Introducción al Análisis de Regresión Lineal

Montgomery • Peck • Vining

3a. edición

CECSA

Introducción al Análisis de Regresión Lineal

Por lo anterior, se ve con frecuencia a t_0 escrito en la forma

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\text{se}(\hat{\beta}_1)} \tag{2.27}$$

Se puede usar un procedimiento parecido para probar hipótesis acerca de la ordenada al origen. Para probar

$$\begin{aligned} H_0&: \beta_0 = \beta_{00} \\ H_1&: \beta_0 \neq \beta_{00} \end{aligned} \tag{2.28}$$

se podría usar el **estadístico de prueba**

$$t_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{MS_{\text{Res}}\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} = \frac{\hat{\beta}_0 - \beta_{00}}{\text{se}(\hat{\beta}_0)} \tag{2.29}$$

en donde $\text{se}(\hat{\beta}_0) = \sqrt{MS_{\text{Res}}(1/n + \bar{x}^2/S_{xx})}$ es el **error estándar de la ordenada al origen**. La hipótesis nula H_0: $\beta_0 = \beta_{00}$ se rechaza si $|t_0| > t_{\alpha/2,\, n-2}$.

2.3.2 Prueba de significancia de la regresión

Un caso especial muy importante de la hipótesis en la ecuación (2.24) es el siguiente:

$$\begin{aligned} H_0&: \beta_1 = 0 \\ H_1&: \beta_1 \neq 0 \end{aligned} \tag{2.30}$$

Estas hipótesis se relacionan con la **significancia de la regresión**. El no rechazar H_0: $\beta_1 = 0$ implica que no hay relación lineal entre x y y. Este caso se ilustra en la figura 2.2. Nótese que eso puede implicar que x tiene muy poco valor para explicar la variación de y y que el mejor estimador para cualquier x es $\hat{y} = \bar{y}$ (Fig. 2.2a), o que la verdadera relación entre x y y no es lineal (Fig. 2.2b). Por consiguiente, si no se rechaza H_0: $\beta_1 = 0$, equivale a decir que **no hay relación lineal entre y y x.**

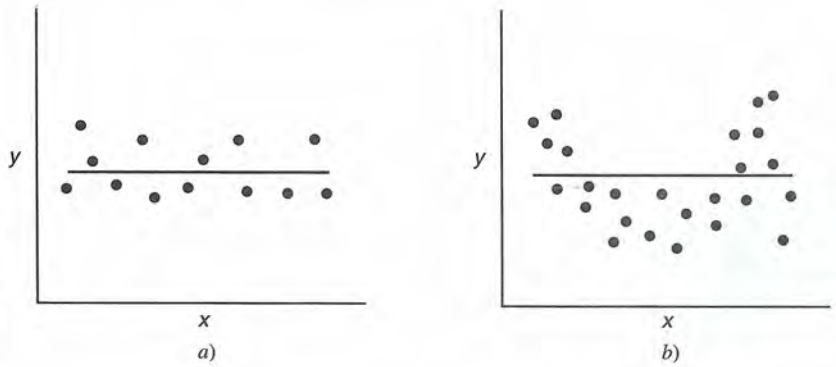


Figura 2.2 Casos en los que no se rechaza la hipótesis H_0: $\beta_1 = 0$.

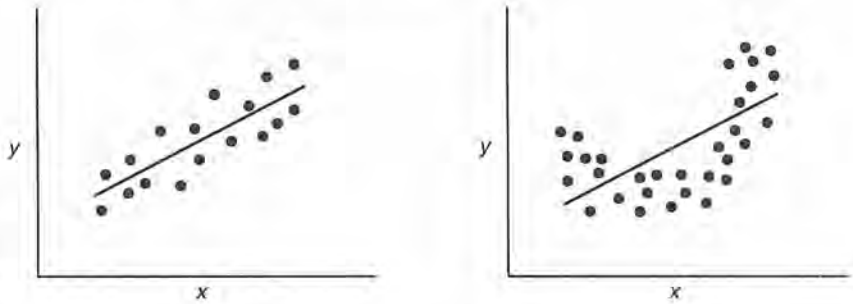


Figura 2.3 Casos en los que no se rechaza la hipótesis H_0: $\beta_1 = 0$.

