, la correlación de segundo orden es un ajuste del parcial de primer orden; que, en este caso, es un ajuste de la correlación

simple de orden cero. En particular, la expresión para calcular correlaciones parciales de segundo orden es la

$$r_{YX/Z,W} = (r_{YX/Z} - r_{YW/Z}\, r_{XW/Z}) / \sqrt{\left(1 - r^2_{YW/Z}\right)\left(1 - r^2_{XW/Z}\right)}$$

$$= (r_{YX/W} - r_{YZ/W}\, r_{XZ/W}) / \sqrt{\left(1 - r^2_{YZ/W}\right)\left(1 - r^2_{XZ/W}\right)} \qquad (10)$$

Entonces, se puede calcular r_{YX_3/X_1X_2}, esto es, de la tabla 6.1.

$$r_{YX_3/X_1X_2} = (r_{YX_3/X_1} - r_{YX_2/X_1} r_{YX_3/X_2X_1}) / \sqrt{\left(1 - r^2_{YX_2/X_1}\right)\left(1 - r^2_{X_3X_2/X_1}\right)}$$

$$= [0.580 - (0.589)(0.988)] / \{[1-(0.589)^2][1 - (0.988)^2]\}^{½} = -0.015$$

De lo anterior se concluye que:

1. La correlación parcial $r_{YX/Z_1Z_2Z_3\ \ Z_P,}$ mide la fuerza de la relación lineal entre dos variables *X* y *Y* mientras se controlan las variables Z_1, Z_2, Z_3, ..., Z_p.
2. El cuadrado de la correlación parcial $r_{YX/Z_1Z_2Z_3\ \ Z_P,}$ mide la proporción de la suma de cuadrados del residual que es tomado en cuenta por la adición de *X* a un modelo de regresión que ya involucra a Z_1, Z_2, Z_3, ..., Z_p, esto es

$$r^2_{YX/Z_1Z_2Z_3\ldots Z_p} =$$

$$\frac{SC\ extra\ debido\ a\ la\ adición\ de\ X\ al\ modelo,que\ ya\ contiene\ Z_1,Z_2,Z_3,..Zp}{SC\ residual\ (usando\ solo\ a\ Z_1,Z_2,Z_3,\ldots\ldots Z_p\ en\ el\ modelo)} \quad (11)$$

3. El coeficiente de correlación parcial $r_{YX/Z_1Z_2\ldots Z_p}$ es un estimador del parámetro poblacional $\rho_{YX/Z_1Z_2\ldots Z_p}$ que es la correlación entre *Y* y *X* en la distribución conjunta de *Y* y *X* dado Z_1, Z_2, Z_3, ..., Z_p. También, el cuadrado de este coeficiente de correlación parcial poblacional está dado por la expresión equivalente

$$\rho^2_{YX/Z_1Z_2Z_3\ldots Z_p} = (\sigma^2_{Y/Z_1Z_2Z_3\ldots Z_p} - \sigma^2_{Y/X,Z_1Z_2Z_3\ldots Z_p}) / \sigma^2_{Y/Z_1Z_2Z_3\ldots Z_p} \quad (12)$$

Donde

$$\sigma^2_{Y/Z_1Z_2Z_3\ldots Z_p}$$

es la varianza de la distribución condicional de Y dado Z_1, Z_2, Z_3... Z_p.

4. El estadístico parcial F; F(*X*/ Z_1, Z_2, Z_3, ..., Z_p) se usa para probar

H_0: $\rho_{YX/Z_1Z_2Z_3\ldots\ldots Z_p} = 0$

5. El coeficiente de correlación parcial (de primer orden) $r_{YX/Z}$ es un ajuste de la correlación de orden cero r_{YX} que toma en cuenta el efecto de la variable control Z. Esto se puede ver en la expresión

$$r_{YX/Z} = (r_{YX} - r_{YZ}\, r_{XZ}) / \sqrt{(1 - r_{YZ}^2)(1 - r_{XZ}^2)} \qquad (13)$$

Las correlaciones parciales de alto orden son calculadas aplicando esta expresión usando parciales de orden más bajos.

6. La correlación parcial $r_{YX/Z}$ puede ser definida como la correlación de los residuales de la regresión de la línea recta de *Y* sobre *X* y de *X* sobre Z.

6.5. Correlación parcial – múltiple

El coeficiente de correlación parcial – múltiple se usa para describir la relación total entre una variable dependiente *Y* y dos o más variables independientes, mientras éste está controlado para las otras variables. Por ejemplo, suponga que se tiene dos variables independientes X_1 y X_2, pero nos interesan los términos al cuadrado de $(X_1)^2$ y $(X_2)^2$,; entonces se tendría un modelo como

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X^2_1 + \beta_{22} X^2_2 + \beta_{12} X_1 X_2 + \varepsilon \qquad (14)$$

Denominaremos a tal modelo como *modelo completo de segundo orden*, ya que incluye todas las posibles variables en términos de segundo orden. Para tal modelo completo, frecuentemente se pregunta si cualquiera de los términos de segundo orden es importante; en otras palabras, si un modelo de primer orden involucrando solamente a X_1 y X_2 es el adecuado. Existen dos formas equivalentes para representar esta cuestión como un problema de prueba de hipótesis. Una manera es probar H_0: $\beta_{11} = \beta_{22} = \beta_{12} = 0$. La otra es probar la hipótesis H_0: $\rho_{Y(X_1^2,X_2^2,X_1X_2)/X_1,X_2} = 0$, donde $\rho_{Y(X_1^2,X_2^2,X_1X_2)/X_1,X_2}$ es la correlación parcial múltiple poblacional de Y con las variables de segundo orden, controlando los efectos de las variables de primer orden. Este parámetro es estimado por la correlación parcial múltiple muestral $r_{Y(X_1^2,X_2^2,X_1X_2)/X_1,X_2)}$. Esta medida describe la contribución múltiple total de adicionar los términos de segundo orden al modelo, luego de los efectos de los términos de primer orden que son parcializados o controlados.

6.6. Ejercicios

6.1 La matriz de correlaciones obtenidas para las variables *Y*, X_1, X_2, y X_3 son:

	Y	X_1	X_2	X_3
Y	1	.7752	.2473	.7420
X_1	.7752	1	-0.1395	.8028
X_2	.2473	-0.1395	1	-0.714
X_3	.7420	.8028	-0.0714	1

a) Basándose en esta matriz, responda: ¿cuál de las variables independientes explica la mayor proporción de la variación total en la variable dependiente Y?

b) Determine las correlaciones parciales r_{YX_2/X_1} y r_{YX_3/X_1}.

c) Pruebe la significancia de r_{YX_2/X_1} usando los resultados del ANVA de la tabla del ejercicio 1 del capítulo 5. Exprese la hipótesis nula en términos de un coeficiente de correlación parcial poblacional.

d) Determine la parcial de segundo orden r_{YX_3/X_1X_2} y pruebe su significancia.

e) Basándose en los resultados de los incisos a – d, responda: ¿cómo ordena las variables independientes por importancia para predecir *Y*? ¿Cuáles de estas variables no son relativamente importantes?

f) Calcule la correlación parcial – múltiple cuadrada $r^2_{YX_3X_2/X_1}$ usando la tabla de ANVA. Pruebe la significancia de esta correlación. ¿Altera este resultado

su selección anterior en (e) de las variables que son incluidas en el modelo de regresión?

Modelo 1			Modelo 2			Modelo 3		
Fuente	**G.L.**	**SC**	**Fuente**	**G. L.**	**SC**	**Fuente**	**G. L.**	**SC**
Regres(X_1)	1	3861.630	Regres(X_1,X_2)	2	4689.684	Regres	3	4889.826
Residual	30	2564.338	Residual	29	1736.285	Residual	28	1536.143

6.2. Usando los resultados del ANVA, que define la regresión entre los resultados del examen de adolescentes de secundaria (Y) con los gastos por alumnos (X_1), porcentaje de profesores con grados de maestrías y doctorados (X_2), y la proporción alumnos/profesor (X_3), pruebe lo siguiente:

a)H_0: $\rho_{YX_3/X_1} = 0$ b) H_0: $\rho_{YX_3X_1/X_3} = 0$ c) H_0: $\rho_{YX_2X_3/X_1} = 0$

b) Basado en estos resultados, y asumiendo que X_1 es un predictor importante de *Y*, ¿cuál variable incluirá en el modelo de regresión?

Fuente	**G.L**	**SC**
Regresión (X_1)	1	18953.08
(X_3/X_1)	1	7010.03
(X_2/X_1, X_3)	1	10.93
Residual	21	2248.23
Total	24	28222.23

6.3. Usando el ANVA del ejercicio 5.7 que define la regresión entre la tasa de mortalidad por cáncer (Y) con el índice de polución del aire (X_1), la edad (X_2) y la fuerza de trabajo empleada en una industria (X_3): pruebe lo siguiente:

a) H_0: $\rho_{YX_2/X_1} = 0$

b)H_0:$\rho_{YX_3X_1X_2}$=0

c) H_0: "La adición de X_2 y X_3 al modelo que contiene X_1 ¿no mejora significativamente la predicción de Y"?

d) Basándose en estos resultados, ¿qué variables considera importantes como predictores de Y? Use un α = 0.01.

e) Establezca el inciso "a" en términos de pruebas equivalentes de correlaciones semiparciales.

Fuente	G.L.	SC
X_1	1	1523.658
X_2/X_1	1	181.743
$X_3/X_1, X_2$	1	130.529
Residual	19	551.723
Total	22	2387.653

6.4. Usando la tabla del ANVA dada en el ejercicio 5.8, que considera la regresión entre la erosión del suelo y tres variables independientes, responda las siguientes cuestiones relacionadas con estos factores.

Stepwise *				Ajustando primero X_1 después X_2 y X_3		
Fuente	**G.L**	**SC**		**Fuente**	**G.L.**	**SC**
Regresión (X_2)	1	667.728		Regresión (X_1)	1	640.425
X_3/X_2	1	5.8228		X_2/X_1	1	32.7819
$X_1/X_3, X_2$	1	6.9405		$X_3/X_1, X_2$	1	7.2844
Residual	7	16.0943		Residual	7	16.0943

*El orden de importancia relativo.

Y X_1 X_2 X_3

$$R = \begin{array}{c} \\ Y \\ X_1 \\ X_2 \\ X_3 \end{array} \begin{pmatrix} 1 & 0.959 & .979 & -0.904 \\ & 1 & 0.951 & -0.819 \\ & & 1 & -0.879 \\ & & & 1 \end{pmatrix}$$

a) Usando la matriz de correlación calcule r_{YX_2/X_1} y r_{YX_3/X_1}.

b) Basándose en los resultados obtenidos en el inciso anterior, ¿qué variables seguirán integrándose al modelo de regresión que ya contiene a X_1?

c) Pruebe H_0: $\rho_{YX_2/X_1} = 0$ usando la prueba de t.

d) Determine la correlación cuadrada parcial múltiple $r^2_{Y(X_2,X_3)/X_1}$ y pruebe H_0: $\rho_{Y(X_2,X_3)/X_1} = 0$.

6.5. Usando los resultados del análisis en el ejercicio 5.9 realice:

a) Pruebe H_0: $\rho_{YX_1} = 0$ y H_0: $\rho_{YX_2} = 0$.

b) Pruebe H_0: $\rho_{YX_1/X_2} = 0$ y H_0: $\rho_{YX_2/X_1} = 0$.

c) Basándose en los resultados de los incisos anteriores, ¿qué variables son incluidas en el modelo de regresión?, y ¿cuál es el orden relativo de importancia?

6.6. Para los datos del ejercicio 6.3, calcule los estimadores de lo siguiente:

a) $r^2_{YX_1}$

b) R^2_{Y/X_1X_2}

c) $R^2_{Y/X_1X_2X_3}$

d) $r^2_{YX_3/X_1X_2}$

e) $r^2_{YX_2/X_1}$

6.7. Considerando los resultados del ejercicio 6.4 responda lo siguiente:

a) Calcule un estimador de $r^2_{YX_1/X_3X_2}$.

b) ¿Cuáles son los dos modelos que son comparados en la prueba si la correlación en el inciso (a) es cero en la población?

c) Calcule un estimador para $r^2_{YX_1}$.

d) ¿Cuál es la diferencia entre el inciso (a) y el (c) acerca de la relación entre las tres variables de predicción?

e) ¿Cuáles son los dos modelos que son comparados en una prueba si la correlación entre (a) y (c) difiere en la población?

Capítulo 7
Confusión e interacción en regresión

7.1. Introducción

El análisis de regresión tiene dos tareas importantes que son: (1) predecir la variable dependiente usando un conjunto de variables independientes y (2) cuantificar la relación de una o más variables independientes con una variable dependiente. Estas tareas difieren porque la primera se orienta a encontrar un modelo que ajusta los datos observados y predice datos futuros tanto como sea posible; mientras que la segunda se refiere a generar un estimador adecuado de uno o más coeficientes de regresión en el modelo. La segunda tarea también es de interés cuando las cuestiones de investigación están relacionadas con la etiología de las enfermedades, lo cual resulta muy complejo para identificar una o más causas determinantes en una enfermedad u otro evento relacionado con la salud.

La confusión e interacción son dos conceptos metodológicos relevantes para alcanzar la segunda tarea; en este capítulo ambos se describen usando terminología de regresión. Se inicia con una revisión general de dichas nociones y después se discuten las formulaciones de regresión para cada uno. Asimismo, se describe un procedimiento popular de regresión, el análisis de

Covarianza (ANCOVA), que puede ser usado para ajustar o corregir los problemas de confusión. Además, se describe ampliamente una estrategia para obtener un mejor modelo de regresión que incorpora la evaluación de ambos conceptos, la confusión y la interacción.

7.2. Confusión e interacción

La confusión e interacción, aún cuando son dos conceptos diferentes, involucran la evaluación de una asociación entre dos o más variables y toma en cuenta los efectos de variables adicionales que pueden afectar tal asociación. La medida de asociación que es seleccionada depende de las características de las variables de interés. Por ejemplo: si ambas variables son continuas, como en el contexto clásico de regresión, la medida de asociación típica es un coeficiente de regresión. Las variables adicionales son referidas como variables extrañas, variables control o covariables. La cuestión esencial relacionada con estas variables es el cómo son incorporadas en el modelo, con el cual la asociación de interés pueda ser estimada.

En términos prácticos, se considera un estudio para evaluar si el nivel de actividad física está asociado con la presión sistólica sanguínea tomando en cuenta la edad. La variable extraña es la edad. Se requiere determinar si se puede ignorar la edad en este análisis y calcular de forma correcta la asociación entre las dos variables. En particular, es necesario preguntar las siguientes cuestiones: (1) ¿es el estimador de la asociación entre las dos

variables significativamente diferente si ignoramos la edad?, y (2) ¿es el estimador de la asociación entre las dos variables significativamente diferente para distintos valores de edad? La primera cuestión está relacionada con la confusión, la segunda con la interacción.

En general, existe la discusión acerca de si diferentes interpretaciones de la relación de interés resultan significativas cuando una variable extraña es ignorada o incluida en el análisis de datos. En la práctica, para valorar una confusión es necesario comparar un estimador crudo de una asociación (que ignora las variables extrañas de interés) y un estimador ajustado de la misma asociación (que toma en cuenta de alguna manera las variables extrañas). Si los estimadores crudo y ajustado son significativamente diferentes, entonces se dice que la confusión está presente y una o más variables extrañas deben ser incluidas en nuestro análisis de datos. Se puede observar que esta definición no requiere de una prueba estadística.

Por ejemplo, usando la ilustración anterior, un estimador crudo de la relación entre las dos variables (ignorando la edad) es dado por el coeficiente de regresión $\hat{\beta}_1$, de la variable independiente (nivel de actividad física) en un modelo de línea recta que predice la variable dependiente (presión sistólica sanguínea) usando la variable independiente. En contraste, un estimador ajustado está dado por el coeficiente de regresión $\hat{\beta}_1$* de la variable independiente, en el modelo de regresión múltiple que predice a la variable dependiente, usando las dos variables (nivel de actividad física y edad). En particular, si el nivel de actividad física se define como una variable dicotómica (de 1 o 0 para alto y bajo nivel de actividad física respec-

tivamente); entonces el estimador crudo es simplemente la diferencia entre la media de la presión sistólica sanguínea en cada grupo de actividad física y el estimador ajustado que representa una diferencia ajustada en estas dos medias de presión sistólica sanguínea que controlan la edad. En general, la confusión está presente si existe cualquier diferencia significativa entre los estimadores crudo y el ajustado. La interacción es la condición donde la relación de interés es distinta a diferentes niveles de las variables extrañas. En contraste con la confusión, al valorar la interacción no se considera a cada uno como estimador crudo o estimador ajustado; en su lugar se enfoca a describir la relación de interés en diferentes valores de las variables extrañas. Por ejemplo, valorando la interacción cuando se describe la relación (niveles de actividad física y presión sistólica sanguínea), se decide si alguna descripción varía según los diferentes valores de edad (si la relación es fuerte en adultos y débil en jóvenes). Si la relación entre ambas variables varía con la edad, entonces se dice que existe una interacción entre edad y el nivel de actividad física de las personas. Para valorar la interacción, se puede emplear una prueba estadística; además de evaluar subjetivamente la significancia de un efecto de interacción estimado. Cuando la confusión e interacción se consideran para el mismo conjunto de datos, el uso de un estimador total (ajustado) tiende a disfrazar cualquier efecto de interacción que pueda estar presente. Por ejemplo, si la asociación entre las dos variables anteriores difiere significativamente de los diferentes valores de edad, el uso de un solo estimador total, tal como lo sería el coeficiente de regresión de la variable independiente en un modelo de regresión múltiple que contenga edad y nivel

de actividad física, oculta esta interacción. Esto ilustra el siguiente principio básico: *la interacción es valorada antes que la confusión, el uso de un estimador (ajustado) que controla la confusión se recomienda solamente cuando no existe una interacción significativa.* En general la confusión y la interacción son dos fenómenos diferentes. Una variable puede manifestar ambos, ninguna de las dos o solamente una de las dos. No obstante, si se encuentra una fuerte interacción, el ajuste para la confusión puede ser inapropiado.

7.3. Interacción en regresión

Es posible describir cómo dos variables independientes pueden interactuar y afectar a una variable dependiente y cómo tal interacción se representa por un modelo de regresión. Para ilustrar el concepto de interacción, considere el siguiente ejemplo: suponga que se desea determinar cómo las variables independientes, temperatura (T) y concentración de catálisis (C), afectan conjuntamente la tasa de crecimiento (Y) de organismos en cierto sistema biológico. Además, suponga que dos niveles de temperatura (T_0, T_1) y dos niveles de concentración de catálisis (C_0, C_1) son estudiados en un experimento en el cual se obtienen observaciones de *Y* para las cuatro combinaciones de temperatura – concentración de catálisis, estos son (T_0, C_0), (T_0, C_1), (T_1, C_0) y (T_1, C_1).

> “En términos estadísticos este experimento se llama experimento factorial completo, ya que las observaciones de *Y* se obtienen de las cuatro combinaciones. La ventaja de un experimento factorial, es que cualquier efecto de interacción que exista puede ser detectado y medido de forma eficiente”.

Ahora, considere dos tipos de gráficas que se basan en dos grupos de datos hipotéticos para el experimento establecido anteriormente. La figura 7.1a, sugiere que la tasa de cambio en la tasa de crecimiento, como una función de la temperatura, es la misma observada del nivel de concentración de catálisis; en otras palabras, la relación entre *Y* y *T* no dependen de ninguna manera de *C*. Es importante puntualizar que no estamos diciendo que *Y* y *C* no están relacionadas, sino que la relación entre *Y* y *T* no varía como una función de *C*; cuando éste es el caso, decimos que *T* y *C* no tienen interacción, no existe el efecto de la interacción *T* x *C*. Esto significa que estamos investigando los efectos de *T* y *C* sobre *Y* independientemente una de otra, y que estamos legítimamente hablando de separar los efectos de *T* y *C* sobre *Y*. Una manera de cuantificar la relación mostrada en la figura 7.1a es con un modelo de regresión de la forma

$$\mu_{Y/T,\,C} = \beta_0 + \beta_1 T + \beta_2 C \qquad (1)$$

Aquí, el cambio en la media de *Y* para un cambio en una unidad de *T* es igual a β_1, del nivel de *C*, en realidad, cambiando los niveles de *C* en el modelo se

tiene solamente el efecto de mover la línea recta relacionando a $\mu_{Y/T,C}$ y *T* sin afectar el valor de la pendiente β_1, como vemos en la figura 7.1a, en particular $\mu_{Y/T,C_0} = (\beta_0 + \beta_2 C_0) + \beta_1 T$ y $\mu_{Y/T,C_1} = (\beta_0 + \beta_2 C_1) + \beta_1 T$. En general, cuando se dice que no hay interacción es sinónimo de paralelismo en el sentido de la curva de respuesta *Y* versus *T* para los valores fijos de *C* que son paralelos; en otras palabras, estas curvas de respuesta tienen la misma forma general, difiriendo una de otra solamente por las constantes aditivas independientes de *T* (figura 7.2).

En contraste, la figura 7.1b muestra la situación donde la relación entre *Y* y *T*, dependen de *C*; en particular, parece que *Y* se incrementa con aumentos en *T* cuando $C = C_0$ pero disminuye con incrementos en *T* cuando $C = C_1$. En otras palabras, el comportamiento de *Y* como una función de la temperatura no puede ser considerado independiente de la concentración de la catálisis. Cuando éste es el caso, decimos que *T* y *C* interactúan, o que existe un efecto de interacción *T* x *C*.

Prácticamente hablando, esto significa que no tiene mucho sentido hablar de los efectos separados de *T* y *C* sobre *Y*, ya que *T* y *C* no operan de forma independiente una de otra en sus efectos sobre *Y*.

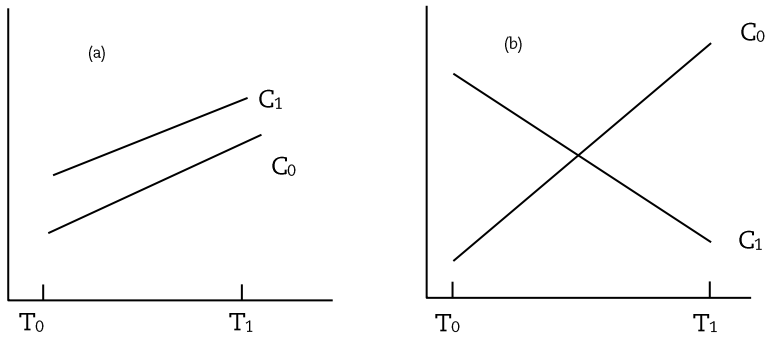


Figura 7.1. (a) No interacción, (b) interacción.

Otra manera de representar matemáticamente un efecto de interacción es considerar un modelo de regresión de la forma

$$\mu_{Y/T,C} = \beta_0 + \beta_1 T + \beta_2 C + \beta_{12} TC \tag{2}$$

Aquí el cambio en el valor medio de *Y* para un cambio en una unidad en *T* es igual a $\beta_1 + \beta_{12}TC$, que claramente depende de los niveles de *C*. En otras palabras, introduciendo un término de producto, tal como, $\beta_{12}TC$, en un modelo de regresión, es una manera de contar el hecho de que los dos factores *T* y *C* no operan de forma independiente uno de otro. En este ejemplo, se tiene que cuando $C = C_0$ el modelo es

$$\mu_{Y/T,C} = (\beta_0 + \beta_2 C_0) + (\beta_1 + \beta_{12} C_0)T \tag{3}$$

Cuando $C = C_1$, entonces,

$$\mu_{Y/T,C} = (\beta_0 + \beta_2 C_1) + (\beta_1 + \beta_{12} C_1)T \quad (4)$$

En particular, la figura 7.1b sugiere que el efecto de interacción β_{12} es negativo, con el efecto lineal ($\beta_1 + \beta_{12} C_0$) de *T* en C_0 siendo positivo y el efecto lineal ($\beta_1 + \beta_{12} C_1$) de *T* en C_1, negativo. Un efecto de interacción negativo es esperado aquí en el ejemplo, ya que la figura 7.1b sugiere que la pendiente de la relación lineal entre *Y* y *T* decrece (va de positivo a negativo), *C* cambia de C_0 a C_1. De esta manera, es posible que β_{12} sea positivo; en este caso el efecto de interacción se manifesta con un gran valor positivo para la pendiente cuando $C = C_1$ más que cuando $C = C_0$.

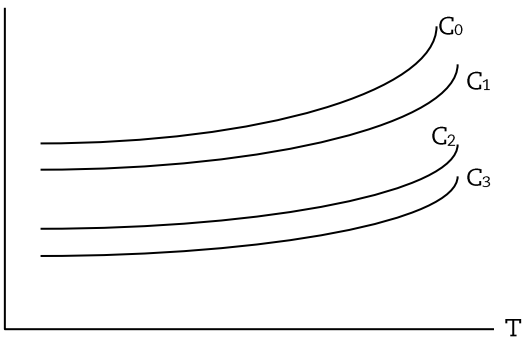


Figura 7.2. No hay efecto de interacción.

7.3.1. Modelando la interacción

Como en la ilustración de la figura 7.1b, que sugiere una interacción entre variables independientes, se puede escribir en forma general en términos de

un modelo de regresión que involucra términos de productos. Desafortunadamente, no hay reglas precisas para especificar tales términos. Por ejemplo; la interacción que involucra tres variables X_1, X_2 y X_3, el modelo sería

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3 + \varepsilon \quad (5)$$

En este modelo los productos de dos factores de la forma $X_i X_j$ se denominan interacciones de primer orden; de la misma manera, los productos de los factores $X_1 X_2 X_3$ se denominan interacciones de tercer orden y productos de mayor orden. Las interacciones de mayor orden son difíciles de interpretar. El modelo anterior (5) no es el modelo más general posible, cuando se consideran las tres variables X_1, X_2 y X_3. Los términos de productos adicionales tales como $X_i X^2_j$, $X_i X^3_j$, $X^2_i X^2_j$ se pueden incluir. Sin embargo, hay un límite del número total de términos. El modelo no puede contener más de n – 1 variables independientes cuando *n* es el número de observaciones en los datos. No puede ser posible ajustar de forma confiable un modelo con más de n – 1 variables; si alguna de las variables (productos de alto orden) está ligeramente correlacionada con otras variables en el modelo, éste es el caso, cuando el modelo contenga varios términos de interacción. Este problema se denomina *colinealidad*.

El modelo anterior puede, por otro lado, ser considerado muy general si se enfoca a las interacciones de interés. Por ejemplo, si el propósito de un estudio es describir la relación entre X_1 y Y controlando las posibles

confusiones y/o efectos de interacción de X_2 y X_3, el siguiente modelo puede ser más eficiente que el modelo anterior

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \varepsilon \quad (6)$$

Los términos X_1X_2 y X_1X_3 describen la interacción entre X_2 y X_3 respectivamente, con X_1. En contraste, el término X_2X_3 que no está incluido en el modelo anterior no mantiene interacción con X_1. Usando pruebas estadísticas para evaluar las interacciones en un modelo de regresión dado, se dispone de un número de opciones. Una aproximación es probar globalmente la presencia de cualquier clase de interacción y si la interacción es significante, identificar los términos de interacción importante usando otras pruebas. Una segunda manera para evaluar interacciones es probar la interacción en una secuencia jerárquica, iniciando con los términos de alto orden procediendo secuencialmente a términos de bajo orden si los términos de mayor orden no son significativos.

7.4. Confusión en regresión

Se ha enfatizado anteriormente que la valoración de la confusión es cuestionable en la presencia de interacción. Entonces, en esta discusión de confusión se asume que no existe interacción.

Se asume el interés por describir la relación entre una variable independiente *T* y una variable aleatoria dependiente *Y*, tomando en cuenta el posible efecto de confusión de una tercer variable *C*. Para valorar la confusión se requiere comparar un estimador crudo de la relación *T – Y*, que ignora el efecto de la variable control *C*, con un estimador de la relación que toma en cuenta esta variable. Esta comparación se puede mostrar en términos de dos modelos, esto es

$$Y = \beta_0 + \beta_1 T + \beta_2 C + \varepsilon \quad (7)$$

$$Y = \beta_0 + \beta_1 T + \varepsilon \quad (8)$$

El supuesto de no interacción de *T* x *C* precisa la necesidad de considerar el término producto de la forma *TC* en este modelo. En el primer modelo la relación entre *T* y *Y* ajustada por la variable *C* que se expresa en términos del coeficiente de regresión (parcial) β_1 de la variable *T*. El estimador de β_1 puede ser denotado por $\hat{\beta}_{1/C}$ obtenido en el ajuste del modelo por mínimos cuadrados, como una medida del efecto ajustado, en el sentido que da un cambio estimado en *Y* por unidad de cambio en *T* tomando en cuenta a *C* (esto es con *C* en el modelo). Un estimador crudo de la relación *T* - *Y* es el coeficiente estimado de *T*, ($\hat{\beta}_1$) basado en el segundo modelo (8), un modelo que no involucra a la variable *C*. Entonces se obtiene la siguiente regla general para evaluar la presencia de confusión, cuando solamente una variable está controlada. La confusión está presente si el estimador del coeficiente (β_1) de la variable de estudio *T* cambia significativamente, cuando la variable *C* es

removida del modelo (primer modelo), esto es si, $\hat{\beta}_{1/C} \neq \hat{\beta}_1$ donde $\hat{\beta}_{1/C}$ denota el estimador ajustado de β_1 usando el primer modelo y $\hat{\beta}_1$ denota el estimador crudo de β_1 usando el segundo modelo.

El signo ≠ (diferente) en la expresión anterior indica que se requiere una decisión subjetiva como si los dos estimadores fueran significativamente diferentes; esto es, se necesita determinar subjetivamente si los dos estimadores describen una interpretación diferente de la asociación *T* – *Y*. Como un ejemplo, considere que *Y* denota la presión sistólica sanguínea, *T* denota el nivel de actividad física y *C* denota la edad. Para un conjunto de datos, suponer que se encontró que

$$\hat{\beta}_{1/C} = 4.1 \quad \text{y} \quad \hat{\beta}_1 = 15.9$$

Entonces, se puede concluir que un cambio de una unidad en *T*, produce un cambio de 15.9 unidades en *Y*, cuando se ignora *C* (la edad), mientras, que cuando *C* es controlada, un cambio de una unidad en *T*, produce solamente un cambio de 4.1 unidades en *Y*; esto es la asociación entre *T* y *Y* es mucho más débil después de controlar *C* (la edad). Por estos resultados se puede decir que *C* (la edad) es un factor de confusión en el análisis. Como otro ejemplo, suponga que

$$\hat{\beta}_{1/C} = 6.2 \quad \text{y} \quad \hat{\beta}_1 = 6.1$$

En este caso, nos inclinaríamos a decir que *C* (la edad) no es factor de confusión porque no existe diferencia significativa entre los estimadores 6.2 y 6.1. Desafortunadamente, un investigador puede hacer comparaciones tales como $\hat{\beta}_{1/C} = 4.1$ *versus* $\hat{\beta}_1 = 5.5$. Cuando se comparan numéricamente tales estimadores, se puede considerar la importancia clínica de la diferencia numérica entre estimadores basados en el conocimiento de las variables involucradas. Por ejemplo, los coeficientes estimados de 4.1 y 5.5, respectivamente, corresponden a las diferencias ajustadas y crudas en presiones sanguíneas medias, entre grupos de alta y baja actividad física; es importante decidir si una diferencia de medias de 5.5 es clínicamente más importante que una diferencia de medias de 4.1. Una aproximación a este problema es controlar cualquier variable (como un factor de confusión) que cambia el efecto del estimador crudo, por alguna cantidad preespecificada por un juicio clínico.

Antes de definir criterios para la confusión, involucrando varias variables, se puede comentar el problema práctico de decidir qué tipo de variables serán consideradas como factores potenciales de confusión. Aunque la respuesta aquí puede ser debatible, se toma la posición de que una lista elegible de variables debe ser construida en base a conocimiento previo y/o investigación, acerca de la relación de la variable dependiente en cada covariable bajo consideración. En particular, se recomienda que solamente las variables conocidas que sean razonablemente predictivas de la variable dependiente sean consideradas como factores potenciales de confusión y/o modificadoras de efectos. En términos epidemiológicos, tales variables son

denominadas como factores de riesgo (Kleinbaum *et al*, 1982). La idea aquí es restringir la atención al control de las variables extrañas, que el investigador anticipa pueden influenciar la relación hipotética entre *T* y *Y*, que son estudiadas. Para desarrollar la lista, los investigadores tienen que hacer una decisión subjetiva.

7.5. Conclusiones

La confusión e interacción son dos conceptos metodológicos pertinentes a la valoración de una relación entre variables dependientes e independientes. La interacción que antecede a la confusión, existe cuando la relación de interés es diferente a diferentes niveles de variables extrañas. En regresión lineal, la interacción se evalúa usando pruebas estadísticas del término producto involucrando variables independientes básicas en el modelo. La confusión no se evalúa con pruebas estadísticas, está presente cuando el efecto de interés difiere, dependiendo de si una variable extraña es ignorada o retenida en el análisis. En términos de regresión, la confusión se evalúa comparando coeficientes de regresión crudos y ajustados de diferentes modelos.

Cuando varios factores potenciales de confusión se consideran, puede ser apreciable identificar factores que no producen confusión, que pueden ser permitidos en el modelo para tener precisión, esto no puede ser posible en algunas situaciones. Cuando existe una fuerte interacción que involucra a ciertas variables extrañas, valorar la confusión para estas variables extrañas

es irrelevante. También, en tal situación, la valoración de la confusión involucra otras variables extrañas, aunque posiblemente, esto es complejo subjetivo. Por lo que la valoración de la confusión no se recomienda cuando se han identificados efectos importantes de interacción.

7.6. Ejercicios

7.1 Considere los resultados del análisis que describen la regresión que involucra dos variables independientes X_1 y X_2 y una variable dependiente Y. Considere que se desea evaluar la relación de X_1 con Y controlando los posibles efectos confundidos de X_2.

Fuente	G.L.	SCM	Parámetro	Estimador	T para H_0
Regresión (X_1)	1	121.000	Intercepto	6.750	5.60(p =.0001)
Residual	14	11.643	β_1	5.500	3.22(p =.0061)
Regresión (X_1, X_2)	2	134.00	Intercepto	5.00	11.80(p =.0001)
Residual	13	1.231	β_1	2.00	3.12(p=.0081)
			β_2	7.00	10.93(p=.0001)

a) Usando un coeficiente de regresión apropiado, como una medida de asociación, determine si existe confusión. Explique.
b) Suponga que la confusión ha sido determinada, haga una comparación de coeficientes de correlación (parcial) crudo y ajustado. ¿Qué conclusiones obtiene?
c) ¿Cuál es el mensaje de este ejemplo?

7.2. Considere los resultados del análisis que describe la regresión que involucra dos variables independientes X_1 y X_2 y una variable dependiente Y (usando diferente conjunto de datos). Los resultados son:

$$\begin{array}{c} \\ Y \\ X_1 \\ X_2 \end{array} \begin{array}{c} \begin{array}{ccc} Y & X_1 & X_2 \end{array} \\ \begin{pmatrix} 1 & .26491 & .92717 \\ & 1 & 0.000 \\ & & 1 \end{pmatrix} \end{array}$$

Matriz de correlaciones

$r_{YX_1/X_2} = 0.707$

Tablas de ANVA y estimador de los parámetros

Fuente	**G.L**	**SCM**	**R^2**	**Parámetro**	**Estimador**	**Valores de t**
Regresión(X_1)	1	8.000	0.0702	Intercepto	8.500	4.04
Residual	6	17.667		β_1	2.000	0.67
Regresión(X_1, X_2)	2	53.000	0.9298	Intercepto	5.000	6.45
Residual	5	1.600		β_1	2.000	2.24
				β_2	7.000	7.83

a) Responda las mismas preguntas del ejercicio 1 del inciso a al c.

b) ¿Qué ilustra este ejemplo en relación al uso de la prueba de hipótesis H_0: $\beta_2 = 0$, para valorar la confusión?

7.3 Se condujo un experimento para analizar cuantitativamente los factores encontrados en una lipoproteína de alta densidad dentro del suero sanguíneo humano. Se asociaron tres variables predictoras con la variable respuesta Y; el colesterol total (X_1) y los triglicéridos totales (X_2) concentrados en la muestra, más la presencia o ausencia de un cierto componente prebeta o SPB (X_3) que fue codificado como 0 en su ausencia y como 1 si estaba presente. Los análisis son mostrados en las siguientes tablas de ANVA.

Fuente	**G.L**	**SC**	**Fuente**	**G.L.**	**SC**
Regresión (X_1)	1	43.2653	Regresión (X_2)	1	21.3397
Residual	40	4567.3835	Residual	40	4592.2793

Fuente	G.L	SC	Fuente	G.L.	SC
Regresión (X_3)	1	735.2054	Regresión (X_1, X_2)	2	135.3820
Residual	40	3878.4136	Residual	39	4478.2369
Fuente	G.L	SC	Fuente	G.L.	SC
Regresión (X_1, X_3)	2	783.1691	Regresión (X_2, X_3)	2	737.8069
Residual	39	3830.450	Residual	39	3875.8122

Fuente	G.L	SC	Fuente	G.L.	SC
Regresión (X_1, X_2, X_3)	3	819.7473	Regresión (X_1, X_3)	2	783.1691
X_1, X_3, X_2/X_1, X_2, X_3	2	74.7443	X_1, X_3/X_1, X_3	1	62.4247
Residual	36	3719.0517	Residual	38	3768.0252

a) Pruebe si X_1, X_2 o X_3 ayudan significativamente en la predicción de Y.

b) Pruebe si X_1, X_2 y X_3 ayudan significativamente para predecir Y.

c) Pruebe si los coeficientes verdaderos de los términos productos X_1X_3 y X_2X_3 son simultáneamente cero en el modelo que contiene a X_1, X_2 y X_3, más estos términos productos. Establezca la hipótesis nula en términos de un coeficiente de correlación múltiple parcial. Si esta prueba no se rechaza, ¿qué se puede concluir acerca de la relación de Y con X_1 y X_2 cuando $X_3 = 1$ comparado cuando $X_3 = 0$?

d) Basado en la información anterior, ¿se puede valorar la confusión de X_1 o X_2 cuando se evalúa la relación de X_3 y Y? Explique.

Capítulo 8
Transformación de los datos

8.1. Introducción

En muchas aplicaciones es apropiado desarrollar el modelo lineal en términos de funciones de los datos observados. En este capítulo se discuten los métodos para determinar la forma funcional de la transformación y el subsecuente análisis. Se tratan los procedimientos más comunes para mejorar las unidades de medidas de los datos y así obtener los mejores modelos ajustados para la predicción de nuevas observaciones de la variable independiente.

8.2. Necesidades de la transformación

El examen de gráficos de residuales, gráficos q ~ q y otros procedimientos diagnósticos pueden sugerir que los supuestos del modelo no se cumplen. Las violaciones más comunes son:

a. La expresión de los valores esperados de la respuesta no es correcta.

b. La varianza no es constante en el rango de los datos.
c. Los datos no están normalmente distribuidos.
d. Puede ser mejor desarrollar el modelo en términos de alguna función de las respuestas.

La determinación de la forma correcta de la función de respuesta requiere que se tengan observados todos los predictores relevantes y que sean usados en la forma funcional correcta. En general, la determinación de la función correcta de los insumos requiere mucha experiencia y esfuerzo. La falta de homogeneidad de varianza es una de las violaciones más comunes. Esto puede ser detectado en las gráficas de residuales, donde la variabilidad se incrementa con la magnitud de la respuesta. Alternativamente, la variabilidad puede depender de una o más variables de insumos (entradas) o alguna función de ellas; y la violación puede ser más difícil de detectar. Las violaciones en el supuesto de distribución se encuentran a veces y el examen de gráficos q ~ q de los residuales puede revelar la naturaleza de la distribución.

El hecho de que nuestras medidas tracen registros de respuestas en una escala no significa que ésta es la mejor escala para la modelación. Por ejemplo, en un estudio de los factores que afectan el kilometraje del gas en automóviles Henderson y Velleman (1981), encontraron que era mejor utilizar la respuesta de galones por milla que las millas por galón de combustible; esto es, el modelo se desarrolla en términos del recíproco de la respuesta. En un ejemplo más detallado, en secciones posteriores veremos

que el uso de un logaritmo de los rendimientos de respuesta mejora el modelo que tiene alguna justificación teórica. Esta y otras transformaciones se usan a menudo para generar un mejor análisis. Frecuentemente las transformaciones son sugeridas por el análisis diagnóstico. Algunos analistas recomiendan aplicar logaritmos si el rango de los datos tiene diferentes órdenes de magnitud. Las transformaciones lineales son convenientes, porque no cambian el ajuste de los datos. Las transformaciones recomendadas en esta sección son no lineales y diseñadas para obtener satisfacción de los supuestos del modelo lineal. De esta forma, no existe garantía de que se pueda hacer esto y se pueden construir otras formas alternativas del modelo.

La transformación de los datos ha sido, por lo general, una alternativa que produce un mejor ajuste o quizás un mejor modelo de predicción. Las transformaciones o reexpresiones de datos se usan para hacer un supuesto más razonable. Se puede tratar de transformaciones para estabilizar varianza y eliminar el problema de la autocorrelación en los datos para una regresión. Aquí se resuelve otro supuesto, que es la forma del modelo. En esta sección, se discuten algunos procedimientos para determinar formas alternativas de modelos o transformaciones y situaciones bajo las cuales el uso de esas transformaciones puede tener éxito. En algunos casos la necesidad de algún tipo de transformación puede ser obvia. Las gráficas de los datos en regresión lineal simple pueden definir una aproximación curvilínea.

8.3. Transformación en el caso de regresión simple

Considere el caso de la regresión lineal simple. Una transformación para cambiar la estructura del modelo se requiere normalmente cuando los datos reflejan curvatura. Una desviación ordinaria del modelo de línea recta es

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (1)$$

Puede ser detectada en las gráficas de residuales, o en una simple gráfica de los datos. Si una gráfica de los datos refleja una tendencia que tiene semejanza curvilínea específica, se puede alterar el modelo para ajustar la tendencia (figura 8.1).

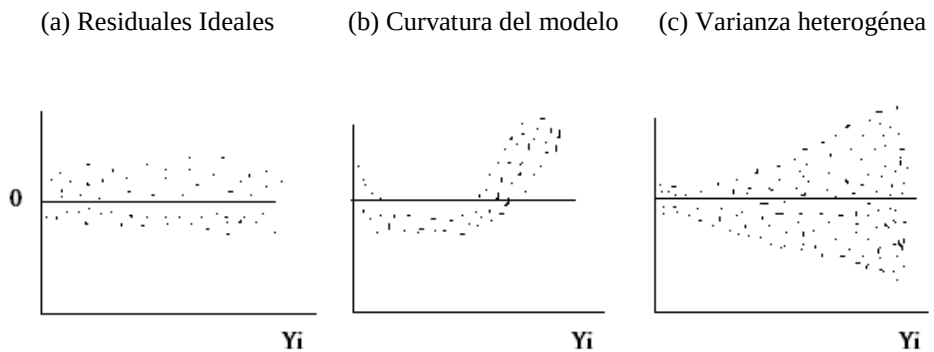


Figura 8.1. Residuales que indican la violación de supuestos.

8.3.1. La parábola

La alteración del primer modelo se considera más que simplemente una transformación, la adición de un término cuadrático al modelo. El modelo es

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \qquad (2)$$

La naturaleza de la gráfica que se produce, y los datos generados por el modelo de la parábola, dependen de los signos y magnitud de los coeficientes β_0, β_1 y β_2. Los ejemplos típicos son (figura 8.2).

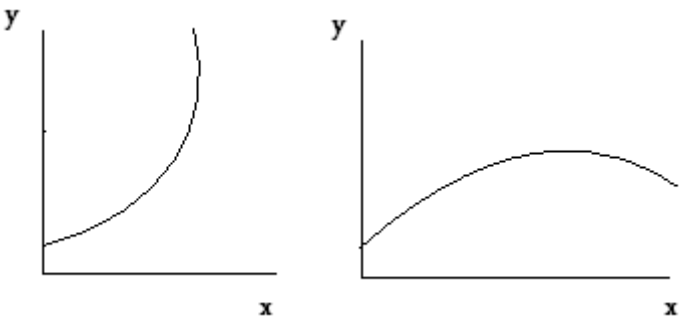


Figura 8.2. (a) Parábola con β_0, β_1 y $\beta_2 > 0$ (b) Parábola con $\beta_0 > 0$, $\beta_1 > 0$ y $\beta_2 < 0$

8.3.2. La hipérbola

Aplicaciones en las áreas de biología, economía y otros campos, permiten el uso de una función hiperbólica que se puede obtener transformando ambas variables; la variable respuesta (Y) y la variable regresora (X). Gráficas

típicas que sugieren el uso de la función de la hipérbola se pueden observar el la figura 8.3. La forma funcional verdadera de la hipérbola es no lineal en los coeficientes del modelo. La ecuación es dada por $y = x/(\alpha + \beta x)$. La forma linearizada involucra la transformación inversa de ambas variables; esto es, una regresión de $\frac{1}{Y}$ contra $\frac{1}{X}$, adoptando así la estructura del modelo

$$\frac{1}{y_i} = \beta_0 + \beta_1 \left(\frac{1}{x_i}\right) + \varepsilon \qquad (3)$$

Es fácil verificar que $\beta_0 = \alpha$ y que $\beta_1 = \beta$. La asíntota que puede ser de interés para el analista se indica por la línea punteada en la figura 8.3. Las curvaturas positivas y negativas se obtienen cuando $\beta_1 > 0$ y $\beta_1 < 0$.

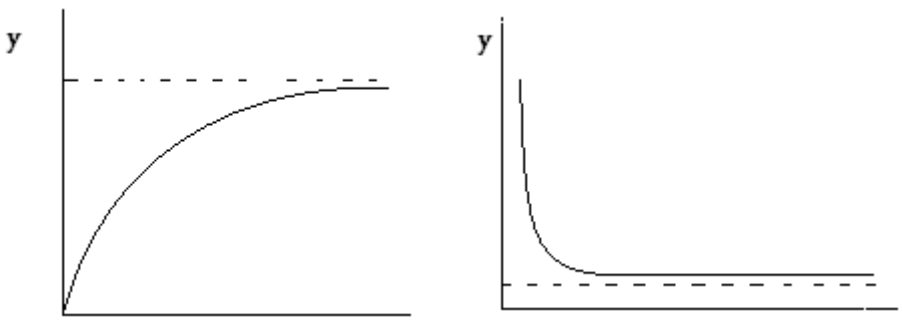


Figura 8.3. (a) Hipérbola con curvatura negativa. (b) Hipérbola con curvatura positiva.