También, si se rechaza H_0: $\beta_1 = 0$, eso implica que x sí tiene valor para explicar la variabilidad de y. Esto se ilustra en la figura 2.3. Sin embargo, rechazar H_0: $\beta_1 = 0$ podría equivaler a que el modelo de línea recta es adecuado (Fig. 2.3a), o que aunque hay un efecto lineal de x se podrían obtener mejores resultados agregando términos polinomiales en x (Fig. 2.3b).

El procedimiento de prueba para H_0: $\beta_1 = 0$ se puede establecer con dos métodos. El primero tan sólo usa el estadístico t en la ecuación (2.26), con $\beta_{10} = 0$, es decir,

$$t_0 = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

La hipótesis de la significancia de la regresión se rechazaría si $|t_0| > t_{\alpha/2,\, n-2}$.

Ejemplo 2.3 *Datos del propelente de reacción*

A continuación se probará la significancia de la regresión en el modelo del propelente de reacción del ejemplo 2.1. El estimado de la pendiente es $\hat{\beta}_1 = -37.15$, y en el ejemplo 2.2 se calculó el estimado de σ^2, que resultó $MS_{\text{Res}} = \hat{\sigma}^2 = 9\,244.59$. El error estándar de la pendiente es

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{MS_{\text{Res}}}{S_{xx}}} = \sqrt{\frac{9\,244.59}{1\,106.56}} = 2.89$$

Por consiguiente, el estadístico de prueba es

$$t_0 = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} = \frac{-37.15}{2.89} = -12.85$$

Si se escoge $\alpha = 0.05$, el valor crítico de t es $t_{0.025,\,18} = 2.101$. Así, se rechazaría H_0: $\beta_1 = 0$ y se llegaría a la conclusión que hay una relación lineal entre la resistencia al corte y la edad del propelente.

Una excelente descripción general de los problemas con valores atípicos se encuentra en Barnett y Lewis [1994]. También véase Myers [1990], para una buena discusión del tema.

Los valores atípicos se deben investigar con cuidado, para ver si se puede encontrar una razón de su comportamiento extraordinario. A veces, los valores atípicos son "malos" y se deben a eventos desacostumbrados, pero explicables. Entre los ejemplos están la medición o el análisis incorrectos, el registro incorrecto de los datos y la falla de un instrumento de medición. Si éste es el caso, el valor atípico se debería corregir (si es posible) o eliminar del conjunto de datos. Es claro que el eliminar valores malos es conveniente, porque los mínimos cuadrados jalan la ecuación ajustada hacia el valor atípico, ya que eso minimiza la suma de cuadrados de residuales, sin embargo, se hace notar que debe contarse con una fuerte evidencia no estadística de que el valor atípico es malo, para entonces descartarlo.

A veces se encuentra que el valor atípico es una observación extraordinaria, pero perfectamente factible. Puede ser peligroso eliminar estos puntos para "mejorar el ajuste de la ecuación", porque puede dar al usuario una sensación falsa de precisión de la estimación o la predicción. A veces se ve que el valor atípico es más importante que el resto de los datos, porque puede controlar muchas propiedades clave del modelo. También, los valores atípicos pueden hacer resaltar inadecuaciones en el modelo, como la falla de tener un buen ajuste con los datos en cierta región del espacio de x. Si el valor atípico es un punto de respuesta especialmente deseable (por ejemplo, bajo costo o alto rendimiento), sería en extremo valioso conocer los valores de los regresores, cuando se observó esa respuesta. Los análisis de identificación y de seguimiento de los valores atípicos con frecuencia dar como resultado mejoras en el proceso, o nuevos conocimientos acerca de factores cuyo efecto sobre la respuesta se desconocía antes.