8.3.3. La función exponencial: transformación de logaritmo natural de Y

En la primera sección se discutió la transformación logarítmica de la respuesta como un mecanismo que puede ser usado para lograr varianzas homogéneas en ciertas situaciones. La transformación puede ser usada también para producir un supuesto para el modelo cuando los datos sugieren una curvatura de un cierto tipo. Si la figura no es una línea recta, pero se parece a las mostradas en la figura 8.4, la estructura verdadera puede ser de la forma $y = \alpha e^{\beta x}$. Entonces un modelo ajustado del tipo

$$\ln (y_i) = \beta_0 + \beta_1 x_i + \varepsilon \qquad (4)$$

puede ser apropiado. Aquí $\beta_0 = \ln \alpha$ y $\beta_1 = \beta$

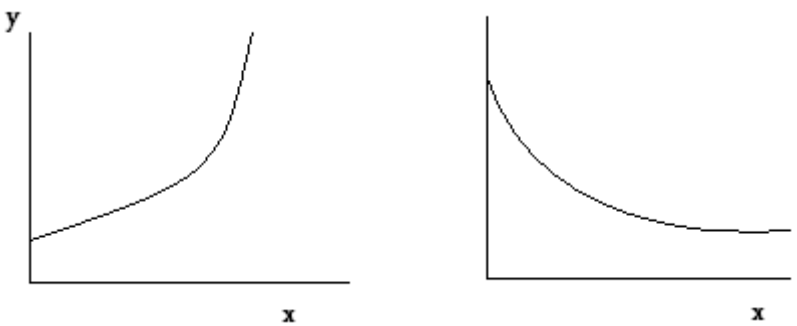


Figura 8.4. (a) Función exponencial en donde $\beta > 0$. (b) Función exponencial en dónde $\beta < 0$.

8.3.4. La función potencia (transformaciones de logaritmo natural de *Y* y *X*)

A veces, la curvatura que ve el analista en la gráfica de *Y* sobre *X* sugiere un procedimiento del tipo $y = \alpha x^{\beta}$, una función de potencia que puede ser linearizada usando la transformación logaritmo de ambas variables. Entonces el modelo ajustado está dado por

$$\ln (Y_i) = \beta_0 + \beta_1 (\ln X_i) + \varepsilon \qquad (5)$$

con los coeficientes β_0 y β_1 estimados por el procedimiento estándar de mínimos cuadrados. La gráfica de la función de potencia depende del signo y la magnitud de la constante β_1, y se puede observar en la figura 8.5 de los diferentes conjuntos de funciones.

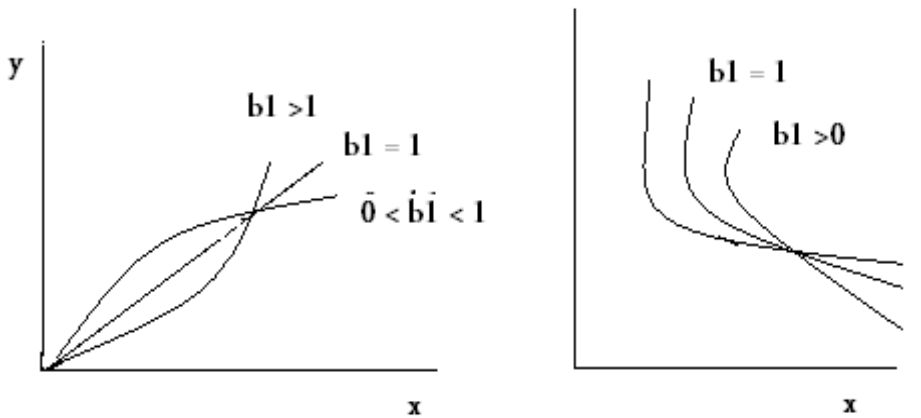


Figura 8.5. (a) Función de potencia. (b) Función de potencia β_1 negativo.

8.3.5. La exponencial inversa la transformación de logaritmo natural de y; transformación inversa de x.

Hay muchos fenómenos científicos que son exponenciales por naturaleza, pero no están contemplados en la familia discutida anteriormente. Por lo general, la porción exponencial del mecanismo es proporcional a la inversa de *X* en lugar de *X*; esto resulta en figuras no parecidas a las presentadas en la figura 8.4. El mecanismo es $y = \alpha e^{\beta/x}$. El modelo ajustado que refleja la transformación es

$$\ln (y) = \beta_0 + \beta_1(\frac{1}{x_i}) + \varepsilon_i \qquad (6)$$

En la figura 8.6 se muestra el tipo de figura del modelo anterior. El propósito de esta sección es proporcionar modelos alternativos que permitan tener más éxito en el ajuste de modelos de predicción que con la regresión lineal simple. En el caso de una sola variable regresora, si la curvatura en una gráfica revela una apariencia similar a la graficada aquí, el usuario empleará la transformación indicada y podrá determinar si los resultados son mejores. El mejoramiento puede tomar la forma de (1) un patrón más atractivo de los residuales, (2) un mejor ajuste y predicción estadística en las variables naturales. En el caso (2), el analista formará residuales en el original de y-unidades por lo que produce un cuadrado medio del error.

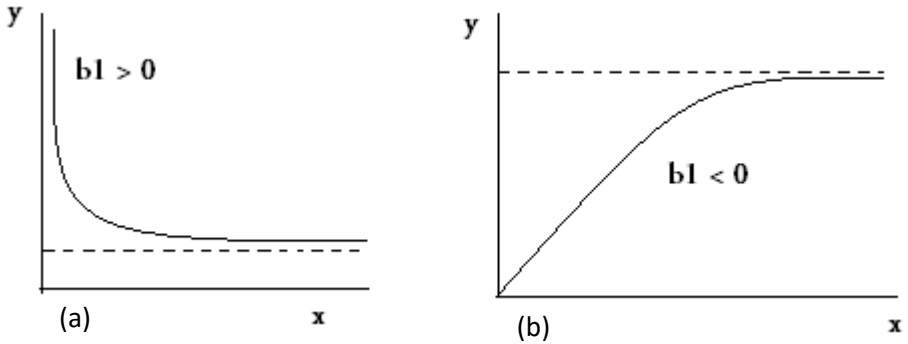


Figura 8.6. (a) Exponencial inversa: uso de la transformación de logaritmo natural de *y* e inversa de *x*. (b) Exponencial inversa: uso de transformación de logaritmo natural de *y*, transformación inversa de *x*.

8.4. Qué pasa con la estructura del modelo transformado

Las transformaciones son parte integral del análisis estadístico de datos. Se intenta formalizarlos aquí para que el usuario entienda su propósito. Sin embargo, el analista de datos debe conocer cuál es la estructura total del modelo, una vez que se ha tomado la decisión de una transformación de los datos. Una transformación puede ser vista como una simple pincelada de lápiz o un cambio en una declaración del modelo de un paquete de cómputo, con la motivación de que los datos son descritos por un nuevo modelo más aproximado. Sin embargo, hay más detalles y el analista debe ser prudente. Se puede ilustrar mejor con un ejemplo: suponga una gráfica de *Y* contra *X*,

que revela una curvatura que se parece a la figura 8.5a, se asume que una función de potencia del tipo $y = \alpha x^{\beta}$, generó los datos y el modelo es

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon \tag{7}$$

Verdaderamente, todas las indicaciones del ajuste resultante puntualizan evidencia de un mejoramiento. Pero, ¿dónde va verdaderamente la transformación y qué hace el analista? La transformación trabaja porque el verdadero mecanismo es la función de potencia $y = \alpha x^{\beta}$. Si ignoramos los errores en el modelo, aplicamos el logaritmo natural a la función de potencia; esto es, $\ln y = \beta_0 + \beta_1 \ln X_i$; sin embargo, si colocamos el término de error aditivo a las dos relaciones, una transformación log del modelo de la función de potencia, no produce la ecuación anterior. El resultado de este desarrollo es que la transformación de los datos altera la estructura de los errores. Mientras la transformación pueda ser estructuralmente ventajosa, es completamente posible que pueda producir una violación del supuesto de varianzas homogéneas o del supuesto de normalidad.

Las desventajas de las transformaciones expuestas aquí no desaniman su uso. Sin embargo, es aconsejable poner mucha atención para hacer rutinas cuidadosas de inspección de residuales estudentizados cuando se ha hecho una transformación. La estimación de los parámetros en modelos no lineales por mínimos cuadrados es más complejo. En ciertas circunstancias, que ocurren regularmente, puede ser preferible no transformar los datos, pero sí asumir un modelo aditivo y aplicar regresión no lineal.

8.5. Ejercicios

8.1 Se realizó un estudio para relacionar el peso y la longitud de peces capturados en una laguna costera del estado de Tabasco. Los pesos y longitud fueron medidos a 24 peces y los datos son:

Observaciones	Log. del peso	Log. de longitud	Observaciones	Log. del peso	Log. de longitud
1	1.973	2.338	13	2.405	2.496
2	1.973	2.367	14	2.626	2.558
3	2.064	2.398	15	2.713	2.559
4	2.152	2.401	16	2.665	2.582
5	2.158	2.412	17	2.737	2.585
6	2.326	2.459	18	2.746	2.600
7	2.262	2.465	19	2.825	2.615
8	2.299	2.473	20	2.892	2.631
9	2.362	2.474	21	2.975	2.639
10	2.338	2.484	22	2.940	2.646
11	2.449	2.484	23	3.025	2.670
12	2.384	2.493	24	2.985	2.781

Podemos conjeturar que el peso del pescado varía con la longitud y se puede aproximar por la siguiente relación: $Y = ax^b$. Donde Y es el peso y X es la longitud.

a) Ajuste una regresión lineal simple a los datos: $Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$.

b) Ajuste el modelo: $\log Y_i = \xi_0 + \xi_1 (\log X_i) + \varepsilon_i$.

c) Compare los dos modelos.

8.2. Se realizó un estudio para determinar las características del crecimiento de las raíces de trigo en presencia de un herbicida, aplicado al suelo. Los datos fueron recolectados del porcentaje de control midiendo el porcentaje del crecimiento observado sin el herbicida, como la variable respuesta.

Concentración de herbicida (X)	0.5	1.0	2.0	8.0	32.0	128.0
% de control (Y)	95.8467	91.6561	81.5142	71.7477	68.7061	35.9895

a) Ajuste un modelo del tipo: $Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i$. Calcule s^2, los residuales PRESS, los residuales ordinarios y la suma de residuales absolutos ordinarios PRESS.

b) Ajuste el modelo ln $(Y_i) = \beta_0 + \beta_1 X_i + \varepsilon_i$. Calcule la misma información que para el inciso (a) para las unidades de respuesta originales.

c) Haga una comparación entre los dos modelos anteriores. ¿Cuál prefiere?

8.3. En un estudio de ingeniería ambiental de una cierta reacción química, la concentración de 18 soluciones preparadas separadamente fueron registradas en diferentes tiempos (tres medidas para cada una de los seis tiempos). El logaritmo natural de la concentración se calculó y se obtuvieron los siguientes resultados:

Solución	Tiempo	Concentrac	Ln(concentrac)	Solución	Tiempo	Concentrac	Ln(concentrac)
1	6	0.029	-3.54	10	12	0.425	-0.856
2	6	0.032	-3.442	11	12	.0384	-0.957
3	6	0.027	-3.612	12	12	0.472	-0.751
4	8	0.079	-2.538	13	14	1.130	0.122
5	8	0.072	-2.631	14	14	1.020	0.020
6	8	0.088	-2.430	15	14	1.249	0.222
7	10	0.181	-1.709	16	16	2.812	1.034
8	10	0.165	-1.802	17	16	2.465	0.902
9	10	0.201	-1.604	18	16	3.099	1.131

a) Grafique en hojas separadas la dispersión de los datos de:

1) La concentración (Y) versus el tiempo (X).

2) El logaritmo natural de la concentración (ln Y) versus el tiempo (X).

b) Calcule por medio de computadora lo siguiente:

1) La ecuación estimada de la línea recta de la regresión (grado 1) de *Y* sobre *X*.

2) La ecuación estimada de la regresión cuadrática (grado 2) de *Y* sobre *X*.

3) La ecuación estimada de la regresión de línea recta (grado 1) de ln *Y* sobre *X*.

4) La gráfica de cada una de estas ecuaciones ajustadas en sus respectivos diagramas de dispersión.

c) Basado en los resultados del análisis, complete la siguiente tabla.

Fuente	**G.L.**	**SC**	**SCM**	**F_0**
Regresión	1			
Residual				
Falta de ajuste	4			
Error puro	12			
Total	**17**			

d) Basado en los resultados del análisis, complete la tabla del ANVA.

Fuente	**G.L.**	**SC**	**SCM**	**F_0**
Regresión				
Grado 1 (X)	1			
Grado2 (X^2/X)	1			
Residual				
Falta de ajuste	3			
Error puro	12			
Total	**17**			

e) Determine y compare las proporciones de variación total en *Y* explicadas por la regresión de línea recta sobre *X* y por la regresión cuadrática sobre *X*.

f) Realice las pruebas de F para la significancia de la regresión de la línea recta de *Y* sobre *X* y para la adecuacia de ajuste de la línea de regresión estimada.

g) Realice la prueba de F total para la significancia de la regresión cuadrática de *Y* sobre *X*, una prueba para la significancia de la adición de X^2 al modelo y una prueba de F para lo adecuado del ajuste del modelo cuadrático estimado.

h) Para la regresión de línea recta de ln*Y* sobre *X*, realice las pruebas de F para la significancia de la regresión completa y para la adecuacia de ajuste del modelo de línea recta estimado.

i) ¿Qué proporción de la variación de ln*Y* es explicada por la línea recta de regresión de ln*Y* sobre *X*? Compare este resultado con los obtenidos en el inciso (e) para la regresión cuadrática de *Y* sobre *X*.

Concentración de Y sobre el tiempo			
Ajuste de Grado 1		**Ajuste de Grado 2**	
R^2 -múltiple	0.732	R^2 -múltiple	0.957
Regresión	Coeficientes	Regresión	Coeficientes
0	-1.9318	0	3.1721
1	0.2459	1	-0.78102
		2	0.046682
Análisis de Varianza			
Fuente	G. L.	SC	SCM
Grado 1	1	12.705	12.705
Grado 2	1	3.9053	3.9053
Falta de ajuste	3	0.51446	0.17149
Error puro	12	0.23248	0.01937
Total	17	17.357	1.021
Concentración de ln Y sobre el tiempo (X)			
R^2 -múltiple	0.996		
Regresión	Coeficientes		
0	-6.2096		
1	0.45117		

Fuente	**G.L.**	**SC**	**SCM**	**F_0**
Grado 1	1	42.746	42.746	
Falta de ajuste	4	0.02783	0.006959	
Error puro	12	0.12247	0.010206	
Total	17	42.896	2.5233	

j) Un supuesto fundamental del análisis es la homocedasticidad.

1) Examine los diagramas de dispersión construidos en el inciso (a) y establezca por qué el tomar el logaritmo natural de la concentración ayuda a cumplir con el supuesto de homogeneidad de varianza.

2) ¿Piensa que la línea de regresión de ln$Y$ sobre X es mejor para describir este conjunto de datos y mejor que la regresión cuadrática de Y sobre X? Explique.

k) ¿Cuál supuesto clave de los datos está en duda si, en lugar de 18 soluciones diferentes, fueran tres; cada una de las cuales fue analizada en seis diferentes tiempos?

Capítulo 9

Regresión polinomial

9.1. Introducción

Esta sección se enfoca a un caso especial de los modelos de regresión múltiple, *el modelo polinomial*, que por lo general es de interés si existe solamente una variable independiente (X) en la relación. Se considera un modelo de línea recta para esta situación, sin embargo, se puede determinar si se logra mejorar significativamente la predicción incrementando la complejidad del modelo ajustado de línea recta. La extensión más simple del modelo de línea recta es el polinomio de segundo orden o parábola, que involucra un término X^2 adicional a X. La adición de un término de alto orden X^2, X^3, etc., es una simple función de una variable simple básica; que puede ser considerado equivalente a adicionar nuevas variables independientes. Entonces, si se denota a X como X_1, a X^2 como X_2, el modelo de segundo orden

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \qquad (1)$$

se convierte en

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (2)$$

En general, los modelos polinomiales son casos especiales del modelo general de regresión múltiple. Sin embargo, puesto que se considera solamente una de las variables independientes, cualquier modelo polinomial puede ser representado por una gráfica curvilínea o una gráfica de dos dimensiones. Cuando existe una sola variable independiente *X*, la tarea fundamental es encontrar una curva que ajuste mejor los datos y que la relación entre *X* y *Y* sea explicada apropiadamente. Porque una curva de alto orden puede ser más apropiada que una línea recta, y por lo general se considera importante ajustar tales curvas.

Primero, se consideran métodos para ajustar y evaluar el modelo de segundo orden (parabólico); después se consideran modelos de mayor orden. Puesto que estos modelos son casos especiales del modelo general de regresión múltiple, el ajuste de estos modelos y los métodos de inferencia son esencialmente los mismos que fueron descritos anteriormente; ya que las variables independientes en un modelo polinomial son funciones de la misma variable básica (*X*) y están inherentemente correlacionadas. Esto genera dificultades de cómputo debido a la colinealidad. Afortunadamente, estas técnicas disponibles, como el uso de polinomios ortogonales y el centrado, que nos ayudan a

remediar tales problemas y estos procedimientos se discuten más adelante. El uso de polinomios ortogonales nos ayuda a simplificar las pruebas de hipótesis.

9.2. Modelos polinomiales

La clase más general de curvas consideradas para describir la relación entre una sola variable independiente *X* y una respuesta llamada *Y*, se denomina como polinomio. Matemáticamente, un polinomio de orden *k* en *X* es una expresión de la forma

$$Y = c_0 + c_1x + c_2x^2 + c_3x^3 + ... + c_kx^k \quad (3)$$

En el cual los *c*'s y *k* son constantes. Hemos considerado el polinomio más simple correspondiente a $k = 1$ (esto es, la línea recta que tiene la forma $Y = c_0 + c_1x$). El polinomio de segundo orden que corresponde a $k = 2$ (esto es la parábola) tiene la forma general $Y = c_0 + c_1x + c_2x^2$.

Pasando de un modelo matemático a un modelo estadístico, como se hizo con el modelo de la línea recta, podemos escribir un modelo parabólico en cada una de las siguientes formas

$$\mu_{Y/X} = \beta_0 + \beta_1X + \beta_2X^2 \quad (4)$$

o el modelo

$$y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

En estas ecuaciones, las letras mayúsculas (Y's, X's) denotan variables estadísticas; β_0, β_1 y β_2 denotan los parámetros desconocidos llamados coeficientes de regresión; $\mu_{Y/X}$ denota la media de *Y* a una *X* dada, y ε denota el componente del error que representa la diferencia entre la respuesta observada *Y* en *X* y la verdadera respuesta promedio $\mu_{Y/X}$ en *X*. Si tentativamente asumimos que un modelo parabólico, que es dado por los modelos anteriores (4), es apropiado para describir la relación entre *X* y *Y*, debemos determinar una parábola estimada que ajuste mejor los datos. Como en el caso de línea recta, esta parábola de mejor ajuste puede ser determinada empleando el método conocido de mínimos cuadrados.

9.3. Procedimiento de mínimos cuadrados para el ajuste de una parábola

Los estimadores mínimos cuadrados de los parámetros β_0, β_1 y β_2 en un modelo parabólico son seleccionados para minimizar la suma de cuadrados de desviaciones de puntos observados que corresponden a la

parábola ajustada (figura 9.1), y sea $\hat{Y}$ quien denota el valor de respuesta predicha en *X*, se puede escribir la parábola estimada como:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2 \qquad (5)$$

La mínima suma de cuadrados obtenida usando esta parábola de mínimos cuadrados es

$$SC_{error} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i - \hat{\beta}_2 X^2{}_i)^2 \qquad (6)$$

Como en el modelo general de regresión, se supone que no es necesario presentar las expresiones precisas para calcular los estimadores de mínimos cuadrados de $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$. Estas expresiones son completamente complejas y se darán sólo para polinomios de mayor orden. Los investigadores parecen no emplear tales modelos de regresión polinomial sin el uso de software estadístico apropiado, que puede desarrollar los cálculos necesarios e imprimir los resultados numéricos.

Para el ejemplo 1 (pág 23) los datos de presión sistólica sanguínea y nivel de actividad física, con la eliminación de un dato lejano (47, 220), los estimadores mínimos cuadrados para los coeficientes de regresión parabólica calculados son:

$\hat{\beta}_0 = 113.41$ $\hat{\beta}_1 = 0.088$ $\hat{\beta}_2 = 0.010$

Entonces, el modelo ajustado es

$$\hat{Y} = 113.41 + 0.088X + 0.010X^2$$

Esta ecuación puede ser comparada con la ecuación de la línea recta para estos mismos datos, con la eliminación del dato lejano (outlier).

$$\hat{Y} = 97.08 + 0.95X$$

Cuando se comparan los dos modelos ajustados es importante notar que los estimadores de β_0 y β_1 son diferentes en los dos modelos, indicando que la estimación de β_2 afecta la estimación de β_0 y β_1 en el modelo cuadrático.

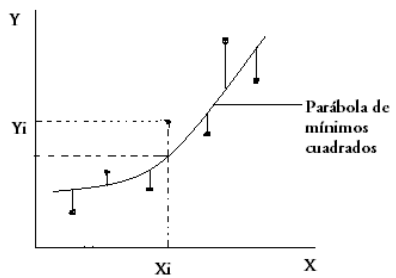


Figura 9.1. Desviaciones de puntos observados de la parábola de mínimos cuadrados.

9.4. Tabla de Análisis de varianza (ANVA) para regresión polinomial de segundo orden

Como se mostró en el caso de regresión lineal simple, los resultados del análisis de un modelo polinomial de segundo orden pueden ser presentados en una tabla. La tabla de ANVA para un ajuste parabólico de los datos de presión sistólica sanguínea y edad se tienen en la tabla 9.1. Los comentarios sobre esta tabla son: primero se realizan pruebas para variables en orden de adición. El orden de las variables naturales sugiere que se haga de la potencia más grande a la más pequeña del predictor o viceversa. De forma consecuente, una prueba para la última variable adicionada en cada término se puede evitar en modelos polinomiales. Usando las pruebas de variables en orden de adición, favorecen la selección del más parsimonioso y relevante modelo posible. Además, todas estas pruebas utilizan el residual del modelo más grande.

Tabla 9.1
Tabla de ANVA para los datos de presión sistólica sanguínea y edad

Fuente		**Grados de libertad**	**Sumas de cuadrados**	**Cuadrados medios**	**F_0**
Regresión	X	1	6110.10	6110.10	68.89
	X^2/X	1	163.30	163.30	1.84
Residual		26	2306.05	88.69	
Total		**28**	**8579.45**		**$R^2 = 0.731$**

9.5. Inferencias asociadas con la regresión polinomial de segundo orden

Hay tres cuestiones de inferencia básica asociadas con la regresión polinomial de segundo orden; que son

a) ¿Es significativa la regresión completa?; es decir, ¿es mayor la variación explicada en *Y* por el modelo de segundo orden que por ignorar *X* completamente?
b) ¿Provee el modelo de segundo orden un mayor poder predictivo que el dado por el modelo de línea recta?
c) Dado que un modelo de segundo orden es más apropiado que uno de línea recta, ¿adicionaremos términos de más alto orden (esto es, X^3, X^4, etc.) al modelo de segundo orden?

9.5.1. Prueba para la regresión total y la fuerza de la relación parabólica total

Para determinar si la regresión total es significativa y probar si en la hipótesis nula H_0: "no existe significancia de la regresión total usando *X* y X^2" (esto es; $\beta_1 = \beta_2 = 0$). El procedimiento de prueba usado para esta hipótesis nula involucra la prueba de F total, como

$$F = \frac{CM_{regresion}}{CM_{residual}} \quad (7)$$

Se compara el valor de este estadístico con un valor crítico apropiado de la distribución de F, con 2 y 26 grados de libertad, en el numerador y denominador respectivamente. Para $\alpha = 0.001$ encontramos que $F = 35.37 > F_{2,26,\ 0.999} = 9.12$, entonces se rechaza la hipótesis nula de la regresión total no significativa ($P < 0.001$). Para obtener una medida cuantitativa de qué tan bien el modelo de segundo orden predice la variable dependiente, podemos usar el coeficiente de correlación múltiple cuadrado (R^2 múltiple). Con r^2 en regresión lineal simple, R^2 representa la proporción de la reducción en la suma de cuadrados del error obtenido usando X y X^2, en lugar del simple predictor Y. La expresión para calcular R^2 está dada por

$$R^2_{\text{(modelo de segundo orden)}} = \frac{SC_{totales} - SC_{error(modelo\ de\ segundo\ orden)}}{SC_{total}} \quad (8)$$

Para el ejemplo anterior (tabla 9.1), $R^2 = 0.731$. La prueba anterior de F indica que la R^2 es significativamente diferente de cero.

9.5.2. Prueba para la adición del término X^2 al modelo

Para responder la segunda cuestión, en relación al incremento del poder predictivo, se debe aplicar una prueba de F parcial de la hipótesis nula H_0: "la adición del término X^2 al modelo de la línea recta, no mejorará significativamente la predicción de *Y*, arriba y abajo del obtenido por el mismo modelo de línea recta" (esto es; $\beta_2 = 0$). Para probar esta hipótesis nula se calcula el estadístico parcial de F

$$F(X^2/X) = \frac{(Sumas\ de\ Cuadrados\ extra\ debido\ a\ la\ adición\ de\ X^2)/1}{Cuadrados\ Medios\ de\ residuales(en\ el\ modelo\ de\ segundo\ orden)} \quad (9)$$

El cual se compara con un valor apropiado del porcentaje de F ($F_{1,26}$, en el ejemplo).

Puesto que $X^2 X$es la última variable adicionada, ésta es una prueba de la última variable adicionada. Alternativamente podemos dividir el coeficiente estimado $\hat{\beta}_2$ por su desviación estándar estimada para formar un estadístico que tiene una distribución de t bajo H_0 con 26 grados de libertad. La información necesaria del ANVA para calcular esta prueba F en nuestro ejemplo, se encuentra en la tabla 9.1. La suma de cuadrados extra para $X^2 X/X$ es de 163.30 y se calcula como la diferencia entre el valor de la suma de cuadrados de regresión para

los modelos de primer y segundo orden. El estadístico parcial F $\{F(X^2 X/X)\}$ se calcula como

$$F = \frac{163.30}{88.69} = 1.84$$

Puesto que $F_{1,26,\ 0.90} = 2.91$ no se rechaza H_0 al nivel de $\alpha = 0.10$. Además, el valor de P, para esta prueba satisface la desigualdad $0.10 < P < 0.25$. Entonces, se puede concluir que cuando se adiciona un término cuadrático al modelo de línea recta no mejora significativamente la predicción. Esta conclusión está justificada por el pequeño incremento en R^2 cuando el término X^2 es adicionado al modelo de línea recta. Puesto que $r^2 = R^2 = 0.712$ para el modelo de línea recta, y $R^2 = 0.731$ para el modelo parabólico, entonces el incremento del R^2 es; $0.731 – 0.712 = 0.019$, que se considera muy pequeño.

9.5.3. Prueba para la adecuacia del modelo de segundo orden

Por los análisis de los datos anteriores se puede ver que el modelo de línea recta es el que ajusta los datos de forma adecuada, la predicción de la respuesta es significativa y es preferible a un modelo parabólico. De forma consecuente, es superfluo evaluar si un modelo de orden más alto que dos es significativamente mejor que un modelo de línea recta.

No obstante, note que cualquier cuestión de adecuacia de un modelo puede ser enfocada (con una prueba de falta de ajuste) para cualquier modelo (de cualquier orden) en un estado dado de análisis. Cualquier prueba de falta de ajuste puede ser caracterizada por una prueba de F parcial o múltiple, para la adición de uno o más términos del modelo bajo estudio.

Ejemplo: suponga que un estudio de laboratorio se diseña para determinar la relación entre la dosis (X) de cierta droga y la ganancia de peso (Y). Por la que ocho animales de laboratorio del mismo sexo, edad y tamaño son seleccionados al azar y asignados aleatoriamente a uno de los ocho niveles de droga. Los datos son:

Dosis droga	1	2	3	4	5	6	7	8
Ganancia de peso	1	1.2	1.8	2.5	3.6	4.7	6.6	9.1

La ganancia de peso se mide en cada animal después de la segunda semana del estudio, periodo durante el cual todos los animales son sometidos al mismo régimen alimenticio y condiciones generales de laboratorio. Se grafican los datos y se observa que el modelo que mejor los representa es un modelo parabólico, más que un modelo de línea recta.

De acuerdo con los cálculos y el análisis de varianza correspondiente para este modelo, tenemos la ecuación de predicción

$$\hat{Y} = 1.13 - 0.41X + 0.17X^2$$

Las tablas de análisis de varianza

a) Para un modelo de línea recta el ANVA es

Fuente	**Grados de libertad**	**Sumas de cuadrados**	**Cuadrados medios**	**F_0**	**El modelo ajustado de línea recta es**
Regresión	1	52.04	52.04	61.95	$\hat{Y}$ = 1.20 + 1.11X
Residual	6	5.03	0.84		R^2 = 0.912
Total	**7**	**57.07**			

b) Para un modelo cuadrático el ANVA es

Fuente		**Grados de libertad**	**Sumas de cuadrados**	**Cuadrados medios**	**F_0**
Regresión	X	1	52.04	52.04	61.95
	X^2	1	4.83	4.83	120.75
Residual		5	0.20	0.04	
Total		**7**	**57.07**	**R^2 = 0.997**	

En el modelo de línea recta, la hipótesis nula es altamente significativa con respecto a la existencia de una fuerte relación entre los niveles de droga y la ganancia de peso en estos animales;. con un R^2 = 0.912. En la segunda tabla de ANVA se puede ver que el término lineal y el cuadrático son altamente significativos y que la diferencia entre los dos coeficientes de correlación, simple y múltiple es significativa; esto es, 0.997 – 0.912 = 0.085, por lo que la inclusión del término X^2 en el modelo sí mejora la predicción de Y de forma significativa.

Por lo tanto, la adición del término X^2 al modelo, mejora la predicción de forma significativa. Se puede esperar que una prueba para la significancia de la regresión total en el modelo de segundo orden presenta una F altamente significativa, esto es

$$F = \frac{Cuadrados\ Medios\ de\ regresión\ del\ modelo\ de\ segundo\ orden}{Cuadrados\ Medios\ de\ residuales\ del\ modelo\ de\ segundo\ orden}$$

$$= \frac{(52.04+4.83)/2}{0.04} = 710.88 \qquad (10)$$

Con esto se puede decir que el modelo de primer orden (línea recta) no es tan bueno como el de segundo. Ahora, se requiere determinar si adicionando términos de más alto orden, el modelo de segundo orden es garantizado. Se puede adicionar el término X^3, al modelo de segundo orden y probar si la predicción mejora significativamente.

c) Tabla de ANVA para un modelo de tercer orden

Fuente		**Grados de libertad**	**Sumas de cuadrados**	**Cuadrados medios**	**F_0**
Regresión	X	1	52.04	52.04	61.95
	X^2	1	4.830	4.83	120.75
	X^3/X, X^2	1	0.140	0.140	10.0
Residual		4	0.056	0.014	
Total		**7**	**57.066**	**R^2 = 0.999**	

El estadístico de F = 10.0 tiene una distribución F con 1 y 4 grados de libertad, bajo H_0: "la adición del término X^3X no es favorable" (esto es,

$\beta_3 = 0$). Entonces, se tiene que $0.025 < P < 0.05$. Este valor de P rechazaría H_0 a un nivel de 0.05, pero no al nivel de 0.025. Esto ayuda a decidir si se incluye el término X^3 o no en el modelo; sin embargo, otros elementos se tienen que tomar en cuenta: (1) el valor de R^2 para el modelo de segundo orden es muy alto de 0.997; (2) el valor de R^2 para el modelo de tercer orden se incrementa a 0.999; (3) la gráfica de las variables sugiere una curva de segundo orden y (4) entonces el modelo más simple es preferible porque es más fácil de interpretar. Por todas estas consideraciones se puede concluir que el modelo de segundo orden es el más apropiado para representar esta relación. También es valioso tener la desviación estándar de los coeficientes de regresión estimados. Éstos son difíciles de calcular manualmente para modelos que involucran dos o más variables predictoras. Sin embargo, por lo general se utilizan programas de computadora que imprimen los valores de los coeficientes y sus errores estándares estimados. Para el modelo de segundo orden de los datos anteriores se tiene que $S_{\beta_1} = 0.141$ y $S_{\beta_2} = 0.015$. Entonces, un intervalo de confianza del 100 $(1 - \alpha)$ % para β_2 será

$$\hat{\beta}_2 \pm t_{5,\, 1-\alpha/2} S_{\beta_2} \qquad (11)$$

$0.17 \pm (2.571)(0.015)$ o $0.13 < \beta_2 < 0.21$

Note que este intervalo no incluye el cero, lo cual está de acuerdo con la tabla de ANVA sobre la importancia de incluir el término X^2 en el modelo de predicción.

9.6. Ajustando y probando modelos de mayor orden

Se ha observado cómo las ideas básicas de regresión múltiple pueden ser aplicadas para ajustar y probar modelos cuadráticos, cúbicos y polinomiales. Estos mismos métodos se generalizan a todos los modelos polinomiales de mayor orden. Sin embargo, varios resultados relacionados necesitan discusión: el uso de polinomios ortogonales y las estrategias para seleccionar un modelo polinomial. El orden de un modelo polinomial depende del problema estudiado y de la cantidad y tipo de datos que se están recolectando. Una consideración, típicamente para estudios en las ciencias biológicas y sociales, es si la relación de regresión puede ser descrita por una función monótona (esto es; una que siempre incrementa o disminuye). Si las funciones monótonas son de interés, por lo general un modelo de segundo o tercer orden es suficiente. Un gran número de valores de predictores y una varianza pequeña del error son necesarios para ajustar modelos confiables de mayor orden que el cúbico.

Una consideración más general es el número de curvatura o los puntos de inflexión en la curva polinomial que se desea ajustar. Por ejemplo, un modelo de primer orden no tiene curvatura, un modelo de segundo orden tiene una sola curvatura y cada término de mayor orden adiciona otra curvatura potencial. En la práctica, ajustar modelos polinomiales de más alto orden que tres, por lo general son modelos que no siempre incrementan o disminuyen las curvaturas. Existe evidencia sustancial teórica y empírica para fundamentar el no empleo de modelos monótonos complicados. La cantidad de datos disponibles limita el orden máximo de un polinomio que se puede ajustar considerando el ejemplo de ganancia de peso. Con ocho valores distintos, un polinomio de orden siete ajustaría perfectamente los ocho puntos, dando una suma de cuadrados del error de cero y un $R^2 = 1$ (esto significa que el modelo ajustado tendría ocho parámetros estimados). Por lo general, el orden máximo del polinomio que se puede ajustar es de uno menos el número de valores de *X*.

El mensaje es que si se considera una mejora al modelo polinomial, se intentan grados más altos para el modelo. Para funciones de una variable es evidente el diagrama de dispersión de los datos pero es menos obvio cuando se consideran varias variables predictoras. La otra motivación para términos polinomiales de alto orden es el resultado conocido de que cualquier función continua puede ser descrita arbitrariamente por un polinomio de suficientemente más alto orden. En

nuestra aplicación, suponga que se tiene una observación en cada uno de los *k* valores distintos de la variable independiente. Un polinomio de grado $k - 1$, que incluye todos los términos de bajo orden, ajusta exactamente los datos. No obstante esto, la tentación de sumar potencias adicionales de la variable independiente, después del segundo y tercero, se debe resistir por varias razones:

1) Dificultad para la interpretación de términos de alto orden.
2) El modelo puede ser inútil para predecir valores intermedios de la variable.
3) El modelo de más alto orden puede dar intervalos de predicción más amplios.

Es completamente posible que el coeficiente de los términos de primer orden no sea significativo cuando se ajusta un modelo de segundo orden, sólo si los datos han sido centrados para remover el efecto de la correlación. En este caso, hay una tentación para omitir el término de primer orden. Esto no se recomienda por las siguientes razones:

1) Hay poca ganancia, ya que se tiene una X observada e incluyendo el término lineal en el modelo requiere un mínimo esfuerzo.

2) En la forma no centrada del modelo, la ecuación ajustada es forzada a ser simétrica en el eje *X*.
3) El modelo ajustado en forma centrada siempre contiene un término lineal.

9.7. Pruebas para la falta de ajuste

Dado que se ajusta un modelo polinomial y los coeficientes de regresión estimados son probados para su significancia, ¿cómo podemos confiar que un modelo de más alto orden que el orden más alto probado, no es necesario? Una prueba de falta de ajuste, puede ser usada para definir esta cuestión. Conceptualmente, una prueba de falta de ajuste relacionada con la evaluación de un modelo más complejo que el considerado primero. Históricamente, el término fue usado algunas veces para describir el procedimiento clásico. La prueba de falta de ajuste clásica puede ser aplicada siempre que se tengan observaciones repetidas en el conjunto de datos. El término repetido significa que una unidad experimental tiene el mismo valor de *X*, de otra unidad experimental. Con *n* observaciones totales, si *d* valores de *X* son distintos, entonces el número de repeticiones es $r = n - d$. Recordar que una curva polinomial de orden $d - 1$ puede pasar exactamente en *d* puntos distintos. Una prueba clásica de falta de ajuste compara el ajuste

de un polinomio de orden *d – 1*, con el ajuste de un modelo polinomial bajo consideración.

Con la disponibilidad de paquetes de cómputo estándar de regresión, el uso de una prueba de F múltiple-parcial para la falta de ajuste puede ser menos fastidiosa para calcular que identificar las observaciones repetidas para calcular de forma directa la suma de cuadrados del error puro. Se recomienda usar polinomios ortogonales para evitar serios errores en el cálculo de la regresión múltiple, cargándolo a la suma de cuadrados de la falta de ajuste.

9.8. Ejercicios

9.1 Se realizó un estudio con parejas en matrimonio con uno o más descendientes para determinar el efecto del ingreso anual del esposo y el tiempo entre el matrimonio y el nacimiento del primer hijo. Los datos se tomaron a partir de una muestra de 25 matrimonios.

Ingreso	Tiempo	Ingreso	Tiempo
5775	16.2	18000	41.25
9800	35.0	13000	44.0
13795	37.2	5400	9.2
4120	9.0	6440	20.0
25015	24.4	9000	40.2
12200	36.75	18180	32.0
7400	31.75	15385	39.2
9340	30.0	1800	39.2
20170	36.0	22400	27.9
22400	30.8	24210	22.3
4608	9.7	5400	11.7
24210	20.0	9340	32.5
19625	38.2		

a) Se ajustó un modelo lineal y los resultados del ANVA son:

Parámetro	Valor	Error estándar
Pendiente	7.13761×10^{-4}	3.5281×10^{-4}
Intercepto	19.62565	5.2129
$S_{Y/X}$	10.49580	
Prueba de F	$F_{1, 18} = 4.09277$	

Usando estos resultados, complete la tabla del ANVA para la regresión de la línea recta del tiempo (Y) sobre ingresos (X).

Fuente	**G.L.**	**SC**	**CM**	**F_0**
Regresión	1			
Falta de ajuste	18			
Error Puro	5			
Total	**24**			

b) Usando los siguientes resultados, complete la tabla del ANVA para la regresión cuadrática del tiempo (Y) sobre el ingreso (X).

Fuente		**G.L.**	**SC**	**CM**	**F_0**
Regresión	Grado 1 (X)	1			
	Grado 2 (X^2/X)	1			
Residual	Falta de ajuste	17			
	Error puro	5			
Total		**24**			

c) Graficar sobre el diagrama de puntos las dos ecuaciones de grado 1 y de grado 2.

Ajuste de grado 1		**Ajuste de grado 2**		**Ajuste de grado 3**	
R^2 – múltiple = 0.153		R^2 – múltiple = 0. 880		R^2 – múltiple = 0.901	
Regresión	Coeficientes	Regresión	Coeficientes	Regresión	Coeficientes
0	20.17655	0	-14.86602	0	-35.29278
1	0.00061	1	0.00787	1	0.01223
		2	-0.00000025	2	-0.0000006
				3	0.000000

La tabla del ANVA es:

Fuente	G.L.	SC	CM
Grado 1	1	442.9141	442.9141
Grado 2	1	2100.08113	2100.08113
Grado 3	1	61.10018	61.10018
Falta de ajuste	16	271.74872	16.10018
Error puro	5	15.20121	3.04024
Total	24	2891.04534	120.46022

d) Calcule y compare los valores de R^2 obtenidos para la línea recta, el ajuste cuadrático y cúbico.

e) Realice las pruebas de F para la significancia de la regresión de la línea recta y para la adecuación del ajuste de este modelo.

f) Realice las pruebas de F para la significancia de la regresión cuadrática, de la adición del término cuadrático en el modelo y de la adecuación del ajuste del modelo cuadrático.

g) ¿Cuál modelo es el más apropiado: el lineal, el cuadrático o el cúbico?

9.2. Considere los datos de la siguiente tabla, donde la edad (X) está relacionada con el tamaño del vocabulario del niño (Y).

Edad	Años	0	0	1	1	1	1	2	2	3	3	4	4	5	6
	Meses	8	10	0	3	6	9	0	6	0	6	0	6	0	6
Vocabulario (Y)		0	1	3	19	22	118	272	446	896	1222	1540	1870	2072	2562

a) Haga un diagrama de dispersión de los datos para las dos variables.
b) Convierta la edad en años (1 año con 6 meses es igual a 1.5 años). Con estos datos, calcule los estimadores mínimos cuadrados de los parámetros de la línea de regresión. Grafique esta línea sobre el diagrama de dispersión.
c) Ajuste un modelo cúbico polinomial para predecir el tamaño del vocabulario como una función de la edad en años. Calcule los estimadores de la regresión.
d) Use las pruebas para las variables adicionadas en orden, determine el mejor modelo.
e) Reporte las pruebas de variables últimas adicionadas. Explique alguna diferencia aquí y en el inciso anterior.
f) Reporte un diagnóstico de colinealidad apropiado para el modelo y evalúelo. Incluir las correlaciones de las predictoras.
g) Grafique los residuales estudentizados versus valores predichos para el mejor modelo basado en los resultados del inciso (d). Construya un histograma o una gráfica de los residuales y coméntelo.

9.3. La siguiente tabla presenta el peso corporal de ratas (gr) y latencia de captura (min) después de una inyección de 40 mg/kg de peso corporal de metrazol.

Latencia	2.30	1.95	2.90	2.30	1.10	2.50	1.30	2.00	1.70	2.00	2.95	1.25	2.05	3.7
Peso	348	372	378	390	392	395	400	409	413	415	423	428	464	468

a) Haga un diagrama de dispersión de los datos considerando la latencia en función del peso.

b) Determine la estimación por mínimos cuadrados de la pendiente y el intercepto de la línea recta de latencia sobre el peso.

c) Pruebe si la pendiente es igual a cero. Use un $\alpha = 0.01$.

d) Pruebe si el intercepto es igual a cero. Use un $\alpha = 0.01$.

e) Grafique la línea de regresión estimada sobre el diagrama de puntos.

f) Ajuste un modelo polinomial cúbico para predecir la latencia como una función del peso menos el peso promedio.

g) Repita los incisos del (d) hasta el (g) del ejercicio 9.3 en este análisis.

h) Comente los resultados.

9.4 La respuesta de la piel de las ratas (Y) a diferentes concentraciones de una vacuna (X) desarrollada recientemente, se midió en un experimento y los resultados del análisis y los datos son:

X	0.5	0.5	1.0	1.0	1.5	1.5	2.0	2.0	2.5	2.5	3.0	3.0
Y	13.9	13.81	14.08	13.99	13.75	13.6	13.32	13.39	13.45	13.53	13.59	13.64

Modelos	Grado 1	$\hat{Y} = 13.986 - 0.1802X$
	Grado 2	$\hat{Y} = 14.270 - 0.6065X + 0.1218X^2$
	Grado 3	$\hat{Y} = 13.362 + 1.680X - 1.3929X^2 + 0.2885X^3$

Grado 1			**Grado 2**			**Grado 3**		
Fuente	G. L.	SC	Fuente	G. L.	SC	Fuente	G. L.	SC
Regresión	1	0.2844	Regresión	2	0.3536	Regresión	3	0.5222
Residual	10	0.3461	Residual	9	0.2769	Residual	8	0.1083

a) Grafique las ecuaciones de línea recta, cuadrática y cúbica sobre el diagrama de dispersión de los datos.
b) Pruebe secuencialmente la significancia del ajuste de la línea recta, la significancia de la adición de X^2 y la significancia de la adición de X^3 en el modelo.
c) ¿Cuál de los tres modelos recomendaría y por qué? (También puede considerar R^2 para cada modelo).

9.5. Una bióloga estudió el efecto de la temperatura de un cierto medio de cultivo sobre el crecimiento de células amnióticas humanas en un tejido de cultivo. Usando el mismo paquete parental, ella condujo un experimento en el cual 5 líneas celulares fueron cultivadas a cuatro temperaturas distintas. Después de 7 días del experimento se obtuvieron los siguientes datos:

Número de células ($x10^{-6}$) después de 7 días (Y)	**Temperatura (X)**	**Número de células ($x10^{-6}$) después de 7 días (Y)**	**Temperatura (X)**
1.13	40	2.30	80
1.20	40	2.15	80
1.00	40	2.25	80
0.91	40	2.40	80
1.05	40	2.49	80
1.75	60	3.18	100
1.45	60	3.10	100
1.55	60	3.28	100
1.64	60	3.35	100
1.60	60	3.12	100

a) Haga un diagrama de dispersión de los datos considerando el número de células en función de la temperatura y comente si piensa que una línea recta ajusta adecuadamente a los datos.

b) Determine la estimación por mínimos cuadrados de la pendiente y el intercepto de la línea recta del número de células sobre la temperatura.

c) Prueber si la pendiente es igual a cero. Use un $\alpha = 0.01$.

d) Pruebe si el intercepto es igual a cero. Use un $\alpha = 0.01$.

e) Grafique la línea de regresión estimada sobre el diagrama de puntos.

f) Determine la tabla de ANVA para la regresión de línea recta de Y sobre X.

g) Determine el R^2 en la tabla de ANVA e interprételo.

El análisis de los datos es:

Medias	**Desviación estándar**
$\bar{X}$ = 70.00	22.9416
$\bar{Y}$ = 2.0450	0.8334
Estimadores	
$\hat{\beta}_1$ = 0.03582	0.00143
$\hat{\beta}_0$ = -0.4627	0.10481
Prueba de hipótesis de $F_{1,18}$ = 630.716	
Correlaciones	
Pearson = 0.9860	

Capítulo 10
Multicolinealidad en regresión lineal múltiple

10.1. Introducción

Un problema común en el modelo lineal con muchas variables regresoras, especialmente en estudios observacionales, es que algunas variables predictoras están altamente correlacionadas entre sí. Por lo general, puede ser que exista una relación aproximadamente lineal entre las variables predictoras. En tales casos, se dice que las variables predictoras son colineales. La presencia de colinealidad puede ser el aspecto más confuso en regresión lineal. El propósito es proporcionar una indicación de las consecuencias de la colinealidad, en el presente capítulo se describen los métodos para detectarla y se sugieren algunos para lidiar con este problema.

Se trata de describir algunos rasgos de un análisis de regresión que pueden presentar problemas que generan estimadores inadecuados de: (1) los coeficientes de regresión, (2) la variabilidad y (3) los valores de p. Estos problemas pueden ser agrupados en uno de dos tipos: colinealidad y escalamiento (incluyendo centrado). La colinealidad se refiere a la relación entre variables independientes (predictoras). El

escalamiento pertenece a unidades donde las variables bajo estudio son medidas y sus diferencias con respecto a las medias. Cierto tipo de problemas de colinealidad (los que involucran términos de regresión polinomial) pueden ser expresados como problemas de escalamiento y pueden ser resueltos de forma fácil.

El término de colinealidad se refiere a la cercanía lineal entre las variables predictoras. En el caso más simple, suponga que dos variables predictoras están ligeramente correlacionadas. Un diagrama de la dispersión de los datos para estas dos variables mostrará los conglomerados de observaciones en una línea recta, ya que las variables están linealmente relacionadas. La colinealidad puede involucrar más de dos variables predictoras. Considere una situación en la cual los datos para tres de las variables predictoras están cerca de un plano de tres dimensiones. En este caso se dice que las tres variables predictoras son colineales. Podemos ver que tal relación existe cuando la correlación simple no indica que las tres variables están relacionadas. Las colinealidades que involucran tres o más variables predictoras son más difíciles de detectar y la consecuencia puede complicarse para su interpretación.

Una definición precisa de colinealidad no es posible ni apropiada, pero el concepto debe quedar claro. La idea de una relación cercana lineal implica que hay una regresión significativa de una de las variables regresoras sobre una o más de las otras. Lo que no es claro es

qué tan sólida debe ser esta relación para que se diagnostique la colinealidad. Sólo si se adopta un criterio basado en el valor de R^2 para la regresión de una variable predictora sobre las restantes, la definición no es única ya que depende de cuál variable regresora se selecciona como la variable respuesta. Entonces con tres variables predictoras X_1, X_2 y X_3 podemos hacer regresiones de X_1 sobre X_2 y X_3 o alternativamente X_2 sobre X_3 y X_1. Los valores de R^2 diferirán de estas dos ecuaciones de regresión ya que en el primer caso estamos midiendo residuales perpendiculares al plano (X_2, X_3) y en el segundo son medidos perpendicular al plano (X_1, X_3). Entonces la misma colinealidad produce dos diferentes valores de este indicador. Sin embargo R^2 es un indicador útil y fácil de obtener y se usa como otro diagnóstico numérico y gráfico para detectar la presencia de colinealidades. La detección de la colinealidad es el primer paso. La multicolinealidad existe cuando las variables regresoras no son independientes y muestran información redundante.

Es imperativo que los usuarios de la metodología de regresión por mínimos cuadrados, no solamente entiendan los efectos de la multicolinealidad sino que aprendan a diagnosticarla. Los procedimientos de estimación alternativos diseñados para combatir la colinealidad, se establecen en una categoría llamada *técnicas de estimación sesgada*. Estas técnicas representan una desviación de los mínimos cuadrados ordinarios. Un economista que intente extraer

inferencias del signo o magnitud de un coeficiente de regresión específico debe saber que los coeficientes mínimos cuadrados pueden ser estimados de forma inapropiada en presencia de la multicolinealidad. Así también, un ingeniero que está interesado en desarrollar una ecuación lineal de predicción será advertido que solo lo puede hacer a través del ajuste de un buen modelo a sus datos ya que la multicolinealidad puede ser un obstáculo para las predicciones de calidad.

10.2. Diagnóstico de la Multicolinealidad

Muchos de los desarrollos que se presentan en esta sección generan cantidades que sirven como un *diagnóstico de la Multicolinealidad*. Estos nos permiten evaluar la magnitud del problema de multicolinealidad. Ilustramos algunos procedimientos con ejemplos y enfatizamos otros procedimientos que ayudan en la detección de la multicolinealidad severa. Los siguientes representan herramientas formales de diagnóstico.

10.2.1. Matriz de correlación simple entre las variables regresoras

El analista por lo general tiene acceso a la matriz de correlación de las variables regresoras, esto es, X'*X*, dónde las columnas de la matriz X* están centradas y escaladas. Estos números indican correlaciones de tipo pareadas. Sin embargo, puntualizamos que multicolinealidad, como su nombre lo indica, involucra asociaciones entre variables regresoras múltiples. Como un resultado, las correlaciones simples, no siempre dimensionan la extensión del problema. Muchos analistas constantemente, examinan valores patrones de correlaciones, valores sobre los cuales se puede acertar de que se tiene una multicolinealidad severa. Definitivamente no hay valores patrones de correlaciones simples y mientras son observados, el analista puede ver que existen asociaciones uno-a- uno, pero no siempre indican la naturaleza o extensión de la multicolinealidad.

10.2.2. Factores de inflación de Varianza (VIF)

Los factores de inflación de varianza (VIF) representan la inflación que cada coeficiente de regresión experimenta sobre lo ideal, esto es, sobre lo que será experimentado, si la matriz de correlación es una matriz de identidad. El factor de inflación de varianza del *i-ésimo* coeficiente es definido por VIF = $(1 - R_i^2)^{-1}$, dónde R_i^2 es el coeficiente de

determinación múltiple de la regresión producido por la regresión de las variables X_i contra las otras variables regresoras, las X_j ($j \neq i$). Es fácil ver que involucra la noción de asociación múltiple. Si R_i^2 está cerca de la unidad, entonces $(VIF)_i$ será completamente grande. Esto ocurre si la *i-ésima* variable regresora tiene una fuerte asociación lineal con las restantes. Los VIF's representan una aproximación considerablemente más productiva para la detección, que lo que hacen los valores de correlación simple. Ellos proveen al usuario una indicación de cuales coeficientes están adversamente afectados y su magnitud. Aunque no se tienen reglas sobre los valores numéricos, se cree por lo general que si cualquier VIF excede a 10, existe razón para considerar la presencia del problema, entonces se considera la posibilidad de eliminar variables o una alternativa para la estimación por mínimos cuadrados, para combatirlo.

10.2.3. Sistema de Eigenvalores de X'X

Sabemos que los eigenvalores (λ_j) y los eigenvectores de la matriz de correlación juegan un papel importante en la multicolinealidad que existe en un conjunto de datos de regresión. Seguramente, la cercanía a cero de los eigenvalores más pequeños es una medida de la fuerza de una dependencia lineal, mientras los elementos del eigenvector normalizado asociado, muestran los pesos de las correspondientes

variables regresoras en la multicolinealidad. De esta manera, los eigenvalores serán unitarios, si las variables definen un sistema ortogonal y esto provee una norma para el analista. Además, el espectrum de los eigenvalores produce otro diagnóstico. La multicolinealidad puede ser medida, en términos de la razón de los eigenvalores más grandes a los más pequeños. La cantidad,

$$\Phi = \lambda_{max} / \lambda_{min} \qquad (1)$$

El cual es llamado el *número condicional de la matriz de correlaciones*. Valores grandes de Φ indican una multicolinealidad severa. Un número condicional excesivamente grande, es una evidencia de que los coeficientes de regresión son inestables (esto es; sujeto a mayores cambios con pequeñas perturbaciones en los datos de las regresoras). Cuando el número condicional de la matriz de correlaciones excede a 1000, será debido al efecto de la presencia de multicolinealidad. El número de eigenvalores cercanos a cero indican el número de colinealidades detectadas entre las variables regresoras. Como un resultado, es difícil establecer un valor umbral, un valor debajo del eigenvalor pequeño, indica una seria colinealidad. En realidad, razones de eigenvalores; esto es; $\Phi_j = \lambda_{max} / \lambda_j$ son más confiables para diagnosticar el impacto de una dependencia, que los mismos eigenvalores λ_j.

El diagnóstico de la multicolinealidad involucra la consideración de muchos elementos, los cuales serán una parte integral de una impresión de computadora en un análisis de regresión. El analista que no es especialista en matemáticas, puede resistir el uso del sistema de eigenvalores para el diagnóstico, pero su utilidad no se puede negar. Las herramientas de diagnóstico son diseñadas como indicadores de la severidad de la multicolinealidad y para determinar si se puede probar una alternativa de mínimos cuadrados. Sin embargo el analista no puede estar seguro de si la técnica de la estimación sesgada para combatir la multicolinealidad proveerá una estimación o predicción mejorada antes que sea probada la técnica. Entonces el diagnóstico no necesariamente es indicador del probable suceso, de estimación alternativa, sino como un indicador de la ineficiencia de los mínimos cuadrados ordinarios. Por otro lado, el tipo de escala usada tiene una influencia en la naturaleza del diagnóstico.

10.3. Alternativas para mínimos cuadrados en caso de Multicolinealidad

Hay muchos procedimientos de estimación diseñados para combatir la multicolinealidad, procedimientos que fueron desarrollados para eliminar la inestabilidad del modelo y para reducir la varianza de los coeficientes

de regresión. Existe una cierta cantidad de controversias alrededor de su uso, por lo que el analista no sentirá que posee una carta en blanco para usar las técnicas en cualquier momento. Pueden ser una parte valiosa del repertorio del analista de datos, particularmente cuando el sujeto de estudio involucra estructura relacionada con variables científicas, que inherentemente solapen su influencia en la variable dependiente.

El punto que el analista ha determinado para el diagnóstico, cuando la multicolinealidad es un problema; por lo general es un beneficio sustancial que puede ser derivado de un intento para eliminar mucha de la multicolinealidad sin recurrir a otras alternativas de mínimos cuadrados. La presencia de la multicolinealidad en el diagnóstico, sugiere para el caso de k variables regresoras, que el ejercicio de modelación involucra menos de k variables. En otras palabras no existe información suficiente en los datos para garantizar la modelación de las k variables. Como un resultado, el analista puede eliminar o ciertamente reducir el efecto de la multicolinealidad removiendo una o más variables regresoras. Cualquiera de los diagnósticos y conociendo el fenómeno modelado, puede sugerir cuales variables son candidatos para eliminarse. En una regresión múltiple, si X_1 y X_2 están ligeramente correlacionadas y X_2 se elimina del modelo, entonces se elimina algo de la multicolinealidad. Sin embargo, el analista debe estar consciente de si o no la calidad de ajuste del modelo ha sido comprometida. La mejora en la capacidad predictiva del modelo

por la eliminación de la variable será reflejada en estadísticos tales como PRESS, la suma de los residuales PRESS y los errores estándares de predicción. Además, el analista se enfocará sobre el factor de inflación de la varianza para los coeficientes restantes.

Una alternativa para la reducción de la multicolinealidad, utilizando la estimación de mínimos cuadrados ordinarios, es realizar transformaciones de las variables regresoras; estas transformaciones reducen la dimensionalidad del sistema regresor, mientras retienen algo del contenido informativo de todas las regresoras. Por ejemplo, en un caso de dos regresoras, X_1 y X_2 que están altamente correlacionadas, redefiniendo una variable $X_1 + X_2$, o quizás, formando razones puede producir un resultado efectivo. Se deberá tener mucho cuidado en la formación de funciones con las variables regresoras, que no tienen sentido en el contexto del problema. El analista puede ser muy escrupuloso para adicionar variables regresoras que están medidas en diferentes unidades.

Ejemplo: Los datos de hospitales de 5 variables regresoras y una dependiente (Y) de 17 sitios hospitalarios de todo el mundo. La variables regresoras son: Y = a las horas –hombres mensualmente, X_1 = al porcentaje de trabajadores incapacitados diariamente X_2 = a la exposición mensual a los rayos X, X_3 = a los días de ocupación de las cama por mes y X_4 = a la población elegible en el área /1000 y X_5 = a la longitud promedio en días de trabajadores encamados. La tarea aquí

es construir un modelo empírico que estime o prediga las necesidades de mano de obra de los hospitales.

Tabla 10.1
Datos del hospital

Sitio	X_1	X_2	X_3	X_4	X_5	Y
1	15.57	246.3	472.92	18.0	4.45	566.52
2	44.02	2048	1339.75	9.5	6.92	696.82
3	20.42	3940	620.25	12.8	4.28	1033.15
4	18.74	6505	568.33	36.7	3.90	1603.62
5	49.20	5723	1497.60	35.7	5.50	1611.37
6	44.92	11520	1365.83	24.0	4.60	1613.27
7	55.48	5779	1687.00	43.3	5.62	1854.17
8	59.28	5969	1639.92	46.7	5.15	2160.55
9	94.39	8461	2872.33	78.7	6.18	2305.58
10	128.02	20106	3655.08	180.5	6.15	3503.93
11	96.00	13313	2912.00	60.9	5.88	3571.89
12	131.42	10771	3921.00	103.7	4.88	3741.40
13	127.21	15543	3865.67	126.8	5.50	4026.52
14	252.90	36194	7684.10	157.7	7.00	10343.81
15	409.20	34703	12446.33	169.4	10.78	11732.17
16	463.70	39204	14098.40	331.4	7.05	15414.94
17	510.22	86533	15524.00	371.6	6.35	18854.45

Es nuestro interés determinar si un subconjunto de las variables regresoras se puede seleccionar aunque presenten poca o una suave colinealidad pero muestren una buena capacidad predictiva. Se ajustaron varios modelos. Aquí observamos los factores de inflación de varianza (VIF), para determinar si, realmente las variables eliminadas reducen sustancialmente la colinealidad. Lo siguiente es una sinopsis de colinealidad para varios subconjuntos de modelos que parecen tener

buenas propiedades predictivas. Se incluye el modelo completo para tomarse como base de comparación.

Análisis de regresión de los datos.

Coeficientes	**VIF**	**MODELO (X_1, X_2, X_3, X_4, X_5)**
β_1	9597.5708	$\hat{y} = 1962.948 - 15.8517X_1 + 0.0559X_2 + 1.5896X_3 - 4.2187X_4 - 394.314X_5$
β_2	7.9406	
$\beta 3$	8933.086	
$\beta 4$	23.2939	
$\beta 5$	4.2798	$R^2 = 0.9908$

Coeficientes	**VIF**	**MODELO (X_1, X_5)**
β_1	1.8199	$\hat{y} = 2537.413 + 37.5336X_1 - 530.116X_5$
β_5	1.8199	
		$R^2 = 0.9840$

Coeficientes	**VIF**	**MODELO (X_3, X_5)**
β_3	1.8195	$\hat{y} = 2585.520 + 1.2324X_3 - 530.933X_5$
β_5	1.8195	
		$R^2 = 0.9948$

Coeficientes	**VIF**	**MODELO (X_1, X_2)**
β_1	5.6605	$\hat{y} = - 96.9009 + 25.0211X_1 - 0.0752X_2$
β_2	5.6605	
		$R^2 = 0.9961$

Coeficientes	**VIF**	**MODELO (X_2, X_3)**
β_2	5.6471	$\hat{y} = - 68.3140 + 0.0749X_2 + 0.8229X_3$
β_3	5.6471	
		$R^2 = 0.9967$

Coeficientes	VIF	MODELO (X_1, X_3, X_5)
β_1	5207.441	$\hat{y} = 2630.478 - 41.0544X_1 + 2.5795X_3 - 529.777X_5$
β_3	5206.324	
β_5	1.8199	
$R^2 = 0.9950$		

Coeficientes	VIF	MODELO (X_2, X_3, X_5)
β_2	7.7373	$\hat{y} = 1523.389 + 0.0530X_2 + 0.9785X_3 - 320.951X_5$
β_3	11.2693	
β_5	2.4929	
$R^2 = 0.9901$		

Coeficientes	VIF	MODELO (X_1, X_2, X_5)
β_1	11.3214	$\hat{y} = 1475.024 + 29.7316X_1 + 0.0534X_2 - 318.140X_5$
β_2	7.7714	
β_5	2.4985	
$R^2 = 0.9894$		

El modelo (X_1, X_3, X_5) es claramente inaceptable desde el punto de vista de la colinealidad. Esto se debe a la fuerte correlación entre X_1 y X_3. Los otros modelos que involucran tres variables regresoras, tienen colinealidad que puede ser considerada marginal, mientras que todos los modelos con dos variables regresoras muestran solamente una muy leve colinealidad. Aparte del modelo (X_1, X_3, X_5) todos los modelos estudiados aquí permanecen como candidatos ya que reducen sustancialmente la colinealidad. Note que la variable X_5, siempre tiene un coeficiente de regresión negativo. Entonces para la interpretación de los coeficientes y además con poca o nada de colinealidad, los modelos (X_1, X_2) y (X_2, X_3) son fuertes competidores. Además estos dos modelos cumplen con otros criterios de selección del mejor modelo.

10.4. Colinealidades evitables

Algunos problemas con los cálculos de regresión no pueden ser fácilmente detectables debido a la presencia de la colinealidad. Es por lo tanto importante, cuando realizamos un análisis de regresión se debe tener cuidado con peligros potenciales. Los siguientes ejemplos fueron seleccionados para mostrar ocurrencias comunes en un análisis de regresión que pueden mostrar colinealidades cercanas (o solo para dependencia lineales exactas entre las variables predictoras). En muchas situaciones, se considera como predictores, las potencias de una variable continua. Un ejemplo es el uso de X_1, X_1^2 y X_1^3, como predictores en un análisis de regresión. Estos predictores se denominan polinomios naturales y el uso cuidadoso de ellos puede mostrar colinealidades cercanas. Una solución a este problema, además del centrado, es usar *polinomios ortogonale*s.

Los problemas de colinealidad, pueden aparecer si algunos valores de datos extremos se incluyen incorrectamente en el conjunto de datos por errores en su colección. Este problema de manejo de datos puede ser detectado usando los métodos para analizar outliers (puntos lejanos del conjunto de interés). Otra vez el lector debe tener precaución para descartar sin resentimientos las observaciones incómodas. El uso de variables dummy puede inadvertidamente introducir colinealidades exactas. Los términos de interacción generalmente crean la atmósfera

para los problemas de colinealidad, especialmente si tales problemas son sobreusados. Por ejemplo, si los predictores son, edad, peso y edad x peso, entonces una colinealidad cercana puede mostrarse debido a la estrecha relación funcional entre el término del producto y los dos predictores básicos, edad y peso. La cantidad de colinealidad introducida dependerá del rango de los pesos y los pesos en los datos y el número de valores repetidos de cada variable. En general, cuando se ajusta un modelo que contiene varios términos de interacción siempre puede aparecer algún problema de colinealidad. Un caso especial muy importante, en que las interacciones no son incómodas, ocurre en los diseños de ANVA con igual o cerca de igual número de observaciones por celda.

Una forma ingeniosa de colinealidad cercana ocurre con los siguientes grupos de predictores; ingreso familiar, educación, número de años laboral y edad. Estas cuatro variables tienden a estar muy positivamente correlacionadas una con otra y es entendible que una de las cuatro es perfectamente predicha por las restantes (alguna combinación lineal). Esto ilustra uno de los más incómodos tipos de colinealidad. Idealmente se puede evitar el problema eliminando una o más de las variables. Si esta solución no es aceptable, entonces se pueden emplear ciertos métodos no muy comunes de análisis (tales como regresión de cordillera).

10.5. Problemas de escalamiento

Una clase general de problemas en el análisis de regresión que puede aparecer, es el de escalamiento impropio de las variables regresoras y de respuesta. Especialmente, tales problemas se relacionan a la pérdida de precisión en los cálculos. Los resultados inadecuados pueden ser grandes que generan coeficientes estimados con los signos equivocados. Un problema de escalamiento puede ocurrir si un predictor tiene un amplio rango de valores. Por ejemplo: el peso corporal de personas adultas en gramos puede ser problemático. Otro ejemplo, un problema puede ocurrir con datos de personas que usan grandes cantidades de vitaminas C, más que las personas promedios. Tales valores medidos propiamente de ingestión de vitaminas C serán outliers (puntos lejanos del conjunto) y su uso presentará problemas en el modelo. Similarmente si la media es grande con pequeña variabilidad, pueden aparecer problemas de cálculos. Muchos problemas de escalamiento se pueden evitar por una validación de los datos y de reescalamiento antes de aplicar el análisis de regresión. Algunos problemas, tales como la ingestión de vitamina C pueden requerir cambios en alguna parte de la estrategia del modelaje.

10.6. Tratamiento para la colinealidad y problemas de escalamiento

Ambos la colinealidad cercana y un mal escalamiento producen problemas numéricos, incluyendo los cálculos inapropiados de: (1) estimadores de los coeficientes de regresión, (2) estimador del error estándar, (3) pruebas de hipótesis estadísticas. En el peor de los casos, el análisis puede cambiar sustantivamente si los datos son procesados en un programa en un orden diferente, o si se eliminan pocas observaciones. Todo esto reduce nuestra confianza en el análisis.

El primer paso para tratar los problemas numéricos, es validar los datos adecuadamente antes de intentar cualquier modelación de regresión. Los procedimientos para esta validación pueden detectar más problemas de escalamiento y sugerir soluciones. Para variables continuas, los rangos numéricos de las variables predictoras serán tan similares como sea posible. Por ejemplo; no medir pesos en gramos y pesos en toneladas. En general, los números que son convenientes para escribir y graficar deben ser los mismos para el análisis. Típicamente estos dan rangos tales como de 1 a 10 o de 10 a 100.

El segundo paso para tratar problemas numéricos es utilizar regresión diagnóstica y el diagnóstico de colinealidad (esto es; eigenvalores e índices condicionales). Estos métodos permiten detectar problemas numéricos pero no necesariamente indican una buena solución. El tercer paso para tratar problemas numéricos, es intentar

eliminar variables redundantes. Un ordenamiento de la importancia de las variables es central en esta tarea. Muchas técnicas sofisticadas involucran formalizar la selección de las variables. Así también el análisis de componentes principales puede ser aplicado al proceso de reducción del número de variables.

10.7. Estrategias de análisis alternativos

Cuando uno o más de los supuestos fundamentales básicos del análisis de regresión no son claramente satisfechos y/o cuando los problemas numéricos son identificados, el analista puede utilizar otra estrategia de análisis. En las subsiguientes secciones se enlistan (1) algunos métodos de análisis alternativos, (2) se menciona ampliamente la regresión lineal que puede ser adecuado en algunas aplicaciones y (3) se describe ampliamente las transformaciones a los datos que pueden permitir el uso de regresión lineal múltiple.

10.7.1. Aproximaciones alternantes

Si el investigador decide que el modelo de regresión usado no ajusta los datos y no puede ser ajustado vía una simple generalización de regresión lineal (tal como mínimos cuadrados pesados) entonces se

pueden usar otros métodos. Si la variable respuesta no puede ser modelada como una función lineal de los parámetros, entonces se usan los métodos de funciones no lineales. Si el problema consiste de distribuciones no normales, entonces se pueden usar los métodos de rangos o de análisis de datos categóricos por ser los más apropiados.

Así también algunos métodos de análisis de rangos pueden ser considerados como un proceso esencialmente de dos pasos. El primer paso es reemplazar los datos originales con rangos apropiados. El segundo es conducir el análisis de regresión lineal de los rangos. El lector debe ser cuidadoso y no presumir que esta aproximación trabaja en cualquier caso particular, pero la discusión es para entender los métodos de rangos.

10.7.2. Generalizaciones de regresión lineal

a) Métodos de medición

Generalizaciones de regresión lineal pueden ser agrupadas en métodos exactos y aproximados. Los métodos exactos tienen procedimientos de estimación y prueba de hipótesis con propiedades conocidas para muestras finitas, mientras que los métodos de aproximación tienen solamente resultados asintóticamente disponibles. Por lo tanto los métodos de aproximación deben ser usados con cuidado en muestras

pequeñas. Los métodos aproximados generalizados incluyen la regresión de componentes principales, la regresión de cordillera y la regresión robusta. Las generalizaciones exactas incluyen, las técnicas multivariadas y mínimos cuadrados pesados exactos.

La regresión de componentes principales y la regresión de cordillera se recomiendan a menudo para tratar los problemas de colinealidad. En el análisis de componentes principales, las variables predictoras originales son reemplazadas por un conjunto de variables no correlacionadas mutuamente, el registro del componente principal. Si es necesario, los componentes asociados con eigenvalores cercanos a cero son eliminados del análisis, lo que elimina el problema de la colinealidad concomitante. La regresión de cordillera, involucra la perturbación de los eigenvalores de la variable predictora original, la matriz de productos cruzados para definirlos diferentes de cero, lo que reduce la cantidad de colinealidad. Ambos métodos proporcionan estimadores de regresión sesgados de los parámetros del modelo de interés. Además los valores de *p* para las pruebas estadísticas pueden ser óptimamente pequeños cuando se usan tales métodos de estimación sesgados.

La regresión robusta involucra pesos o transformaciones de los datos, para minimizar el efecto de observaciones extremas. La tarea es hacer el análisis más robusto (menos sensible) a cualquier observación

y también menos sensible a los supuestos básicos del análisis de regresión.

Los métodos multivariados incluyen, la regresión múltiple multivariada, el análisis multivariado de varianza, análisis discriminante, correlación canónica y el análisis de curvas de crecimiento, entre otras. En todos estos procedimientos, se puede contar la no independencia entre las observaciones ya que estas independencias son explícitamente modeladas. Los tipos de variables predictoras y de respuesta que pueden ser usadas en estos métodos difieren principalmente en que tipos de hipótesis son probadas y como se modela la no independencia.

b) Análisis de mínimos cuadrados pesados

El método de análisis de mínimos cuadrados pesados es una modificación del procedimiento estándar del análisis de regresión y se usa cuando un modelo de regresión es ajustado a un conjunto de datos para los cuales el supuesto de homogeneidad de varianza y/o independencia no se cumple. Cuando la varianza de *Y* varía para diferentes valores de las variables independientes, el análisis de mínimos cuadrados pesados puede ser usado, siempre que estas varianzas sean conocidas (esto es; σ_i^2 para la *i-ésima* observación de *Y*) o puedan ser consideradas de la forma $\sigma_i^2 = \sigma^2/W_i$, dónde los pesos ($W_i$) son conocidos.

10.7.3. Transformaciones

Las tres razones principales para usar la transformación de datos son: (1) para estabilizar la varianza de la variable dependiente si el supuesto de homocedasticidad es violado, (2) para normalizar (esto es; para transformar la distribución normal) la variable dependiente si el supuesto de normalidad es notablemente violado y (3) para linealizar el modelo de regresión si los datos originales sugieren un modelo que no es lineal, en todos los coeficientes de regresión y/o las variables originales (dependiente o independientes). Es una fortuna que la misma transformación por lo general ayuda, para acompañar las primeras dos tareas y algunas veces también la tercera, más que lograr una tarea a costa de los otros dos.

Una mejor discusión de las propiedades de varias transformaciones puede ser encontrada en Armitage, 1971; Draper y Smith, 1981; Neter *et al*.,1983). Además, Box y Cox (1964); (Carroll y Ruppert, 1984) describieron una aproximación para hacer una investigación exploratoria de una familia de transformaciones. Por lo que se considera útil describir un poco las transformaciones más usadas.

a) La transformación logarítmica (Y' = log *Y*) es usada para estabilizar la varianza de *Y*, si se incrementa marcadamente con incrementar a *Y*, para normalizar la variable dependiente

si la distribución de los residuales para *Y* es positivamente sesgada y linealizar el modelo de regresión si la relación de *Y* para alguna variable independiente sugiere un modelo con incrementos consistentes de la pendiente.

b) La transformación de raíz cuadrada (Y' = $\sqrt{Y}$). Se usa para estabilizar la varianza si es proporcional a la media de *Y*. Esto es apropiado si la variable dependiente tiene una distribución de Poisson.

c) La transformación recíproca (Y' = $\frac{1}{Y}$). Se usa para estabilizar la varianza si es proporcional a la cuarta potencia de la media de *Y*, indica que existe un enorme incremento de la varianza sobre algunos valores umbrales de *Y*. Esta transformación minimiza el efecto de grandes valores de *Y* puesto que para estos valores, los valores Y' son cercanos a cero y grandes incrementos en *Y*, causan solamente pequeños decrementos en Y'.

d) La transformación cuadrada (Y' = Y^2). Se usa para estabilizar la varianza, si esta decrece con la media de *Y*, para normalizar la variable dependiente si la distribución de los residuales para *Y* es negativamente sesgado y para linealizar el modelo si la relación original con alguna variable independiente es curvilínea inclinada hacia abajo.

e) La transformación arc seno (Y' = arcseno $\sqrt{Y}$ = sen $^{-1}\sqrt{Y}$). Se usa para estabilizar la varianza si *Y* en una proporción o razón.

10.8. Un punto importante

Todas las técnicas en este capítulo involucran el chequeo de la validez de los supuestos y estimadores de un análisis de regresión. Con dirección al final, el análisis de outlier y la reducción de la colinealidad sugieren eliminar todas las observaciones o variables predictoras para mejorar la calidad del modelo. Naturalmente, las observaciones y variables que son desiguales en el modelo ajustado son eliminadas. Todas las técnicas hacen una sobreestimación de la variabilidad y generan valores de *p* que son óptimamente pequeños. La pérdida potencial asociada con el mejoramiento en el ajuste del modelo es que genera resultados sesgados. Los métodos discutidos en este capítulo son seguros para usarse en muestras grandes y son menos confiables en muestras pequeñas. Desafortunadamente, la influencia potencial de una sola observación es más grandes que en muestras pequeñas. Esto es un fuerte argumento contra el uso de muestras pequeñas en análisis de regresión.

En algunos casos un problema puede ser resuelto solamente evaluando una segunda muestra de datos. Picard y Cook (1984) discutieron el beneficio del optimismo en la selección de un modelo de regresión. Ellos recomendaron, considerar las muestras divididas. En un diseño de muestras divididas, parte de los datos se usa para el análisis exploratorio y los restantes se usan para confirmar la validez y confiabilidad de los resultados exploratorios.

10.9. Ejercicios

10.1 Se realizó un estudio con 19 pacientes asmáticos en una ciudad tropical de México, considerando la variable: volumen de expiración forzada por segundo (Y) como variable dependiente y la edad (X_1), el sexo (X_2), la estatura (X_3) y el peso (X_4). Los datos son:

	Edad	Sexo	Estatura	Peso	Y
1	24	M	175	78.0	4.7
2	36	M	172	67.6	4.3
3	28	F	171	98.0	3.5
4	25	M	166	65.5	4.0
5	26	F	166	65.0	3.2
6	22	M	176	65.5	4.7
7	27	M	185	85.5	4.3
8	27	M	171	76.3	4.7
9	36	M	185	79.0	5.2
10	24	M	182	88.2	4.2
11	26	M	180	70.5	3.5
12	29	M	163	75.0	3.2
13	33	F	180	68.0	2.6
14	31	M	180	65.0	2.0
15	30	M	180	70.4	4.0
16	22	M	168	63.0	3.9
17	27	M	1686	91.2	3.0
18	46	M	178	67.0	4.5
19	36	M	173	62.0	2.4

a) Ajuste un modelo de Y en función de la estatura, el peso y la edad.

b) Realice pruebas para las variables que se adicionan como las últimas para todas las predictoras y una prueba para el intercepto.

c) Defina factores de inflación de varianza para cada predictor.

d) Defina la matriz de correlaciones incluyendo todas las variables predictoras y la respuesta.

e) Defina los eigenvalores, los índices condicionales y el número de condición para la matriz de correlación (incluyendo el intercepto).

f) Defina los eigenvalores, los índices condicionales y los números de condición para la matriz de productos cruzados escalados (incluyendo el intercepto)

g) Defina los residuales (preferiblemente estudentizados) y los valores de punto de apoyo.

10.2 Crear una nueva variable FEMENINO para los datos del problema anterior, donde FEMENINO = 1, si el sexo es Femenino y FEMENINO = 0 si es masculino. Repetir los incisos del (a) hasta (g) del problema anterior adicionando la variable predictor FEMENINO.

10.3 Crear tres nuevas variables de interacción para los datos del problema 10.2.

X_6 = Femenino x Edad

X_7 = Femenino x Estatura

X_8 = Femenino x Peso

Repita del inciso (a) hasta (g) del problema anterior adicionando la variable Femenino y las tres variables de interacción definidas en este problema como predictoras.

h) Explique cómo la proporción de la variable Femenino dificulta el problema anterior y este. Sugiera una solución para analizar estos datos y una para futuras investigaciones.

10.4 Freund (1979) realizó un estudio para analizar la evaporación diaria del suelo (Y) como una función de las siguientes variables predictoras: la temperatura máxima del aire (X_1), temperatura mínima del aire (X_2), la temperatura promedio del aire (X_3), la temperatura máxima del suelo (X_4), la temperatura mínima del suelo (X_5), la temperatura promedio del suelo integrada (X_6), la humedad relativa máxima (X_7), la humedad relativa mínima diaria (X_8), la curva de humedad integrada bajo la curva (X_9) y el viento total diario (X_{10}). La tabla del ANVA y las pruebas de los coeficientes son:

Fuente	**G.L**	**SC**	**SCM**	**F_0**	**Variable**	$\widehat{\boldsymbol{\beta}}$	**t**	**VIF**
Regresión	10	8159.83	815.98	19.27	X_1	0.5011	0.88	8.828
Residual	35	1482.27	42.35		X_2	0.3041	0.39	8.887
Total	45	9642.11			X_3	0.09219	0.42	22.21
					X_4	2.232	2.22	39.29
					X_5	0.2049	0.19	14.08
					X_6	0.7426	-2.12	52.36
					X_7	1.110	0.98	1.9981
					X_8	0.7514	1.54	25.38
					X_9	-0.5563	-3.44	24.12
					X_{10}	0.00892	0.97	1.985

a) Calcule el valor de R^2 y realice una prueba de significancia usando un $\alpha = 0.01$.

b) Calcule el valor de R^2 para cada predictor.

c) ¿Cuáles variables están implicadas como productoras de colinealidad?

d) Usa tus conocimientos de evaporación, explicar porqué algunos coeficientes tienen signos inversos, poniendo atención especial a los que tienen valores de t extremos.

10.5. En un esfuerzo por hacer un modelo para las compensaciones en el año 2008, se seleccionaron 33 empresas del ramo hotelero y sus datos sobre compensaciones, ventas, utilidades y empleo. Los datos son

Empresa	Comp (y)	Ventas (X_1)	Utilidad (X_2)	Empleo (X_3)	Empresa	Comp (y)	Ventas (X_1)	Utilidad (X_2)	Empleo (X_3)
1	450	4600.6	128.1	48000	18	225	578.9	63.3	4139
2	387	9255.4	783.9	55900	19	254	966.8	42.8	6255
3	368	1526.2	136.0	13783	20	208	591.0	48.5	10605
4	277	1683.2	179.0	27765	21	518	4933.1	310.6	65392
5	676	2752.8	231.5	34000	22	406	7613.2	491.6	89400
6	454	2205.8	329.5	26500	23	332	3457.4	228.0	55200
7	507	2384.6	381.8	30800	24	340	545.3	54.6	7800
8	496	2746.0	237.9	41000	25	698	22869.8	3011.3	337117
9	487	1434.0	222.3	25900	26	306	2361.0	203.0	52000
10	383	470.6	63.7	8600	27	613	2614.1	201.0	50500
11	311	1508.0	149.5	21075	28	302	1013.2	121.3	18625
12	271	464.4	30.0	6874	29	540	4560.3	194.6	97937
13	524	9329.3	577.3	39000	30	293	855.7	63.4	12300
14	498	2377.5	250.7	34300	31	528	4211.6	352.1	71800
15	343	1174.3	82.6	19405	32	456	5440.4	655.2	87700
16	354	409.3	61.5	3586	33	417	1229.9	97.5	14600
17	324	724.7	90.8	3905					

Considere el modelo: $Y_i = \beta_0 + \beta_1 \ln X_{1i} + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + \varepsilon_i$

a) Ajuste el modelo anterior con los datos.

b) Calcule la matriz de correlación de las variables regresoras.

c) Calcule los eigenvalores de la matriz de correlación.

d) Calcule los factores de inflación de varianza para los coeficientes del modelo anterior.

e) Valore la dimensión de la multicolinealidad en este problema.

f) Calcule los factores de inflación de varianza, los eigenvalores de la matriz de correlación y la proporción de la descomposición de la varianza.

g) Use la regresión de cordillera (ridge) para la estimación de los coeficientes. Use *k* sobre la base de la estabilidad de los coeficientes.

h) Use la regresión ridge (cordillera) con los valores de *k* determinados considerando PR (Ridge).

i) Estime los coeficientes usando la regresión ridge con una *k* seleccionada y considerando el criterio C_k.

Capítulo 11
Seleccionando la mejor ecuación de regresión

11.1. Introducción

Suponga una variable respuesta Y, y un grupo de k variables predictoras X_1, X_2, X_3, ... X_k. Se pretende determinar el mejor subconjunto de las k predictoras y el correspondiente modelo de regresión mejor ajustado para describir la relación entre Y y las X's. ¿Qué entendemos exactamente por "mejor"?, pues depende en parte de la tarea total en la modelación. Dos tareas son importantes que se describen a continuación. Una es encontrar un modelo que proporcione la mejor predicción de Y dado X_1, X_2, X_3, ..., X_k, para alguna observación o paquete de nuevas observaciones. En la práctica enfatizamos la estimación de la regresión de Y sobre las X's, o sea, $E(Y/X_1, X_2, X_3, ..., X_k)$ que expresan la media de Y como una función de las predictoras. Usando esta tarea, podemos decir que el mejor modelo es confiable si predice bien una nueva muestra. Los detalles del modelo pueden ser pequeños sin consecuencia, tal como la inclusión de cualquier variable o la magnitud o signos de sus coeficientes de regresión. Por ejemplo, considerando una muestra de presión sistólica sanguínea, la tarea puede

ser simplemente predecir la presión sanguínea, como una función de variables demográficas, como edad, raza y género. No se puede cuidar qué variables están en el modelo o cómo están definidas, pero el modelo final obtenido debe proporcionar la mejor predicción posible.

Además de la cuestión de predicción se tiene la cuestión de validez, esto es, obtener estimadores válidos para uno o más coeficientes de regresión en un modelo y hacer la inferencia acerca de los parámetros de interés correspondientes. La tarea aquí es cuantificar la relación entre una o más variables independientes de interés y la variable dependiente, controlando las otras variables. Cómo un ejemplo, puede describir la relación de la presión sistólica sanguínea y edad, controlando *raza y género*. En este caso, se está enfocando a coeficientes de regresión que involucran edad (incluyendo funciones de edad), sólo que raza y género pueden permanecer en el modelo para propósito de control, sus coeficientes de regresión no son de interés.

En este capítulo se enfocan las estrategias para seleccionar el mejor modelo, cuando la tarea principal del análisis es la predicción.

11.2. Pasos en la selección de la mejor ecuación de regresión

En la selección de la mejor ecuación de regresión, se recomiendan los siguientes pasos:

1) Especificar el máximo modelo que se puede construir.
2) Especificar el criterio para la selección del modelo.
3) Especificar una estrategia para aplicar el criterio.
4) Conducir el análisis especificado.
5) Evaluar la confiabilidad del modelo seleccionado.

De acuerdo a los pasos anteriores la idea principal es encontrar el mejor predictor de *Y* que se puede convertir en acciones concretas. Cada paso nos asegura confiabilidad y reduce el trabajo. La especificación del máximo modelo obliga al analista a (1) establecer claramente la tarea del análisis, (2) reconocer las limitaciones de los datos y (3) describir los rangos explícitamente de los modelos plausibles. Todos los programas de computadoras, para seleccionar modelos requieren que se especifique un máximo modelo. Haciendo esto, el analista considera todos los conocimientos científicos disponibles. De esta manera, especificando el criterio de selección del modelo y la estrategia para aplicar este criterio simplifica y facilita el proceso de análisis. Finalmente, si la tarea es predicción o validez, la confiabilidad del modelo seleccionado se debe demostrar.

11.2.1. Paso 1: especificando el máximo modelo

El máximo modelo se define como el modelo más grande considerado en cualquier punto del proceso de selección del modelo. Todos los otros posibles modelos, pueden ser creados eliminando variables predictoras del máximo modelo. Un modelo creado eliminando predictoras del máximo modelo se denomina una *restricción* del máximo modelo.

En todo este capítulo se hace el supuesto muy importante que el máximo modelo con *k* variables o alguna restricción con $p \leq k$ variables es el modelo correcto para la población. Una implicación importante de este supuesto es que la correlación múltiple cuadrada poblacional para el máximo modelo es denominada $\rho^2 (Y/X_1, X_2, X_3, ..., X_k)$ que no es más grande que la del modelo correcto (que puede tener pocas variables); como un resultado, adicionando más predictoras al modelo correcto no se incrementa la correlación múltiple cuadrada poblacional para el modelo correcto.

Para ilustrar esto considere los siguientes datos. Suponga que se tiene una muestra aleatoria de 12 niños, que son atendidos en una clínica. Se miden las variables Peso (Y), altura (X_1) y edad (X_2) para cada niño.

Niños	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Y	64	71	53	67	55	58	77	57	56	51	76	68
X_1(altura)	57	59	49	62	51	50	55	48	42	42	61	57
X_2(Edad)	8	10	6	11	8	7	10	9	10	6	12	9

Un posible modelo (no necesariamente el máximo) puede ser,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \quad (1)$$

Este modelo solo permite considerar una relación lineal entre la respuesta *Y* y las dos predictoras (altura y edad). No obstante, la naturaleza de crecimiento sugiere, que la relación aunque monótona, en ambos, altura y edad, puede ser no lineal. Esto implica que al menos un término cuadrático es necesario incluir en el modelo. Puesto que altura y edad están ligeramente correlacionadas y el tamaño de la muestra es muy pequeño, se considera solamente $(\text{edad})^2$. ¿Serán consideradas las interacciones entre edad y altura? ¿Serán consideradas las transformaciones de las predictoras y la variable respuesta? Las limitaciones de los datos nos permiten definir el máximo modelo cómo,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (2)$$

Con X_1 = altura, X_2 = Edad, y $X_3 = (\text{Edad})^2$

Existen muchas razones para seleccionar un máximo modelo. El más importante es el de evitar cometer el Error Tipo II (falso o negativo). En un análisis de regresión, un Error Tipo II, corresponde a omitir una variable predictora que tiene un coeficiente de regresión verdaderamente diferente de cero en la población. Otras razones para considerar un máximo modelo, son los de incluir:

a) Todas las predictoras básicas concebibles.
b) Potencias de predictoras básicas de alto orden [$(edad)^2$, $(altura)^2$].
c) Otras transformaciones de predictoras [log (edad), 1/altura].
d) Interacciones entre predictoras (edad x altura) incluyendo interacciones de más alto orden y,
e) Todas las posibles variables de "control", así como sus potencias e interacciones.

Recordar que un modelo sobreestimado no introduce sesgos en los coeficientes de regresión poblacionales estimados si se cumplen los supuestos de la regresión. Sin embargo, se debe tener cuidado, en que la sobreestimación está relacionada con la colinealidad nociva. También la subestimación introduce sesgos en los coeficientes de regresión estimados.

Hay algunas razones importantes para trabajar con un primer modelo máximo. Con una tarea de predicción, la necesidad de una fuerte confiabilidad se justifica para un primer modelo máximo y con una tarea de validez, se concentra en pocas variables importantes. En cada caso, se desea evitar un error tipo I (falso, positivo). En un análisis de regresión, un error tipo I corresponde a incluir una variable predictora que tiene un coeficiente de regresión poblacional de cero. Prácticamente las predictoras no importantes, pero estadísticamente significativas, pueden confundir la interpretación de los resultados de la regresión; los términos de interacciones complejas son particularmente penosas en este sentido. La muestra de datos particular que se analiza, impone ciertas restricciones sobre la selección del máximo modelo. En general, se establece que con un tamaño de muestra más pequeño, el máximo modelo será más pequeño. La idea general es que un número grande de observaciones independientes, se necesitan para estimar confiablemente un número importante de coeficientes de regresión. Esta noción ha establecido varias guías en relación al tamaño del máximo modelo. La restricción básica, es la de que los grados de libertad del error son positivos. Simbólicamente se requiere,

Grados de libertad del error = $n - k - 1 > 0$, que es equivalente a la restricción $n > k + 1$

Como siempre, *n* es el número de observaciones y *k* el número de variables predictoras, dando *k + 1* coeficientes de regresión

(incluyendo el intercepto). Con grados de libertad del error negativos el modelo tiene al menos una colinealidad perfecta, consecuentemente, con estimadores únicos de los coeficientes y las varianzas no se pueden calcular. Con grados de libertad del error cero ($n = k + 1$), los estimadores únicos de los coeficientes se pueden calcular, pero no los de las varianzas [recordar que $\hat{\sigma}^2$ = (Suma de Cuadrados del error) /grados de libertad del error]. Además, si la correlación poblacional es cero, entonces, $R^2 = 1.0$, cuando los grados de libertad del error es igual a cero, refleja el hecho de que el modelo ajusta exactamente los datos observados. Este es el ejemplo más extremo del sesgo positivo en R^2. En tal situación, se ha intercambiado *n* valores de *Y* para n coeficientes de regresión estimados.

La cuestión entonces, parece que es, cuántos grados de libertad del error son necesarios. Algunas reglas simples se pueden sugerir para asegurar que los estimadores para un solo modelo son confiables, pero esas reglas son inadecuadas, cuando se considera una serie de modelos.

El requerimiento más débil es un mínimo de aproximadamente 10 grados de libertad del error, esto es, $n - k - 1 \geq 10$ o puede ser, $n \geq 10 + k + 1$. Otra regla para manejar la regresión que se ha sugerido, es tener al menos 5 (o 10) observaciones por predictor, o sea se requiere que $n \geq 5k$ (o $n \geq 10k$).

Asumir, por ejemplo, que se considera un máximo modelo involucrando 30 variables predictoras. Para tener 10 grados de libertad

del error, se requiere una muestra de tamaño 41, mientras que $n > 5k$, demanda un tamaño de muestra de 150 observaciones.

Otra restricción sobre el máximo modelo, está relacionada a la cantidad de variabilidad presente en el valor predictor, considerándolos individualmente o conjuntamente. Si un predictor tiene el mismo valor para todos los sujetos, entonces obviamente no puede ser usado en cualquier modelo. Por ejemplo, considere la variable Género de niños (varón = 1, mujer = 0) como un predictor candidato, para el ejemplo del peso. Si todos los sujetos son varones, entonces el Género = 1, para todos los sujetos, la varianza muestral de la variable es 0.0 y es perfectamente colineal, con la variable intercepto. Claramente, si todos los sujetos son de un solo género, entonces las comparaciones entre género no se pueden hacer.