Se han propuesto diversas pruebas estadísticas para detectar y rechazar los valores atípicos. Por ejemplo, véase Barnett y Lewis [1994]. Stefansky [1971, 1972] ha propuesto una prueba aproximada para identificar puntos atípicos, basada en el residual máximo normado $|e_i|/\sqrt{\Sigma_{i=1}^{n} e_i^2}$, cuya aplicación es bastante fácil. Ejemplos de esta prueba y de otras referencias relacionadas se encuentran en Cook y Prescott [1981], Daniel [1976] y Williams [1973], además véase también el apéndice C.9. Si bien esas pruebas pueden tener utilidad para identificar valores atípicos, no se debe creer que los puntos identificados con ellas deben eliminarse en forma automática. Como se ha dicho, esos puntos pueden ser claves importantes que contienen información valiosa.

El efecto de los valores atípicos sobre el modelo de regresión se puede comprobar con facilidad eliminándolos y volviendo a ajustar la ecuación de regresión. Se podrá encontrar que los valores de los coeficientes de regresión, o de los estadísticos de resumen como t, F o R^2, y que el cuadrado medio de residuales pueden ser muy sensibles a los valores atípicos. Los casos en los que un porcentaje relativamente pequeño de los datos tiene un gran impacto sobre el modelo podrán no ser aceptables para el usuario de la ecuación de regresión. En general, uno se siente más cómodo suponiendo que una ecuación de regresión es válida si no es muy sensible a unas pocas observaciones. Se preferiría que la relación de regresión estuviera embebida en todas las observaciones, y no sólo fuera un artificio de unos pocos puntos.

Ejemplo 4.7 *Datos del propelente de reacción*

La figura 4.11 muestra la gráfica de probabilidad normal y la gráfica de valores R de Student, en función de las $\hat{y}_i$ predichas para los datos del propelente de reacción presentados en el ejemplo 2.1. Se ve que hay dos residuales negativos grandes, muy apartados del resto (observaciones 5 y 6, en la Tabla 2.1). Esos puntos son atípicos potenciales; tienden a producir

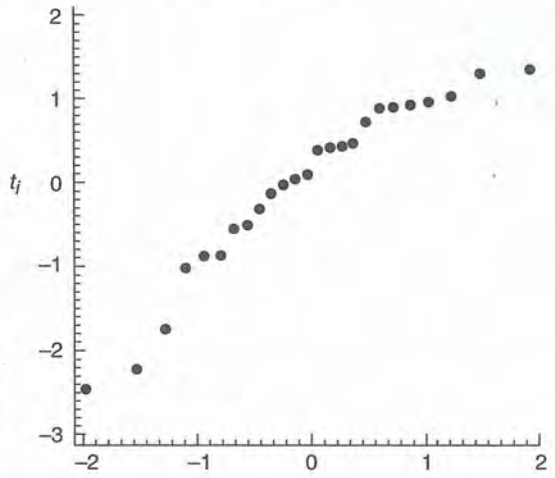


Figura 5.9 Gráfica de probabilidad normal de los residuales del modelo transformado para los datos del molino de viento.

Los estadísticos de resumen para este modelo son $R^2 = 0.9800$, $MS_{\text{Res}} = 0.0089$ y $F_0 = 1\,128.43$ (el valor P es < 0.0001).

Los valores ajustados y los residuales correspondientes al modelo transformado están en la columna B de la tabla 5.6. La figura 5.8 muestra una gráfica de los residuales R de Student del modelo transformado, en función de $\hat{y}$. En esta gráfica no se advierte problema grave alguno de desigualdad de varianza. La gráfica de probabilidad normal, que muestra la figura 5.9, parece indicar que los errores provienen de una distribución de colas más gruesas que la normal (nótense las ligeras curvaturas, hacia arriba y hacia abajo, en los extremos). Como no hay fuerte indicación de inadecuación del modelo, se llega a la conclusión que el modelo transformado es satisfactorio.

5.4 MÉTODOS ANALITICOS PARA SELECCIONAR UNA TRANSFORMACIÓN

Si bien en muchos casos las transformaciones se seleccionan en forma empírica, se pueden aplicar técnicas más formales y objetivas para ayudar a especificar una transformación adecuada. Esta sección describirá e ilustrará procedimientos analíticos para seleccionar transformaciones, tanto de la variable de respuesta como de las variables regresoras.

5.4.1 Transformaciones de y: el método de Box-Cox

Supóngase que se debe transformar y para corregir la no normalidad y/o la varianza no constante. Una clase útil de transformaciones es la de la **transformación de potencia** y^λ, donde λ es un parámetro que se debe determinar. Por ejemplo, $\lambda = \frac{1}{2}$ quiere decir usar y como respuesta. Box y Cox [1964] indican cómo se pueden estimar en forma simultánea los parámetros del modelo de regresión y λ, con el método de la máxima posibilidad.

Al imaginarse la transformación de potencia y^λ surge una dificultad cuando $\lambda = 0$: cuando λ tiende a cero, y^λ tiende a la unidad. Es obvio que es un problema, porque no tiene sentido tener todos los valores de respuesta iguales a uno. Un método para resolver esta dificultad (lo llamaremos discontinuidad en $\lambda = 0$) es usar $(y^\lambda - 1)/\lambda$ como variable de respuesta. Con esto se resuelve el problema de discontinuidad, porque cuando λ tiende a cero

$(y^{\lambda} - 1)/\lambda$ tiende al límite ln y. Sin embargo, sigue habiendo un problema, porque cuando cambia λ, los valores de $(y^{\lambda} - 1)/\lambda$ cambian en forma dramática, por lo que sería difícil comparar los estadísticos de resumen de modelos con distintos valores de λ.

El procedimiento correcto es usar

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda \dot{y}^{\lambda - 1}}, & \lambda \neq 0 \\ \dot{y} \ln y, & \lambda = 0 \end{cases} \tag{5.1}$$

en las que $\dot{y} = \ln^{-1}[(1/n)\Sigma_{i=1}^{n} \ln y_i]$ es el promedio geométrico de las observaciones, y ajustar el modelo

$$\mathbf{y}^{(\lambda)} = \mathbf{X}\beta + \varepsilon \tag{5.2}$$

por mínimos cuadrados (o por máxima posibilidad). Sucede que el divisor $\dot{y}^{\lambda - 1}$ se relaciona con el **jacobiano** de la transformación que convierte la variable de respuesta y en $y^{(\lambda)}$. Es, de hecho, un factor de escala que asegura que las sumas de cuadrados de residuales sean comparables para modelos con distintos valores de λ.

Procedimiento de cálculo

El estimado de λ por máxima posibilidad corresponde al valor de λ para el cual es mínima la suma de cuadrados de residuales del modelo ajustado, $SS_{\text{Res}}(\lambda)$. Este valor de λ se suele determinar ajustando un modelo a $y^{(\lambda)}$ para diversos valores de λ, graficando la suma de cuadrados de residuales $SS_{\text{Res}}(\lambda)$ en función de λ y viendo el valor de λ que minimiza $SS_{\text{Res}}(\lambda)$ en la gráfica. En general, son suficientes de 10 a 20 valores de λ para estimar el valor óptimo. Se puede hacer una segunda iteración con una malla más fina, si se desea. Como se vio arriba, **no** se puede seleccionar λ sólo comparando **en forma directa** las sumas de cuadrados de residuales de las regresiones de y^{λ} respecto a x, porque para cada λ, la suma de cuadrados de residuales se mide en una escala distinta. La ecuación 5.1 escala las respuestas de tal modo que las sumas de cuadrados de residuales se pueden comparar en forma directa. Se recomienda al analista usar valores sencillos de λ, porque es probable que la diferencia práctica en los ajustes para $\lambda = 0.5$ y $\lambda = 0.596$ sea pequeña, pero es más fácil interpretar el primer valor.