Similarmente considerar la variable Género, Raza (blanco = 0, negro = 1) y su interacción (género x raza). Si las mujeres negras no son representadas en los datos, entonces el efecto de la interacción con 1 grado de libertad no se puede estimar. Si una celda raza-sexo tiene empate, entonces los coeficientes de interacción estimada pueden ser inestables (tienen una gran varianza).

Los términos polinomiales (X_2, X_3) y otras transformaciones, ameritan consideraciones particulares cuando se está especificando el máximo modelo. Para el ejemplo de peso, se desea considerar, Edad, $(\text{Edad})^2$ y exp (Edad) como posibles predictoras. La colinealidad puede

generar resultados inestables y a menudo no interpretables en un modelo ajustado. Consecuentemente se intenta reducir, tal colinealidad como sea posible. El centrado, si es aplicable, ayuda a incrementar la calidad del modelo, se pueden multiplicar las variables por varias constantes, para producir aproximaciones de igualdad de varianzas para todas las predictoras.

Si el problema de la colinealidad no se puede superar, entonces probablemente se puede: (1) conducir un análisis separado para cada forma de X; por ejemplo, un análisis usando X y X^2, separando el análisis cuando se use exp (X), (2) eliminar alguna variable, o (3) imponer una estructura (un orden fijo de pruebas) en el procedimiento de prueba. Si no se hace nada en un caso de una severa colinealidad, entonces los coeficientes de regresión estimados en el mejor modelo pueden ser ligeramente inestables (tienen alta varianza) y posiblemente están completamente lejos del valor verdadero del parámetro.

11.2.2. Paso 2: especificando los criterios para la selección de un modelo

El segundo paso en la selección del mejor modelo, consiste en especificar la selección de criterios. Una selección de criterios es un índice que puede ser calculado para cada candidato a modelo y usarlo para compararlos entre sí. Entonces, dando una selección de criterios, los modelos

candidatos pueden ser ordenados del mejor al peor. Esto ayuda a automatizar el proceso de selección del "mejor" modelo. Como se puede ver este proceso de selección de criterios, puede no encontrar el mejor modelo en un sentido global. Sin embargo, el uso de una selección específica de criterios puede sustancialmente reducir el trabajo involucrado en encontrar un buen modelo. Obviamente, la selección de criterios está relacionada a la tarea del análisis. Por ejemplo, si la tarea es predicción confiable de futuras observaciones, la selección de criterios será algo liberal para evitar la pérdida de predictores usados.

> *"*Esto nos permite notar la distinción entre (1) diferencias numéricas, (2) diferencias estadísticamente significativas, y (3) diferencias científicamente importantes. Las diferencias numéricas pueden o no corresponder a diferencias significativas o científicamente importantes. En este caso, con pruebas sensitivas, las diferencias significativas pueden o no ser diferencias importantes. Finalmente, con pruebas insensibles, las diferencias importantes pueden ser no significativas*".*

Se han sugerido muchos criterios de selección para seleccionar el mejor modelo; por ejemplo: Hocking (1976), revisó 8 candidatos. Consideró 4 criterios: R_p^2, F_p, CM_{error} (p) y C_p. Antes de discutir estos criterios, se debe definir la notación necesaria para entenderlos. Los cuatro criterios

intentan comparar dos modelos de ecuaciones: el máximo modelo con *k* predictoras y un modelo restringido con *p* predictoras ($p \leq k$). El máximo modelo es,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p + \beta_{p-1} X_{p-1} + ... + \beta_k X_k + \varepsilon \quad (3)$$

Y el modelo reducido (una reducción del máximo modelo) es,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p + \varepsilon \quad (4)$$

Sea la SCE (k) la suma de cuadrados del error para el modelo de *k* variables, SCE (p), la suma de cuadrados del error para el modelo de *p* variables. Así también, SCY = $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ la suma de cuadrados totales corregidos de la respuesta *Y*.

Por lo general, p = k – 1, que es el caso cuando se evalúa la adición o eliminación de una sola variable. Se puede asumir que las *k* – *p* variables en consideración para la adición o la eliminación son denotadas por X_{p+1}, X_{p+2}, ..., X_k , por conveniencia en la notación.

La correlación múltiple cuadrada muestral R^2, es un candidato natural para decidir cuál modelo es el mejor y es por lo tanto el primer criterio discutido. El R^2 múltiple para el modelo de las p- variables es:

$$R_p^2 = R^2(Y/X_1, X_2, X_3, ... X_p) = 1 - [SCE\ (p)]\ /SCY \quad (5)$$

Desafortunadamente, R^2_p, tiene tres características potencialmente engañosas. Primero, tiende a sobreestimar ρ_p^2, el valor poblacional. Segundo, al adicionar predictoras, no se puede disminuir nunca R_p^2. En realidad, la adición de variables no incrementará R_p^2, al menos ligeramente. Finalmente, R_p^2 siempre es más grande en el modelo máximo, aunque se puede obtener un mejor modelo eliminando alguna variable. Tal modelo reducido puede ser mejor, porque sacrifica una cantidad despreciable de su fuerza predictiva, simplificando sustancialmente el modelo. Otro criterio razonable, para seleccionar el mejor modelo, es la prueba estadística de F, para comparar el modelo completo con el restringido. Este estadístico F_p, puede ser expresado en términos de las sumas de cuadrados de los errores, como:

$$F_p = \frac{[SCE(p) - SCE(k)]/(k-p)}{SCE(k)/(n-k-1)} = \frac{[SCE(p) - SCE(k)]/(k-p)}{CME(k)} \qquad (6)$$

Este estadístico puede ser comparado con una distribución de F con $k - p$ y $n - k - 1$ grados de libertad. El criterio de la prueba de F_p, es si SCE (p) – SCE (k), la diferencia entre la suma de cuadrados del residual para el modelo de las p-variables y la suma de cuadrados de residual para el modelo máximo con k-variables, es diferente significativamente de cero. Si F_p, no es significativo, podemos usar el modelo más pequeño (p-variables) y obtener la misma habilidad predictiva como la lograda con el

modelo completo. Una regla razonable, para la selección de variables puede ser el retener p variables, si ambos F_p no es significativo y p es lo más pequeño posible. Un caso especial usado con mucha frecuencia, de F_p, ocurre si $p = k - 1$, en que F_p es una prueba de H_0: $\beta_k = 0$ en el modelo completo.

El tercer criterio considerado para seleccionar el mejor modelo es la varianza del error estimada para el modelo de las p-variables, esto es:

$$\text{CME (p)} = \frac{SCE(p)}{n-p-1} \tag{7}$$

Considerando un modelo cualquiera, se ha simbolizado este estimador como S^2. La cantidad CME (p) es una selección atractiva para un criterio de selección, ya que se desea encontrar un modelo con una varianza residual pequeña.

Un candidato menos obvio, para un criterio de selección, que involucra la SCE (p) es el estadístico C_p de Mallow's,(Mallow's, 1973) dado por:

$$C_p = \frac{SCE(p)}{CME(k)} - [n - 2(p + 1)] \tag{8}$$

El criterio C_p, ayuda a decidir cuantas variables se pueden considerar en el mejor modelo, ya que obtiene un valor de aproximadamente *p + 1* si el CME (p) es cercanamente igual al CME (k) (si el modelo correcto es de tamaño *p*). Conociendo el tamaño correcto del modelo, nos ayuda a la selección del mejor.

Los criterios F_p, R_p^2, CME (p) y el C_p están íntimamente relacionados. Por ejemplo; la prueba F_p puede ser expresada en términos de correlaciones cuadradas múltiples, esto es:

$$F_p = \frac{[R_k^2 - R_p^2]/(k-p)}{[1-R_k^2]/(n-k-1)} \tag{9}$$

Y el estadístico C_p es una función simple del estadístico F_p:

$$C_p = (k - p)\, F_p + (2p - k + 1) \tag{10}$$

¿Por qué considerar más de un criterio? La razón es que un solo criterio no siempre es correcto. En la práctica, las alternativas pueden establecer la diferencia en la selección del modelo. Un aspecto importante de esta discusión es una demostración de las limitantes del estadístico R_p^2 como el único criterio para seleccionar el modelo. Se favorece al estadístico C_p, ya que tiende a simplificar la decisión sobre las variables que se retienen en el modelo final.

11.2.3. Paso 3: especificando la estrategia para la selección de variables

El tercer paso para seleccionar el mejor modelo es especificar la estrategia para seleccionar las variables. Tal estrategia, está relacionada con la determinación de cuántas y qué variables estarán en el modelo. Tradicionalmente, tales estrategias han sido enfocadas para decidir si se adiciona una sola variable al modelo (un método de selección forward), o si una sola variable será eliminada del modelo (un método de selección backward). Como las computadoras son más poderosas, los métodos para considerar más de una variable por paso son prácticos.

11.2.3.1. Procedimiento de todas las regresiones posibles

Por razones prácticas, el procedimiento de todas las posibles regresiones es preferido sobre cualquier otra estrategia de selección de variables. Es el único método garantizado para definir el modelo que tiene la más grande R_p^2, el más pequeño CME (p), etc. Esta estrategia no es siempre usada por el número tan grande de cálculos necesarios si el número de variables *k* en el modelo es grande. El procedimiento de todas las regresiones requiere que ajustemos cada ecuación de regresión posible asociada con cada posible combinación de las *k* variables independientes. En este ejemplo, se requiere ajustar siete modelos

correspondientes a los siguientes siete grupos de variables independientes: (1) Altura, (2) Edad, (3) $(Edad)^2$, (4) Altura y Edad, (5) Altura y $(Edad)^2$, (6) Edad y $(Edad)^2$, y (7) Altura, Edad y $(Edad)^2$. Para k variables independientes el número de modelos que serán ajustado es $2^k - 1$; por ejemplo si $k = 10$, entonces $2^{10} - 1 = 1{,}023$ modelos.

Una vez que todos los $2^k - 1$ modelos han sido ajustados, entonces, se reúnen los modelos ajustados en grupos que involucren de 1 a k variables y se ordenan los modelos dentro de cada grupo de acuerdo a algún criterio (R_p^2, CMerror (p), o C_p).

De los datos del ejemplo, los resultados de este procedimiento de todas las regresiones son dados en la Tabla 11.1. De la tabla, los líderes (en términos de valores de R^2_p) en cada uno de los grupos involucran, uno, dos y hasta tres variables, esto es,

Grupo de una variable: altura con $R_1^2 = 0.6630$
Grupo de dos variables: altura, Edad con $R_2^2 = 0.7800$
Grupo de tres variables: altura, Edad y $(edad)^2$ con $R_3^2 = 0.7802$

De los tres modelos (modelos 1, 4 y 7 respectivamente en la Tabla 11.1) se puede observar claramente que el modelo 4 que involucra altura y Edad se puede seleccionar ya que el valor de R^2, prácticamente es el mismo que el del modelo 7 y es mucho más alto que el del modelo 1. Entonces la selección del mejor modelo de regresión, basado en el

procedimiento de todas las regresiones posibles, usando el criterio con R_p^2 es:

$$\hat{Y} = 6.553 + 0.722 \text{ (Altura)} + 2.050 \text{ (Edad)}$$

Otros aspectos de la Tabla 11.1 que se pueden comentar son: considerar los estadísticos F parciales. Para una variable dada en un modelo, el estadístico de F parcial asociado valora la contribución que esta variable tiene en la predicción de *Y*, abajo y arriba de las contribuciones de las otras variables en el modelo. Esto es, para el modelo 4, que involucra a X_1 y X_2 la F parcial para X_2 es $F(X_2/X_1) = 4.785$; pero para el modelo 7, que incluye X_1, X_2 y X_3, la F parcial es $F(X_2/X_1, X_3) = 0.140$.

Tabla 11.1

Resumen de resultados del procedimiento de todas las regresiones posibles

Modelo	No. Var	Var	Coeficientes estimados				Estadístico Parcial de F			Prueba de F	R^2_p	CME (p)	Cp
			$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	X_1	X_2	X_3				
1	1	X_1	6.19	1.07			19.7			19.7	.663	29.93	4.27
2	1	X_2	30.57		3.64			14.6		14.6	.593	36.18	6.83
3	1	X_3	45.99			0.21			14.3	14.3	.587	36.63	7.01
4	2	X_1,X_2	6.553	0.72	2.05		7.66	4.78		15.9	.780	21.71	2.01
5	2	X_1,X_3	15.11	0.72		0.11	7.60		4.56	15.6	.776	21.07	2.14
6	2	X_2,X_3	32.40		3.20	0.02		0.11	0.00	6.55	.593	40.20	8.83
7	3	X_1,X_2,X_3	3.438	0.72	2.77	-0.04	6.83	0.14	0.01	9.47	.780	24.40	4.00

Las F parciales, son pruebas de variables –adicionadas al final, basadas cada una en los cuadrados medios del error (CME) de los modelos correspondientes ajustados. Tales pruebas deben ser tratadas con mucho cuidado. Cualquier prueba basada en un modelo con menos términos que el modelo correcto será sesgado, quizás sustancialmente, ya que la prueba involucra el uso de términos de errores sesgados. De la misma manera muchas pruebas son calculadas para las cuales la tasa del Error Tipo I, es más alta que la tasa nominal α.

Las pruebas de F totales, también son afectadas, por el estimador de la varianza del error usado. En la selección de variables, este estimador es importante, porque las pruebas sesgadas aplicadas en el algoritmo stepwise pueden parar el proceso prematuramente y perder predictores importantes. Si el cuadrado medio del error (CM_{error}) del modelo más grande, el número 7, fue usado en el denominador de cada prueba de F, cada uno de los estadísticos resultantes será un estadístico de F_p. Por ejemplo para el modelo 4, se tiene:

$$F_p = \frac{[R_k^2 - R_p^2]/(k-p)}{[1-R_k^2]/(n-k-1)} = \frac{0.7802-0.7800)/(3-2)}{(1-0.7802)/(12-3-1)} = 0.007$$

Por lo general, esta prueba es una prueba F parcial-múltiple. Los valores pequeños de F indican que la habilidad predictiva del máximo modelo y del modelo 4 no difiere significativamente.

En la Tabla 11.1, se presentan los valores C_p. Se espera que el valor de C_p se aproxime a $p + 1$ si se considera el modelo correcto o si se considera un modelo más grande que contiene el modelo correcto. Si los predictores importantes son omitidos, C_p será más grande que $p + 1$. También si $F_p < 1$, entonces $C_p < (p + 1)$; esto puede ocurrir cuando el valor de R_p^2 está cercano a R_k^2. Se prefieren modelos con valores C_p no tan lejanos de $p + 1$. Por lo tanto, para un modelo de una variable C_p es comparado con el valor 2.0, para dos variables, con el valor de 3.0 y así sucesivamente. Para este ejemplo, ningún modelo de una variable tiene un C_p cercano a 2.0. Para modelos con tres variables el C_p es exactamente 4.0. El modelo completo con *k* predictores se garantiza que tiene un C_p, exactamente igual a $k + 1$. El modelo de dos variables con Edad y $(Edad)^2$ tiene un C_p mucho mayor de 3.0, mientras los modelos 4 y 5 tienen valores C_p cercanos al valor C_p mínimo posible de:

$$\frac{(n-k-1)CMerror(k)}{CMerror(k)} - [n - 2(p + 1)] = (2p - k + 1) \qquad (11)$$

Que es igual a 2 cuando $k = 3$ y $p = 2$. Note que éste bajo límite de C_p es obtenido cuando $F_p = 0$ y esto puede ser negativo. Tales valores son mejores que el valor de $p + 1 = 3.0$.

El procedimiento de todas las regresiones se presenta ya que se prefiere en situaciones prácticas. Esto tiene la distinción de ser el único

método garantizado para encontrar el mejor modelo, en el sentido de que cualquier criterio de selección será numéricamente optimizado por la muestra bajo estudio. Naturalmente, esto no garantiza que se ha encontrado el modelo correcto. En realidad, en muchas situaciones, se pueden tener varios candidatos de modelos para definir el mejor, con diferentes criterios de selección, esto sugiere, que hay diferentes modelos mejores. De esta manera, tales encuentros pueden variar de muestra a muestra, aunque todas las muestras sean seleccionadas de la misma población. Consecuentemente la selección del mejor modelo puede variar de muestra a muestra.

Como se mencionó anteriormente, el algoritmo de todas las regresiones posibles, por lo general no es práctico ya que se deben definir 2^k -1 modelos, con *k* predictores candidatos que se evalúan. Por lo que muchos métodos han sido sugeridos para llegar a una aproximación del procedimiento de todas las posibles regresiones. Estos métodos no se garantizan para encontrar el mejor modelo. No obstante, estos métodos se discuten a continuación ya que se puede recoger toda la información en los datos necesarios para seleccionar el mejor modelo.

11.2.3.2. Procedimiento de Eliminación Backward

En este procedimiento se define lo siguiente.

Paso 1. Determinar la ecuación de regresión ajustada, que contenga todas las variables independientes. En nuestro ejemplo:

$$\hat{Y} = 3.438 + 0.724(\text{Altura}) + 2.777(\text{Edad}) - 0.042(\text{Edad})^2$$

La tabla del análisis de varianza correspondiente es:

Fuente	Grados de libertad	Sumas de Cuadrados	Cuadrados medios	F	R^2
Regresión	3	693.06	231.01	9.47**	0.7802
Residual	8	195.19	24.40		
Total	11	888.25			

Paso 2. Calcular el estadístico F – parcial para cada variable en el modelo, como si fuera la última variable a entrar en el modelo (Tabla 11.1)

Variable	F – parcial*
Altura	6.83
Edad	0.140
$(\text{Edad})^2$	0.010

*basado en 1 y 8 grados de libertad

Paso 3. Nos enfocamos en el valor de la prueba de F-parcial más bajo (esto es F_L). En el ejemplo, $F_L = 0.010$ para la variable $(Edad)^2$.

Paso 4. Comparamos este valor F_L con un valor crítico preseleccionado (F_C) de la distribución de F. (1) Si $F_L < F_C$, remover del modelo la variable bajo consideración, recalcular la ecuación de regresión para las restantes variables y repetir el paso 2, 3 y 4. (2) Si $F_L > F_C$, adoptar la ecuación de regresión completa calculada. En el ejemplo, si se trabaja con un nivel del 10%, entonces, $F_L = 0.010 < F_{1,8,0.90} = F_C = 3.46$. Por lo tanto removemos $(Edad)^2$ del modelo y recalculamos la ecuación usando solamente Altura y Edad. Se obtiene lo siguiente,

$$\hat{Y} = 6.553 + 0.722(\text{Altura}) + 2.050(\text{Edad})$$

Con su tabla del análisis de varianza correspondiente.

Fuente	**Grados de libertad**	**Sumas de Cuadrados**	**Cuadrados medios**	**F**	**R^2**
Regresión	2	692.82	346.41	15.95**	0.7800
Residual	9	195.43	21.71		
Total	11	888.25			

Con la variable $(Edad)^2$ fuera del modelo, las F's parciales son, 7.665 para Altura y 4.785 para Edad. Por lo que la nueva $F_C = F_{1,9,90} = 3.36$,

es menor que 4.785. Por lo tanto la F parcial para Edad es significativa y se detiene aquí el modelo, que es el mismo modelo al que se llega al usar el procedimiento de todas las regresiones posibles.

11.2.3.3. Procedimiento de selección forward

Este procedimiento se aplica de la siguiente manera.

Paso 1. Seleccionar como primer variable a entrar en el modelo aquella que esté ligeramente más correlacionada con la variable dependiente y ajustar la ecuación de línea recta asociada. Para este ejemplo, se tienen las siguientes correlaciones: r_{YX_1} = 0.814, r_{YX_2} = 0.770 y r_{YX_3} = 0.767. Entonces la primera variable a entrar en el modelo es X_1 = Altura.

La ecuación de línea recta ajustada es:

$$\hat{Y} = 6.190 + 1.073 \text{ (Altura)}$$

El análisis de varianza correspondiente es:

Fuente	Grados de libertad	Sumas de Cuadrados	Cuadrados medios	F	R^2
Regresión	1	588.92	588.92	19.67**	0.6630
Residual	10	299.33	29.93		
Total	11	888.25			

Si el estadístico F no es significativo en la tabla, detener y concluir que las variables independientes no son predictores importantes. Si el estadístico F es significativo, se incluye esta variable en el modelo y se procede al paso 2.

Paso 2. Se calcula el estadístico de F-parcial, correspondiente con cada variable restante basado en una ecuación de regresión que contiene esta variable y la variable inicialmente seleccionada. Esto es:

F parcial de $(X_2/X_1) = 4.785$
F parcial de $(X_3/X_1) = 4.565$

Paso 3. Se observa a la variable con mayor estadístico F parcial. Para estos datos, la variable Edad, tiene el estadístico F parcial más grande de 4.785.

Paso 4. Se prueba la significancia del estadístico F parcial asociado con la variable seleccionada en paso 3. (a) Si esta prueba es significativa,

adicionar la nueva variable al modelo de regresión, (b) Si esta prueba no es significativa, use en el modelo sólo la variable considerada en el paso 1. En nuestro ejemplo, la F parcial para Edad es significativa al nivel del 10% ($F_{1,9,.90} = 3.36$), adicionamos edad y ajustamos un modelo de dos variables,

$$\hat{Y} = 6.553 + 0.722 \text{ (Altura)} + 2.050 \text{ (Edad)}$$

Paso 5. En cada uno de los pasos subsecuentes, determinar el estadístico de F- parcial para las variables que no están en el modelo y adicionar al modelo las variables que tienen un valor de F- parcial más grande si es estadísticamente significativa. En cualquier paso, si el F- parcial más grande no es significativo, entonces no se incluirán más variables al modelo y el proceso ha terminado. En este ejemplo, se ha adicionado altura y edad al modelo. Se puede ver si se adiciona $(\text{Edad})^2$ al modelo. La F- parcial para $(\text{Edad})^2$, controlando altura y edad es dada por:

$$\text{Parcial de } F(X_3/X_1,X_2) = 0.010$$

Este valor no es estadísticamente significativo, porque $F_{1,8,.90} = 3.46$. Otra vez se ha llegado al mismo modelo de dos variables seleccionado en el método anterior.

11.2.3.4. Procedimiento de selección stepwise

La regresión stepwise es una versión modificada de la regresión forward, que permite una revaloración, en cada paso, de las variables incorporadas al modelo, en pasos anteriores. Una variable que está en el modelo en pasos anteriores, puede ser superflua en el último estadío, debido a su relación con otras variables que están en el modelo. Para verificar esta posibilidad, en cada paso es realizada una prueba de F-parcial para cada variable presente en el modelo, como si fuese la variable más reciente que está en el modelo, independientemente de su punto de entrada al modelo. Esta variable con el estadístico F- parcial más pequeño no significativo (si existe tal variable) es removida y el modelo es reajustado con las variables restantes. Se obtienen las F-parciales y son examinadas similarmente y así sucesivamente. El proceso total continúa hasta que ninguna variable puede ser adicionada o removida del modelo.

Para este ejemplo, el primer paso, como en el procedimiento de selección forward, será adicionar la variable altura al modelo, ya que tiene el más alto coeficiente de correlación con *Y*, continuando, adicionamos la Edad al modelo ya que tiene la correlación parcial más alta significativa con *Y* más que $(\text{edad})^2$, controlando la altura. Para ver si $(\text{edad})^2$ puede ser adicionada al modelo, se puede observar la F-parcial de altura, dado por $F(X_1/X_2) = 7.665$, que excede a $F_{1,9,.90} = 3.36$.

Entonces no se remueve la variable altura del modelo. Se continúa verificando para ver si se puede adicionar $(edad)^2$, la respuesta es no, ya que se ha tratado antes con esta situación.

La tabla del análisis de varianza que resume los resultados obtenidos del ejemplo anterior es:

Fuente		**Grados de libertad**	**Sumas de cuadrados**	**Cuadrados medios**	**F**	**R^2**
Regresión	X_1	1	588.92	588.92	19.67**	0.7800
	X_2/X_1	1	103.90	103.90	4.79	
Residual		9	195.43	21.71		
Total		11	888.25			

La tabla de análisis de varianza que considera a todas las variables es:

Fuente		**Grados de libertad**	**Sumas de cuadrados**	**Cuadrados medios**	**F**	**R^2**
Regresión	X_1	1	588.92	588.92	19.67**	0.7802
	X_2/X_1	1	103.90	103.90	4.79	
	$X_3/X_1,X_2$	1	0.24	0.24	0.01	
Residual		8	195.19	21.71		
Total		11	888.25			

11.2.3.5. Métodos Chunkwise

Los métodos para seleccionar una variable, descritos anteriormente, pueden ser generalizados en una forma muy útil. La idea básica es que cualquier método de selección en que una sola variable es adicionada o eliminada puede ser generalizado para adicionar o eliminar un grupo de variables. Considerar, por ejemplo, usando eliminación backward, construir un modelo de la variable respuesta nivel de colesterol. Asumir que tres grupos de predictores están disponibles, demográficos (género, raza, edad y sus interacciones), antropométricos (altura, peso y sus interacciones) y dieta (cantidad de cinco tipos de alimentos) se tiene un total de 6 + 3 + 5 = 14 predictores. Los tres grupos de variables constituyen los denominados Chunks (gruesos), grupos de predictores que están lógicamente relacionados y son de igual importancia como predictores candidatos.

Varios métodos posibles de Chunkwise están disponibles. La selección entre ellos depende de (1) las preferencias del analista por el método backward, forward u otra estrategia de selección y (2) la extensión que el analista puede hacer con grupos de variables (formar chunks) y ordenar los grupos en importancia. En muchas aplicaciones, existe un orden a priori entre los chunks. Para este ejemplo, el investigador desea considerar las variables dietas solamente controlando las variables demográficas y antropométricas. La

imposición de un orden en este estilo, ayuda en la simplificación del análisis, por lo general incrementa la confiabilidad e incrementa la posibilidad de encontrar un modelo significativamente plausible. Se ilustra el uso de los métodos de prueba de chunkwise, describiendo una estrategia de eliminación backward. Se puntualiza que otras estrategias se pueden preferir, dependiendo de la situación. Una aproximación al método de eliminación backward para probar chunkwise, requiere un chunk de variables especificadas para construir el modelo, las variables en el chunk especificado son candidatas para ser eliminadas. Asumir que el grupo de variables dieta, es el primer chunk considerado para ser eliminado.

Entonces, en este ejemplo, todas las variables demográficas y antropométricas son forzadas a quedarse en el modelo. Si por ejemplo, la prueba F múltiple-parcial del chunk de las variables dietas, no es significativo, entonces la entrada del chunk puede ser eliminada. Si este conjunto es significativo, entonces, al menos una de las variables en este chunk-dieta puede ser retenida. De esta manera, el método más simple chunkwise, suma o borra todas las variables en un chunk conjuntamente. Sin embargo, una aproximación más sensible, es manipular variables únicas dentro de un chunk significativo mientras guarda los otros chunks. Si se asume que el chunk dieta es importante, entonces se debe decidir cuál de las variables dietas deben ser retenidas como predictoras importantes. Un segundo paso razonable, será

requerir todas las variables demográficas y la importancia individual de las variables dietas será retenida, mientras se considera el segundo chunk de variables antropométricas para ser eliminadas. El paso final para este ejemplo de tres chunk, requiere las variables individuales seleccionadas de los primeros dos chunks para retenerlos en el modelo, mientras las variables del tercer chunk (demográficas) son candidatas para ser eliminadas. Los métodos de selección forward y stepwise para una variable, pueden ser generalizados para su uso en la prueba de chunkwise.

Los métodos chunkwise para la selección de variables, pueden tener ventajas importantes sobre los métodos de selección de una variable. Primero, los métodos chunkwise, incorporan en el análisis conocimiento científico y la preferencia de grupos de variables. Según-do, el número de posibles modelos que son evaluados es redu-cido. Si una prueba chunk no es significativa y el chunk de variables completo es borrado, entonces no se hacen las pruebas en variables individuales. En muchas situaciones tales pruebas de significancia para grupos son más efectivas y confiables que probando variables de forma individual.

11.2.4. Paso 4: conduciendo al análisis

Una vez que se tiene especificado el máximo modelo, el criterio para la selección de un modelo y la estrategia para aplicar el criterio, se debe

conducir el análisis de acuerdo a lo planeado. Obviamente esto será hecho con algún tipo de software de cómputo apropiado. La bondad de ajuste del modelo seleccionado se debe examinar. También se puede aplicar el análisis de residuales para demostrar que el modelo seleccionado es razonable para el manejo de los datos.

11.2.5. Paso 5: evaluando la confiabilidad con muestras divididas

Una vez que se ha seleccionado el mejor modelo para una muestra de datos particular, no se tiene la seguridad de que el modelo se puede aplicar confiablemente a otras muestras. En un sentido se pregunta la cuestión," ¿Se podrán generalizar estas conclusiones?" Si el modelo seleccionado predice bien para otras muestras de la población de interés, se dice que el modelo es confiable. Aquí se discuten métodos para evaluar la confiabilidad de un modelo. Muchos de los métodos aceptados para evaluar la confiabilidad de un modelo involucran alguna forma de aproximación de una muestra dividida. Se discuten tres aproximaciones para evaluar la confiabilidad del modelo: el estudio de seguimiento, el análisis de muestras divididas y la muestra extendida. La manera más precisa para valorar la confiabilidad de un modelo seleccionado, es conducir un nuevo estudio y probar el ajuste del modelo seleccionando nuevos datos. Sin embargo, esta aproximación es muy costosa y algunas veces complicada. La cuestión entonces,

aparece como si un solo estudio puede alcanzar las dos tareas de encontrar el mejor modelo y valorar su confiabilidad.

Un análisis de muestras divididas, intenta lograr ambas tareas en un solo estudio. Para el ejemplo del colesterol, el análisis de muestras divididas más simple, procede de la siguiente manera; primero, se realiza una asignación aleatoria a todas las observaciones a uno de los dos grupos, el grupo de entrenamiento o el grupo extendido. Esto se hace antes de conducir cualquier análisis. Los sujetos pueden ser agrupados en estratos basados en una o más variables categóricas importantes. Para el ejemplo del colesterol, se pueden definir estratos apropiados para varias combinaciones de género-raza. Si estos estratos son importantes, un esquema de estratos específicos de muestras divididas se puede usar. Con este método, todos los sujetos son asignados aleatoriamente dentro de un estrato a un entrenamiento o el grupo extendido, tal asignación se hace separadamente para cada estrato. La tarea de asignación aleatoria del estrato específico es para asegurar que los dos grupos de observaciones (entrenamiento y extendido) son igualmente representativos de la población de interés.

Una alternativa para la división aleatoria estratificada es un esquema de asignación de apareamiento. Con la asignación apareada, se encuentran pares de sujetos que son tan similares los cuales se asignan a un miembro del par a la muestra de entrenamiento y el otro a la muestra extendida. Desafortunadamente, la diferencia entre los

resultados de los grupos de entrenamiento y extendido tienden a ser mucho menores a las diferencias correspondientes entre las muestras subsecuentes seleccionadas aleatoriamente. Esta tendencia produce una opinión optimista no realista de la confiabilidad del modelo, que por lo general se desea evitar. Cualquiera de las dos alternativas se recomienda, como el segundo paso en un análisis de muestras divididas. El primero es conducir, la selección del modelo separadamente para cada uno de los dos grupos de datos. Típicamente, ninguna diferencia en las variables predictoras seleccionadas, por los dos procesos de selección se toma como una indicación de no confiabilidad. En la práctica los dos modelos obtenidos, siempre difieren en algo, que es la razón principal de que el método de selección del modelo, tenga una reputación de ser no confiable.

Las aproximaciones anteriores para valorar la confiabilidad son tan rigurosas, cuando se trata de predecir. Un buen modelo predictivo puede (1) predecir muy bien en cualquier muestra, como lo hace en la muestra analizada y (2) pasar todas las pruebas de la regresión diagnóstica para lo adecuado del modelo aplicado en cualquier muestra. Estos comentarios nos permiten considerar una segunda aproximación de un análisis de muestras divididas para valorar la confiabilidad. Esta segunda aproximación intenta responder la cuestión "¿El modelo seleccionado predice bien en una nueva muestra?" Con esta segunda aproximación se inicia conduciendo un proceso de construcción de

modelo, usando los datos para el grupo de entrenamiento. Suponer que la ecuación ajustada obtenida es:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + ... + \hat{\beta}_p X_p \quad (12)$$

Para el ejemplo de altura de los niños página 199, la ecuación de regresión es:

$$\hat{Y} = 6.553 + 0.722 \text{ (Altura)} + 2.050\text{(Edad)}$$

Se han considerado, varias estrategias de selección de variables, cuando la tarea principal del estudio es la predicción. En contraste, se recuerda que las estrategias de validez- orientadas están relacionadas, con la generación de estimadores válidos de uno o más coeficientes de regresión en un modelo. Esencialmente, un estimador válido, es aquel que refleja la verdadera relación en la población bajo estudio. Sin embargo, las muestras de datos bajo consideración deben reflejar adecuadamente la población de interés y los supuestos del modelo y análisis que se están usando deben ser satisfechos razonablemente.

Para ser un poco más específicos, una estrategia de selección de variables con una tarea de validez, primero involucra determinar los efectos de interacción importantes, seguido por una evaluación de la

confusión, que es casual de los resultados de la valoración de interacción en el análisis.

11.3. El estadístico PRESS

En algunas circunstancias que involucran la construcción de modelos no es muy práctico el dividir los datos para los propósitos de validación. Se puede ser generoso con el dilema de los investigadores de los gastos en la recolección de los datos. Ciertamente, la división de los datos o validación con datos nuevamente colectados no se pueden acompañar en todas las aplicaciones. Un criterio muy interesante e importante, que se puede usar como una forma de validación, mucho mejor que la división de los datos es el estadístico PRESS. El interés se centra en la generación de errores de predicción del tipo $r_j = \hat{y}(x_j) - y(x_j)$, donde ambos miembros del lado derecho de la ecuación son independientes. Estos son los verdaderos errores de predicción. Considerar un grupo de datos en que se retienen o se apartan de la primera observación de la muestra y usamos las restantes $n - 1$ observaciones para estimar los coeficientes de un modelo. La primera observación es reemplazada y la segunda observación es retenida con coeficientes estimados. Se puede remover cada observación una a la vez y el modelo es ajustado *n* veces. La respuesta eliminada es estimada en cada vez, obteniéndose como

resultados *n* errores de predicción o residuales PRESS $y_i - \hat{y}_{i,-i} = e_{i,-i}$ (i = 1, 2, ..., n). Estos residuales PRESS son los errores de predicción verdaderos con $\hat{y}_{i,-i}$ independientes de y_i, en esta manera la observación y_i no fue usada simultáneamente para ajustar y valorar el modelo, esta es la verdadera prueba de la validación. La predicción $\hat{y}_{i,-i}$ es la función de regresión evaluada en $x = x_i$, pero y_i fue apartada y no fue usada en la obtención de los coeficientes. Por lo que se tiene:

$$\hat{y}_{i,-i} = x'_i \, b_{-i} \tag{13}$$

Donde b_{-i} es el grupo de coeficientes calculado sin el uso de la *i-ésima* observación. Entonces cada modelo tendrá *n* residual PRESS asociados con él y la suma de cuadrados de predicción PRESS se define cómo:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2 = \sum_{i=1}^{n} (e_{i,-i})^2 \tag{14}$$

Para seleccionar el mejor modelo, se favorece al modelo con el PRESS más pequeño. El PRESS es importante ya que se tiene información en la forma de *n* validaciones en la cual la muestra ajustada para cada uno es de tamaño *n* – 1. El primer vistazo a la impresión del PRESS los lectores ofrecen probable resistencia al criterio, debido a la complejidad de los cálculos. Los residuales PRESS podrán ser usados para generar

otro estadístico R^2 que refleja capacidades de predicción. Este estadístico está dado por:

$$R^2_{pred} = 1 - PRESS / \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad (15)$$

Los residuales individuales PRESS pueden ser valuados completamente aparte de su papel en el cálculo del estadístico PRESS en la ecuación anterior. Dan medidas separadas de la estabilidad de la regresión y pueden ayudar al analista a separar los puntos u observaciones que tienen una influencia bastante grande sobre los resultados de la regresión. Por ejemplo, suponga un residual ordinario de 17.75 gm/pulg^2, es establecido en un punto particular de un experimento en el cual la fuerza de una envoltura es la respuesta. Pero suponga que los residuales PRESS = 850.92 gm/pulg^2. Esta gran diferencia entre los residuales PRESS y los ordinarios implican que los puntos en cuestión son una observación de mayor influencia en la construcción de la regresión. La implicación es que el uso de los puntos resulta en una fuerte atracción de los $\hat{y}_i$ ajustados y las observaciones y_i. Se observa esta condición en el caso de los tres sitios separados en los datos del hospital. La fina distinción entre los residuales ordinarios y los errores de predicción o residuales PRESS, sugieren que estos tres sitios están ligeramente influenciados. Los residuales PRESS no son el único mecanismo para detectar observaciones de alta influencia y no

son necesariamente el mejor. Como fue indicado anteriormente, el cálculo de PRESS es relativamente simple y no requiere corridas de regresiones repetidas por el analista. Se pueden calcular los residuales PRESS de los residuales ordinarios, esto es:

$$e_{i,-1} = \frac{e_i}{(1-h_{ii})} = \sum_{i=1}^{n} \left(\frac{e_i}{(1-h_{ii})} \right)^2$$

donde

$$h_{ii} = \frac{1}{n} + (x_i - \bar{x})^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (16)$$

Otra realidad que hace esta declaración quizás no entendible es que h_{ii} es una medida de la distancia cuadrada estandarizada, la distancia del punto; x_{1i}, x_{2i}, ..., x_{ki} a los puntos x_1, x_2, x_3, ..., x_k. Para el caso en que k = 1, h_{ii} se puede escribir:

$$h_{ii} = \frac{1}{n} + (x_i x - \bar{x})^2 / \sum_{i=1}^{n} (x_i x - \bar{x})^2 \qquad (17)$$

Un punto que es extremo en la dirección de cualquier *x´*s es donde la predicción es relativamente pobre y donde existe una tendencia de producir residuales PRESS grandes.

11.4. Ejemplo usando SAS

Estos datos provienen de un estudio realizado sobre el tiempo de vida de organismos y cinco variables independientes. Provienen de una muestra de 54 pacientes que esperan una cirugía de hígado. Es un estudio observacional en el cual la variable respuesta es el tiempo de sobrevivencia *Y* del paciente. El objetivo del estudio es obtener una ecuación usando la información obtenida en el diagnóstico pre-operatorio. Antes de la operación los datos fueron obtenidos de cuatro variables que son posibles predictoras. Estas variables son:

CLOT = un registro del índice de coagulación de la sangre.
PROG = Un índice de pronóstico que incluye la edad del paciente.
ENZ = Una prueba de la función enzimática
LIV= Un registro de la prueba funcional del hígado.
TIEM = Tiempo de supervivencia del paciente.
VITIEM = Logaritmo de la variable TIEM.
El modelo es: $VITIEM = \xi_0 + \beta_1 CLOTC + \beta_2 PROGC + \beta_3 ENZC + \beta_4 LIVC + \varepsilon_i$
La respuesta el tiempo (TIEM), es el número de días que vivirá después de la operación.

CLOT	PROG	ENZ	LIV	TIEM	VITIEM	CLOT	PROG	ENZ	LIV	TIEM	VITIEM
6.7	62	81	2.59	200	2.3010	11.2	76	90	5.59	574	2.7589
5.1	59	66	1.70	101	2.0043	5.2	54	56	2.71	72	1.8573
7.4	57	83	2.16	204	2.3096	5.8	76	59	2.58	178	2.2504
6.5	73	41	2.01	101	2.0043	3.2	64	65	0.78	71	1.8513
7.8	65	115	4.30	509	2.7067	8.7	45	23	2.52	58	1.7634
5.8	38	72	1.42	80	1.9031	5.0	59	73	3.50	116	2.0645
5.7	46	63	1.91	80	1.9031	5.8	72	93	3.30	295	2.4698
3.7	68	81	2.57	127	2.1038	5.4	58	70	2.64	115	2.0607
6.0	67	93	2.50	202	2.3054	5.3	51	99	2.60	184	2.2648
3.7	76	94	2.40	203	2.3075	2.6	74	86	2.05	118	2.0719
6.3	84	83	4.13	329	2.5172	4.3	8	119	2.85	120	2.0792
6.7	51	43	1.86	65	1.8129	4.8	61	76	2.45	151	2.1790
5.8	96	114	3.95	830	2.9191	5.4	52	88	1.81	148	2.1703
5.8	83	88	3.95	330	2.5185	5.2	49	72	1.84	95	1.9777
7.7	62	67	3.40	168	2.2253	3.6	28	99	1.30	75	1.8751
7.4	74	68	2.40	217	2.3365	8.8	86	88	6.40	483	2.6840
6.0	85	28	2.98	87	1.9395	6.5	56	77	2.85	153	2.1847
3.7	51	41	1.55	34	1.5315	3.4	77	93	1.48	191	2.2810
7.3	68	74	3.56	215	2.3324	6.5	40	84	3.00	123	2.0899
5.6	57	87	3.02	172	2.2355	4.5	73	106	3.05	311	2.4928
5.2	52	76	2.85	109	2.0374	4.8	86	101	4.10	398	2.5999
3.4	83	53	1.12	136	2.1335	5.1	67	77	2.86	158	2.1987
6.7	26	68	2.10	70	1.8451	3.9	82	103	4.55	310	2.4914
5.8	67	86	3.40	220	2.3424	6.6	77	46	1.95	124	2.0934
6.3	59	100	2.95	276	2.4409	6.4	85	40	1.21	125	2.0969
5.8	61	73	3.50	144	2.1584	6.4	59	85	2.33	198	2.2967
5.2	52	86	2.45	181	2.2577	8.8	78	72	3.20	313	2.4955

Usando el procedimiento de eliminación stepwise del SAS

```
data regres;
input clot prog enz liv tiem vitiem;
```

```
clot2 = clot*clot;
enz2 = enz*enz;
liv2 = liv*liv;
cards;

6.7  62  81  2.59  200  2.3010
51.  59  66  1.70  101  2.0043
7.4  57  83  2.16  204  2.3096
6.5  73  41  2.01  101  2.0043
7.8  65  115 4.30  509  2.7067
5.8  38  72  1.42  80   1.9031
5.7  46  63  1.91  80   1.9031
3.7  68  81  2.57  127  2.1038
6.0  67  93  2.50  202  2.3054
3.7  76  94  2.40  203  2.3075
6.3  84  83  4.13  329  2.5172
6.7  51  43  1.86  65   1.8129
5.8  96  114 3.95  830  2.9191
5.8  83  88  3.95  330  2.5185
7.7  62  67  3.40  168  2.2253
7.4  74  68  2.40  217  2.3365
6.0  85  28  2.98  87   1.9395
3.7  51  41  1.55  34   1.5315
7.3  68  74  3.56  215  2.3324
5.6  57  87  3.02  172  2.2355
5.2  52  76  2.85  109  2.0374
3.4  83  53  1.12  136  2.1335
6.7  26  68  2.10  70   1.8451
5.8  67  86  3.40  220  2.3424
6.3  59  100 2.95  276  2.4409
5.8  61  73  3.50  144  2.1584
5.2  52  86  2.45  181  2.2577
```

5.2	76	90	5.59	574	2.7589
5.2	54	56	2.71	72	1.8573
5.8	76	59	2.58	178	2.2504
3.2	64	65	0.78	71	1.8513
8.7	45	23	2.52	58	1.7634
5.0	59	73	3.50	116	2.0645
5.8	72	93	3.30	295	2.4698
5.4	58	70	2.64	115	2.0607
5.3	51	99	2.60	184	2.2648
2.6	74	86	2.05	118	2.0719
4.3	8	119	2.85	120	2.0792
4.8	61	76	2.45	151	2.1790
5.4	52	88	1.81	148	2.1703
5.2	49	72	1.84	95	1.9777
3.6	28	99	1.30	75	1.8751
8.8	86	88	6.40	483	2.6840
6.5	56	77	2.85	153	2.1847
3.4	77	93	1.48	191	2.2810
6.5	40	84	3.00	123	2.0899
4.5	73	106	3.05	311	2.4928
4.8	86	101	4.10	398	2.5999
5.1	67	77	2.86	158	2.1987
3.9	82	103	4.55	310	2.4914
6.6	77	46	1.95	124	2.0934
6.4	85	40	1.21	125	2.0969
6.4	59	85	2.33	198	2.2967
8.8	78	72	3.20	313	2.4955

```
proc reg;
model vitiem = clot prog enz liv tiem clot2 enz2
liv2/selection=stepwise;
run;
```

LOS RESULTADOS

Model: MODEL1

Dependent Variable: vitiem

Stepwise Selection: Step 1

Variable liv Entered: R-Square = 0.5274 and C(p) = 874.0664

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.09535	2.09535	58.04	<.0001
Error	52	1.87742	0.03610		
Corrected Total	53	3.97277			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr>F
Intercept	1.69555	0.07184	20.11291	557.08	<.0001
Liv	0.18601	0.02442	2.09535	58.04	<.0001

Stepwise Selection: Step 2

Variable enz Entered: R-Square = 0.6865 and C(p) = 565.0059

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.72733	1.36366	55.84	<.0001
Error	51	1.24544	0.02442		
Corrected Total	53	3.97277			

Model: MODEL1

Dependent Variable: vitiem

Stepwise Selection: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.38825	0.08450	6.59186	269.93	<.0001
Enz	0.00565	0.00111	0.63198	25.88	<.0001
Liv	0.13920	0.02209	0.96983	39.71	<.0001

Stepwise Selection: Step 3

Variable prog Entered: R-Square = 0.8827 and C(p) = 183.3416

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3.50682	1.16894	125.44	<.0001
Error	50	0.46595	0.00932		
Corrected Total	53	3.97277			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.94209	0.07144	1.62039	173.88	<.0001
Pprog	0.00790	0.00086356	0.77949	83.64	<.0001
Enz	0.00700	0.00070195	0.92694	99.47	<.0001
Liv	0.08187	0.01502	0.27706	29.73	<.0001

All variables left in the model are significant at the 0.1500 level.

Model: MODEL1

Dependent Variable: vitiem

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr>F
1	liv	1	0.5274	0.5274	874.066	58.04		<.0001
2	enz	2	0.1591	0.6865	565.006	25.88		<.0001
3	prog	3	0.1962	0.8827	183.342	83.64		<.0001

En este análisis de datos usando el software estadístico SAS, se puede definir el modelo de regresión de acuerdo a los indicadores de calidad que se han presentado a través de este libro y evaluar su capacidad de predicción de nuevos valores para las variables predictoras.