Una vez seleccionado un valor de λ, el analista queda libre para ajustar el modelo usando a y^{λ} como variable de respuesta si $\lambda \neq 0$. Si $\lambda = 0$, se usa en y como variable de respuesta. Es totalmente admisible usar $y^{(\lambda)}$ como respuesta para el modelo final; este modelo tendrá una escala diferente y un origen trasladado en comparación del que usa y^{λ} (o que usa ln y). De acuerdo con nuestra experiencia, la mayor parte de los ingenieros y los científicos prefieren usar y^{λ} (o ln y) como respuesta.

Un intervalo de confianza aproximado para λ

También se puede determinar un intervalo de confianza aproximado para el parámetro de transformación λ. Este intervalo de confianza puede servir para seleccionar el valor definitivo de λ. Por ejemplo, si $\hat{\lambda} = 0.596$ es el valor que minimiza la suma de cuadrados de residuales y si $\lambda = 0.5$ está en el intervalo de confianza, sería preferible usar la transformación

En esta sección se desarrollarán los mínimos cuadrados ponderados para el modelo de regresión múltiple. Se comenzará considerando un caso un poco más general que concierne a la estructura de los errores del modelo.

5.5.1 Mínimos cuadrados generalizados

Las suposiciones que se suelen hacer acerca del modelo de regresión lineal $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ son que $E(\varepsilon) = \mathbf{0}$ y que Var $(\varepsilon) = \sigma^2\mathbf{I}$. Como se ha observado, a veces esas premisas son irrazonables, por lo que ahora se examinará qué modificaciones se necesitan para este procedimiento de mínimos cuadrados ordinarios, cuando Var $(\varepsilon) = \sigma^2\mathbf{V}$, siendo $\mathbf{V}$ una matriz conocida de $n \times n$. Este caso tiene una interpretación fácil: si $\mathbf{V}$ es diagonal pero con elementos diagonales distintos, las observaciones $\mathbf{y}$ son **no correlacionadas**, pero tienen **varianzas desiguales**, mientras que si algunos de los elementos fuera de la diagonal de $\mathbf{V}$ son distintos de cero, las observaciones son **correlacionadas**.

Cuando el modelo es

$$\begin{aligned} y &= \mathbf{X}\beta + \varepsilon \\ E(\varepsilon) &= \mathbf{0},\ \text{Var}(\varepsilon) = \sigma^2\mathbf{V} \end{aligned} \tag{5.13}$$

el estimador de mínimos cuadrados ordinarios $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ ya no es adecuado. Se resolverá este problema transformando el modelo en un nuevo conjunto de observaciones que satisfagan las premisas estándar de mínimos cuadrados. A continuación se usarán mínimos cuadrados ordinarios con los datos transformados. Como $\sigma^2\mathbf{V}$ es la matriz de covarianza de los errores, $\mathbf{V}$ debe ser no singular y positiva definida, y en consecuencia existe una matriz $\mathbf{K}$, no singular y simétrica, de $n \times n$, tal que $\mathbf{K}'\mathbf{K} = \mathbf{K}\mathbf{K} = \mathbf{V}$. A menudo se le llama **raíz cuadrada** de $\mathbf{V}$ a la matriz $\mathbf{K}$. En forma típica, σ^2 se desconoce, y en ese caso $\mathbf{V}$ representa la estructura supuesta de las varianzas y covarianzas entre los errores aleatorios, aparte de una constante.

Se definen las nuevas variables

$$\mathbf{z} = \mathbf{K}^{-1}\mathbf{y}, \quad \mathbf{B} = \mathbf{K}^{-1}\mathbf{X} \quad \text{y} \quad \mathbf{g} = \mathbf{K}^{-1}\varepsilon \tag{5.14}$$

por lo que el modelo de regresión $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ se transforma en $\mathbf{K}^{-1}\mathbf{y} = \mathbf{K}^{-1}\mathbf{X}\beta + \mathbf{K}^{-1}\varepsilon$, es decir

$$\mathbf{z} = \mathbf{B}\beta + \mathbf{g} \tag{5.15}$$

Los errores en este modelo transformado tienen valor esperado cero, esto es, $E(\mathbf{g}) = \mathbf{K}^{-1}E(\varepsilon) = \mathbf{0}$. Además, la matriz de covarianza de $\mathbf{g}$ es