11.5. Ejercicios

11.1 El departamento de pesca y vida silvestre de la Universidad de Virginia en Estados Unidos, realizó un experimento para estudiar el efecto de las características de la corriente sobre la biomasa de peces. Se consideraron las siguientes variables regresoras; x_1 = promedio de la profundidad, x_2 = área de cobertura de la corriente, x_3 = área de cobertura del dosel, x_4 = área $\geq$ 25 cm de profundidad. La variable respuesta *y* = biomasa de peces. Los datos son:

Observ.	***y***	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$
1	100	14.3	15.0	12.2	48.0
2	388	19.1	29.4	26.0	152.2
3	755	54.6	58.0	24.2	469.7
4	1288	28.8	42.6	26.1	485.9
5	230	16.1	15.9	31.6	87.9
6	0	10.0	56.4	23.3	6.9
7	551	28.5	95.1	13.0	192.9
8	345	13.8	60.6	7.5	105.8
9	0	10.7	35.2	40.3	0.0
10	348	25.9	52.0	40.3	116.6

a) Calcular s^2, Cp, PRESS y la suma de los residuales PRESS absolutos para el modelo que considera las cuatro variables con la respuesta *Y*.

b) Calcular s^2, Cp, PRESS y la suma de residuales PRESS absolutos para el modelo que contiene las variables x_1, x_2, x_4 y la respuesta *Y*.
c) Compare lo apropiado de los modelos en los incisos anteriores para la predicción de la biomasa de peces.

11.2 Mostrar que $R^2_{pred} < R^2$. Donde $R^2_{pred} = 1 - PRESS/ \sum_{i=1}^{n}(Y_i - \bar{Y})^2$

11.3 En secciones anteriores se ha demostrado que la diagonal HAT cumple la siguiente desigualdad. $\frac{1}{n} \leq h_{ii} \leq 1$; para el modelo de regresión lineal múltiple con un interceptor. Construir un conjunto de datos y su correspondiente modelo en el cual las diagonales HAT sean todos 1.0.

11.4. Los datos de la siguiente tabla contienen información sobre la Edad, Sexo (1 = hombre, 2 =femenino), índice de problemas de trabajo (PT), el índice de problemas maritales (PM) y el índice de depresión (ID) para una muestra de 40 pacientes de una clínica psiquiátrica. Para cada sexo separadamente determine (usando un $\alpha = 0.10$) el mejor modelo de regresión que relacione a PT y a PM, controlando la Edad, usando el siguiente procedimiento secuencial: Iniciar el modelo con edad, (2) usar todas las regresiones posibles de las restantes dos variables independientes, PT y PM, y (3) determinar si el término de interacción (PT x PM) se puede adicionar al modelo. Compare y discuta los resultados obtenidos para cada sexo.

No.	Edad	Sexo	PT	PM	ID	No.	Edad	Sexo	PT	PM	ID
1	45	2	90	70	69	21	28	1	85	30	194
2	35	1	90	75	75	22	37	1	90	9	294
3	32	2	70	32	35	23	29	1	80	14	94
4	32	2	80	30	73	24	29	1	70	24	126
5	39	2	85	55	86	25	31	1	80	21	192
6	25	2	85	6	161	26	29	1	60	11	232
7	22	1	75	20	202	27	29	1	70	10	184
8	30	2	70	63	91	28	23	2	80	10	238
9	49	2	75	4	113	28	44	2	78	19	112
10	47	1	84	12	68	30	28	1	70	22	141
11	48	1	64	11	109	31	32	2	70	21	108
12	49	2	85	7	92	32	36	2	74	77	87
13	45	2	80	8	80	33	22	2	78	67	33
14	41	2	80	15	82	34	46	2	70	25	73
15	45	2	82	6	156	35	21	1	70	14	168
16	59	2	72	5	198	36	34	1	80	17	218
17	42	2	70	17	170	37	27	2	80	18	175
18	35	1	70	29	188	38	31	2	80	42	126
19	31	2	70	80	82	39	19	2	75	36	135
20	45	1	70	126	37	40	27	1	75	58	179

11.5. Para los datos del ejercicio 5.7 del capítulo 5, encontrar (usando un $\alpha = 0.05$) el mejor modelo de regresión que relacione la tasa de homicidios Y con el tamaño de la población X_1, el porcentaje de familias con ingresos menores a \$ 5, 000.00 (X_2) y la tasa de desempleo (X_3).

Use: a). la eliminación stepwise, b) la eliminación backward y c) todas las posibles regresiones.

11.6. Para los datos del ejercicio 11.4, definir el mejor modelo de regresión usando; a) la eliminación stepwise.

b) la eliminación forward.

c) todas las posibles ecuaciones de regresión.

11.7. Los siguientes datos son de un estudio de la tasa de reforestación (Y), la tasa de crecimiento de la población (X_1) y el producto interno bruto per cápita (X_2) para 50 ciudades del país.

a) Ajustar un modelo de regresión lineal de *Y* sobre X_1 y X_2.

b) Calcule los residuales y grafíquelos contra los valores predichos. Comente sobre la necesidad de una transformación.

POB	PIB	DEFOR	POB	PIB	DEFOR	POB	PIB	DEFOR
25.1	940	1.0	23.3	310	1.7	22.6	21220	0.90
28.5	310	2.6	22.6	670	0.0	26.8	820	0.10
28.1	570	0.20	33.8	660	2.4	28.8	1610	4.6
22.2	2240	0.40	19.9	260	0.20	26.0	1310	0.40
25.2	190	0.30	17.7	580	0.50	25.4	820	0.70
25.9	120	0.30	28.0	950	5.9	25.2	290	0.20
25.4	890	0.40	41.1	390	0.70	20.3	320	2.1
22.9	310	0.10	25.1	80	1.2	28.5	440	0.20
23.3	1460	1.70	31.5	490	2.2	25.1	280	0.40

25.9	1180	1.0	27.9	320	1.2	21.0	790	2.4
26.3	1430	3.90	22.9	1890	1.2	28.5	340	0.70
10.4	710	0.10	26.8	2270	1.2	35.2	230	1.1
31.7	1350	2.3	27.6	150	0.80	29.6	4140	0.40
27.7	140	0.20	23.3	170	3.9	21.1	190	0.60
16.5	4000	0.10	34.5	920	2.7	29.4	190	0.20
32.4	360	0.90	33.3	860	4.0	33.0	640	1.2
29.1	1130	2.0	27.4	380	0.20			

El análisis de los datos en SAS es.

Number of observations 50

The GLM Procedure

Dependent Variable: defor

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5.77955898	2.88977949	1.67	0.1995
Error	47	81.39664102	1.73184343		
Corrected Total	49	87.17620000			

R-Square	Coeff Var	Root MSE	defor Mean
0.066297	103.2963	1.315995	1.274000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Crepo	1	5.77423081	5.77423081	3.33	0.0742
Priper	1	0.00532817	0.00532817	0.00	0.9560

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Crepo	1	5.71180270	5.71180270	3.30	0.0757
Priper	1	0.00532817	0.00532817	0.00	0.9560

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-.4771711046	0.99345450	-0.48	0.6332
Crepo	0.0661769738	0.03643969	1.82	0.0757
Priper	0.0000035016	0.00006313	0.06	0.9560

Empleando el procedimiento de eliminación de variables stepwise, tenemos.

Model: MODEL1

Dependent Variable: defor

Stepwise Selection: Step 1

Variable crepo Entered: R-Square = 0.0662 and C(p) = -0.4112

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5.77423	5.77423	3.40	0.0712
Error	48	81.40197	1.69587		
Corrected Total	49	87.17620			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.46542	0.96048	0.39821	0.23	0.6302
crepo	0.06590	0.03571	5.77423	3.40	0.071

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr>F
1	crepo		1	0.0662	0.0662	-0.4112	3.40	0.0712

En este caso que el modelo no ajusta las variables debido al valor de los R^2 en ambas situaciones, es necesario analizar cada variable y hacer una transformación de acuerdo a sus unidades de medida.

Capítulo 12
Diseños de superficies de respuesta

12.1. Introducción

La metodología de superficie de respuesta (MSR) es, como su nombre sugiere, una colección de herramientas para ajustar una superficie a un conjunto de datos y determinar los niveles óptimos de los factores. De esta manera la forma de la superficie se determina por el modelo que se está ajustando y los valores de la respuesta. Se puede hacer una distinción entre la superficie ajustada y la superficie original, que por lo general no coinciden, tal como el estimador de un parámetro raras veces es igual al valor de dicho parámetro. Una superficie ajustada, es en esencia una estimación de la superficie verdadera, ya que si se repite un experimento, la superficie ajustada difiere ligeramente de la primera superficie ajustada.

Típicamente es difícil ajustar un modelo de segundo orden completo para determinar la combinación óptima de niveles de factores; esto es, el modelo con términos lineales y cuadráticos en cada factor y

todas las interacciones de dos factores. Antes de hacer esto, es posible ajustar un modelo de primer orden y utilizar el método de pendiente ascendente o descendente para igualar a cero en una región de operación óptima y después usar un diseño para establecer un modelo de segundo orden que caracterice la región de interés.

Entonces, la aproximación estándar es usar un procedimiento de tres estadíos o pasos. El procedimiento se puede frustrar si se utiliza un diseño de resolución III (un diseño de desperdicio), en el primer paso, en la presencia de interacciones significativas. Si se sospecha la presencia de las interacciones, se puede usar un diseño de resolución IV en el primer paso ya que el subconjunto de factores incorrectos no está identificado en este paso. Otro posible impedimento, discutido por Steinberg y Bursztyn (2001) y Cheng y Wu (2001), es que por lo general no siempre es posible, o al menos práctico, conducir experimentos secuenciales, aunque la necesidad de hacerlo ha sido enfatizado por décadas por Box (1975), en particular. Un componente clave de la metodología de superficie de respuesta (MSR) es la respuesta a diseños de superficie. Snee (1985) dio una lista de las propiedades deseables para los diseños de superficie de respuesta, esta lista apareció primero en Box y Draper (1975) y finalmente en Box y Draper (1987). Las 14 propiedades enlistadas son muy generales, sin embargo, se aplican esencialmente a cualquier diseño experimental. Uno de estos objetivos, “la construcción secuencial de diseños de alto

orden a partir de diseños más simples" se discute más adelante. De forma similar, Anderson y Whitcomb (2004) publicaron una lista, que incluye propiedades generales que cualquier tipo de diseño experimental debe poseer, además de algunas propiedades específicas, tales como, "comportarse bien cuando los errores ocurren en los conjuntos de factores (las *x*'s) y ser insensible para las observaciones desordenadas (solitarias).

En este capítulo, se examinan diseños de superficie de respuestas tradicionales y no tradicionales y se consideran los avances que han ocurrido en los últimos años, para ubicarse lejos de los diseños tradicionales y considerar otros tipos de diseños, especialmente el diseño uniforme (UD, por sus siglas en inglés) que se discutirá más adelante. Los diseños que se han usado tradicionalmente son los diseños centrales compuestos, que permiten el ajuste de modelos de segundo orden, por lo que se puede ver a los diseños de superficie de respuestas como una extensión de los diseños de experimentos conocidos. Se puede también relacionar a un diseño central compuesto como un alargamiento de un diseño factorial puesto que parte del diseño es un factorial completo o fraccional.

Se puede considerar también una propuesta de Cheng y Wu (2001) y otros, para ubicarse lejos de la manera tradicional en que los diseños de superficie de respuesta han sido usados. Típicamente, se ha usado un diseño de dos niveles para identificar factores importantes,

cuando estos factores son investigados en un diseño de superficie de respuesta el investigador está interesado por lo general en ajustar un modelo de segundo orden y quizás también en determinar los puntos o región de condiciones óptimas de operación. Las aplicaciones publicadas de la metodología de superficie de respuesta se han realizado en la industria química y de alimentos, aunque como lo estableció Myers (1999), el interés por la MSR se ha dispersado a los campos biológico, biomédico y biofarmacéutico. Una muestra pequeña de recientes publicaciones de aplicaciones son: Ghadge y Raheman (2006) quienes usaron MSR para la optimización de un proceso en la producción de biodiesel; Moberg *et al.* (2005) usaron MSR en una aplicación de masa espectrométrica; Huang Lu *et al.* (2006) usaron MSR en una aplicación microbiológica para determinar los niveles óptimos de cuatro agentes protectores y Kim y Akoh (2005) usaron MSR para la modelación de una acidosis lipasa catalizada en hexano. Otra aplicación fue la de Tuck *et al.* (1993) quienes aplicaron la MSR en la industria militar.

Aunque existe un gran número de aplicaciones de la MSR, no existen muchos investigadores que trabajen en esta área y lo poco que hacen es incipiente por su pequeño número. Esto está relacionado a la relativa escases de libros sobre la MSR, el primero de estos fue el de Myers (1971) con la segunda edición y ahora la tercera edición de Myers y Montgomery (2002). El trabajo original sobre MSR, fue dado

por el artículo clásico de Box y Wilson (1951) y Box y Draper (1987) que es otro libro bien conocido sobre la MSR. Más recientemente, Khuri (2003) acotó la aproximación de la modelación contemporánea y edición de diseños en la MSR.

12.2. Experimentación con superficie de respuesta: ¿Un diseño o más de uno?

Se puede asumir que existen 20 factores que están relacionados a la variable respuesta y se desea determinar la mejor combinación de niveles de los factores, que están relacionados a la variable respuesta para optimizar los valores de esta variable. El primer paso, obviamente será determinar cuál de los 20 factores son realmente importantes. Típicamente esto será realizado con algún tipo de diseño de desperdicio, que tiene un número pequeño de corridas (combinaciones) relativas al número de factores y para el cual los efectos principales están confundidos con las interacciones de dos factores. De esta manera estos diseños de desperdicio trabajan solamente si los efectos principales disminuyen las interacciones, como sucede a menudo.

Si ciertas interacciones de dos factores, son significativas, se puede falsamente identificar los efectos principales con los cuales están confundidos, que son significativos. La consecuencia de esto, es que

también muchos factores pueden ser analizados en un diseño de superficie de respuesta subsecuente, lo que resulta en un diseño ineficiente. Claramente es mucho mejor, si los factores de desperdicio y la superficie de respuesta ajustada y la optimización son realizadas con un solo experimento, teniendo una alta probabilidad de identificar los efectos reales con los factores de desperdicios. Tal estrategia ha sido propuesta por Cheng y Wu (2001) y Bursztyn y Steinberg (2001), entre otros. El uso de esta estrategia perpetúa la proyección de un espacio factor amplio a un espacio factor más pequeño, denominado, el espacio de factores que son importantes. Esto significa que el diseño usado con factores de desperdicio, deben tener al menos tres niveles ya que el diseño se proyectará a un espacio de factor más pequeño del mismo diseño.

Box y Wilson (1951) criticaron el uso de diseños 3^{k-p}, como diseños de segundo orden, puntualizando que no existen diseños 3^{3-p} útiles, por lo que se tiene que usar un 3^3. Ellos puntualizaron que un diseño 3^{5-1} (81 combinaciones) es necesario para ajustar un modelo de segundo orden con 21 parámetros. Esto es, hay grados de libertad disponibles para la estimación de términos de alto orden que no se usan y los 59 grados de libertad que están disponibles para estimar la varianza del error son más que suficientes. Como puntualizaron Cheng y Wu (2001) sobre la aseveración de Box y Wilson que es necesario un diseño 3^4 para ajustar un modelo de segundo orden, con cuatro factores

y 15 parámetros es incorrecto ya que sería acompañado de cualquier diseño 3_V^{4-1}. Al respecto, Cheng y Wu (2001) puntualizaron que los argumentos de Box y Wilson para no usar este diseño se cumple solo cuando k = 3 y k = 5.

Además, Cheng y Wu (2001) puntualizaron que el argumento del tamaño ineficiente de la corrida para los diseños 3^{k-p} claramente no cumple, cuando el diseño está "haciendo doble trabajo" sirviendo como un diseño de desperdicio y como un diseño para ajustar un modelo de segundo orden. Por ejemplo, asumir que un diseño 3^{10-7} es usado y se establece que hay cinco factores importantes. El diseño proyectado debe tener el mismo número de puntos-diseño, tal como sería un 3^{5-2}. Cheng y Wu (2001) dieron un ejemplo de un diseño 3^{6-3}, proyectado en un diseño 3^{5-2}. Es necesario tener en mente que los diseños tales como un 3^{10-7} y 3^{6-3} son diseños de resolución III, tal que si existen las interacciones, la aproximación será indeterminada. Además, un diseño 3^{5-2}, tiene la misma resolución, las interacciones no son estimadas, esto es, los términos cuadráticos puros son ajustados y no los términos cuadráticos mixtos.

No existe un diseño 3_V^{k-p} con 27 corridas (combinaciones) y solo el diseño 3_V^{k-p} con 81 corridas (combinaciones) es el diseño 3_V^{5-1}, que se puede usar.

Es necesario tener en mente que la idea de usar un solo diseño, con al menos tres niveles, sería frustrante; si el diseño no cubre la región

óptima y además la superficie ajustada que cubre es la misma que la superficie verdadera. Esto sería un supuesto muy arriesgado. Cheng y Wu (2001) usaron el término *diseño elegible proyectado* para designar un diseño de segundo orden. Ya que el número y la identidad de los factores importantes son desconocidos antes de conducir el experimento, por lo que es deseable usar un diseño que tiene un número relativamente grande de diseños elegibles proyectados. Por ejemplo, asumir que se examinan 12 factores. Por lo que existen $\binom{12}{4} = 495$ diseños proyectados en cuatro factores, pero no todos estos son diseños elegibles.

En general, la idea de considerar un diseño 3_V^{k-p} para hacer un doble trabajo, no es muy práctico ya que el diseño tendría puntos sólo para el diseño proyectado de resolución V y todos los efectos de segundo orden son estimados cuando el modelo de segundo orden se ajusta para caracterizar la superficie de respuesta y establecer la combinación óptima de los niveles de los factores. Sin embargo, se pueden hacer algunas comparaciones. Suponga que se tienen 10 factores que son analizados, de los cuales se sabe que cinco son importantes, así como ciertas interacciones entre los cinco factores. Se usa un diseño 3_{IV}^{10-6} y se proyecta a un diseño 3_V^{5-1}, asumiendo que se han identificado correctamente los cinco factores más importantes. Por lo que se puede ajustar un modelo de segundo orden con estos cinco

factores, con 81 corridas (combinaciones), asumiendo que el diseño 3_V^{5-1} cubre la región óptima. Si un diseño 2_{IV}^{10-k} con el número más bajo de puntos-diseño fuera usado, se tiene un diseño 2_{IV}^{10-5}, que se proyecta en un diseño 2^5 para cualquiera de los cinco factores. Convirtiendo esto en un diseño compuesto central que requiere 10 corridas más, por los puntos axiales y se considera que 8 puntos centrales se pueden usar para un total de 50 corridas (combinaciones), por lo que se tiene 31 corridas (combinaciones) menos que con el diseño de tres niveles. Entonces, el último no es competitivo con dos niveles, la aproximación de diseño en el primer paso.

De esta forma, ambas aproximaciones caen aparte si la región óptima no es cubierta por el primer diseño, pero la aproximación con el diseño de dos niveles puede ser superior si se emplea el método de la pendiente ascendente/descendente, además del uso de los dos diseños, dado que la región en la cual el diseño de segundo orden fue usado debe ser rechazada completamente sin mucha experimentación junto con la senda de pendiente ascendente/descendente. De esta manera, la región óptima por lo general es desconocida y parece que no se cubre con un diseño de dos niveles a menos que los niveles del factor se extiendan prácticamente a los límites de la región de operación. Puesto que el número de experimentos que son necesarios a través de una senda ascendente/descendente considerada varían entre experimentos y son desconocidos antes de emprender cualquier experimentación y es casi

imposible comparar la aproximación de un diseño establecido, con la aproximación secuencial estándar.

Si la superficie de respuesta es compleja con picos y valles múltiples, puede ser difícil determinar condiciones de operación óptimas con cada aproximación. Para tales superficies, un método de experimentación continua puede ser empleado tal como la Operación Evolutiva (EVOP), en la investigación de la región óptima; entonces se usa la aproximación estándar de la Operación Evolutiva (Box, 1957) o cualquiera de las aproximaciones de diseño descritas aquí, para determinar las condiciones óptimas de la región óptima que ha sido punteada.

12.3. ¿Cuál diseño?

La cuestión de cuál diseño usar en MSR, fue promovido en la literatura por Tang *et al.* (2004), quienes reportaron que los resultados teóricos de Fang y Mukerjee (2000) mostraron que los éxitos de los diseños factoriales en la exploración de superficies de respuestas se debe al hecho de que los diseños, cubren uniformemente los espacios-diseños, más que a las propiedades combinatorias u ortogonales de los mismos. Al respecto Tang *et al.* (2004) describieron la aplicación sucesiva de un diseño uniforme, en el cual creyeron que tiene considerable potencia en

ciertas situaciones relativas en diseños de superficie de respuesta tradicionales.

En el mismo sentido de diseños alternativos, Hardin y Sloane (1991) discutieron la generación de cálculos de los diseños de superficie de respuesta para regiones esféricas. Establecieron que "los mejores diseños tienen corridas (combinaciones) repetidas al centro y puntos dispersos sobre toda la superficie de la esfera". Similarmente, Cox y Reid (2000) establecieron que cuando las relaciones entre la variable respuesta y los factores es ligeramente no-lineal, "un diseño de espacio-rellenado puede ser usado para explorar la naturaleza de la superficie de respuesta". Por otro lado, Mee (2004) recomendó el uso de diseños D-óptimos e I-óptimos a tres niveles, para regiones de diseño esféricos que involucran tres o más factores. Además, McDaniel y Ankenman (2000) encontraron que una versión de la MSR tradicional trabajó muy bien cuando el objetivo es hacer pequeños cambios en los factores en la investigación de un máximo o un mínimo. Asimismo, esto se puede hacer cuando se usa la Operación Evolutiva (EVOP) y se logran condiciones óptimas (Box y Draper, 1969).

También se pueden usar otros diseños más tradicionales de la MSR, según Giesbrecht y Gumpertz (2004), quienes establecieron en un capítulo sobre los diseños de superficie de respuesta que "en la práctica es más común encontrar serias restricciones que límites para la investigación de regiones irregulares". Esto es una evidencia que mueve afuera los

diseños tradicionales de la MSR y usa cualquier diseño uniforme o diseños completamente similares a los diseños uniformes en la región de aceptabilidad u operatividad. Por lo que se revisan los diseños de superficie de respuesta tradicionales y se retoma la discusión de los diseños uniformes.

12.4. Diseños de superficie de respuesta clásicos *versus* alternativos

Con la tecnología de cómputo actual, se puede fácilmente ver gráficas de datos en tres dimensiones. Por ejemplo, la Figura 12.1, es la gráfica de la superficie de dos factores usando los datos de Vásquez y Martin (1998) cuyo objetivo fue "optimizar el crecimiento de la levadura (*Phaffia rhodozyma*) en fermentación continua, usando turba hidrolisada como substrato". Los dos factores en el experimento fueron: la tasa de dilución en los niveles de 0.13, 0.23 y 0.33 y PH en los niveles de 5, 7 y 9.

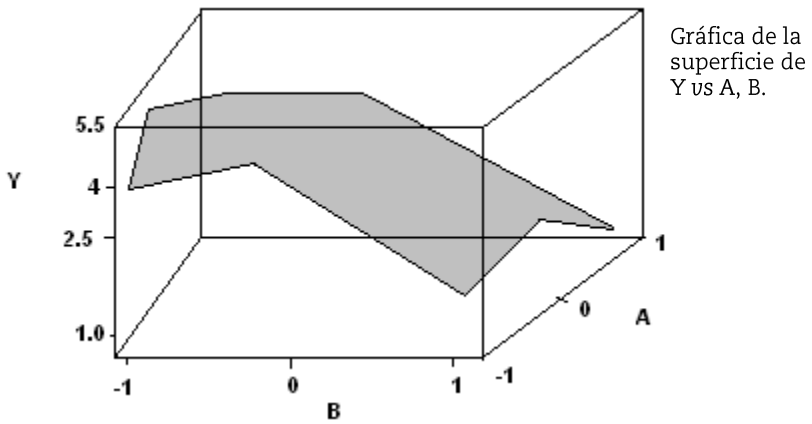


Figura 12.1. Gráfica de la superficie de los datos de Vázquez y Martin (1998).

Esta gráfica fue construida utilizando la versión 14 de MINITAB, con los datos de un diseño 3^2 en donde los 9 puntos se conectaron con segmentos de línea recta para formar la superficie. La necesidad de un término cuadrático en cada factor es obvio, observando el pliegue en la superficie de cada uno de estos., la interacción también se puede intuir a partir del doblez en la superficie (ésta tendría que ser baja de color si sólo los efectos principales fueran importantes, la gráfica sugiere que el término lineal y cuadrático más el término de interacción explican toda la variabilidad en *Y*, por lo que el modelo con estos términos tiene un $R^2 = 0.9984$).

La superficie de respuesta verdadera, no siempre es como la mostrada en la Figura 12. 1 y es muy irregular con picos y valles. Se tiene en mente que la superficie de la Figura 12.1 es obtenida

conectando los puntos con segmentos de línea recta, lo que resulta diferente de una superficie ajustada la cual se obtiene ajustando un modelo a los datos; sin embargo, tal superficie es también aleatoria en el sentido de que los coeficientes son sucesos de variables aleatorias. Obviamente la única manera de obtener una muy buena aproximación a la naturaleza de la superficie de respuesta entre los puntos extremos del diseño es teniendo un número de puntos muy grande de una fina parrilla de esta región.

A menos que la experimentación sea muy barata, como sucede en muchos experimentos computacionales, el costo de obtener el número de puntos diseño necesario para una aproximación cercana a la superficie de respuesta es por lo general prohibitivo. La tarea es aproximar la superficie de respuesta tan cercana como sea posible, de acuerdo a los objetivos y hacerlo de una manera económica, no es una tarea fácil.

Asumir que hay dos factores de interés y considere los arreglos del diseño mostrados en las Figuras 12.2 ,12.3 y 12.4.

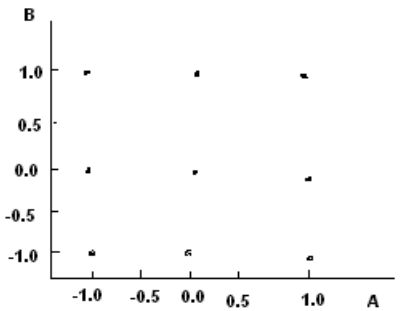


Figura 12.2. Un diseño 3^2.

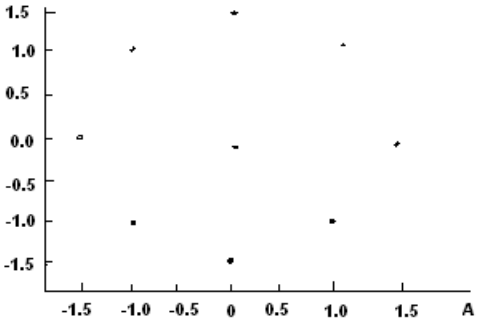


Figura 12.3. Diseño Central compuesto para dos factores

En la Figura 12.2, se muestra el arreglo de un diseño 3^2 que usó Vásquez y Martin (1998). Aunque la región parece estar uniformemente cubierta, una debilidad es que existe solamente un punto en el interior del diseño. Este es un diseño pobre, especialmente si los valores de los factores en las unidades originales se han seleccionado como valores extremos más que como valores que están dentro de los extremos. Con tales valores extremos el punto-central puede ser el único punto-diseño que es siempre un punto usado en la práctica o en proximidad a puntos usados en la práctica (de esta manera, aquí el término "punto central "es usado para referirse al punto del diseño que está en el centro del diseño, que es diferente a los puntos centrales que son adicionados por decir a un diseño 2^k).

Para la Figura 12.3, el arreglo es de un diseño central compuesto. Este diseño también tiene nueve únicos puntos diseño y no difiere fuertemente del diseño 3^2. Como el diseño factorial 3^2, el diseño

central compuesto tiene solamente un punto en el interior del mismo. Este diseño puede ser generado por MNITAB, usando los valores por default para los puntos axiales, que son los puntos próximos a los puntos del diseño 2^2.

La Figura 12.4, es un Diseño Uniforme (UD) usando JMP (nota: este diseño, el hipercubo latino y los diseños esféricos que son mostrados en las Figuras 12.6 y 12.7 son diseños aleatorios, construidos con JMP, este software se usa para construir los diseños, entonces los diseños sucesivamente generados del mismo tipo y con los mismos parámetros de diseño, son diferentes).

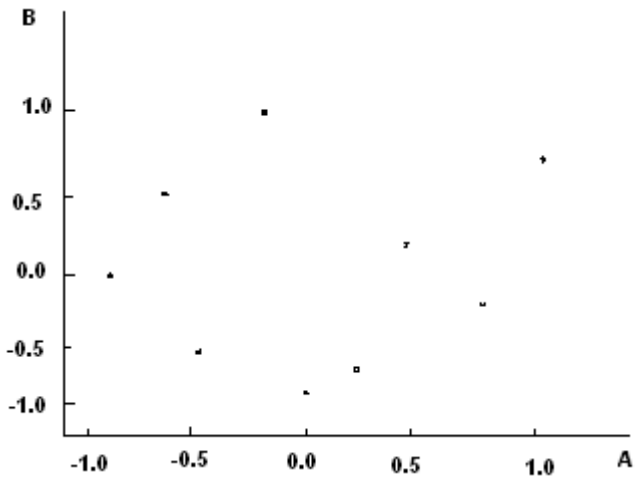


Figura 12.4. Diseño uniforme para dos factores.

A diferencia de los primeros dos diseños, este tiene una configuración irregular, aunque existe mejor cobertura del diseño interior en

comparación a los dos primeros. Recordar que la superficie de la Figura 12.1 fue construida conectando con segmento de línea recta los nueve puntos del diseño, por lo tanto, existen no linealidades muy serias en el interior del diseño, entonces un UD proporcionaría una mejor oportunidad para detectarlos, aunque es probable que se prefieran unos puntos interiores más. El diseño central compuesto de la Figura 12.3, tiene 13 corridas experimentales ya que el punto central está repetido cinco veces. Entonces, un buen ejercicio consistirá en comparar un UD con 13 puntos con el diseño central compuesto.

La escala de los dos diseños es la misma para facilitar una buena comparación y obviamente ayuda si los diseños se muestran en una misma gráfica. Puesto que el rango fue - $\sqrt{2}$ a $\sqrt{2}$ para cada factor en el diseño central compuesto, este rango también fue usado para el UD. La comparación se puede observar en la Figura 12.5 y no existe comparación en términos de cobertura ya que un UD genera esencialmente los mismos grados de cobertura de la periferia del espacio del diseño, como lo hace un diseño central compuesto, pero tiene mejor cobertura de la región interior que el diseño central compuesto. Esto es debido, primeramente al hecho de que existen 13 puntos de diseño distintos para un UD y únicamente nueve para el diseño central compuesto.

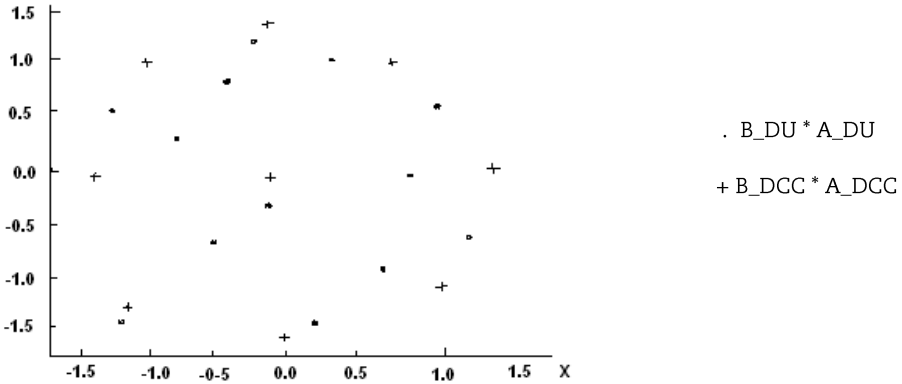


Figura 12.5.Comparación de un diseño central compuesto y un diseño uniforme.

Algunos usuarios del UD pueden omitir su uso por el desigual espaciamiento de los puntos diseño y por el hecho de que no existe un solo UD para un número dado de factores y puntos diseño. Existen otros tipos de diseños de espacios rellenados. La Figura 12.6 muestra la comparación del diseño esférico y el diseño central compuesto.

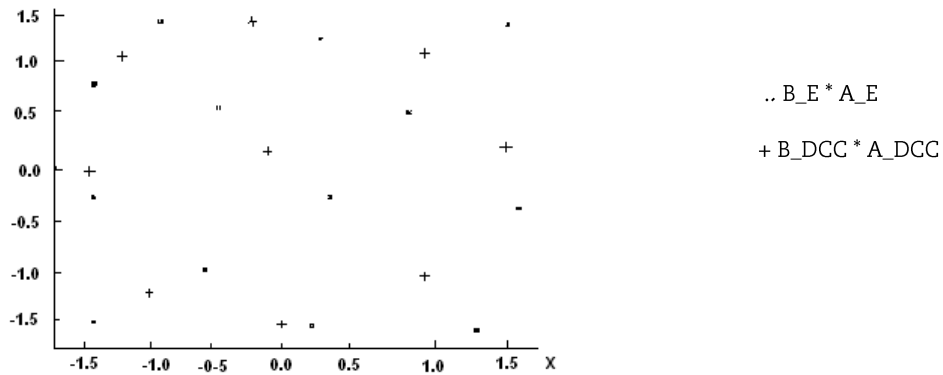


Figura 12.6. Comparación del diseño esférico y el diseño central compuesto.

Finalmente, la Figura 12.7 muestra la comparación de un diseño hipercubo latino con el diseño central compuesto. Es posible de cada una de las últimas tres gráficas, que un diseño de espacio-rellenado provee mejor cobertura de la parte interior del espacio del diseño que un diseño central compuesto. Existe un precio que se paga, aunque es pequeño. Especialmente, los diseños de espacios-rellenados, no son ortogonales para la estimación de los efectos principales, aunque la falta de ortogonalidad no es grande.

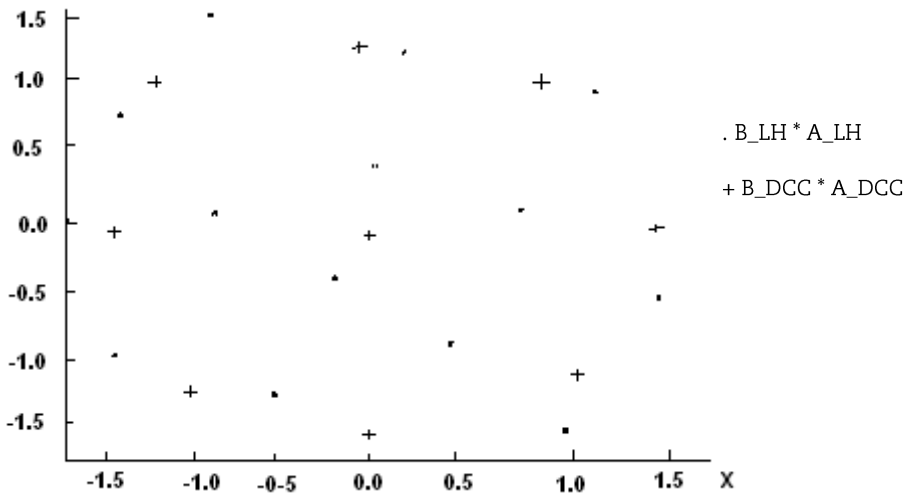


Figura 12.7. Comparación de un diseño hipercubo latino con un diseño central compuesto.

Por ejemplo, la correlación entre la columna de A y B para un diseño hipercubo latino en particular que fue generado es – 0.066, 0.064, con

el diseño esférico y de – 0.041 para el UD. La capacidad para detectar no linealidades más que equivalencias depende de la ortogonalidad. En el espíritu de los diseños de espacios rellenados, el software Design-Expert, puede generar diseños de superficies de respuesta, que dispersan los puntos uniformemente sobre la región experimental. Estos no son diseños aleatorios ya que solamente ciertos puntos son generados aleatoriamente y los otros puntos tienen valores de coordenadas fijos, de 0, 1 y -1.

12.5. Propiedades deseables de los diseños de superficie de respuesta

Hay muchas clases de diseños de experimentos en la literatura y hay muchos criterios en los cuales se basan los diseños experimentales, seguramente existen muchos paquetes de cómputo que generan diseños óptimos basados en criterios especiales de interés del investigador.
Sin embargo, es importante para el lector revisar un grupo de propiedades que se toman en cuenta cuando se selecciona un diseño de superficie de respuesta. Algunas de las características importantes son:

a) El modelo se ajusta muy bien a los datos.
b) Proporciona suficiente información que permite probar la falta de ajuste.
c) Permite construir modelos secuenciales de mayor orden.
d) Proporciona un estimador del error experimental puro.
e) Es insensible (robusto) a la presencia de outliers (puntos lejanos) en los datos.
f) Es robusto a errores de control de los niveles del diseño.
g) Es eficiente en costos.
h) Permite utilizar bloques al realizar el experimento.
i) Permite verificar el supuesto de homogeneidad de varianza.
j) Proporciona una buena distribución de la Var $\hat{y}(x)/\sigma^2$.

Por lo que se puede observar que no todas estas propiedades, son requeridas en todas las experiencia con la MSR. Sin embargo, a muchas de ellas se les debe dar la seriedad debida en cada ocasión en la que se diseñen experimentos. Muchas de las propiedades son auto-explicativas. El punto (e) es importante si se espera la presencia de outliers en los datos. Asimismo, el punto (j) es muy importante, por mantener la estabilidad de la varianza de predicción.

Hay varios propósitos al inicio de la introducción de la lista de las 10 características. De primordial importancia es un recordatorio al lector de que el diseño de experimento no necesariamente es fácil.

Seguramente, pocos de los 10 puntos pueden ser importantes y todavía el investigador no pone la atención debida a la magnitud de su importancia. Algunos puntos causan conflictos con otros. Como resultado, hay abusos que siempre ocurren cuando se diseña un experimento.

12.5.1. Región operativa, región de interés y adecuacia del modelo

Anteriormente se ha discutido la selección de rangos de las variables diseño, lo que hace una introducción de la *región de interés* en las variables de diseño. Como se ha establecido, se asume que la función verdadera que relaciona a *y* con las variables de diseño es desconocida. En realidad,

$$E(y) = f(x, \theta) \tag{1}$$

Y podemos establecer que en la región de interés, R(x), f tiene una buena aproximación con un polinomio de bajo orden. Esto de esta manera, conduce a la noción de funciones de respuesta de primer y segundo orden. Son las aproximaciones (justificadas por una expansión de una serie de Taylor de f) en la región definida R.

Es importante que el lector entienda que mientras que el diseño experimental se pueda definir en R, la región R puede cambiar de

experimento a experimento (por ejemplo; en un experimento de pendiente ascendente). Sin embargo, existen regiones secundarias denominadas "regiones de operación", O (x). Esta es la región en la cual el equipo, sistema electrónico, proceso químico, drogas, entre otros trabajan por lo que es teóricamente posible hacer el experimento y observar valores de respuesta. En algunos casos los puntos del diseño pueden ser tomados fuera de R (x). Obviamente si se toman datos muy fuera de R(x) entonces la adecuación de nuestra representación de la superficie de respuesta es cuestionada. A través de la forma de la región de interés u operación es importante para el lector entenderla ya que el conocimiento de R(x) u O(x) en una situación dada puede ser idealista. La región R cambia y a veces O(x) no es verdaderamente conocida, hasta que se conoce el proceso.

La Figura 12.8 proporciona al lector una muestra pictórica de O y R. En este caso con dos experimentos diferentes, tenemos R y R' ambos son mejor definidos que O. La región R es la "mejor suposición" en donde se encuentra el óptimo. Es también la región en la cual sentimos que $\hat{y}(x)$ es la mejor predictor o un buen estimador de f (x, θ). Sin embargo, el usuario constantemente debe ser cuidadoso de la adecuacia del modelo en casos en donde $\hat{y}(x)$ no sea una buena aproximación de f (x, θ).

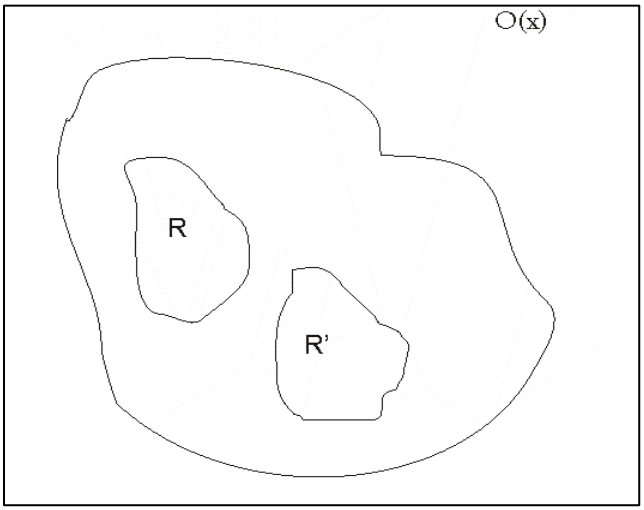


Figura 12.8. Región de operación y región de interés.

12.5.2. Diseños de experimentos para modelos de primer orden

Se puede considerar una situación en la cual se conducen n corridas experimentales en k variables diseño; x_1, x_2, x_3, ..., x_k y una sola respuesta *y*. Se postula el siguiente modelo:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \quad \text{donde } i = 1, 2, 3, ..., n \tag{2}$$

Como un resultado un modelo ajustado es dado por:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + ... + \hat{\beta}_k x_{ik} \tag{3}$$

Los estimadores de los parámetros β_i son calculados por el método de mínimos cuadrados. Obviamente, el criterio de varianza más intuitivo involucra la varianza de los coeficientes de regresión – esto es, Var (β_i), donde i = 1, 2, 3, ... k. Se asume que cada variable de diseño está codificada en unidades de diseño y que los rangos de los niveles de diseño son tales que $x_j \in [-1, 1]$. Este ingrediente clave en la selección de un diseño que minimiza la varianza se le denomina *ortogonalidad.*

12.5.2.1. El diseño ortogonal de primer orden

Los términos; diseño ortogonal, arreglo ortogonal y plan de efectos principales ortogonales recibieron una gran atención en las décadas de 1950 y 1960. Un diseño ortogonal de primer orden es aquél que tiene una matriz diagonal en el producto de *X'X*. La definición anterior implica que las columnas de la matriz *X* son mutuamente ortogonales. En otras palabras si es:

$$X = [1, x_1, x_2, x_3, \ldots x_k] \quad (4)$$

Donde x_j es la *j-ésima* columna de *X*, cuando un diseño ortogonal de primer orden es tal que $x'_i x_j = 0$ para toda $i \neq j$ y de esta manera $1'x_j = 0$ para j =1, 2, 3 ... k. Es claro que si dos columnas son ortogonales los niveles de las dos variables correspondientes son linealmente

independientes. La implicación es que los papeles de las dos variables son evaluados independientemente una de otra. Esto subraya las virtudes de la ortogonalidad. En situaciones de primer orden, un diseño ortogonal en que los rangos de estas variables son tomados como base de resultados extremos para minimizar la Var (β_i) por observación.

Para el modelo de primer orden y una muestra n fija, si $x_j \in [-1, 1]$ para j = 1, 2, 3 ... k, entonces la Var (β_i / σ^2) para i = 1, 2, 3 ... k es minimizada si el diseño es ortogonal y todos los niveles de x_i en el diseño son ± 1 para i = 1, 2, 3, ... k.

Entonces los elementos de la diagonal de $(X'X)^{-1}$ son minimizados [recordar que la Var (β) = $\sigma^2(X'X)^{-1}$] haciendo ceros la diagonal principal de $X'X$ y forzando las diagonales de $X'X$ para que sean lo más grande posible. La noción de independencia lineal y los niveles de las variables al extremo ± 1 como un grupo de condiciones deseables por lo general son intuitivos para el lector.

Ejemplo 12.1 Los planes factoriales de dos niveles y fracciones de resolución ≥ III, en realidad minimizan la varianza de todos los coeficientes básicos por observación. Considere la fracción de un factorial 2^4 (resolución IV) con una relación definida ABCD = I, entonces la matriz *X* está dada por:

$$\begin{pmatrix} & x_1 & x_2 & x_3 & x_4 \\ 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \quad (5)$$

De la que se puede determinar que $X'X = 8\ I_5$ y $(X'X)^{-1} = \frac{1}{8} I_5$. Evaluando esto, se puede decir que ningún otro diseño con 8 corridas (combinaciones) experimentales puede generar una varianza más pequeña que $\sigma^2/8$. Es comprensible para el lector que un factorial 2^4 completo tenga una varianza de $\sigma^2/16$ para todos los coeficientes. En otras palabras si el tamaño del diseño es doble la varianza de los coeficientes es la mitad de la normal. Sin embargo, ambos se consideran diseños óptimos sobre una base por observación.

Ejemplo 12.2. Suponga una situación en la cual se consideran tres variables de diseño y el investigador desea usar 8 corridas experimentales pero también desea repetir el diseño. Una fracción de $\frac{1}{2}$ (resolución

III) es usada con cada uno de los puntos del diseño repetido. La matriz *X* para el diseño considerando la relación ABC = I es:

$$\begin{pmatrix} & x_1 & x_2 & x_3 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad (6)$$

Se puede observar que las columnas son ortogonales y todos los niveles están a ± 1.

Entonces el diseño es de "varianza óptima" para el modelo:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 \quad (7)$$

Todos los coeficientes en el modelo anterior son de varianza mínima sobre todos los diseños con tamaño de muestra n = 8. En realidad la matriz de varianza- covarianza está dada por:

$$\text{Var}\,\hat{\beta} = \frac{1}{8}\sigma^2\, I_4 \qquad (8)$$

Es interesante observar que para la situación descrita en el ejemplo anterior, un factorial 2^3 resulta con la misma matriz de varianza-covarianzas. Aunque los dos diseños son diferentes desde el punto de vista de varianza, son equivalentes para un modelo de primer orden con tres variables de diseño. Obviamente en el caso de una $\frac{1}{2}$ fracción no se tienen grados de libertad para la falta de ajuste, mientras que el factorial 2^3, posee cuatro grados de libertad para la falta de ajuste que son atribuibles a x_1x_2, x_1x_3, x_2x_3 y $x_1x_2x_3$. Por el otro lado, la $\frac{1}{2}$ fracción repetida permite cuatro grados de libertad para el error de repetición (error puro) y de esta manera, el 2^3 completo no posee grados de libertad para el error de repetición. Como se puede ver los dos diseños ortogonales son de varianzas equivalentes por lo que se pueden usar en circunstancias diferentes.

Es claro para el lector que el uso de factoriales de dos niveles o fracciones para un modelo de primer orden, la ortogonalidad se obtiene con diseños de resolución III. En el caso de resolución III, las columnas x_i en la matriz X serán alias con las columnas x_jx_k. Sin embargo, ninguna columna x_i será alias con ninguna columna x_j. En el caso de factoriales 2^k o fracciones regulares; cualesquiera dos columnas en la matriz X que no son alias, son verdaderamente ortogonales.

La noción de ortogonalidad y su relación con diseños de resolución, sugieren que la varianza de los diseños óptimos está disponible si el modelo de respuesta ajustado contiene términos de primer orden y uno o más términos de interacción.

12.5.2.2. Método de pendiente ascendente y/o descendente

Este método, propuesto por Box y Wilson (1951), es usado para tratar de identificar la región, tan exactamente definida como sea posible, que contiene las condiciones óptimas de operación. La metodología para establecer una "senda con pendiente ascendente o descendente" ha estado disponible por décadas, pero con las capacidades modernas de cómputo no es necesario pensar en construir literalmente la senda. Un ejemplo simple para ilustrar la idea básica.

Considerar que existen dos factores de interés y por motivos de ilustración, se presume que la relación entre ellos y la variable respuesta $Y = 2 + 0.40X_1 + 0.30X_2$, y los dos niveles restringidos están en el rango (- 1, 1) como en un diseño de dos niveles. Por lo que se logra que la variable respuesta sea maximizada. ¿Cuál es la senda de pendiente ascendente cuando la combinación de los niveles de factor es (0, 0)?

Observándolo intuitivamente, parece que la senda está en el cuadrante en donde X_1 y X_2 son positivos. Lo que no puede ser obvio, es la relación entre X_1 y X_2 a través de la senda. Ya que el coeficiente

de X_1 es más grande que el coeficiente de X_2, por lo que el cambio en X_2 será expresado en términos del cambio en X_1 y será una función de la relación entre los coeficientes.

Esta relación puede ser fácilmente determinada si se piensa en incrementos iguales en *Y*, iniciando con $Y = 2$, ya que es el valor de *Y* cuando $(X_1, X_2) = (0, 0)$. Se puede pensar en incrementos de 1.0 para *Y*, recordando que *Y* no puede exceder el valor de 2.7 dadas las restricciones de X_1 y X_2.

Si fijamos el incremento de X_1, ΔX_1, en 0.1, podemos resolver para ΔX_2, determinando la dirección de la pendiente ascendente, que es simplemente λ (0.4, 0.3) obtenido de los coeficientes de la ecuación establecida, también se puede obtener usando cálculos directos. Esto es, $\Delta X_1 = 0.1$, $\Delta X_2 = \frac{3}{4} \Delta X_1 = 0.075$.

Otra manera de ver esto es reconocer que para un valor dado de *Y* tal como $Y = 2$, la pendiente de la línea cuando X_2 se define como una función de X_1 es – 1.33. La pendiente de una línea perpendicular para esta línea es el recíproco negativo de esta pendiente, que da 0.75. Un cambio en X_1 de 0.1 resultará en un cambio en X_2 de 0.075.

La línea de la pendiente asciende, si las escalas son las mismas, entonces serán perpendicular a la línea de la Figura 9. Obviamente el máximo ocurre en el punto (1, 1) y no es necesario el método de pendiente ascendente para decir que esto sólo es un ejemplo.

Box y Draper (1987) presentaron un método para determinar si la senda de pendiente ascendente es determinada precisamente lo suficiente. El método involucra la construcción de un cono de confianza cerca de la dirección de la pendiente ascendente o descendente.

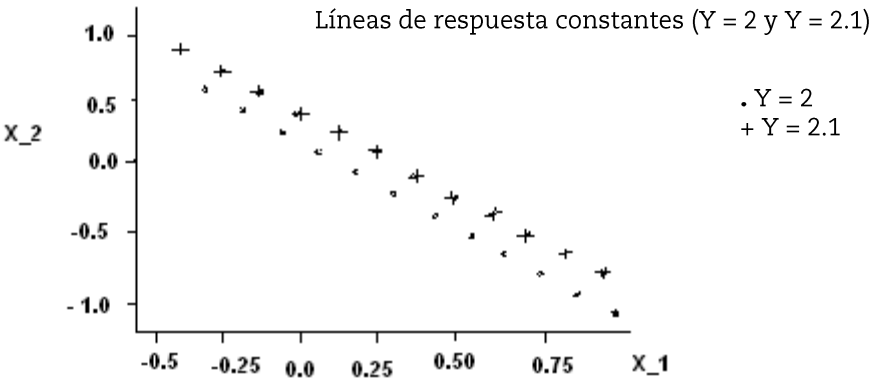


Figura 12.9. Líneas de respuestas constantes para el ejemplo.

El cono está dado por la desigualdad,

$$\sum_{i=1}^{k} b_i^2 - \frac{\left(\sum_{i=1}^{k} b_i X_i\right)^2}{\sum_{i=1}^{k} X_i^2} \leq (\mathrm{k}-1)\, S_b^2 \mathrm{F}_{\alpha,\, \mathrm{k}-1,\, \mathrm{n}-\mathrm{p}} \qquad (9)$$

Donde los b_i denotan los coeficientes de los k (lineales) términos del modelo y S_b^2 denota la varianza común de los coeficientes. Cualquier punto con coordenadas (x_1, x_2, ..., x_k) que satisfacen esta desigualdad,

están en la región de confianza, dado que $\sum_{i=1}^{k} b_i X_{i\ i} > 0$ para pendiente ascendente y $\sum_{i=1}^{k} b_i X_{ii} < 0$ para pendiente descendente. Box y Draper (1987) dieron una medida como buena de una dirección de investigación que es por la fracción de direcciones que son excluidas. De esta manera se desea que la fracción esté lo más cercana a 1 como sea posible. La cercanía de la fracción es a 1, la más grande confianza en la dirección que es usada. De esta manera se puede ver esto en alta dimensión, pero Myers y Montgomery (1995) aplicaron la metodología de Box y Draper (1987) para un ejemplo con k = 2 para calcular la fracción mientras que Sztendur y Diamond (2002) extendieron la aproximación de Box y Draper para ocultar varianzas heterogéneas, diseños no-lineales y modelos lineales generalizados. Usaron efectos condicionales para llegar a la senda de pendiente ascendente.

Si la trayectoria está bien definida, obviamente dependerá de la magnitud del error estándar de los coeficientes de factor, que por lo general serán desconocidos. El valor de R^2 también puede ser usado como un indicador general ya que un valor bajo de R^2 para el modelo ajustado no provee ninguna confianza de que la dirección ha sido bien determinada y valores de R^2 bajos a moderados por lo general corresponden a un cono de confianza relativamente amplio.

Una de las críticas del método de la pendiente ascendente es que la dirección es de escala dependiente. Considere el ejemplo; si X_1 está en pulgadas y es dividido entre 12 para convertirlo a pies, el nuevo

coeficiente será de 0.40 x 12 = 4.8 y resulta una dirección diferente ya que ésta se encuentra determinada por la relación entre los coeficientes del modelo.

Esto ha motivado recientes trabajos de métodos de escala-independiente para pendiente ascendente/descendente de acuerdo a Kleijnen, *et al* (2004). Mientras del Castillo discutió reglas de escalas para el método de pendiente ascendente y sugirió que la experimentación se detiene cuando los valores de la función objetivo seleccionada (que puede ser R^2) es inferior que el valor anterior. Entonces, la experimentación se puede detener cuando R^2 muestra una disminución. Esto indica que un modelo de primer orden no es completamente adecuado para la región que ha sido admitida y es necesario un modelo de segundo orden para representar la experimentación. De esta manera se desea ver el incremento o disminución de la respuesta promedio a través de la senda, el cual depende de la búsqueda de un máximo y un mínimo.

Por lo que se tienen dos situaciones: la necesidad de orientarse a un modelo de segundo orden y una indicación de que se está siguiendo la senda derecha (al menos una buena senda). Nicolai *et al* (2004) sugirieron usar una simple prueba-t en la respuesta de experimentos sucesivos para determinar si los movimientos en las respuestas promedios están en la dirección deseada.

De esta forma, si el método de la pendiente ascendente es usado o no como parte de la aproximación estándar de tres pasos, depende de

que la identificación de los factores importantes y el ajuste del modelo sean aplicadas con el mismo diseño. Esto dependerá de varios factores incluyendo los costos de correr experimentos múltiples y lo práctico de hacerlo si los investigadores creen o no conocer la región de respuesta óptima.

12.6. Diseños para ajustar modelos de segundo orden

Antes de abordar el amplio tema de diseños experimentales para modelos de segundo orden, se revisa para el lector algunos requerimientos mínimos de diseño y la filosofía de MSR que motiva uno o dos de la clase de diseño que se presentan en detalle. Variables screening es una de las fases esenciales de la MSR que ha sido presentada en secciones anteriores. Aquí los diseños factoriales de dos niveles y las fracciones, juegan un papel importante. Cualquier movimiento secuencial que sea necesario es acompañado con un diseño de primer orden. De esta manera, pueden ser ejemplos donde región-deseada vía pendiente ascendente (descendente) y/o las variables screening no se requieren. Sin embargo, la posibilidad de cada uno será incluida en el plan secuencial total. En algunos puntos los investigadores están interesados en ajustar una superficie de respuesta de segundo orden en las variables diseño; x_1, x_2, x_3, . . ., x_k. Este análisis

de superficie de respuesta puede involucrar la optimización de procesos. Se puede conducir más movimientos secuenciales con el uso del análisis canónico o ridge (cordillera). Pero, a pesar de la forma del análisis, el propósito del diseño experimental es que permite al usuario ajustar un modelo de segundo orden como:

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i}^{k} \sum_{j}^{k} \beta_{ij} x_i x_j + \varepsilon (10)$$

Este modelo contiene 1 + 2k + k (k – 1)/2 parámetros. Este es por lo menos el número de puntos de distintos diseños y al menos tres niveles de cada variable. Estos representan las condiciones mínimas pero se deben de tomar en cuenta las 10 características enlistadas al inicio del punto (1). En las siguientes secciones se discutirá la importancia de las propiedades de diseño para los modelos de segundo orden y clases específicas de diseños de segundo orden. Asimismo, se puede recordar que en los diseños de primer orden la propiedad dominante es la ortogonalidad. En el caso de diseños de segundo orden la ortogonalidad deja de ser tan importante así como la estimación de los coeficientes, mientras se torna importante la varianza de predicción escalada N Var $\hat{y}/s^2$. Esto viene del hecho de que a menudo existe menos relación con las variables que están en el modelo y se enfatiza sobre la calidad de $\hat{y}$ (x) como una predicción o más como un estimador para E [y (x)]. De

esta manera se puede pensar formalmente en la clase de diseños centrales compuestos.

12.6.1. Los diseños centrales compuestos

Las clases de diseños centrales compuestos (DCC) fueron comentados anteriormente de manera informal. El DCC es sin lugar a dudas la clase más popular de los diseños de segundo orden y fue introducido por Box y Wilson (1951). Mucha de la motivación de los DCC se desarrolla de su uso en la experimentación secuencial. Involucra el uso de diseños factoriales de dos niveles o fracciones (de resolución V) combinado con los siguientes 2k puntos axiales o estrellas.

El diseño involucra, F, puntos factoriales, 2k puntos axiales y n_c puntos centrales. La naturaleza secuencial del diseño es muy obvia. Los puntos factoriales representan un diseño de varianza óptima para un modelo de primer orden o un tipo de modelo de primer orden + las interacciones de dos factores. Esto es:

x_1	x_2	…	x_k
-α	0	…	0
α	0	…	0
0	-α	…	0
0	α	…	0
.	.		.
.	.	…	.
.	.		.
0	0	…	-α
0	0	…	α

(11)

Las corridas centrales (puntos centrales) proporcionan información acerca de la existencia de curvatura en el sistema. Si se encuentra la curvatura en el sistema la adición de puntos axiales permite la estimación eficiente del término cuadrático puro.

Mientras la génesis de este diseño se derivó de la experimentación secuencial, el DCC es un diseño muy eficiente en situaciones para un paquete de experimentos de superficie de respuesta no secuenciales. En efecto los tres componentes del diseño juegan un papel importante y algunas veces diferentes.

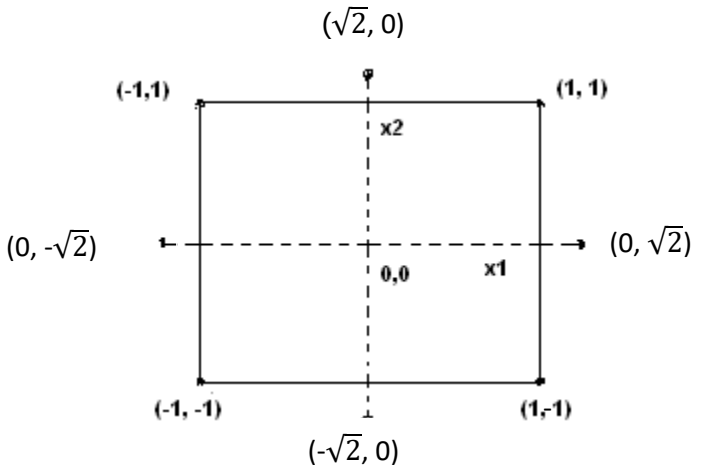
1) Las fracciones de resolución V contribuyen en una mejor manera en la estimación de términos lineales e interacciones de dos factores. La varianza es óptima para estos términos.

Los puntos factoriales son los que contribuyen a la estimación de los términos de la interacción.

2) Los puntos axiales contribuyen en gran manera a la estimación de términos cuadráticos. Sin los puntos axiales, solamente se pueden estimar la suma de los términos cuadráticos $\sum_{i=1}^{k} \beta_{ii}$. Los puntos axiales no contribuyen a la estimación de los términos de interacción.
3) Las corridas centrales (puntos centrales) proveen una estimación del error (error puro) interno y contribuyen en la estimación de términos cuadráticos.

Las áreas de flexibilidad en el uso de diseños centrales compuestos reside en la selección de α, la distancia axial y n_c el número de corridas centrales. La selección de estos dos parámetros es muy importante. La selección α depende de la gran extensión de la región de operación y la región de interés. La selección de n_c por lo general tiene un impacto en la distribución de N Var $\hat{y}/s^2$, en la región de interés. Las Figuras 12.10 y 12.11 muestran los DCC para k =2 y k = 3. Para el caso de k =2, el valor de α, la distancia axial es $\sqrt{2}$. Para el caso de k = 3, el valor de α es $\sqrt{3}$. Note que para k = 3, los puntos axiales van a través de las seis caras a una distancia de $\sqrt{3}$ del origen. Note que para k = 2 el diseño representa ocho puntos igualmente espaciados en un circulo, más las corridas centrales (puntos centrales). Para k = 3, el diseño representa 14

puntos sobre una esfera común, más las corridas centrales (puntos centrales)



cuando k = 2 y α = $\sqrt{2}$

Figura 12.10. Diseño central compuesto.

Los valores de la distancia axial generalmente varían de 1.0 a $\sqrt{k}$, la asignación de todos los puntos axiales en la cara del cubo o hipercubo, obteniéndose que todos los puntos están en una esfera común. Hay veces que dos o más corridas centrales son necesarias y cuando una o dos son suficientes. Se discute la selección de α y n_c con un ejemplo numérico de uso de un DCC.

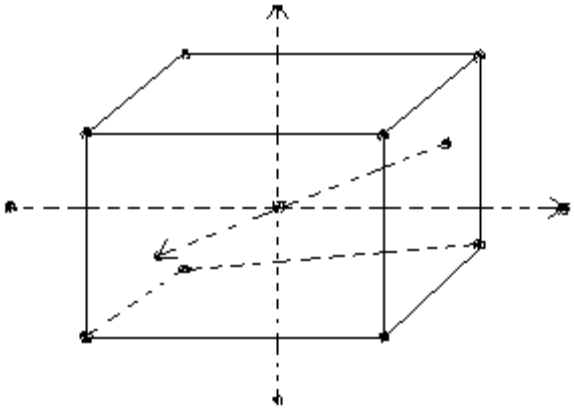


Figura 12.11. Diseño central compuesto para k = 3 y $\alpha = \sqrt{3}$.

Ejemplo 12.3. Se condujo un experimento para estudiar la superficie de respuesta relacionada con la resistencia a la rotura del embalaje en gr/pulg^2, con la temperatura (x_1), temperatura de enfriamiento(x_2) y el porcentaje del polietileno aditivo (x_3). La definición del diseño fue:

$$x_1 = \frac{°F - 255}{30} \qquad x_2 = \frac{°F - 55}{9}$$

$$x_3 = \frac{\%\ polietilieno - 1.1}{0.60}$$

Se establecieron cinco niveles de cada factor en el diseño. Los niveles naturales y codificados son los siguientes:

Factores	-1.682	-1.000	0.000	1.000	1.682
x_1	204.5	225	255	285	305.5
x_2	39.9	46	55	64	70.1
x_3	0.09	0.50	1.1	1.7	2.11

Los datos con los valores de los factores codificados son:

x_1	x_2	x_3	y
-1	-1	-1	6.6
1	-1	-1	6.9
-1	1	-1	7.9
1	1	-1	6.1
-1	-1	1	9.2
1	-1	1	6.8
-1	1	1	10.4
1	1	1	7.3
-1.682	0	0	9.8
1.682	0	0	5.0
0	-1.682	0	6.9
0	1.682	0	6.3
0	0	-1.682	4.0
0	0	1.682	8.6
0	0	0	10.1
0	0	0	9.9
0	0	0	12.2
0	0	0	9.7
0	0	0	9.7
0	0	0	9.6

Note que k = 3, el DCC contiene un $\alpha = 1.682$ y una $n_c = 6$ corridas centrales.

Para su análisis usamos el software SAS.

```
data proceso;
input tie tem catal acti;
```

```
cards;
-1      -1      -1      6.6
1       -1      -1      6.9
-1      1       -1      7.9
1       1       -1      6.1
-1      -1      1       9.2
1       -1      1       6.8
-1      1       1       10.4
1       1       1       7.3
-1.682  0       0       9.8
1.682   0       0       5.0
0       -1.682  0       6.9
0       1.682   0       6.3
0       0       -1.682  4.0
0       0       1.682   8.6
0       0       0       10.1
0       0       0       9.9
0       0       0       12.2
0       0       0       9.7
0       0       0       9.7
0       0       0       9.6

run;
proc sort;
by tie tem catal acti;
run;
proc rsreg data=proceso out=proceso;
model acti = tie tem catal/lackfit residual;
ridge max;
run;
```

```
proc rsreg data=proceso out=super1;
model acti= tie tem catal/actual lackfit residual predict press;
ridge radius=0 to 20 by 1 maximun minimum;
run;
```

Los resultados son.

The RSREG Procedure

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
Tie	0	1.682000
Tem	0	1.682000
Catal	0	1.682000

Response Surface for Variable acti

Response Mean	8.150000
Root MSE	1.089003
R-Square	0.8557
Coefficient of Variation	13.3620

Type I

Regression	DF	Sum of Squares	R-Square	F Value	Pr > F
Linear	3	30.961289	0.3768	8.70	0.0039
Quadratic	3	36.189446	0.4404	10.17	0.0022
Crossproduct	3	3.160000	0.0385	0.89	0.4801
Total Model	9	70.310735	0.8557	6.59	0.0034

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	5	6.899265	1.379853	1.39	0.3630
Pure Error	5	4.960000	0.992000		
Total Error	10	11.859265	1.185926		

Este análisis de varianza presenta cinco grados de libertad para la falta de ajuste, representando la contribución del término cuadrático. La prueba de F para la falta de ajuste no es significativa

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	10.164923	0.444151	22.89	<.0001	10.164923
tie	1	-1.103626	0.294667	-3.75	0.0038	-1.856299
tem	1	0.087185	0.294667	0.30	0.7734	0.146646
catal	1	1.020424	0.294667	3.46	0.0061	1.716353
tie*tie	1	-0.759633	0.286812	-2.65	0.0244	-2.149097
tem*tie	1	-0.350000	0.385021	-0.91	0.3847	-0.990193
tem*tem	1	-1.042406	0.286812	-3.63	0.0046	-2.949097
catal*tie	1	-0.500000	0.385021	-1.30	0.2232	-1.414562
catal*tem	1	0.150000	0.385021	0.39	0.7050	0.424369
catal*catal	1	-1.148446	0.286812	-4.00	0.0025	-3.249097

La estimación por mínimos cuadrados nos da la función de superficie de respuesta de segundo orden.

$$\hat{y} = 10.165 - 1.1036x_1 + 0.0871x_2 + 1.0120x_3 - 0.760x^2_1 - 1.042x^2_2 - 1.148x^2_3 - 0.350x_1x_2 - 0.500x_1x_3 + 0.150x_2x_3$$

The RSREG Procedure

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
Tie	4	27.934611	6.983653	5.89	0.0106
Tem	4	16.929052	4.232263	3.57	0.0467
Catal	4	35.416312	8.854078	7.47	0.0047

The RSREG Procedure

Canonical Analysis of Response Surface Based on Coded Data

Factor	Critical Value Coded	Uncoded
Tie	-0.600860	-1.010647
Tem	0.154878	0.260505
Catal	0.405041	0.681278

Predicted value at stationary point: 11.081564

Este punto se calcula de acuerdo a la siguiente expresión.

$$x_s = -\tfrac{1}{2}\,\hat{B}^{-1}\beta = \begin{pmatrix} -1.011 \\ 0.260 \\ 0.681 \end{pmatrix}$$

Lo cual sustituyéndolo en la ecuación de segundo orden nos da: $\hat{y}$ (x_s) = 11.08

	Eigenvectors		
Eigenvalues	tie	tem	catal
-1.590113	0.837364	-0.368155	-0.404084
-3.160626	0.305117	0.928117	-0.213313
-3.596551	0.453569	0.055328	0.889502

Stationary point is a maximum.

The RSREG Procedure

Estimated Ridge of Maximum Response for Variable acti

Coded	Estimated	Standard	Uncoded Factor Values		
Radius	Response	Error	tie	tem	catal
0.0	10.164923	0.444151	0	0	0
0.1	10.399535	0.442699	-0.126974	0.015213	0.109258
0.2	10.597963	0.439213	-0.259379	0.039688	0.210506
0.3	10.761424	0.436369	-0.395151	0.071247	0.305616
0.4	10.890745	0.438707	-0.533026	0.108215	0.396021
0.5	10.986501	0.452302	-0.672231	0.149344	0.482787
0.6	11.049096	0.483602	-0.812286	0.193715	0.566701
0.7	11.078824	0.537622	-0.952891	0.240647	0.648351
0.8	11.075899	0.616726	-1.093851	0.289629	0.728179
0.9	11.040482	0.720877	-1.235042	0.340276	0.806520
1.0	10.972698	0.848770	-1.376381	0.392288	0.883635

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
Tie	0	1.682000
Tem	0	1.682000
catal	0	1.682000

Response Surface for Variable acti

Response Mean	8.150000
Root MSE	1.089003
R-Square	0.8557
Coefficient of Variation	13.3620
Sum of Squared Residuals	11.859264833
Predicted Residual SS (PRESS)	59.746491516

Type I

Regression	DF	Sum of Squares	R-Square	F Value	Pr > F
Linear	3	30.961289	0.3768	8.70	0.0039
Quadratic	3	36.189446	0.4404	10.17	0.0022
Crossproduct	3	3.160000	0.0385	0.89	0.4801
Total Model	9	70.310735	0.8557	6.59	0.0034

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	5	6.899265	1.379853	1.39	0.3630
Pure Error	5	4.960000	0.992000		
Total Error	10	11.859265	1.185926		

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	10.164923	0.444151	22.89	<.0001	10.164923
Tie	1	-1.103626	0.294667	-3.75	0.0038	-1.856299
tem	1	0.087185	0.294667	0.30	0.7734	0.146646
catal	1	1.020424	0.294667	3.46	0.0061	1.716353
tie*tie	1	-0.759633	0.286812	-2.65	0.0244	-2.149097
tem*tie	1	-0.350000	0.385021	-0.91	0.3847	-0.990193
tem*tem	1	-1.042406	0.286812	-3.63	0.0046	-2.949097
catal*tie	1	-0.500000	0.385021	-1.30	0.2232	-1.414562
catal*tem	1	0.150000	0.385021	0.39	0.7050	0.424369
catal*catal	1	-1.148446	0.286812	-4.00	0.0025	-3.249097

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
Tie	4	27.934611	6.983653	5.89	0.0106
Tem	4	16.929052	4.232263	3.57	0.0467
Catal	4	35.416312	8.854078	7.47	0.0047

Canonical Analysis of Response Surface Based on Coded Data

Critical Value

Factor	Coded	Uncoded
Tie	-0.600860	-1.010647
Tem	0.154878	0.260505
Catal	0.405041	0.681278

Predicted value at stationary point: 11.081564

Eigenvectors

Eigenvalues	tie	tem	catal
-1.590113	0.837364	-0.368155	-0.404084
-3.160626	0.305117	0.928117	-0.213313
-3.596551	0.453569	0.055328	0.889502

Stationary point is a maximum.

The RSREG Procedure

Estimated Ridge of Minimum Response for Variable acti

Coded	Estimated	Standard	Uncoded Factor Values		
Radius	Response	Error	tie	tem	catal
0	10.164923	0.444151	0	0	0
1.000000	5.133441	0.848713	0.306666	0.337630	-1.618977
2.000000	-6.469373	3.192262	-0.325797	0.498917	-3.310806
3.000000	-25.192266	7.232741	-1.055178	0.514737	-4.907521
4.000000	-51.086599	12.907182	-1.803961	0.482027	-6.463695
5.000000	-84.164565	20.207727	-2.559349	0.427479	-7.999693
6.000000	-124.430621	29.132469	-3.317671	0.361196	-9.524235
7.000000	-171.886795	39.680761	-4.077524	0.287858	-11.041649
8.000000	-226.534147	51.852332	-4.838266	0.209943	-12.554324
9.000000	-288.373286	65.647054	-5.599565	0.128883	-14.063687
10.000000	-357.404593	81.064860	-6.361235	0.045570	-15.570646
11.000000	-433.628309	98.105713	-7.123162	-0.039415	-17.075802
12.000000	-517.044600	116.769587	-7.885276	-0.125675	-18.579572
13.000000	-607.653584	137.056470	-8.647528	-0.212928	-20.082253
14.000000	-705.455345	158.966352	-9.409886	-0.300972	-21.584063
15.000000	-810.449946	182.499226	-10.172326	-0.389655	-23.085165
16.000000	-922.637434	207.655088	-10.934831	-0.478862	-24.585685
17.000000	-1042.017846	234.433934	-11.697388	-0.568505	-26.085719
18.000000	-1168.591210	262.835762	-12.459989	-0.658512	-27.585344
19.000000	-1302.357527	292.860565	-13.222625	-0.748830	-29.084621
20.000000	-1443.316865	324.508354	-13.985290	-0.839412	-30.583601

Estimated Ridge of Maximum Response for Variable acti

Coded	Estimated	Standard	Uncoded Factor Values		
Radius	Response	Error	tie	tem	catal
0	10.164923	0.444151 0	0	0	
1.000000	10.972698	0.848770	-1.376381	0.392288	0.883635
2.000000	8.528616	3.192558	-2.791280	0.953665	1.617336
3.000000	2.891029	7.233331	-4.204643	1.546846	2.321715
4.000000	-5.931517	12.908111	-5.616449	2.151260	3.015553
5.000000	-17.936514	20.209040	-7.027282	2.760927	3.704414
6.000000	-33.122962	29.134211	-8.437498	3.373475	4.390530
7.000000	-51.490380	39.682979	-9.847305	3.987772	5.074971
8.000000	-73.038511	51.855072	-11.256829	4.603211	5.758315
9.000000	-97.767204	65.650364	-12.666148	5.219437	6.440901
10.000000	-125.676365	81.068786	-14.075316	5.836228	7.122942
11.000000	-156.765931	98.110302	-15.484369	6.453439	7.804578
12.000000	-191.035856	116.774885	-16.893332	7.070970	8.485905
13.000000	-228.486119	137.062526	-18.302223	7.688751	9.166989
14.000000	-269.116691	158.973213	-19.711057	8.306730	9.847881
15.000000	-312.927559	182.506939	-21.119844	8.924870	10.528617
16.000000	-359.918709	207.663701	-22.528592	9.543142	11.209226
17.000000	-410.090131	234.443494	-23.937307	10.161523	11.889728
18.000000	-463.441818	262.846316	-25.345994	10.779995	12.570142
19.000000	-519.973765	292.872167	-26.754658	11.398546	13.250479
20.000000	-579.685966	324.521043	-28.163301	12.017162	13.930752

Los eigenvalores de la matriz $\hat{B}$ son:

$\lambda_1 = -0.562$; $\lambda_2 = -1.271$ y $\lambda_3 = -1.117$

por lo que la función en forma canónica es;

$$\hat{y} = 11.08 - 0.562w_1^2 - 1.271w_2^2 - 1.117w_3^2$$

Entonces se concluye que el punto estacionario es una media máxima estimada de la fuerza a la ruptura del embalaje.

Por lo general se tiene que $\alpha = \sqrt[4]{F}$ donde F es el número de puntos del factorial ($F = 2^k$ si es un factorial completo). Es importante notar que la rotabilidad es obtenida simplemente usando α, no obstante del número de corridas centrales. En la Tabla 1 se dan valores de α de un diseño rotable para varios valores del número de variables de diseño. Note que para k = 2 y k = 4 el DCC contiene 8 y 24 puntos (aparte de las corridas centrales) respectivamente, que son equidistantes del centro del diseño. Para k= 3, el α = 1.682 corresponde al DCC usado en el ejemplo anterior. Para k = 2, 3 y 4 los DCC rotables son exactamente o muy cercanos a un diseño esférico, en que todos los puntos están exactamente (para k = 3) a una distancia de $\sqrt{k}$ del centro del diseño.

Tabla 1. Valores de α para un diseño central compuesto rotable.

k	F	N	α
2	4	8 + nc	1.414
3	8	14 + nc	1.682
4	16	24 + nc	2.000
5	32	42 + nc	2.378
5(½ rep)	16	26 + nc	2.000
6	64	76 + nc	2.828
6(½ rep)	32	44 + nc	2.378
7	128	142 + nc	3.364
7(½ rep)	64	78 + nc	2.828

La propiedad de rotabilidad es un intento para producir un cierto sentido de estabilidad de N Var $\hat{y}$ /σ^2. El sentido es mantener constante a N Var $\hat{y}$ /σ^2 en la esfera. Sin embargo, la presencia de un diseño rotable no implica estabilidad en la región del diseño. En realidad un diseño esférico (todos los puntos tienen un mismo radio) usado para ajustar un modelo de segundo orden tiene una N Var $\hat{y}$ /σ^2 infinita ya que el diseño es singular, esto es, (X'X) es una matriz singular. El uso de corridas centrales provee una estabilidad razonable de N Var $\hat{y}$ /σ^2 en la región de diseño, por lo que es recomendable integrar algunas corridas centrales en DCC ya que es benéfico. El uso de un DCC rotable o semi-rotable con pocas corridas centrales no es práctico ni recomendable. Una ilustración de este punto es dado en las Figuras 12.12 y 12.13. La Figura 12.12 muestra los contornos

de N Var $\frac{\hat{y}}{\sigma^2}$ para un DCC con k = 2 ($\alpha = \sqrt{2}$) y una corrida central. La Figura 12.13 muestra los contornos con un $\alpha = \sqrt{2}$ y cinco corridas centrales. Se puede notar que el diseño en la Figura 12.13 es preferible. Note también que el criterio involucra un peso para N, lo que significa que el diseño en la Figura 12.13 tiene un peso más grande. No obstante esto, el diseño con n_c = 1 corridas tiene una varianza de predicción escalada al centro del diseño que es 3.5 veces más grande que la del diseño con n_c = 5. Por lo anterior, se puede decir que los diseños esféricos o semiesféricos requieren de 3 - 5 corridas centrales para evitar un severo desbalance en la N Var $\hat{y}$ /σ^2 en la región de diseño. El mensaje que transmiten las dos figuras anteriores se puede extender para valores más grandes de k.

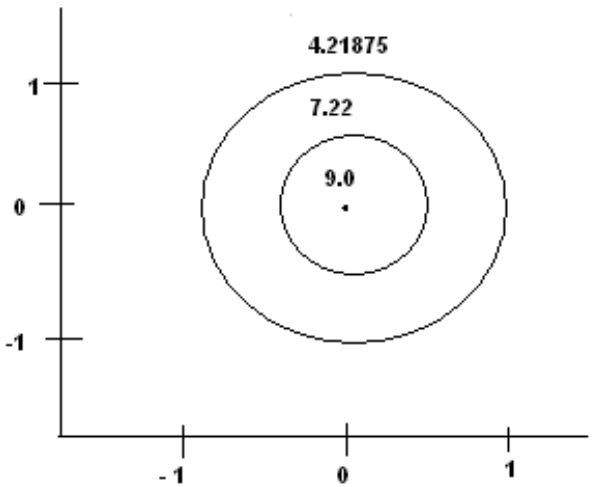


Figura 12.12. N Var $\frac{\hat{y}}{\sigma^2}$ para k = 2, $\alpha = \sqrt{2}$, n_c = 1

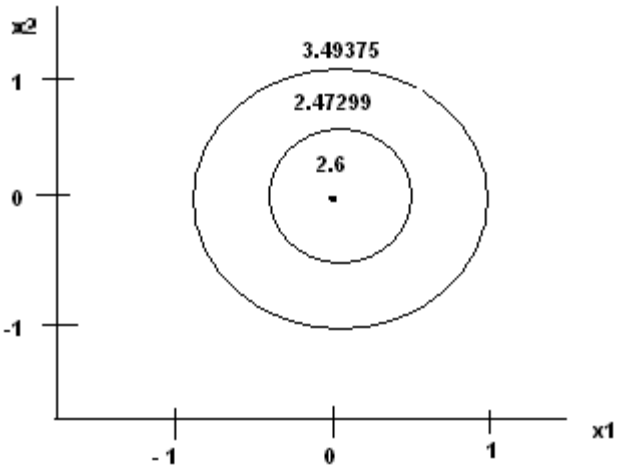


Figura 12.13. N Var $\frac{\widehat{y}}{\sigma^2}$ para k = 2, α= $\sqrt{2}$, n_c = 5

Los DCC juegan un papel muy importante en la MSR desde un punto de vista histórico y operativo. Sin embargo, es importante para el analista entender que no es necesario tener una rotación exacta en un diseño de segundo orden. En realidad, si la región de diseño deseada es esférica, el DCC, que es más efectivo desde el punto de vista de varianza, es usar un α = $\sqrt{k}$ y de 3 a 5 puntos centrales. Este diseño no es necesariamente rotable pero es semi-rotable. La recomendación está basada en la estabilidad y el tamaño de N Var $\frac{\widehat{y}}{\sigma^2}$ en la región esférica del diseño. Por ejemplo para k = 3 el α = $\sqrt{3}$ no produce un diseño rotable. Sin embargo, la pérdida de la rotabilidad es trivial y el valor grande de α con los valores rotables de α = 1.682, resulta en un diseño

que es ligeramente preferible. Evidencias numéricas se obtienen en otras referencias como Box y Draper (1987), Kuri y Cornell (1987), Lucas (1976), Giovannitti-Jensen y Myers (1989) y Myers *et al* (1992b).

Hay muchas situaciones prácticas en las cuales el científico o investigador especifica rangos de las variables del diseño y esos rangos son estrictos. Esto es, la región de interés y la de operación son las mismas y la región del diseño es un cubo. Por ejemplo: en un estudio experimental diseñado para el estudio de organismos en crecimiento, las variables diseño y sus rangos son; porcentaje de glucosa (2%, 4%), porcentaje de levadura (0.4, 0.6) y tiempo en hr (30, 60). Suponga que el interés es construir un modelo de superficie de respuesta y el biólogo está interesado en predecir el crecimiento del organismo dentro y en el perímetro de la región cuboidal producida por el cubo. Además por razones biológicas, no se puede hacer el experimento fuera del cubo, aunque la experimentación a los extremos de la región es permisible y en realidad deseable. Para este escenario, que ocurre frecuentemente en muchas áreas científicas, se sugiere un diseño central compuesto en que las 8 esquinas del cubo están centradas y escaladas a $(\pm 1, \pm 1, \pm 1)$ y α no puede exceder a 1.0. El diseño final es dado (en forma codificada) por:

x_1	x_2	x_3
-1	-1	-1
1	-1	-1
-1	1	-1
-1	-1	1
1	1	-1
1	-1	1
1	1	1
-1	0	0
1	0	0
0	-1	0
0	1	0
0	0	-1
0	0	1
0	0	0

Donde el punto (0, 0, 0) al centro del diseño, indica un vector de corridas centrales.

La Figura 12.14 muestra el diseño, denominado por lo general la cara del cubo central, ya que los puntos axiales ocurren en el centro de las caras, más que fuera de éstas, como el caso de una región esférica.

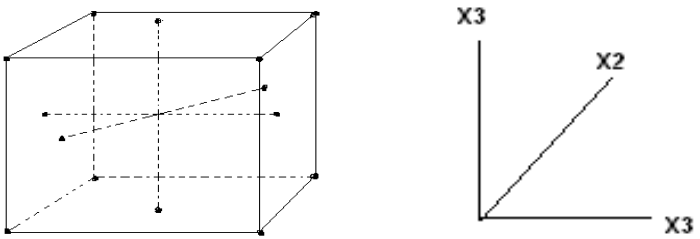


Figura 12.14. Cara del cubo central (DCC con α = 1.0) para k = 3.

En el caso de la región de diseño cuboidal, la cara central del cubo es un diseño efectivo de segundo orden. Cuando se encuentra que es una región cuboidal natural, es importante que los puntos estén, "empujados al extremo" de la región experimental. Esto produce una distribución más atractiva de N Var $\frac{\widehat{y}}{\sigma^2}$. Es importante que la región sea cubierta de una forma simétrica; la cara central del cubo define esto. De esta forma, el diseño no es rotable. Sin embargo, la rotabilidad o la semi-rotabilidad no es una prioridad importante, cuando la región de interés es claramente cuboidal. La rotación (o la semi-rotabilidad) es una opción útil que viene de diseños esféricos o semi-esféricos; estos diseños son ciertamente apropiados para regiones de interés esféricas o regiones de operación y son menos apropiados con regiones cuboidales.

La cara central del cubo es un diseño útil para la porción factorial. Aquí una fracción de resolución V es útil para la porción factorial. La recomendación para corridas centrales es completamente

diferente a la de los diseños esféricos. En el caso de diseños esféricos, las corridas centrales son una necesidad para evaluar una distribución razonable de N Var $\frac{\widehat{y}}{\sigma^2}$, con n_c = 3 a 5 obteniéndose buenos resultados. En el caso cuboidal (con α = 1.0) uno o dos puntos centrales son suficientes para producir una estabilidad razonable de N Var $\frac{\widehat{y}}{\sigma^2}$ /. La sensibilidad de N Var $\frac{\widehat{y}}{\sigma^2}$, al número de corridas centrales para el diseño esférico se puede observar en las Figuras 12.12 y Figura 12.13. Es claro que no son necesarias muchas corridas para estabilizar la varianza de predicción. En realidad, una corrida central es completamente suficiente para estabilizarla aunque con n_c = 2 se mejora la estabilidad. Aunque se ha dicho mucho aquí del impacto de las corridas centrales sobre N Var $\frac{\widehat{y}}{\sigma^2}$, para los diseños esféricos y cuboidales, se puede notar que corridas centrales múltiples o repeticiones de puntos exteriores en muchos casos pueden ser requeridos para tener suficientes grados de libertad para el error puro.

Como se ha indicado en esta sección, el DCC cuboidal es apropiado cuando una región de interés cuboidal y una región cuboidal de operación son claras para el investigador. Claramente el problema puede sugerir rangos en los factores con los cuales se pueden definir las “esquinas” donde se establecen los puntos factoriales. Cuando la cuestión relacionada a la región del diseño depende sobre si los puntos axiales fuera del rango son científicamente permisibles y se pueden

incluir en la región de interés. Considere la Figura 12.15. El área sombreada forma el cubo, pero después de alguna deliberación se determina que el área no sombreada no está solamente dentro del área de operación sino que el investigador está interesado en predecir la respuesta en esta área, así como en el área sombreada. La región de interés es una esfera que está circunscrita a un cubo. Un diseño esférico (un DCC con $\alpha = \sqrt{k}$ apropiado).

Una segunda situación puede ocurrir mucho muy parecida a la Figura 12.16. Los rangos son seleccionados en las variables de diseño, pero como la planificación desarrollada del experimento determina que varios de los vértices del cubo definido por los rangos no son científicamente permisibles, esto es, están fuera de las regiones de operación y de interés (en el caso de un producto alimenticio, los niveles altos de flúor, mantequilla y tiempo de cocimiento) es conocido para producir un producto inaceptable. Como un resultado las esquinas sombreadas en la Figura 12.16 son raspadas y la región del diseño es formada de la región no sombreada. Aquí, la esfera está inscrita dentro de la región cuboidal formada de la selección de rangos.

Es importante para principiante en diseño experimental entender que en muchas situaciones la región de interés (o quizás solo la región de operación) no es clara. Es muy difícil, por lo general, disponer de un comité de científicos o ingenieros que establezcan los rangos de los factores interesantes y permisibles. Por lo general, se cometen errores y

se adoptan ajustes para futuros diseños. La confusión con respecto al tipo de diseño nunca será una excusa para no usar los diseños. Usando una región esférica cuando es naturalmente más cuboidal, proveería información importante que entre otras cosas permite la selección de regiones para experimentos futuros.

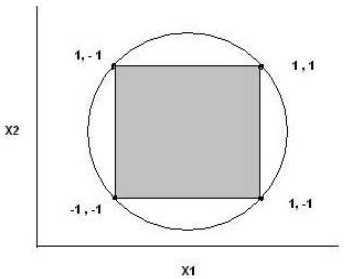


Figura 12.15. Región esférica (esfera circunscrita)

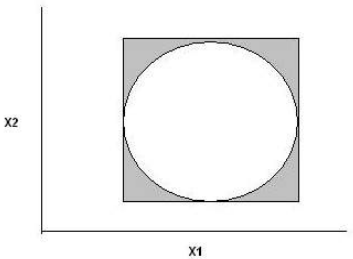


Figura 12.16. Región esférica (esfera inscrita)

La estructura básica del diseño consiste de tres secciones: a) los puntos de un diseño 2^{k-p}, con k denotando el número de factores y p usualmente pero no necesariamente igual a cero, b) los puntos axiales y c) los puntos centrales. Todas las secciones son variable en el sentido de que un DCC es presentado con la porción factorial del diseño compuesto de los puntos en el diseño factorial 2^k, se puede usar un factorial fraccional. MINITAB, por ejemplo, puede ser usado para construir un DCC completo (usando el diseño 2^k en la parte factorial) para k entre 2 y 7,

una mitad de un DCC para k entre 5 y 8 y un cuarto de un DCC para k igual a 8 o 9, considerándose fracciones de medio y cuarto del DCC, respectivamente. Específicamente, para k =5 la mitad del DCC tendría 32 puntos que consisten de los 16 puntos del factorial, 10 puntos axiales y 6 puntos centrales. Esta es una alternativa útil para el DCC completo cuando k = 5, ya que la parte factorial del DCC es de resolución V. (El DCC, tiene la misma resolución que la parte factorial, para la estimación de los efectos principales e interacciones, porque los puntos axiales y los centrales no están involucrados en la estimación de las interacciones).

Asimismo, el número de puntos factorial se determina por la selección de un factorial completo o fraccional, el número de puntos axiales siempre es dos veces el número de factores. No existe una regla formal y rápida para seleccionar el número de puntos centrales, pero algunos estudios han mostrado que el mejor diseño de superficie de respuesta tiene puntos centrales repetidos. De esta forma los puntos centrales ayudan en la investigación de curvaturas en la región alrededor del centro. Estos puntos también ayudan a estabilizar la Var ($\hat{Y}$) y alrededor del centro de la región de interés.

Un DCC en dos factores se muestra en la Figura 12.3 y se compara con otros diseños en una comparación gráfica. Los puntos del diseño son:

No.	A	B
1	-1	-1
2	1	-1
3	-1	1
4	1	1
5	- 1.414	0
6	1.414	0
7	0	-1.414
8	0	1.414
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0

Los puntos axiales son (α, 0), (- α, 0), (0, α) y (0, - α) para dos factores, los valores de α se deben seleccionar. Los valores de α para el diseño de arriba es $\sqrt{2}$, que es el valor por default en MINITAB. Note que si α = 1 y se usa el conjunto de puntos del factorial completo, el resultado es un diseño 3^2, cuando estos puntos son combinados con los puntos centrales. De otra manera, el diseño tiene 5 niveles para cada factor, más que suficientes para estimar los efectos lineales y cuadráticos, aunque las columnas de efectos cuadráticos no son ortogonales unos con otros. De esta manera, sólo los puntos del factorial, son usados para la estimación de la interacción AB, ya que solo los primeros cuatro elementos en la columna obtenida al multiplicar A x B son diferentes de cero.

El objetivo es por lo general, no crear un diseño 3^2, pero además resolver para α, como para tener un diseño rotable. El último es un

diseño para el que la Var ($\hat{Y}_X$) es la misma para todos los puntos *X* (donde *X* denota el conjunto de puntos coordenados) que son equidistantes desde el centro del diseño, al punto (0, 0) donde hay dos factores, independiente en la dirección del centro. Es bien conocido que esta condición se encuentra para $\alpha = \sqrt[4]{F}$ cuando F denota el número de puntos en la parte factorial. Entonces $(4)^{1/4} = \sqrt{2} = 1.414$ cuando k = 2 y la parte factorial contiene el factorial completo. Note que todos los puntos después de los puntos centrales son equidistantes del origen con esta selección de α.

Ya que los puntos axiales (que también son llamados puntos estrella ya que forman una configuración de estrella cuando son combinados con los otros puntos) son ortogonales para cualquier selección de α y los puntos del factorial también son ortogonales, por esto el diseño es ortogonal para la estimación de los efectos lineales de los factores, no obstante del número de puntos centrales que son usados. Los efectos cuadráticos de los factores no son ortogonales entre ellos porque no pasan, a menos que las columnas que representan los efectos cuadráticos tengan algunos números negativos, pero no puede pasar puesto que el cuadrado de un número positivo o negativo es un número positivo. Las columnas para los efectos cuadráticos son ortogonales para los efectos lineales y para los efectos de interacción.

Hablando de "efectos" que fueron usados de una manera general en los párrafos anteriores, se puede orientar la cuestión para un DCC.