$$\begin{aligned} \text{Var}(\mathbf{g}) &= \{[\mathbf{g} - E(\mathbf{g})][\mathbf{g} - E(\mathbf{g})]'\} \\ &= E(\mathbf{g}\mathbf{g}') \\ &= E(\mathbf{K}^{-1}\varepsilon\varepsilon'\mathbf{K}^{-1}) \\ &= \mathbf{K}^{-1}E(\varepsilon\varepsilon')\mathbf{K}^{-1} \\ &= \sigma^2\mathbf{K}^{-1}\mathbf{V}\mathbf{K}^{-1} \\ &= \sigma^2\mathbf{K}^{-1}\mathbf{K}\mathbf{K}\mathbf{K}^{-1} \\ &= \sigma^2\mathbf{I} \end{aligned} \tag{5.16}$$

Por lo anterior, los elementos de $\mathbf{g}$ tienen media cero y varianza constante, y no están correlacionados. Como los errores $\mathbf{g}$ en el modelo (5.15) satisfacen las premisas acostumbradas, se pueden aplicar los mínimos cuadrados ordinarios. La función de mínimos cuadrados es

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{g}'\mathbf{g} = \boldsymbol{\varepsilon}'\mathbf{V}^{-1}\boldsymbol{\varepsilon} \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned} \tag{5.17}$$

Las ecuaciones de mínimos cuadrados son

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\,\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{5.18}$$

y la solución de esas ecuaciones es

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{5.19}$$

Aquí, la $\hat{\boldsymbol{\beta}}$ se llama **estimador de mínimos cuadrados generalizado** de $\boldsymbol{\beta}$.

No es difícil demostrar que $\hat{\boldsymbol{\beta}}$ es un estimador insesgado de $\boldsymbol{\beta}$. La matriz de covarianza de $\hat{\boldsymbol{\beta}}$ es

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{B}'\mathbf{B})^{-1} = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \tag{5.20}$$

En el apéndice C.10 se demuestra que $\hat{\boldsymbol{\beta}}$ es el mejor estimador lineal insesgado de $\boldsymbol{\beta}$. El análisis de varianza en términos de los mínimos cuadrados generalizados se resume en la tabla 5.8.

5.5.2 Mínimos cuadrados ponderados

Cuando los errores $\boldsymbol{\varepsilon}$ no están correlacionados, pero tienen varianzas desiguales de modo que la matriz de covarianza de $\boldsymbol{\varepsilon}$ sea, por ejemplo

$$\sigma^2\mathbf{V} = \sigma^2 \begin{bmatrix} \frac{1}{w_1} & & & 0 \\ & \frac{1}{w_2} & & \\ & & \ddots & \\ 0 & & & \frac{1}{w_n} \end{bmatrix}$$

TABLA 5.8 Análisis de varianza para mínimos cuadrados generalizados

Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0
Regresión	$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{B}'\mathbf{z}$ $= \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$	p	SS_R/p	MS_R/MS_{Res}
Error	$SS_{Res} = \mathbf{z}'\mathbf{z} - \hat{\boldsymbol{\beta}}'\mathbf{B}'\mathbf{z}$ $= \mathbf{y}'\mathbf{V}^{-1}\mathbf{y}$ $-\mathbf{y}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$	$n - p$	$SS_{Res}/(n-p)$	
Total	$\mathbf{z}'\mathbf{z} = \mathbf{y}'\mathbf{V}^{-1}\mathbf{y}$	n		

al procedimiento de estimación se le suele llamar **mínimos cuadrados ponderados**. Sea $\mathbf{W} = \mathbf{V}^{-1}$. Como $\mathbf{V}$ es una matriz diagonal, $\mathbf{W}$ también es diagonal y sus elementos diagonales, **pesos**, o **factores de ponderación**, son $w_1, w_2, \ldots, w_n$. De acuerdo con la ecuación (5.18), las ecuaciones normales de mínimos cuadrados ponderados son

$$(\mathbf{X}'\mathbf{W}\mathbf{X})\,\hat{\beta} = \mathbf{X}'\mathbf{W}\mathbf{y}$$

Éste es un análogo, en regresión múltiple, de las ecuaciones normales de mínimos cuadrados ponderados para la regresión lineal simple, de la ecuación (5.12). En consecuencia,

$$\hat{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

es el **estimador ponderado de mínimos cuadrados**. Nótese que las observaciones con varianzas grandes tienen menos peso que las de varianzas pequeñas.