Si se considera el DCC para dos factores dado anteriormente en esta sección, se observa que las observaciones en los cinco niveles de cada factor ocurren con diferente frecuencia. Esto es, se tiene solamente una observación en cada punto axial, dos observaciones en cada punto del factorial y cinco puntos centrales. Calculando los efectos condicionales y usando solamente los puntos potencialmente útil del factorial.

Myers y Montgomery (1995) hicieron la siguiente pregunta: "¿qué tan importante es la rotabilidad?" Su visión es que un diseño próximo a rotarse será suficientemente bueno y lo mismo se dirá probablemente de la ortogonalidad. Se puede notar que los estimadores de efectos cuadráticos como un grupo están menos correlacionados unos con otros en un diseño de espacio-rellenado que con un DCC, aunque esto es equivalente por el hecho de que los estimadores de efectos cuadráticos están correlacionados con los estimadores de los efectos de interacción y con los estimadores de efectos lineales y por lo tanto los últimos dos están correlacionados unos con otros.

Aunque la rotabilidad es una propiedad deseable de un diseño de superficie de respuesta, se pierde algo de vigor cuando pensamos en cómo la ecuación se ajusta al resultado de la experimentación. Una vez que los niveles del factor óptimo han sido determinados, indudablemente es interesante obtener un intervalo de predicción para la respuesta. Por lo general, no es necesario un segundo intervalo de predicción para una combinación de niveles del factor tal que el punto

es la misma distancia del centro al punto que representa las condiciones óptimas. Además si uno o más niveles del factor son alterados ligeramente, por cualquier razón, de la combinación de niveles del factor que se supone óptimo, el punto que representa la combinación de niveles es usado en la misma distancia del centro del diseño al punto óptimo. Sólo si la distancia fuera la misma, no existe garantía que las dos varianzas de los valores ajustados será la misma ya que la rotabilidad se aplica solamente a puntos usados en el diseño y no a todos los puntos en la región cubierta por el diseño que tienen la misma distancia del centro a los puntos diseño.

Se puede observar una aplicación de un DCC en la siguiente sección, se orienta la cuestión hacia la posibilidad de que un DCC genere muchos puntos de diseño. Para cuatro factores, un DCC tendrá 25 puntos si existe sólo un punto central (16 + 1 + 8), además un diseño 3^{4-1} tiene 27 puntos no muy diferentes y de esta manera los DCC tienen más de un punto central. Wu y Hamada (2000) mostraron que todos los efectos principales pueden ser estimados, además de la capacidad para estimar cada interacción de dos factores usando dos grados de libertad (en lugar de cuatro grados de libertad que tiene cada interacción). Por lo que 20 grados de libertad son usados para estimar los efectos, de los 26 que están disponibles.

Se puede notar que el número de efectos que son estimados en un modelo de segundo orden completo es $k + k + \binom{k}{2} = (k^2 + 3k)/2$, así

como el número de puntos en el DCC es $2^k + 2k + c$, si se usa un factorial completo, cuando c denota el número de puntos centrales. Esto dará suficientes grados de libertad, para estimar todos los efectos; en realidad, el diseño es algo inútil en este aspecto, por ejemplo, se tienen 20 efectos que son estimados para k = 5, además se tendrán 43 puntos de diseño, si sólo se usa un punto central, lo cual no se recomienda. Además, no necesitamos factores con 5 niveles para estimar efectos de segundo orden.

Se puede reducir considerablemente el número de puntos de diseño usando siempre un diseño 2^{k-1} o un diseño 2^{k-2} en la porción factorial.

12.6.2. Variaciones en DCC

Los diseños centrales compuestos (DCC) mencionados en la sección anterior se refieren a los diseños centrales compuestos estándar. Hay variaciones del diseño que algunas veces se usan. Por ejemplo; la región de operación se puede definir por los puntos factoriales, los otros puntos en el espacio del diseño tienen las coordenadas de $\pm \sqrt{2}$. Los puntos principales (estrellas) tendrán que ser "acortados" para conformar esta región, que es, en valor absoluto más grande de cualquier coordenada y no excede a 1. Para dos factores, esto significa que el factorial y los puntos estrella constituyen un diseño 3^2, excepto para el hecho de que

el punto (0, 0) se pierda, que se puede considerar no perdido del diseño completo ya que el diseño contiene puntos centrales. En general, este diseño se denomina un *cubo central de lado* ya que los puntos axiales están en el centro de la superficie en la región cuboidal. Se puede denominar a este diseño el cubo central de lado para diferenciarlo del diseño central compuesto estándar.

Un cubo central de lado tiene sentido si la región de interés es cuboidal; si la región de interés es esférica se puede usar el diseño central compuesto estándar. Este también se puede usar si la rotabilidad es importante ya que el cubo central de lado obviamente no es rotable puesto que los puntos estrella no tienen las coordenadas que son necesarias para la rotabilidad.

Si se desea una región esférica, entonces las coordenadas no pueden exceder a 1 en valor absoluto, una solución es "reducir la escala" del diseño, tal que los puntos estrella estén dentro de la región restringida y los puntos del factorial tengan coordenadas más pequeñas. Esto puede ser acompañado empezando con el diseño central compuesto regular y dividiendo completamente por α. Por ejemplo, para $k = 2$ los nueve puntos distintos son:

A	B
- 0. 707	- 0.707
0. 707	- 0.707
- 0. 707	0.707
0.707	0.707
1	0
- 1	0
0	1
0	- 1
0	0

Esto se llama por lo general Diseño Compuesto Central Inscrito, con la notación DCCI, usado en algunas situaciones para distinguir este diseño del diseño central compuesto y diseño central compuesto de lado. De esta manera el DCCI es rotable ya que es una versión escalada de un diseño rotable.

12.6.3. Diseños compuestos pequeños

Hay varias maneras de construir diseños compuestos que tienen menos corridas (combinaciones) que un diseño central compuesto. Estos diseños fueron propuestos por Hartley (1959) que usó un diseño de resolución III para la porción factorial, con otros métodos propuestos por Ghosh y Al –Sabah (1996) y Draper y Lin (1990), entre otros. Además de que son económicos, los diseños compuestos pequeños tienen su valor en la característica de que la correlación entre los

estimadores de los efectos cuadráticos se reduce, así como el número de puntos factoriales.

12.6.3.1. Diseños Draper – Lin

Trabajando para producir diseños compuestos con menos puntos de diseño que los DCC, que resultan de usar un factorial completo o un factorial fraccional para la porción factorial incluye a Westlake (1965) quién usó fracciones irregulares del sistema 2^k. Draper (1985) utilizó columnas de los diseños de Plackett – Burman, para producir diseños con el mismo número de factores pero con menos puntos de diseño. Draper y Lin (1990) progresaron en el grupo de diseños, encontrando diseños para valores más grandes de k y diseños que Draper (1985) consideró que no existían en una versión singular en una primera instancia. Los diseños de Draper- Lin no son rotables. Otra deficiencia es que los coeficientes de términos del mismo orden, tal como coeficientes de términos lineales no son estimados con la misma precisión. Draper y Lin (1990) no discutieron la selección de α para estos diseños pero Croarkin y Tobias (2002) indicaron que α es seleccionado entre $\sqrt[4]{F}$ y $\sqrt{k}$.

12.6.3.2. Analizando la superficie ajustada

Una vez que se ha seleccionado el diseño, el experimento ha sido conducido, los datos han sido analizados y el modelo se ha ajustado, el siguiente paso es analizar la superficie ajustada ya que la tarea principal de un estudio de superficie de respuesta es la optimización. Esto es, el investigador está interesado en determinar la combinación de factores que maximice o minimice la respuesta esperada, dependiendo de los objetivos.

La naturaleza de la superficie debe ser caracterizada para que el investigador pueda determinar; (1) si la respuesta máxima o mínima se comprende dentro de la región experimental y (2) si el máximo o mínimo se comprende dentro de la región experimental ¿cuánto cambia la respuesta para movimientos pequeños en diferentes direcciones?

En general esto no facilita la determinacion del máximo o el mínimo en una aplicación práctica ya que es necesario utilizar métodos más sofisticados. Un problema que ha sido tratado negligentemente por el método usado es que los coeficientes de los términos cuadrados en un modelo de segundo orden no son independientes ya que los términos cuadrados no son independientes. Por ejemplo considere el modelo,

$$Y = 70 + 2A + 3B - A^2 - 2B^2 - AB \qquad (12)$$

Existe obviamente un mínimo global no finito, suponer que A = 0 y B $\rightarrow \infty$, resulta que Y $\rightarrow - \infty$. Esto parece como que existe un máximo, no obstante que el máximo no es obvio. Por lo que es necesario un análisis de la superficie ajustada para determinar el o los puntos estacionarios. Con sólo dos factores se puede hacer esta determinación usando cálculo y resolviendo el sistema de ecuaciones. Esto es:

$$\frac{\partial Y}{\partial A} = 2 - 2A - B = 0$$

$$\frac{\partial Y}{\partial B} = 3 - 4B - A = 0$$

Resolviendo estas ecuaciones para A y B se tiene que A = $\frac{5}{7}$ = 0.714 y B = $\frac{4}{7}$ = 0.751. Esto produce una Y = 71.5714, que puede ser tomado como el "máximo verdadero".

Los datos que fueron generados usando el modelo anterior y añadiendo el término de error $\varepsilon \sim N(0, 0.6)$ y los valores de diseño de A y B son de un diseño central compuesto (DCC) con cinco puntos centrales y un $\alpha = \sqrt{2}$. La ecuación ajustada es:

$$\hat{Y} = 69.810 + 1.973A + 2.985B - 1.159A^2 - 1.908B^2 - 1.414AB \quad (13)$$

Con un $R^2 = 0.992$. Note que los coeficientes de A^2 y AB son estimados con un gran porcentaje de error a pesar de que R^2 es completamente alto. La correlación entre los estimadores de los términos cuadráticos es – 0.130; los estimadores de los términos de error no están correlacionados.

La superficie ajustada de la ecuación anterior está dada por la Figura 12.17. Un análisis de esta superficie ajustada produce:

$$\frac{\partial \hat{Y}}{\partial A} = 1.\ 973 - 2.\ 318A - 1.414B = 0$$

$$\frac{\partial \hat{Y}}{\partial B} = 2.985 - 3.\ 816B - 1.\ 414A = 0$$

Si se resuelven estas ecuaciones se obtiene que A = 0. 483 y B = 0. 603. Los valores de A difieren notablemente de los valores de A obtenidos en la primera ecuación. El valor de $\hat{Y}$ en este punto es de 71.19 no tan lejano del máximo verdadero de 71.57 de la primera ecuación.

De la Figura 12.17 se observa que el máximo ocurre cuando A y B están entre 0 y 1 y si se observa detenidamente la figura, se puede ver que el máximo ocurre cuando A y B están cercanos a 0.50. Se puede observar el punto (0. 483, 0. 603).

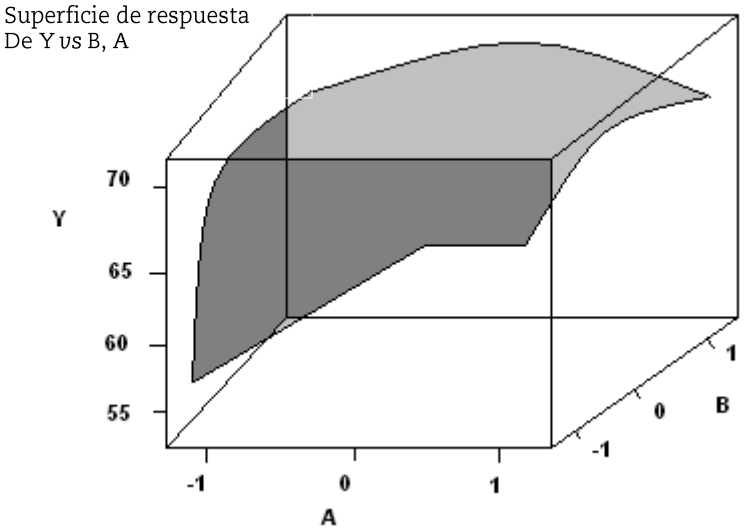


Figura 12.17. Superficie de respuesta para la ecuación $\hat{Y}$.

12.6.3.3. Caracterización de puntos estacionarios

Es fácil determinar si un punto estacionario es un máximo o un mínimo, cuando se tienen dos factores, ya que la superficie se puede graficar fácilmente; este no es el caso cuando se tienen más de dos factores. Consecuentemente se tiene la necesidad de determinar si un punto estacionario es un máximo, un mínimo o un punto silla.

Esto se observa en los eigenvalores de la matriz de coeficientes de los términos de segundo orden. Sea W que denota la matriz de eigenvalores. De la ecuación anterior se tiene:

$$\begin{bmatrix} -1.159 & -1.414 \\ -1.414 & -1.908 \end{bmatrix} \quad (14)$$

Note que los coeficientes de AB ocupan la diagonal principal. Ambos eigenvalores de la matriz son negativos. Ya que los eigenvalores son negativos, el punto debe ser un máximo. Si todos los eigenvalores son positivos, entonces el punto es un mínimo y una mezcla de signos significa que el punto es un punto silla.

Se puede ilustrar un punto silla. Considere la ecuación,

$$Y = 70 + 2A + 3B - A^2 + 3B^2 - AB \quad (15)$$

Es una pequeña modificación de la primera ecuación con el coeficiente de B^2 positivo. Esto causa que los signos de los eigenvalores sean mixtos. Usando la ecuación anterior se generan los datos que resultan en una superficie ajustada que tiene la semejanza a una silla. Por lo que generando datos con errores aleatorios que tienen la misma distribución se obtiene la ecuación ajustada,

$$\hat{Y} = 70.133 + 2.0672A + 2.988B - 0.9301A^2 + 2.9032B^2 - 0.9590AB \quad (16)$$

Con los estimadores de los parámetros que difieren ligeramente de los correspondientes parámetros.

La superficie ajustada para esta ecuación es la Figura 12.18.

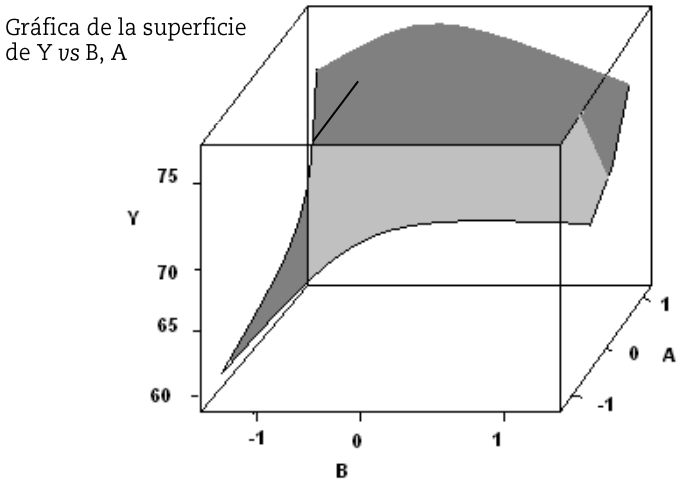


Figura 12.18 Superficie ajustada para la ecuación $\hat{Y}$.

La silla puede que no sea tan obvia en la Figura 12.18, pero se observa aparente cuando los contornos de respuesta constantes son mostrados en la Figura 12.19.

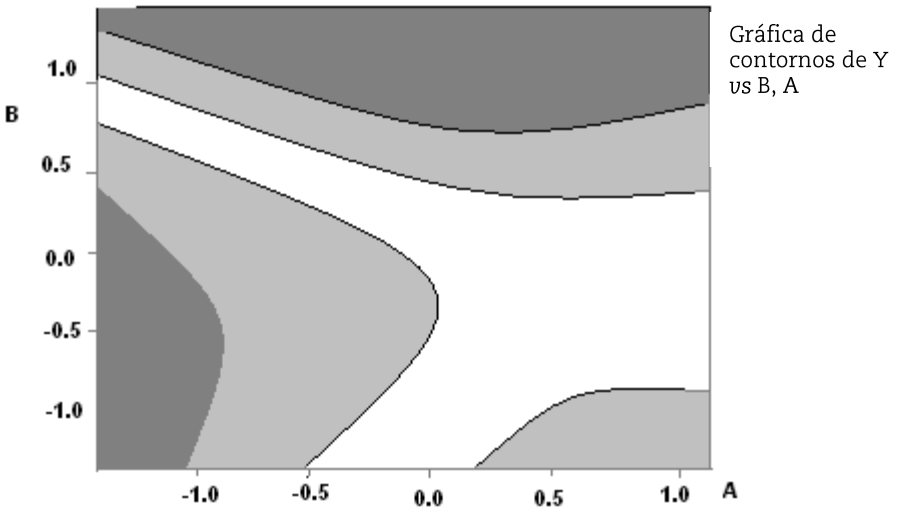


Figura 12.19. Contornos de la respuesta constante de la ecuación $\hat{Y}$.

La Figura 12.19 no es una silla perfecta, pero si se observa con atención parece una silla torcida con un lado perdido. Cabe señalar que en este caso la respuesta incrementa al iniciarse en el centro y moverse hacia la derecha. Además, la respuesta disminuye si se mueve a la izquierda.

12.6.3.4. Regiones de confianza de puntos estacionarios

Es difícil determinar niveles de factores óptimos para maximizar o minimizar la respuesta, es útil saber cómo se puede determinar bien el óptimo y cómo la respuesta es sensible a ligeros cambios en los niveles del factor lejos del óptimo. Tal como se construyen los intervalos de

confianza para los parámetros en la aplicación de procedimientos estadísticos básicos, también es deseable tener una región de confianza del punto óptimo. Desafortunadamente, no es práctico intentar calcular una región de confianza del punto óptimo sin un software apropiado.

12.6.3.5. Análisis Ridge (cordillera)

Las gráficas de contornos son completamente usadas cuando existe un pequeño número de factores, pero su uso es muy limitado cuando existe un gran número de factores. El análisis ridge propuesto por Hoerl (1959) es una técnica que puede ser usada para determinar la combinación óptima de los niveles de factores para cualquier número de ellos. El objetivo es determinar una senda hacia la región óptima usando una aproximación de optimización restringida. (Note que esto es diferente del método de la senda de pendiente ascendente/descendente). Se determina de un modelo ajustado de primer orden, mientras que al análisis ridge se aplica por lo general a modelos de segundo orden.

Específicamente, el óptimo está sujeto a las restricciones que $\sum_{i=1}^{k} X_i^2 = R^2$. El método de los multiplicadores de Langrange se utiliza para determinar que las X_i son las coordenadas del punto óptimo sujeto a las restricciones, aunque Peterson, *et al* (2002) proponen un método que no requiere el uso de los multiplicadores de Lagrange. Su método

y la metodología de Gilmour y Draper (2003) son discutidos y debatido en Peterson, *et al* (2004). Las bandas de confianza de la senda hacia la respuesta óptima fueron introducidas por Carter *et al.* (1986) con una mejor aproximación dada por Peterson (1993).

Considere la senda creciente mostrada en la Figura 12.20, ya que el punto óptimo parece que cae fuera de la región experimental; es útil conocer la senda para seguir hacia el punto óptimo con la intención de que la futura experimentación se haga a lo largo de la senda. Se puede hacer una conjetura áspera de la senda de la Figura 12.20, pero si se tienen tres factores en lugar de dos, se requieren gráficas de contornos múltiples y se tiene que combinar la información de estas gráficas.

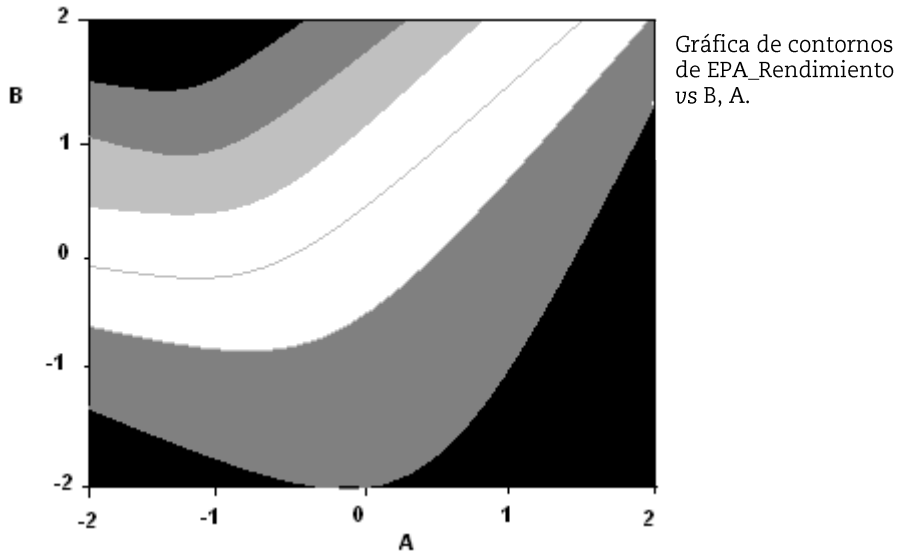


Figura 12.20. Ejemplo de una gráfica de contornos.

12.6.3.6. Análisis Ridge con factores de ruido

Para este análisis, Peterson y Kuhn (2005) generaron una aproximación para conducir un análisis ridge y optimizar una superficie de respuesta con la presencia de variables de ruido, que extendió el trabajo de Peterson (1993). Los autores mostraron cómo construir, una senda óptima para la raíz del cuadrado medio del error de un proceso de interés. Ellos repitieron el mensaje de Hoerl (1985) de que las gráficas de contornos no son suficientes para entender superficies de respuesta de altas dimensiones, especialmente cuando existen variables de ruido. Recomendaron el uso de gráficas trazadas de ridge (cordilleras) sobre fijas para indicar cómo el proceso medio, difiere del valor de interés cuando existe movimiento a lo largo de la senda ridge (cordillera).

12.6.3.7. Condiciones óptimas y regiones de operatibilidad.

Es completamente posible que los métodos para determinar condiciones de operación óptima no sean aplicados debidos a la región y de sus condiciones de operación factible que no son hiperrectangular. Cuando esto ocurre, se pueden emplear diferentes métodos. Giesbrecht y Gumpertz (2004) discutieron esto en detalle, relativo a SAS PROC OPTEX, que permite la selección de un subconjunto específico de

puntos, con los otros puntos dispersos uniformemente sobre la región factible.

12.7. Resumen

Aproximaciones clásicas y modernas de la MSR han sido presentadas en este apartado. Históricamente, la aproximación estándar ha sido usar un procedimiento de tres pasos: (1) Use un diseño de dos niveles, como un diseño parrilla, para identificar los factores importantes, (2) conducir los experimentos a lo largo de una senda de pendiente ascendente /descendente, en un esfuerzo por identificar la región de respuesta óptima y (3) una vez que la región está localizada, use un diseño de superficie de respuesta para caracterizar la naturaleza de la superficie (considerando que se ajusta a un modelo de segundo orden) para identificar los factores óptimos.

Tal aproximación se obtiene si hay interacciones significativas, como cuando el falso grupo de factores es identificado en el primer paso. Para guardar otra vez esta posibilidad, especialmente si se sospecha que hay interacciones, un diseño de resolución V de tres niveles se puede usar y para los factores parrilla ajustar un modelo de segundo orden, con el diseño proyectado a otros factores que se supone son significativos.

Ejemplo 12.4 Un proceso químico que convierte 1,2-propanidol a 2,5-dimetylpiperazina es el objeto de un experimento para determinar las condiciones óptimas (las condiciones de máxima conversión), estudiándose los siguientes factores. Para su análisis susamos el software SAS.

x_1= (cantidad de NH_3 – 102) /51

x_2 = (Temperatura – 250) /20

x_3 = (cantidad de H_2O – 300) /200

x_4 = (presión de hidrógeno – 850) /350

```
data proceso;
input nitro temp agua presi conver;
cards;

-1    -1    -1    -1    58.2
1     -1    -1    -1    23.4
-1    1     -1    -1    21.9
1     1     -1    -1    21.8
-1    -1    1     -1    14.3
1     -1    1     -1    6.3
-1    1     1     -1    4.5
1     1     1     -1    21.8
-1    -1    -1    1     46.7
1     -1    -1    1     53.2
-1    1     -1    1     23.7
1     1     -1    1     40.3
```

```
-1      -1      1       1       7.5
1       -1      1       1       13.3
-1      1       1       1       49.3
1       1       1       1       20.1
0       0       0       0       32.8
-1.4    0       0       0       31.1
1.4     0       0       0       28.1
0       -1.4    0       0       17.5
0       1.4     0       0       49.7
0       0       -1.4    0       49.9
0       0       1.4     0       34.2
0       0       0       -1.4    31.1
0       0       0       1.4     43.1

run;
proc sort;
by nitro temp agua presi;
run;
proc rsreg data=proceso out=rproceso;
model conver = nitro temp agua presi/lackfit residual;
ridge max;
run;
```

The RSREG Procedure

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
Nitro	0	1.400000
Temp	0	1.400000
Agua	0	1.400000
Presi	0	1.400000

Response Surface for Variable conver

Response Mean	29.752000
Root MSE	14.067809
R-Square	0.6622
Coefficient of Variation	47.2836

Type I

Regression	DF	Sum of Squares	R-Square	F Value	Pr > F
Linear	4	2088.648735	0.3566	2.64	0.0972
Quadratic	4	519.802536	0.0887	0.66	0.6357
Crossproduct	6	1270.318750	0.2169	1.07	0.4398
Total Model	14	3878.770021	0.6622	1.40	0.3004

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	10	1979.032379	197.903238	.	.
Pure Error	0	0	.		
Total Error	10	1979.032379	197.903238		

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	40.198215	8.321708	4.83	0.0007	40.198215
Nitro	1	-1.511044	3.151968	-0.48	0.6420	-2.115462
temp	1	1.284137	3.151968	0.41	0.6923	1.797791
agua	1	-8.738956	3.151968	-2.77	0.0197	-12.234538
presi	1	4.954819	3.151968	1.57	0.1470	6.936747
nitro*nitro	1	-6.332399	5.035510	-1.26	0.2371	-12.411502
temp*nitro	1	2.193750	3.516952	0.62	0.5467	4.299750
temp*temp	1	-4.291583	5.035510	-0.85	0.4140	-8.411502
agua*nitro	1	-0.143750	3.516952	-0.04	0.9682	-0.281750
agua*temp	1	8.006250	3.516952	2.28	0.0461	15.692250
agua*agua	1	0.019642	5.035510	0.00	0.9970	0.038498
presi*nitro	1	1.581250	3.516952	0.45	0.6626	3.099250
presi*temp	1	2.806250	3.516952	0.80	0.4435	5.500250
presi*agua	1	0.293750	3.516952	0.08	0.9351	0.575750
presi*presi	1	-2.505869	5.035510	-0.50	0.6295	-4.911502

La función ajustada de segundo orden es,

$$\hat{y} = 40.198 - 1.511x_1 + 1.284x_2 - 8.739x_3 + 4.955x_4 - 6.332x_1^2 + 2.194x_1x_2 - 4.292x_2^2 - 0.144x_1x_3 + 8.006x_2x_3 + 0.0196x_3^2 + 1.581x_1x_4 + 2.806x_2x_4 + 0.\ 294x_3x_4 - 2.506x_4^2$$

Entonces la respuesta en el punto estacionario en $\hat{y}_s = 43.52$

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
Nitro	5	475.789167	95.157833	0.48	0.7832
Temp	5	1405.197804	281.039561	1.42	0.2973
Agua	5	2548.592316	509.718463	2.58	0.0952
Presi	5	705.437292	141.087458	0.71	0.6278

The RSREG Procedure

Canonical Analysis of Response Surface Based on Coded Data

Critical Value

Factor	Coded	Uncoded
Nitro	0.189062	0.264687
Temp	0.738318	1.033646
Agua	0.207556	0.290578
Presi	1.191401	1.667961

Predicted value at stationary point: 43.524455

	Eigenvectors			
Eigenvalues	nitro	temp	agua	presi
5.103843	0.074140	0.528240	0.826418	0.180275
-4.232252	0.215098	0.137387	-0.307108	0.916811
-11.776317	0.768774	0.456751	-0.285763	-0.344535
-14.791284	0.597681	-0.702471	0.375577	0.090850

Stationary point is a saddle point.

Esto indica que un análisis ridge revela candidatos razonables para las condiciones de operación con sus restricciones de que pueden caer dentro o en los límites del diseño experimental. En este caso los puntos factoriales están a una distancia de dos unidades del centro del diseño. Las coordenadas para los candidatos razonables son:

$$x_1 = -0.1308 \; x_2 = -0.7861 \; x_3 = -1.8281 \text{ y } x_4 = 0.1514$$

con un radio de dos unidades. La respuesta estimada está dada por $\hat{y}_s = 64.61$ y un error estándar en este punto de $S_{\hat{y}} = 16.543$. Notar que el error estándar empieza a crecer después de un radio de 1.4. Recuerde que la distancia axial en el diseño es de 1.4 y los puntos factoriales están a una distancia de 2.0. En realidad en este caso uno se siente obligado a recomendar condiciones en radios de 1.4 o 1.5 donde la respuesta predicha es más pequeña pero el error estándar es considerablemente más pequeño.

Ejemplo 12.5.Considerar el siguiente ejemplo con tres variables de diseño: el tiempo de reacción, temperatura y el porcentaje de catálisis. Las variables respuesta son; el porcentaje de conversión y la actividad termal. El programa de análisis usando SAS es:

data proceso;

input tie tem catal cover acti;

cards;

-1	-1	-1	74	53.2
1	-1	-1	51	62.9
-1	1	-1	88	53.4
1	1	-1	70	62.6
-1	-1	1	71	57.3
1	-1	1	90	67.9
-1	1	1	66	59.8
1	1	1	97	67.8
-1.682	0	0	76	59.1
1.682	0	0	79	65.9
0	-1.682	0	85	60
0	1.682	0	97	60.7
0	0	-1.68	2 55	57.4
0	0	1.68	281	63.2
0	0	0	81	59.2
0	0	0	75	60.4
0	0	0	76	59.1
0	0	0	83	60.6
0	0	0	80	60.8
0	0	0	91	58.9

```
run;
proc sort;
by tie tem catal cover acti;
run;
proc rsreg data=proceso out=rproceso;
model cover = tie tem catal/lackfit residual;
ridge max;
run;
proc rsreg data=proceso out=super1;
model cover= tie tem catal/actual lackfit residual predict press;
ridge radius=0 to 20 by 1 maximun minimum;
run;
proc rsreg data=proceso out=rproceso;
model acti = tie tem catal/lackfit residual;
ridge max;
run;
proc rsreg data=proceso out=super1;
model acti= tie tem catal/actual lackfit residual predict press;
ridge radius=0 to 20 by 1 maximun minimum;
run;
```

Los resultados del análisis son:

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
Tie	0	1.682000
Tem	0	1.682000
Catal	0	1.682000

Response Surface for Variable cover

Response Mean	78.300000
Root MSE	4.716774
R-Square	0.9199
Coefficient of Variation	6.0240

Type I

Regression	DF	Sum of Squares	R-Square	F Value	Pr > F
Linear	3	763.060957	0.2747	11.43	0.0014
Quadratic	3	601.284484	0.2164	9.01	0.0034
Crossproduct	3	1191.375000	0.4288	17.85	0.0002
Total Model	9	2555.720441	0.9199	12.76	0.0002

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	5	56.479559	11.295912	0.34	0.8691
Pure Error	5	166.000000	33.200000		
Total Error	10	222.479559	22.247956		

Parameter	DF	Standard Estimate	Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	81.090931	1.923744	42.15	<.0001	81.090931
Tie	1	1.028390	1.276285	0.81	0.4391	1.729751
tem	1	4.040343	1.276285	3.17	0.0101	6.795856
catal	1	6.203724	1.276285	4.86	0.0007	10.434664
tie*tie	1	-1.833557	1.242263	-1.48	0.1707	-5.187360
tem*tie	1	2.125000	1.667631	1.27	0.2314	6.011889
tem*tem	1	2.938238	1.242263	2.37	0.0396	8.312640
catal*tie	1	11.375000	1.667631	6.82	<.0001	32.181286
catal*tem	1	-3.875000	1.667631	-2.32	0.0425	-10.962856
catal*catal	1	-5.191487	1.242263	-4.18	0.0019	-14.687360

La función de segundo orden ajustado a los datos de conversión es:

$$\hat{y} = 81.09 + 1.0284x_1 + 4.043x_2 + 6.2037x_3 - 1.8336x_1^2 + 2.9382x_2^2 - 5.1915x_3^2 + 2.1250x_1x_2 + 11.375x_1x_3 - 3.8750x_2x_3$$

The RSREG Procedure

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
Tie	4	1134.162370	283.540593	12.74	0.0006
Tem	4	503.674234	125.918559	5.66	0.0121
Catal	4	2069.453363	517.363341	23.25	<.0001

Canonical Analysis of Response Surface Based on Coded Data

	Critical Value	
Factor	Coded	Uncoded
Tie	-0.604961	-1.017545
Tem	-0.315222	-0.530204
Catal	-0.189892	-0.319399

Predicted value at stationary point: 78.505882

	Eigenvectors		
Eigenvalues	tie	tem	catal
9.639512	-0.139545	0.942203	-0.304600
6.575268	0.797732	0.289209	0.529133
-27.776860	-0.586643	0.169152	0.791983

Stationary point is a saddle point.

Estimated Ridge of Maximum Response for Variable cover

Coded	Estimated	Standard	Uncoded Factor Values		
Radius	Response	Error	tie	tem	catal
0.0	81.090931	1.923744	0	0	0
0.1	82.303904	1.917452	0.056971	0.100639	0.122136
0.2	83.538483	1.902353	0.149510	0.211124	0.215030
0.3	84.871185	1.890037	0.250711	0.327487	0.290719
0.4	86.327493	1.900168	0.351821	0.450441	0.354943
0.5	87.918563	1.959061	0.449282	0.581060	0.409629
0.6	89.651105	2.094639	0.540983	0.720151	0.455198
0.7	91.530323	2.328625	0.625363	0.868107	0.491510
0.8	93.560935	2.671250	0.701228	1.024819	0.518327
0.9	95.747503	3.122355	0.767789	1.189638	0.535584
1.0	98.094485	3.676284	0.824724	1.361443	0.543531

The RSREG Procedure

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
Tie	0	1.682000
Tem	0	1.682000
Catal	0	1.682000

Response Surface for Variable cover

Response Mean	78.300000
Root MSE	4.716774
R-Square	0.9199
Coefficient of Variation	6.0240
Sum of Squared Residuals	222.47955866
Predicted Residual SS (PRESS)	676.28123633

Type I

Regression	DF	Sum of Squares	R-Square	F Value	Pr > F
Linear	3	763.060957	0.2747	11.43	0.0014
Quadratic	3	601.284484	0.2164	9.01	0.0034
Crossproduct	3	1191.375000	0.4288	17.85	0.0002
Total Model	9	2555.720441	0.9199	12.76	0.0002

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	5	56.479559	11.295912	0.34	0.8691
Pure Error	5	166.000000	33.200000		
Total Error	10	222.479559	22.247956		

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	81.090931	1.923744	42.15	<.0001	81.090931
Tie	1	1.028390	1.276285	0.81	0.4391	1.729751
Tem	1	4.040343	1.276285	3.17	0.0101	6.795856
Cata	1	6.203724	1.276285	4.86	0.0007	10.434664
tie*tie	1	-1.833557	1.242263	-1.48	0.1707	-5.187360
tem*tie	1	2.125000	1.667631	1.27	0.2314	6.011889
tem*tem	1	2.938238	1.24226	32.37	0.0396	8.312640
catal*tie	1	11.375000	1.667631	6.82	<.0001	32.181286
catal*tem	1	-3.875000	1.667631	-2.32	0.0425	-10.962856
catal*catal	1	-5.191487	1.242263	-4.18	0.0019	-14.687360

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
Tie	4	1134.162370	283.540593	12.74	0.0006
Tem	4	503.674234	125.918559	5.66	0.0121
Catal	4	2069.453363	517.363341	23.25	<.0001

Canonical Analysis of Response Surface Based on Coded Data

Critical Value

Factor	Coded	Uncoded
Tie	-0.604961	-1.017545
Tem	-0.315222	-0.530204
Catal	-0.189892	-0.319399

Predicted value at stationary point: 78.505882

Eigenvectors

Eigenvalues	tie	tem	catal
9.639512	-0.139545	0.942203	-0.304600
6.575268	0.797732	0.289209	0.529133
-27.776860	-0.586643	0.169152	0.791983

Stationary point is a saddle point.

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
Tie	0	1.682000
Tem	0	1.682000
Catal	0	1.682000

Response Surface for Variable acti

Response Mean	60.510000
Root MSE	1.763165
R-Square	0.8918
Coefficient of Variation	2.9138

Type I

Regression	DF	Sum of Squares	R-Square	F Value	Pr > F
Linear	3	244.139993	0.8498	26.18	<.0001
Quadratic	3	10.056754	0.0350	1.08	0.4019
Crossproduct	3	1.993750	0.0069	0.21	0.8846
Total Model	9	256.190497	0.8918	9.16	0.0009

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	5	27.434169	5.486834	7.51	0.0226
Pure Error	5	3.653333	0.730667		
Total Error	10	31.087503	3.108750		

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	59.849766	0.719110	83.23	<.0001	59.849766
Tie	1	3.583007	0.477085	7.51	<.0001	6.026618
Tem	1	0.254601	0.477085	0.53	0.6052	0.428238
Catal	1	2.229832	0.477085	4.67	0.0009	3.750578
tie*tie	1	0.834790	0.464367	1.80	0.1024	2.361724
tem*tie	1	-0.387500	0.623373	-0.62	0.5481	-1.096286
tem*tem	1	0.074837	0.464367	0.16	0.8752	0.211724
catal*tie	1	-0.037500	0.623373	-0.06	0.9532	-0.106092
catal*tem	1	0.312500	0.623373	0.50	0.6270	0.884101
catal*catal	1	0.057164	0.464367	0.12	0.9045	0.161724

La función de respuesta de segundo orden para la variable actividad está dada por:

$$\hat{y} = 59.85 + 3.5830x_1 + 0.25460x_2 + 2.2298x_3 + 0.83479x_1^2 + 0.07484x_2^2 + 0.05716x_3^2 - 0.38750x_1x_2 - 0.3750x_1x_3 + 0.3125x_2x_3$$

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
Tie	4	186.602821	46.650705	15.01	0.0003
Tem	4	2.948590	0.737148	0.24	0.9110
Catal	4	68.750484	17.187621	5.53	0.0130

Canonical Analysis of Response Surface Based on Coded Data

Critical Value

Factor	Coded	Uncoded
tie	-2.239518	-3.766869
tem	-4.022702	-6.766185
catal	-1.334673	-2.244920

Predicted value at stationary point: 49.737172

Eigenvectors

Eigenvalues	tie	tem	catal
2.503967	0.967278	-0.244425	-0.068037
0.533638	0.214345	0.643762	0.734593
-0.302433	0.135754	0.725139	-0.675088

Stationary point is a saddle point

Estimated Ridge of Maximum Response for Variable acti.

Coded Radiu	Estimated Response	Standard Error	Uncoded Factor Values tie	tem	catal
0.0	59.849766	0.719110	0	0	0
0.1	60.577850	0.716758	0.144944	0.008588	0.084905
0.2	61.340962	0.711113	0.294283	0.014009	0.162376
0.3	62.140694	0.706508	0.447419	0.016216	0.232755
0.4	62.978506	0.710292	0.603774	0.015254	0.296453
0.5	63.855729	0.732300	0.762814	0.011239	0.353934
0.6	64.773554	0.782970	0.924061	0.004332	0.405681
0.7	65.733047	0.870421	1.087095	-0.005277	0.452181
0.8	66.735148	0.998480	1.251557	-0.017385	0.493905
0.9	67.780688	1.167091	1.417145	-0.031787	0.531300
1.0	68.870395	1.374138	1.583607	-0.048283	0.564784

Coding Coefficients for the Independent Variables

Factor	Subtracted off	Divided by
tie	0	1.682000
tem	0	1.682000
catal	0	1.682000

Response Surface for Variable acti

Response Mean	60.510000
Root MSE	1.763165
R-Square	0.8918
Coefficient of Variation	2.9138
Sum of Squared Residuals	31.087502811
Predicted Residual SS (PRESS)	214.50054499

Type I

Regression	DF	Sum of Squares	R-Square	F Value	Pr > F
Linear	3	244.139993	0.8498	26.18	<.0001
Quadratic	3	10.056754	0.0350	1.08	0.4019
Crossproduct	3	1.993750	0.0069	0.21	0.8846
Total Model	9	256.190497	0.8918	9.16	0.0009

Residual	DF	Sum of Squares	Mean Square	F Value	Pr > F
Lack of Fit	5	27.434169	5.486834	7.51	0.0226
Pure Error	5	3.653333	0.730667		
Total Error	10	31.087503	3.10875		

Parameter	DF	Estimate	Standard Error	t Value	Pr > |t|	Parameter Estimate from Coded Data
Intercept	1	59.849766	0.719110	.23	<.0001	59.849766
tie	1	3.583007	0.477085	7.51	<.0001	6.026618
tem	1	0.254601	0.477085	0.53	0.6052	0.428238
catal	1	2.229832	0.477085	4.67	0.0009	3.750578
tie*tie	1	0.834790	0.464367	1.80	0.1024	2.361724
tem*tie	1	-0.387500	0.623373	-0.62	0.5481	-1.096286
tem*tem	1	0.074837	0.464367	0.16	0.8752	0.211724
catal*tie	1	-0.037500	0.623373	-0.06	0.9532	-0.106092
catal*tem	1	0.312500	0.623373	0.50	0.6270	0.884101
catal*catal	1	0.057164	0.464367	0.12	0.9045	0.161724

Factor	DF	Sum of Squares	Mean Square	F Value	Pr > F
tie	4	186.602821	46.650705	15.01	0.0003
tem	4	2.948590	0.737148	0.24	0.9110
catal	4	68.750484	17.187621	5.53	0.0130

Canonical Analysis of Response Surface Based on Coded Data

Critical Value

Factor	Coded	Uncoded
tie	-2.239518	-3.766869
tem	-4.022702	-6.766185
catal	-1.334673	-2.244920

Predicted value at stationary point: 49.737172

Eigenvectors

Eigenvalues	tie	tem	catal
2.503967	0.967278	-0.244425	-0.068037
0.533638	0.214345	0.643762	0.734593
-0.302433	0.135754	0.725139	-0.675088

Stationary point is a saddle point.

The RSREG Procedure

Estimated Ridge of Minimum Response for Variable acti

Coded	Estimated	Standard	Uncoded Factor Values		
Radius	Response	Error	tie	tem	catal
0	59.849766	0.719110	0	0	0
1.000000	54.098618	1.374222	-1.109921	-0.162139	-1.253360
2.000000	49.976093	5.168808	-1.569762	0.081801	-2.974164
3.000000	46.276617	11.710653	-1.597487	0.997132	-4.681440
4.000000	42.361802	20.898136	-1.462415	2.172146	-6.197508
5.000000	38.001146	32.718620	-1.276700	3.404441	-7.583398
6.000000	33.113897	47.168983	-1.071868	4.647460	-8.893856
7.000000	27.666142	64.248142	-0.858212	5.890554	-10.158342
8.000000	21.641204	83.955641	-0.639850	7.131296	-11.392943
9.000000	15.029895	106.291260	-0.418709	8.369367	-12.607038
10.000000	7.826718	131.254879	-0.195798	9.605004	-13.806446
11.000000	0.028174	158.846435	0.028305	10.838555	-14.994964
12.000000	-8.368076	189.065884	0.253249	12.070351	-16.175177
13.000000	-17.363656	221.913202	0.478806	13.300678	-17.348910
14.000000	-26.959730	257.388368	0.704825	14.529772	-18.517485
15.000000	-37.157158	295.491379	0.931201	15.757828	-19.681886
16.000000	-47.956585	336.222221	1.157856	16.985001	-20.842861
17.000000	-59.358506	379.580888	1.384735	18.211422	-22.000988
18.000000	-71.363310	425.567376	1.611796	19.437195	-23.156722
19.000000	-83.971304	474.181683	1.839008	20.662408	-24.310424
20.000000	-97.182733	525.423805	2.066344	21.887133	-25.462388

12.8 Ejercicios

12.1 Se realizó un experimento en donde se usó un factorial 3^2 y los resultados son los siguientes.

A	B	Y
-1	-1	10
-1	0	20
-1	1	30
0	-1	20
0	0	20
0	1	20
1	-1	30
1	0	20
1	1	10

a) Construya e interprete la gráfica de la superficie.
b) ¿Sugiere la gráfica que algo está sesgado? Explique.

12.2 Usando un software apropiado (MINITAB o JMP) construya un diseño uniforme para cuatro factores en 16 corridas (combinaciones) usando (-2, 2) el rango de los factores. Compare estos resultados con los correspondientes DCC (rotables) de 16 corridas para cuatro factores. Comente.

12.3 Considere que un investigador usa un DCC para tres factores, utilizando un diseño 2^3 para la parte factorial y cinco puntos centrales.

Los punto centrales fueron usados para estimar σ, que fue mucho más pequeño que el estimador obtenido de todos los grados de libertad obtenido para obtener el estimador. Explique una posible razón de esta discrepancia.

12.4 En su artículo, Allen *et al.* (2000) describieron un escenario en el cual los ingenieros decidieron usar un método de superficie de respuesta para seleccionar los parámetros del diseño, en la ausencia de "modelos de ingeniería adecuados". Un DCC fue esquematizado ya que requerían 25 corridas y la administración estuvo dispuesta a garantizar recursos para desarrollar 12 corridas experimentales. En su lugar fue usado un diseño de bajo costo con 14 corridas y sus características son:

Corrida	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	-.5	1	-1	1	0.0	0.0	-.5	-1	1	-1	0.0	.5	.5	.5
B	-1	1	1	-1	0	1	-1	0	1	1	0	-.5	-.5	-.5
C	-.5	-1	1	-.5	0.0	0.0	-.5	0.0	-1	-1	-1	.5	.5	.5
D	1	1	1	-.5	0.0	0.0	-.5	0.0	-1	-1	-1	.5	.5	.5

a) ¿Es este un número suficiente de puntos del diseño para ajustar un modelo de segundo orden con cuatro factores? Explique.

b) ¿Cuáles son las propiedades del diseño? Específicamente, ¿hay algunos efectos que pueden ser estimados ortogonalmente? ¿Hay algunos efectos para los cuales la correlación entre los efectos estimados es indeseable? Explicar.

c) ¿Recomendaría usted que este diseño sea usado? Explicar.

12.5 En el artículo citado en el ejercicio anterior el autor también dio los resultados de un estudio con 12 corridas experimentales con el mismo diseño anterior. Los resultados fueron:

Hileras	1	2	3	4	5	6	7	8	9	10	11	12
A	1.25	2.0	1.0	2.0	1.5	1.5	1.25	1.0	2.0	1.0	1.5	1.75
B	1.7	2.1	2.1	1.7	1.9	2.1	1.7	1.9	2.1	2.1	1.9	1.8
C	12.5	10.0	20.0	12.5	10.0	15.0	20.0	15.0	20.0	10.0	15.0	17.5
D	10.0	10.0	10.0	6.25	7.50	7.50	6.25	7.50	5.0	5.0	5.0	8.75
Y	55.95	101.76	101.23	52.93	59.93	80.54	60.87	72.02	102.7	51.36	59.42	81.94

a). Note que los valores de diseño en las unidades originales no están codificadas. Convertirlas a unidades codificadas.

b) ¿Puede un modelo de segundo orden completo ser ajustado a estos datos? Si no, ¿cómo podemos procesarlos para determinar el estimador de los efectos?

c) Los autores ajustaron un modelo con 10 términos ¿cómo determinarías los efectos significativos? ¿Usarías un ANVA u otra aproximación?

d) Dos de los términos del modelo ajustado por los autores fueron A y A^2A. Calcular la correlación entre los términos establecidos para A. Ahora restar la media de A de cada valor de A y calcular la correlación entre los valores medios centrados y los cuadrados de estos valores. Comentar ¿Sería recomendado hacer esto?

e) ¿Recomendaría que un diseño diferente sea usado, que fuera más económico que este? Explicar.

12.6 Enliste los puntos del diseño para un DCC inscrito para cuatro factores ¿Es rotable el diseño? Explicar. Si no, ¿el diseño se podría hacer rotable? Explique.

12.7 ¿Cuál sería la motivación para hacer un diseño CCI en lugar de un DCC estándar?

12.8 Explique porqué una gráfica de superficie construida de un DCC con cuatro factores por lo general no representa la superficie verdadera.

12.9 Asuma que k = 3 y se construye una región de confianza sobre un punto que aparentemente es un óptimo. Explicar porque es llamada una región en lugar de un intervalo ¿Qué influenciaría el tamaño de la región?

12.10 Si estuvieras determinando los factores significativos de un gran grupo para determinar la combinación óptima de niveles de estos factores, para maximizar la respuesta, ¿Iniciaría con un diseño de tres niveles o con un diseño de dos niveles? Explicar.

12.11 Explique porqué el método de pendiente ascendente no se puede aplicar a modelos ajustados que contienen un término de interacción.

12.12 Asuma que un diseño Draper-Lin para cuatro factores es usado en un estudio y los valores de las respuestas son: 10.2, 10.7, 11.5, 12.6, 13.4, 12.2, 15.1, 10.9, 11.9, 11.3, 12.2, 14.1, 15.8, 13.6, 13.9, 15.0. Ajuste el modelo de segundo orden y compare los errores estándares de los estimadores ¿Es la diferencia en precisión de alguna consecuencia?

12.13 Considere un DCC completo para 5 factores con 48 puntos: 32 puntos del factorial, 10 puntos axiales y 6 puntos centrales. Calcular la correlación entre las columnas de efectos cuadráticos, hacer lo mismo para una mitad del DCC para el mismo número de factores con 32 puntos: 16 del factorial, 10 puntos axiales y 6 puntos centrales. Compare los dos conjuntos de coeficientes de correlación y comente.

12.14 Considere un experimento con 5 factores.

a) Obtenga una ½ fracción apropiada del factorial 2^5 para ajustar un modelo de primer orden.

b) Para el diseño seleccionado en el inciso anterior bosqueje la tabla del ANVA asumiendo que cada punto del diseño se repite dos veces.

c) Suponga que es necesario correr el experimento en bloques de tamaño 8. Escriba un plan apropiado y defina el modelo lineal asociado y bosqueje la tabla del ANVA.

12.15 Considere un diseño simple con 4 factores.

a) Defina explícitamente la matriz D del modelo del diseño.

b) Bosqueje la tabla del ANVA con r =2 repeticiones para cada punto del diseño.

12.16 Considere un DCC para un experimento con 5 factores. Muestre que con una ½ fracción de resolución V del diseño 2^5 como la parte factorial se pueden estimar todos los efectos lineales, cuadrático y lineal x lineal.

12.17 Los siguientes datos fueron recolectados para examinar una superficie de respuesta:

Observ.	1	2	3	4	5	6	7	8	9	10	11
X_1	-1	1	-1	1	-2	2	0	0	0	0	0
X_2	-1	-1	1	1	0	0	-2	2	0	0	0
Y	52	45	48	55	47	52	36	44	49	54	42

a) Ajuste un modelo de segundo orden a los datos y establezca el punto estacionario.

b) Determine los eigen valores de B y describa la superficie.

c) Use el método del análisis ridge (cordillera) para definir la respuesta y su localización como una función del radio de un circulo cerca del origen.

Bibliografía

Acton, F. S. (1959). *Analysis of straight –line data*. Nueva York. Dover.

Aitkin, M. A. (1974). *Simultaneous inference and the choice of variable subsets in multiple linear regression*. Technometrics. 16: 221 -227.

Aktar, M. y Prescott (1986). *Response surface designs robust to missing observations. Communications in Statistics: Simulation and Computation* 15: 354 -363.

Allen, D. M. (1971). *The prediction sum of squares as a criterion for selecting predictor variables*. Reporte Técnico No. 23. Departamento de Estadística, Universidad de Kentucky.

Allen, D. M. (1971). *Mean square error of prediction as a criterion for selecting variables*. Technometrics 13: 469 – 475.

Allen, D. M. (1974). *The relationship between variables selection and data augmentation and a method for prediction*. Technometrics 16: 125 -127.

Allen, T., L. Yu, y Bernshteyn (2000). *Low-cost response surface methods applied to the design of plastic fasteners*. Quality Engineering 12(4): 583 -591.

Anderson, M. J. y P. J. Whitcomb (2004). *RSM Simplified: Optimizing processes using response surface methods for design of experiments*. University Park. IL: Productivity Press.

Andrews, D. F. (1971). *A note on the selection of data transformation*. Biometrika 58: 249 -254.

Ankenman, B. E., H. Liu, A. F. Karr, y J. D. Picka (2002). *A class of experimental designs for estimating a response surface and variance components*. Technometrics 44(1): 45-54.

Armitage, P. (1971). *Statistical Methods in Medical Research*. Oxford: Blackwell Scientific Publications. U.K.

Atkinson, A. C. (1985). *Plots, transformations and regression*. Oxford: Clarendon Press.

Barnett, V. D. (1970). *Fitting straight lines – the linear functional relationship with replicated observations*. Applied Statistics 19: 135 – 144.

Bartlett, M. S. (1947). *The use of transformations*. Biometrics 3: 39 -52.

Bartlett, M. S. (1949). *Fitting a straight line when both variables are subject to error*. Biometrics 5: 164 – 180.

Beale, E. M. L. (1970). *Note on procedures for variable selection in multiple regression*. Technometrics 12: 909 – 914.

Beale, E. M. L., y P. C. Hutchinson (1974). *Note on constrained optimum regression*. Applied Statistics 23: 208 -210.

Bendel, R. B., y A. A. Afifi. (1977). *Comparison of stopping rules in forward "stepwise" regression*. Journal of the American Statistical Association 72: 46 – 53.

Benignus, V. A., Muller, K. E., Barton, C. N. y Bittekofer, J. A. (1981). *Toluene levels in blood and brain of rats ans after respiratory exposure*. Toxicology an Application. Pharmacology 61: 326 -334.

Berk, K. N. (1978). *Comparing subset regression procedures*. Technometrics 20: 1-6.

Berenblut, I. I., y G. I. Webb (1973). *A new test for autocorrelation errors in the linear regression model*. Journal of the Statistical Society, Serie B 35: 33 – 50.

Berkson, J. (1950). *Are there two regressions*? Journal of the American Statistical Association 45: 164 – 180.

Bisgaard, S. (1997). *Why three-level designs are not so useful for technological applications*. Quality Engineering 9(3): 545-550.

Blalock, H. M., Jr., ed. (1971). *Causal models in the Social Sciences*. Chicago: Aldine Publishing Company.

Bliss, C. I. (1936). *The Size Factor in Action of Arsenic upon Silkworms Larvae*. Journal Experimental Bilboar. 13: 95 -110.

Bowden, D. C. (1970). *Simultaneous confidence bands for linear regression models*. Journal of the American Statistical Association 65: 413 – 421.

Box, G. E. P. (1966). *Use and abuse of regression*. Technometrics 8: 625 -629.

Box, G. E. P. (1957). *Evolutionary operation: A method dor increasing industrial productivity*. Applied Statistics 6(2): 81 – 101.

Box, G. E. P. y P. W. Tidwell (1962). *Transformations of the independent variables*. Technometrics 4: 531- 550.

Box, G. E. P. (1999 – 2000). *The invention of the composite design*. Quality Engineering 12(1): 119 – 122.

Box, G. E. P. y D. R. Cox (1964). *An analysis of transformations (with discussion)*. Journal of the Royal Statistical Society, Series B 26: 211 – 246.

Box, G. E. P. y D. W. Behnken (1960). *Some new three-level designs for ths study of quantitative variables*. Technometrics 2: 455 – 475.

Box, G. E. P. y N. R. Draper (1969). *Evolutionary Operation*. New York: Wiley.

Box, G. E. P. y N. R. Draper (1975). *Robust designs*. Biometricka 62: 347 – 352.

Box, G. E. P. y N. R. Draper (1982). *Measures of lack of fit for response surface designs and predictor variable transformations*. Technometrics 24: 1 – 8.

Box, G. E. P. y N. R. Draper (1987). *Empirical model building and response surfaces*. New York: Wiley.

Box, G. E. P. y J. S. Hunter (1957). *Multifactor experimental designs for exploring response surfaces*. Annals of Mathematical Statistics 28: 195 -241.

Box, G. E. P. y K. B. Wilson (1951). *On the experimental attainment of optimum conditions*. Journal of the Royal Statistical Society, Series B 13: 1 – 45.

Brown, P. J. (1977). *Centering and scaling in ridge regression*. Technometrics 19: 35 – 36.

Bursztyn, D. y D. M. Steinberg (2001). *Rotation designs for experiments in high bias situations*. Journal of Statistical Planning and Inference 97: 399 – 346.

Carter, W. H., V. M. Chinchilli, R. H Myers, y E. D. Campbell (1986). *Confidence intervals and an improved ridge analysis of response surfaces*. Technometrics 28: 339 -346.

Carroll, R. J., y D. Ruppert (1985). *Transformations in regression: A robust analysis*. Technometrics 27: 1 – 12.

Carroll, R. J., y D. Ruppert (1984). *Power transformations when fitting theoretical models to data*. Journal American Statistics Association. 79: 321 – 328.

Chapman, R. E. y K. Masinda (2003). *Response surface designed experiment for door closing effort*. Quality Engineering 15(4): 581 - 585.

Chatterjee, S. y A. S. Hadi (1988). *Sensitivity Analysis in linear regression*. New York.

Cheng, S. W. y C. F. J. Wu (2001). *Factor screening and response surface exploration*. Statistica Sinica 11: 553 -580.

Conniffe, D. y J. Stone (1973). *A critical view of ridge regression*. Statistic 22: 181 - 187.

Cox, D. R. (1960). *Regression analysis when there is priori information about supplementary variables*. Journal of the Royal Statistical Society Serie B 22: 172 -176.

Cox, D. R. (1970). *The Analysis of Binary Data*. Londres: Methuen and Co. Ltd.

Cox, D. R. y N. Reid (2000). *The Theory of the Design of Experiments*. Boca Raton, FL. CRC Press.

Croarkin, C. y P. Tobias, eds (2002). NIST/SEMATECH e-Handbook of Statistical Methods, joint effort of the national institute of standards and technology and international SEMATECH.

Davies, R. B. y B. Hutton (1975). *The effects of errors in the independent variables in linear regression*. Biometrika 62: 383 – 391.

Del Castillo, E. (1997). *Stopping rules for steepest ascent in experimental optimization*. Communications in statistics: simulations an computation 26(4): 1599 -1615.

Del Castillo, E. y S. Cahya (2001). *A tool for computing confidence regions on the stationary point of a response surface*. The American Statistician 55(4): 358 – 365.

Doehlert, D. H. (1970). *Uniform shell designs*. Applied Statistics 19: 231 -239.

Draper, N. R. y R. C. Van Nostrand. (1979). *Ridge regression and James Stein estimators; Review and comments*. Technometrics 21: 451 – 466.

Draper, N. R. y H. Smith (1981). *Applied regression analysis*. Segunda Edición. Nueva York. John Wiley and Sons.

Draper, N. R. (1985). *Small composite designs*. Technometrics 27: 173 – 180.

Draper, N. R. y J. A. John (1998). *Response surface designs where levels of some factors are difficult to change*. Australian and New Zeland Journal of Statistics 40: 487 -495.

Draper, N. R. y D. K. J. Lin (1990). *Small response-surface designs*. Technometrics 32: 187 – 194.

Durbin, J. (1970). *An alternative to the bounds test for testing for serial correlation in least squares regression*. Econometrica 38: 422 -429.

Durbin, J. (1969). *Test for serial correlation in regression analysis based on the periodogram of least squares residuals*. Biometrika 56: 1 – 15.

Durbin, J. y G. S. Watson (1950). *Testing for serial correlation in least squares regression, I*. Biometrika 37: 409 -428.

Durbin, J. y G. S. Watson (1951). *Testing for serial correlation in least squares regression, II*. Biometrika 38: 1 - 19.

Durbin, J. y G. S. Watson (1971). *Testing for serial correlation in least squares regression, III*. Biometrika 58: 159 – 178.

Edmondson, R. N. (1991). *Agricultural response surface experiments based on four-level factorial designs*. Biometrics 47: 1435 – 1448.

Ezekiel, M. y K. A. Fox (1959). *Methods of correlation and regression analysis*. Nueva York. John Wiley and Sons.

Fang, K. T. y D. K. J. Lin (2003). *Uniform experimental designs and their applications in industry*. In Handbook of statistics, Vol. 22, Chsp. 4. Amsterdam: Elsevier Science B. V.

Fang, K. T. y R. Mukerjee (2000). *A connection between uniformity and aberration in regular fractions of two –level factorials*. Biometrika 87: 193 -1989.

Folks, J. L. (1967). *Straight line confidence regions for linear models*. Journal of the American Statistical Association 62: 1365 – 1374.

Foraythe, A. B., L. Engelman, R. Jennrich, y P. R. A. May (1973). *A stopping rule for variable selection in multiple regression*. Journal of the American Statistical Association 68: 75 – 77.

Freund, R. J. (1979). *Multicollinearity etc., Some New Examples*. Procedings of the statistical Computing Section, American Statistics Association, pág. 111 – 112.

Furnival, G. M. (1971). *All possible regressions with less computation*. Technometrics 13: 403 -408.

Galton, Sir Francis (1885). *Regression towards mediocrity in heredity stature*. Journal of Anthropological Institute 15: 246 -263.

Geisser, S. (1975). *The predictive sample reuse method with applications*. Journal of the American Statistical Association 70: 320 – 328.

Ghadge, S. V. y H. Raheman (2006). *Process optimization for biodiesel production from mahua (Madhuca indica) oil using response surface methodology*. Bioresource Technology 97(3): 379 – 384.

Ghosh, S. y W. S. Al-Sabah (1996). *Efficient composite designs with small number of runs*. Journal of statistical planning and inference 53(1): 117 – 132.

Giesbrecht, F. G. y M. L. Gumpertz (2004). *Planning, construction and statistical analysis of comparative experiments*. Hoboken. NJ: Wiley.

Gilmour, S. G. (2006). *Response surface desings for experiments in bioprocessing*. Biometrics 62: 323 – 331.

Gilmour, S. G. y N. R. Draper (2003). *Confidence intervals around the ridge of optimal response on fitted second-order-response surface*. Technometrics 45: 333-339.

Gilmour, S. G. y N. R. Draper (2004). *Response*. Technometrics 46: 358.

Gioivannitti – Jensen, A. y R. H. Myers (1989). *Graphical assessment of the prediction capability of response surface designs*. Technometrics 31: 159 – 171.

Goldberger, A. S. (1961). *Stepwise least-squares: residual analysis and specification*. Journal of the American Statistical Association 56: 998 -1000.

Gorman, J. W., y R. J. Toman (1966). *Selection of variables for fitting equations to data*. Technometrics 8: 27 – 51.

Graybill, F. A. (1976). *Theory and Application of the Linear Model*. Boston, Massachusetts. Duxbury Press.

Green, L. W. (1970). *Manual for scoring socioeconomic status for research on Health Behaviors*. Public Healt Rept. 85: 815 -827.

Gunst, R. F. y R. L. Mason (1979). *Some considerations in the evaluation of alternate prediction equations*. Technometrics 21: 55 - 63.

Gunst, R. F. y R. L. Mason (1980). *Regression analysis and its applications*. A data oriented approach. Nueva York. Marcel Dekker.

Hahn, G. J. (1977). *Fitting regressions models with no intercept term*. Journal of Quality Technology 9: 56-60.

Hahn, G. J. (1973). *The coefficient of determination exposed*. Chemical Technology 9(2): 56 - 61.

Hahn, G. J. (1972). *Simultaneous prediction intervals for a regression model*. Technometrics 14: 203 -214.

Halperin, M. (1961). *Fitting of straight lines and predictions when both variables are subject to error*. Journal of American Statistical Association 56: 657 – 669.

Hardin, R. H. y N. J. A. Sloane (1991). *Computer-generated minimal (and larger) response surface designs: (I) The sphere*.

Hartley, H. O. (1959). *Smallest composite designs for response surfaces*. Biometrics 15: 611 – 624.

Hedayat, A. y D. S. Robson (1970). *Independence stepwise residuals for testing homoscedasticity*. Journal of the American Statistical Association 65: 1573 – 1581.

Hemmerle, W. J. (1975). *An explicit solution for generalizated ridge regression*. Technometrics 17: 309 - 314.

Henderson, H. V. y P. F. Velleman (1981). *Building multiple regression models interactively*. Biometrics 37: 391 - 411.

Hill, A. B. (1971). *Principles of Medical Statistics*, Novena Edición. Nueva York: Oxford University Press.

Hinkley, D. V. (1975). *On power transformations to symmetry*. Biometrika 62: 101 - 111.

Hoagling, D. C., y R. Welch (1978). *The hat matrix in regression and ANOVA*. American statistician 32: 17 – 22.

Hoerl, R. W. (1985). *Ridge analysis 25 years later*. The American statistician 39: 186-192.

Hocking, R. R., y R. N. Leslie (1967). *Selection of the best subset in regression analysis*. Technometrics 9: 531 – 540.

Hocking, R. R. (1972). *Criteria for selection of a subset regression: which one should be used?* Technometrics 14: 967 – 970.

Hocking, R. R. (1974). *Misspecification in regression*. American Statistician 28: 39 - 40.

Hocking, R. R. (1976). *The analysis and selection of variables in linear regression*. Biometrics 32: 1 – 51.

Huang, L. Z., Y. Lu, Y. Yuan, F: Lü, y X. Bie (2006). *Optimization of a protective medium for enhancing the viability of freeze-dried Lactobacillus delbrueckii subsp. Bulgaricus based on response surface methodology*. Journal of Industrial Microbiology and Biotechnology 33(1): 55 – 61.

John, J. A. y N. R. Daper (1980). *An alternative family of transformations*. Applied Statistics 29: 190 -197.

Kennard, R. W. (1971). *A note on the C_p statistic*. Technometrics 13: 899 – 900.

Khuri, A. I. (2003). *Current modeling and design issues in response surface methodology: GLMs and models with block effects*. In Handbook of Statistics, Vol. 22, Chap.6. Amsterdam: Elsevier Science B.V.

Khuri, A. I., ed. (2005). *Response surface methodology and related topics*. Washington, D.C. World Scientific.

Khuri, A. I. y J. Cornell (1996). *Response surface*. Design and Analysis, segunda edición. Nueva York: Marcel Dekker.

Kim, B. H. y C. C. Akoh (2005). *Modeling of lipase-catalyzed acidolysis of sesame oil and caprylic acid by response surface methodology: optimization of reaction conditions by considering both acyl incorporation and migration.* Journal of Agricultural and Food Chemistry 53(20): 8033 – 8037.

Kleijnen, J. P. C., D. den Hertog, y E. Angün (2004). *Response surface methodology´s steepest ascent and step size revisited*. European Journal of Operational Research 159: 121 -131.

Kleinbaum, D. G., Kupper, L. L., y Morgenstern, H. (1982). *Epidemiologic Research*. Belmont, Calif. Lifetime Learning Publications.

Krasker, W. S. (1980). *Estimation in linear regression models with disparate data points*. Econometrica 48: 1333 – 1346.

Kvalseth, T. O. (1985). *Cautionary note about R^2*. American Statistician 39: 279 -285.

Lewis, T., y Taylor, L. R. (1967). *Introduction to Experimental Ecology*. Nueva York. A. P

Lindley, D. V. (1968). *The choice of variables in multiple regression*. Journal of the Royal Statistical Society Serie B 30: 31 -53.

Lowe, C. W. (1974). *Evolutionary operation in action*. Applied Statistics 23(2): 218-226.

McCabe, G. P. (1984). *Principal variables*. Technometrics 26: 137 -144.

Mc Daniel, W. R. y B. E. Ankenman (2000). *A response surface test bed*. Quality and Reliability Engineering International 16: 363 – 372.

Mallows, C. L. (1973). *Some comments on C_p*. Technometrics 15: 661 – 675.

Mantel, N. (1970). *Why step-down procedures in variable selection*. Technometrics 12: 621 – 625.

Marquardt, D. W. y R. D. Snee (1975). *Ridge regression in practice*. American Statistician 29: 3 -20.

Mays, D. P. y S. M. Easter (1997). *Optimal response surface designs en the presence of dispersion effects*. Journal of Quality Technology 29: 59 -70.

Mee, R. W. (2001). *Noncentral composite designs*. Technometrics 43(1): 34 -43.

Mee, R. W. (2004). *Optimal three-level designs for response surface in spherical experimental regions*. Reporte Técnico 2004-3, Departamento de Estadísticas, Operaciones y Ciencias de la Administración, Universidad de Tennessee.

Mobert, M. K., K. E. Markides y D. Bylund (2005). *Multi-parameter investigation of tandem mass spectrometry in a linear ion trap using response surface modeling*. Journal of Mass Spectrometry 40(3): 317 -324.

Montgomery, D. C., y E.A. Peck (1982). *Introduction to linear regression analysis*. Nueva York. John Wiley and Sons.

Morrison, d. F. (1976). *Multivariate Statistical Methods*. Tercera edición. Nueva York.

Mosteller, F., y Tukey, J. W. (1967). *Data analysis and regression*. Reading, Mass: Addison-Wesley Publishing Company, Inc.

Myers R. H. (1971). *Response surface Methodology*. Boston: Allyn and Bacon.

Myers, R. H. (1999). *Response surface methodology – current status and future directions*. Journal of Quality Technology 31(1): 30 -44.

Myers, R. H. y D. C. Montgomery (1995). *Response surface methodology: Process and Product Optimization using Designed Experiments*. Nueva York: Wiley.

Myers, R. H. y D. C. Montgomery (2002). *Response Surface Methodology: Process and Product Optimization using Designed Experiments*. Segunda edición. Nueva York: Wiley.

Myers, R. H., D. C. Montgomery, y G. G. Vining, C. M. Borror and S. M. Kowalski (2004). *Response surface methodology: A retrospective and literature survey*. Journal of Quality Technology 36(1): 53 – 77.

Nagasawa, S., Osano, S., y Kondo, K (1964). *An analytical method for evaluating the susceptibility of Fish species to an Agricultural Chemical*. Japón. Journal Applied Entomology Zool. 8: 118 -122.

Neter, J., Wasserman, W., y Kutner, M. H. (1983). *Applied linear regression models*. Homewood, Ill.; Richard D. Irwin.

Nicolai, R. P., R. Dkker, N. Piersma y G. J. Van Oortmarssen (2004). *Automated response surface methodology for stochastic optimization models with known variance. In proceedings of the 2004 Winter Simulation Conference*, pp. 491 – 499. The society for computer simulation international, San Diego, Cal.

Notz, W. (1982). *Minimal point second order designs*. Journal of Statistical Planning and Inference 6: 47 -58.

Obenchain, R. L. (1977). *Classical F-test and confidence regions for ridge regression*. Technometrics 19: 429 – 439.

Obenchain, R. L. (1978). *Good and optimal ridge estimators*. Annals of Statistics 6: 1111 – 1121.

Olkin, I. (1967). *Correlation Revised. In Improving experimental design and statistical analysis*, ed. Julian C. Stanley. Chicago.

Olkin, I. y Siotani, M. (1964). *Asymptotic distribution functions of a correlation matrix. Stanford University Laboratory for Quantitative research in education*, Report No. 6, Stanford, California.

Packer, P. E. (1951). *An approach to Watershed protection criteria*. J. Forestry 49: 638 -644.

Peterson, J. J. (1993). *A general approach to ridge analysis with confidence intervals*. Technometrics 35: 204 -214.

Peterson, J. J., S. Cahya y E. del Castillo (2002). *A general approach to confidence regions for optimal factor levels of response surfaces*. Biometrics 58: 422 – 431.

Peterson, J. J., S. Cahya y E. del Castillo (2004). *Letter to the editor*. Technometrics 46(3): 355 -357.

Peterson, J. J. y A. M. Kuhn (2005). *Ridge analysis with noise variables*. Technometrics 47(3): 274 -283.

Picard, R. R. y Cook, R, D. (1984) *Cross- validations of regression models*. Journal American Statistics Association 79: 575 – 583.

Pope, P. T., y J. T. Webster (1972). *The use of an F-statistic in stepwise regression procedures*. Technometrics 14: 327 - 340.

Rao, P. (1971). *Some notes on misspecification in multiple regression*. The American Statistician 25: 37 -39.

Rencher, A. C., y F. C. Pun (1980). *Inflation of R^2 in best subset regression*. Technometrics 22: 49 -53.

SAS Institute Inc. 1985. *SAS User´s Guide Statistics*, Version 5. Cary, North Carolina.

Schatzoff, M., R. Tsao y S. Fienberg (1968). *The efficient calculation of all possible regressions*. Technometrics 10: 769 -779.

Schreiner, H. R., Gerogine, R. C., y Lawrie, J. A. (1962). *New Biological effects of the gases of the helium group*. Science 136: 653 -654.

Searle, S. R. (1971). *Linear models*. Segunda Edición. Nueva York: John Wiley and Sons.

Seber, G. A. F. (1977). *Linear regression analysis*. Nueva York: John Wiley and Sons.

Shapiro, S. S. y Wilks, M. B. (1965). *An Analysis of Variance Test for Normality (complete samples).* Biometrika 52: 591 - 611.

Silvey, S. D. (1969). *Multicollinearity and imprecise estimation*. Journal of the Royal Statistical Society Series B 31: 539 – 552.

Smith, G. y Campbell, F (1980). *A critique of some ridge regression methods*. Journal American. Statist. Assoc. 75: 74 – 81.

Snee, R. (1977). *Validation of regression models: Methods and Examples*. Technometrics 19: 415 – 428.

Snee, R. D. (1985). *Computer –aided design of experiments- some practical experiences*. Journal of Quality Technology 17(4): 222 -236.

Sprent, P. (1971). *Parallelism and concurrence in linear regression*. Biometrics 27: 440-444.

Steinberg, D. M. y D. Bursztyn (2001). *Discussion of Factor Screening and response surface exploration by S. W. Cheng and C. F. J. Wu.* Statistica Sinica 11: 596 -599.

Stevens, J. P. (1984). *Outliers and influential data points in regression ridge*. Psychology Bull. 95: 334-344.

Stigler, S. M. (1981). *Gauss and the invention of least squares*. Annals of Statistics 99: 465-474.

Suich, R. y G.C. Derringer (1977). *Is the regression equation adequate – one criterion*. Technometrics 19: 213-216.

Sztendur, E. M. y N. T. Diamond (2002). *Extensions to confidence region calculations for the path of steepest ascent*. Journal of Quality Technology 34(3): 289-296.

Tang, M., J. Li, L. Y. Chan y D. K. J. Lin (2004). *Application of uniform design in the formation of cement mixtures*. Quality Engineering 16(3): 461-474.

Timm, N. H. (1975). *Multivariate analysis with applications in education and psychology*. Monterrey, California. Brooks.

Thompson, S. J. (1972). *The doctor-patient relationship and outcomes of pregnancy*. Ph. D. dissertation. Department of Epidemiology, University of North Carolina.

Tuck, M. G., S. M. Lewis y J.I.L. Cottrell (1993). *Response surface methodology and Taguchi: A quality improvement study from the milling industry*. Journal of the Royal Statistical Society, Series C 42: 671-676.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass. Addison – Wesley. P. C.,I.

Vázquez, M. y A. M. Martin (1998). *Optimization of Phaffia rhodozyma continuous culture through response surface methodology*. Biotechnology and Bioengineering 57(3): 314-320.

Webster, J. T., R. F. Gunst y R. L. Mason (1974). *Latent root regression analysis*. Technometrics 16: 513-522.

Weisberg, S. (1980). *Applied Linear Regressions*. Nueva York: John Wiley and Sons.

Westlake, W. J. (1965). *Composite designs based on irregular fractions of factorials*. Biometrics 21: 324-335.

White, J. W. y R. F. Gunst (1979). *Latent root regression: large sample analysis*. Technometrics 21: 481-488.

Wu, C. F. J. y Y. Ding (1998). *Construction of response surface designs for qualitative and quantities factors*. Journal of Statistical Planning and Inference 71: 331-348.

Wu, C. F. J. y M. Hamada (2000). *Experiments: Planning, Analysis and Parameter Design Optimization*. Nueva York: Wiley.

Yoshida, M. (1961). *Ecological and physiological researches on the wireworm, Melanotus caudex Lewis*. Iwata. Shizuoka Pref., Japón.

ANEXOS

Anexo A. Puntos porcentuales de la distribución t

A / v	.40	.25	.10	.05	.025	.01	.005	.0025	.001	.0005
1	.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	.289	.816	1.886	2.920	4.303	6.965	9.925	14.089	23.326	31.598
3	.277	.765	1.636	1.353	3.182	4.541	5.841	7.453	10.213	12.924
4	.271	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.267	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.265	.727	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.263	.711	1.415	1.895	2.365	2.998	3.499	4.019	4.785	5.408
8	.262	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.261	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.260	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	.260	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.259	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.259	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.258	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.258	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.258	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.257	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.257	.688	1.330	1.736	2.101	2.552	2.878	3.197	3.610	3.922
19	.257	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.257	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.257	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.256	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.256	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.256	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.256	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	.256	.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	.256	.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	,256	.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	.256	.683	1.311	10699	2.045	2.462	2.756	3.038	3.396	3.659
30	.256	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.255	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.254	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.254	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.253	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

v= grados de libertad

Anexo B. Puntos porcentuales de la distribución χ^2

ν \ α	0.995	0.990	0.975	0.950	0.500	0.050	0.025	0.010	0.005
1	0.00	0.00	0.00	0.00	0.45	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	1.39	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	2.37	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	3.36	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	4.35	11.07	12.38	15.09	16.75
6	0.68	0.87	1.24	1.64	5.53	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	6.35	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	7.34	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	8.34	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	9.34	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	10.34	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	11.34	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	12.34	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	13.34	23.68	26.12	29.14	31.32
15	4.60	5.23	6.27	7.26	14.34	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	15.34	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	16.34	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	17.34	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	18.34	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	19.34	31.41	34.17	37.57	40.00
25	10.52	11.52	13.12	14.61	24.34	37.65	40.65	44.31	46.93
30	13.79	14.95	16.79	18.49	29.34	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	39.34	55.76	59.34	63.69	66.77
50	27.99	29.11	32.36	34.76	49.33	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	59.33	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	69.33	90.53	95.02	100.42	104.22
80	551.17	53.54	57.15	60.39	79.33	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	89.33	113.14	118.14	124.12	138.30
100	67.33	70.06	74.22	77.93	99.33	124.34	129.56	135.81	140.17

ν = grados de libertad

Anexo C. Áreas bajo la curva normal

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0009	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	01379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	06517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	07580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.7023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.99857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Anexo D. Puntos porcentuales de la distribución F

$F_{0.05,v1,v2}$

V_1 / V_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	234.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84

21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.21	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.55	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	7.52	1.46	1.39	1.32	1.22	1.00

Anexo E. Puntos porcentuales de la distribución F

$F_{0.025,v1,v2}$

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.112	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09

21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

Anexo F. Puntos porcentuales de la distribución F

$F_{0.01,v1,v2}$

v_1 / v_2	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999.5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.00	26.50	226.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
12	9.33	6.91	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4014	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.26	3.21	3.13	3.05	2.96	2.87
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.94	2.75	2.66	2.57
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42

21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.31	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.90	2.75	2.60	2.52	2.44	2.34	2.26	2.17	2.06
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00	2.87	2.73	2.57	2.49	2.41	2.33	2.23	2.14	2.03
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	1.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

Tabla G. Valores de $\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$

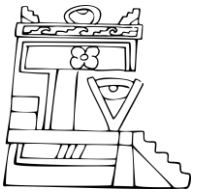
r	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.000	0.000	0.0010	0.0020	0.0030	0.0040	0.0050	0.0060	0.0070	0.0080	0.0090
0.010	0.0100	0.0110	0.0120	0.0130	0.0140	0.0150	0.0160	0.0170	0.0180	0.0190
0.020	0.0200	0.0210	0.0220	0.0230	0.0240	0.0250	0.0260	0.0270	0.0280	0.0290
0.030	0.0300	0.0310	0.0320	0.0330	0.0340	0.0350	0.0360	0.0370	0.0380	0.0390
0.040	0.0400	0.0410	0.0420	0.0430	0.0440	0.0450	0.0460	0.0470	0.0480	0.0490
0.050	0.0501	0.0511	0.0521	0.0531	0.0541	0.0551	0.0561	0.0571	0.0581	0.0591
0.060	0.0601	0.0611	0.0621	0.0631	0.0641	0.0651	0.0661	0.0671	0.0681	0.0691
0.070	0.0701	0.0711	0.0721	0.0731	0.0741	0.0751	0.0761	0.0771	0.0781	0.0791
0.080	0.0802	0.0812	0.0822	0.0832	0.0842	0.0852	0.0862	0.0872	0.0882	0.0892
0.090	0.0902	0.0912	0.0922	0.0933	0.0943	0.0953	0.0963	0.0973	0.0983	0.0993
0.100	0.1003	0.1013	0.1024	0.1034	0.1044	0.1054	0.1064	0.1074	0.1084	0.1094
0.110	0.1105	0.1115	0.1125	0.1135	0.1145	0.1155	0.1165	0.1175	0.1185	0.1195
0.120	0.1206	0.1216	0.1226	0.1236	0.1246	0.1257	0.1267	0.1277	0.1287	0.1297
0.130	0.1308	0.1318	0.1328	0.1338	0.1348	0.1358	0.1368	0.1379	0.1389	0.1399
0.140	0.1409	0.1419	0.1430	0.1440	0.1450	0.1460	0.1470	0.1481	0.1491	0.1501
0.150	0.1511	0.1522	0.1532	0.1542	0.1552	0.1563	0.1573	0.1583	0.1593	0.1604
0.160	0.1614	0.1624	0.1634	0.1644	0.1655	0.1665	0.1676	0.1686	0.1696	0.1708
0.170	0.1717	0.1727	0.1737	0.1748	0.1758	0.1768	0.1779	0.1789	0.1799	0.1810
0.180	0.1820	0.1830	0.1841	0.1851	0.1861	0.1872	0.1882	0.1892	0.1903	0.1913
0.190	0.1923	0.1934	0.1944	0.1954	0.1965	0.1975	0.1986	0.1996	0.2007	0.2017
0.200	0.2027	0.2038	0.2048	0.2059	0.2069	0.2079	0.2090	0.2100	0.2111	0.2121
0.210	0.2132	0.2142	0.2153	0.2163	0.2174	0.2184	0.2194	0.2205	0.2215	0.2228
0.220	0.2237	0.2247	0.2258	0.2268	0.2279	0.2289	0.2300	0.2310	0.2321	0.2331
0.230	0.2342	0.2353	0.2363	0.2374	0.2384	0.2395	0.2405	0.2416	0.2427	0.2437
0.240	0.2448	0.2458	0.2469	0.2480	0.2490	0.2501	0.2511	0.2522	0.2533	0.2543
0.250	0.2554	0.2565	0.2575	0.2586	0.2597	0.2608	0.2618	0.2629	0.2640	0.2650
0.260	0.2661	0.2672	0.2682	0.2693	0.2704	0.2715	0.2726	0.2736	0.2747	0.2758
0.270	0.2769	0.2779	0.2790	0.2801	0.2812	0.2823	0.2832	0.2844	0.2855	0.2866
0.280	0.2877	0.2888	0.2898	0.2909	0.2920	0.2931	0.2942	0.2953	0.2964	0.2975
0.290	0.286	0.2997	0.3008	0.3019	0.3029	0.3040	0.3051	0.3062	0.3073	0.3084
0.300	0.3095	0.3106	0.3117	0.3128	0.3139	0.3150	0.3161	0.3172	0.3183	0.3195
0.310	0.3206	0.3217	0.3228	0.3239	0.3250	0.3261	0.3272	0.3283	0.3294	0.3305
0.320	0.3317	0.3328	0.3339	0.3350	0.3361	0.3372	0.3384	0.3395	0.3406	0.3417
0.330	0.3428	0.3439	0.3451	0.3462	0.3473	0.3484	0.3496	0.3507	0.3518	0.3530
0.340	0.3541	0.3552	0.3564	0.3575	0.3586	0.3597	0.3609	0.3620	0.3632	0.3643

0.350	0.3654	0.3666	0.3677	0.3689	0.3700	0.3712	0.3723	0.3734	0.3746	0.3757
0.360	0.3769	0.3780	0.3792	0.3803	0.3815	0.3826	0.3838	0.3850	0.3861	0.3873
0.370	0.3884	0.3802	0.3907	0.3919	0.3931	0.3942	0.3954	0.3966	0.3977	0.3989
0.380	0.4001	0.4012	0.4024	0.4036	0.4047	0.4059	0.4071	0.4083	0.4094	0.4106
0.390	0.4118	0.4130	0.4142	0.4153	0.4165	0.4177	0.4189	0.4201	0.4213	0.4225
0.400	0.4236	0.4248	0.4260	0.4272	0.4284	0.4296	0.4308	0.4320	0.4332	0.4344
0.410	0.4356	0.4368	0.4380	0.4392	0.4404	0.4416	0.4429	0.4441	0.4453	0.4465
0.420	0.4477	0.4489	0.4501	0.4513	0.4526	0.4538	0.4550	0.4562	0.4574	0.4587
0.430	0.4599	0.4611	0.4623	0.4636	0.4648	0.4660	0.4673	0.4685	0.4697	0.4710
0.440	0.4722	0.4735	0.4747	0.4760	0.4772	0.4784	0.4797	0.4809	0.4822	0.4835
0.450	0.4847	0.4860	0.4872	0.4885	0.4897	0.4910	0.4923	0.4935	0.4948	0.4961
0.460	0.4973	0.4986	0.4999	0.5011	0.5024	0.5037	0.5049	0.5062	0.5075	0.5088
0.470	0.5101	0.5114	0.5126	0.5139	0.5152	0.5165	0.5178	0.5191	0.5204	0.5217
0.480	0.5230	0.5243	0.5256	0.5279	0.5282	0.5295	0.5308	0.5321	0.5334	0.5347
0.490	0.5361	0.5374	0.5387	0.5400	0.5413	0.5427	0.5440	0.5453	0.5466	0.5480

Tabla G. Valores de $\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$

r	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009
0.500	0.5493	0.5506	0.5520	0.5533	0.5547	0.5560	0.5573	0.5587	0.5600	0.5614
0.510	0.5627	0.5641	0.5654	0.5668	0.5681	0.5695	0.5709	0.5722	0.5736	0.5750
0.520	0.5763	0.5777	0.5791	0.5805	0.5818	0.5832	0.5846	0.5860	0.5874	0.5888
0.530	0.5901	0.5915	0.5929	0.5943	0.5957	0.5971	0.5985	0.5999	0.6013	0.6027
0.540	0.6042	0.6056	0.6070	0.6084	0.6098	0.6112	0.6127	0.6141	0.6156	0.6170
0.550	0.6184	0.6198	0.6213	0.6227	0.6241	0.6256	0.6270	0.6285	6299	0.6314
0.560	0.6328	0.6343	0.6358	0.6372	0.6387	0.6401	0.6416	0.6431	0.6446	0.6460
0.570	0.6475	0.6490	0.6505	0.6620	0.6535	0.6550	0.6565	0.6579	0.6594	0.6610
0.080	0.6625	0.640	0.6655	0.6670	0.6685	0.6700	0.6715	0.6731	0.6746	0.6761
0.590	0.6777	0.6792	0.6807	0.6823	0.6838	0.6854	0.6869	0.6885	0.6900	0.6916
0.600	0.6931	0.6974	0.6963	0.6978	0.6994	0.7010	0.7026	0.7042	0.7057	0.7073
0.610	0.7089	0.7105	0.7121	0.7137	0.7153	0.7169	0.7185	0.7201	0.7218	0.7234
0.620	0.7250	0.7266	0.7283	0.7299	0.7315	0.7332	0.7348	0.7364	0.7381	0.7398
0.630	0.7414	0.7431	0.7447	0.7464	0.7481	0.7497	0.7514	0.7531	0.7548	0.7565
0.640	0.7582	0.7599	0.7616	0.7633	0.7650	0.7667	0.7684	0.7701	0.7718	0.7736
0.650	0.7753	0.7770	0.7788	0.7805	0.7823	0.7840	0.7858	0.7875	0.7893	0.7910
0.660	0.7928	0.7946	0.7964	0.7981	0.7999	0.8017	0.8035	0.8053	0.8071	0.8089
0.670	0.8107	0.8126	0.8144	0.8162	0.8180	0.8199	0.8217	0.8236	0.8254	0.8273
0.680	0.8291	0.8310	0.8328	0.8347	0.8366	0.8385	0.8404	0.8423	0.8442	0.8461
0.690	0.8480	0.8499	0.8518	0.8537	0.8556	0.8576	0.8595	0.8614	0.8634	0.8653
0.700	0.8673	0.8693	0.8712	0.8732	0.8752	0.8772	0.8792	0.8812	0.8832	0.8852
0.710	0.8872	0.8892	0.8912	0.8933	0.8953	0.8973	0.8994	0.9014	0.9035	0.9056
0.720	0.9076	0.9097	0.9118	0.9139	0.9160	0.9181	0.9202	0.9223	0.9245	0.9266
0.730	0.9287	0.9309	0.9330	0.9352	0.9373	0.9395	0.99417	0.9439	0.9461	0.9483
0.740	0.9505	0.9527	0.9549	0.9571	0.9594	0.9616	0.9639	0.9661	0.9684	0.9707
0.750	0.9730	0.9752	0.9775	0.9799	0.9822	0.9845	0.9868	0.9892	0.9915	0.9939
0.760	0.9962	0.9986	1.0010	1.0034	1.0058	1.0082	1.0108	1.0130	1.0154	1.0179
0.770	1.0203	10.228	1.0253	1.0277	1.0302	1.0327	1.0352	1.0378	1.0403	1.0428
0.780	1.0454	1.0479	1.0505	1.0531	1.0557	1.0583	1.0609	1.0635	1.0661	1.0688
0.790	1.0714	1.0741	1.0768	1.0795	1.0822	1.0849	1.0875	1.0903	1.0931	1.0958

0.800	1.0986	1.1014	1.1041	1.1070	1.1098	1.1127	1.1155	1.1184	1.1212	1.1241
0.810	1.1270	1.1299	1.1329	1.1358	1.1388	1.1417	1.1447	1.1477	1.1507	1.1538
0.820	1.1568	1.1599	11630	1.1660	1.1692	1.1723	1.1754	1.1786	1.1817	1.1849
0.830	1.1870	1.1913	1.1946	1.1979	1.2011	1.2044	1.2077	1.2111	1.2144	1.2178
0.840	1.2212	1.2246	1.2280	1.2315	1.2349	1.2384	1.2419	1.2454	1.2490	1.2526
0.850	1.2561	1.2598	1.2634	1.2670	1.2708	1.2744	1.2782	1.2819	1.2857	1.2895
0.860	1.2934	1.2972	1.3011	1.3053	1.3089	1.3129	1.3168	1.3209	1.3249	1.3290
0.870	1.3331	1.3372	1.3414	1.3456	1.3498	1.3540	1.3583	1.3626	1.3670	1.3714
0.880	1.3758	1.3802	1.3847	1.3892	1.3938	1.3984	1.4030	1.4077	1.4124	1.4171
0.890	1.4219	1.4268	1.4316	1.4366	1.4415	1.4465	1.4516	1.4566	1.4618	1.4670
0.900	1.4722	1.4775	1.4828	1.4883	1.4937	1.4992	1.5147	1.5103	1.5169	1.5217
0.910	1.5275	1.5334	1.5393	1.5453	1.5513	1.5574	1.5636	1.55698	1.5762	1.5825
0.920	1.5890	1.5956	1.6022	1.6089	1.6157	1.6226	1.6296	1.6366	1.6438	1.6510
0.930	1.6584	1.6659	1.6734	1.6811	1.6888	1.6967	1.7047	1.7129	1.7211	1.7295
0.940	1.7380	1.7467	1.7555	1.7645	1.7736	1.7828	1.7923	1.8019	1.8117	1.8216
0.950	1.8318	1.8421	1.8527	1.8635	1.8745	1.8857	1.8972	1.9090	1.9210	1.9333
0.960	1.9459	1.9588	1.9721	1.9857	1.9996	2.0140	2.0287	2.0439	2.0595	2.0756
0.970	2.0923	2.1095	2.1273	2.1457	2.1649	2.1847	2.2054	2.2269	2.2494	2.2729
0.980	2.2976	2.3223	2.3507	2.3796	2.4101	2.4426	2.4774	2.5147	2.5550	2.5988
0.990	2.6467	2.6995	2.7587	2.8257	2.9031	2.9945	3.1063	3.2504	3.4534	3.8002
0.9999	4.9517									
0.99999	6.103									



Difusión y Divulgación
Científica y Tecnológica

Esta obra se terminó de imprimir el 30 de agosto de 2015, con un tiraje de 300 ejemplares, en los talleres de M. A. Impresores, S. A. de C. V.; Avenida Hierro Número 3; Colonia Ciudad Industrial; Villahermosa, Tabasco, México. El cuidado estuvo a cargo de los autores y del Departamento Editorial de Publicaciones No Periódicas de la Dirección de Difusión y Divulgación Científica y Tecnológica de la UJAT.