Los estimados por mínimos cuadrados ponderados se pueden obtener con facilidad con un programa ordinario de cómputo de mínimos cuadrados. Si se multiplica cada uno de los valores observados de la i-ésima observación (incluyendo el 1 de la ordenada al origen) por la raíz cuadrada del peso de esa observación, se obtendrá entonces un conjunto de datos transformados.

$$\mathbf{B} = \begin{bmatrix} 1\sqrt{w_1} & x_{11}\sqrt{w_1} & \cdots & x_{1k}\sqrt{w_1} \\ 1\sqrt{w_2} & x_{21}\sqrt{w_2} & \cdots & x_{2k}\sqrt{w_2} \\ \vdots & \vdots & & \vdots \\ 1\sqrt{w_n} & x_{n1}\sqrt{w_n} & \cdots & x_{nk}\sqrt{w_n} \end{bmatrix}, \quad \mathbf{z} = \begin{bmatrix} y_1\sqrt{w_1} \\ y_2\sqrt{w_2} \\ \vdots \\ y_n\sqrt{w_n} \end{bmatrix}$$

Ahora si se aplican los mínimos cuadrados ordinarios a esos datos transformados, se obtendrá

$$\hat{\beta} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'\mathbf{z} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$$

que es el estimado de β por mínimos cuadrados ponderados.

5.5.3 Algunos asuntos prácticos

Para usar mínimos cuadrados ponderados se deben conocer los pesos w_i. A veces se puede recurrir a la experiencia o conocimientos anteriores, o a la información de un modelo teórico, para determinar los pesos. Véase un ejemplo de este método en Weisberg [1985]. También, el análisis de residuales puede indicar que la varianza de los errores puede ser una función de uno de los regresores, por ejemplo, $\text{Var}(\varepsilon_i) = \sigma^2 x_{ij}$, de modo que $w_i = 1/x_{ij}$. En algunos casos, en realidad y_i es un promedio de n_i observaciones en x_i, y si todas las observaciones **originales** tienen varianza constante σ^2, entonces la varianza de y_1 es $\text{Var}(y_i) = \text{Var}(\varepsilon_i) = \sigma^2/n_i$, y los pesos se escogerían como $w_i = n_i$. A veces, la fuente principal de error es la del error de medición, y distintas observaciones se miden con distintos instrumentos de precisión desigual (pero bien estimada). En ese caso los pesos se podrían elegir inversamente proporcionales a las varianzas del error de medición. En muchos casos prácticos se podrán adivinar los pesos, hacer el análisis para después volver a estimar los pesos con base en los resultados. Pueden ser necesarias varias iteraciones.

Sebert, Montgomery y Rollier [1998] investigan el uso de **análisis de conglomerados** para determinar el conjunto de observaciones influyentes en regresión. El análisis de conglomerado es una técnica de varias variables para determinar grupos de observaciones semejantes. El procedimiento consiste en definir una medida de similaridad entre las observaciones para después aplicar una serie de reglas y clasificar las observaciones en grupos, con base entre sus similaridades interobservacionales. Usan un procedimiento de conglomerado de un solo eslabón (véase Johnson y Wichern [1992] y Everitt [1993]) aplicado a los residuales y a los valores ajustados por mínimos cuadrados para agrupar $n - m$ observaciones en un grupo "limpio" y un grupo potencialmente influyente de m observaciones. Este grupo se evalúa en subconjuntos de tamaño $1, 2, \ldots, m$, con la versión de la medida de distancia de Cook para observaciones múltiples. Los autores informan que este procedimiento es muy efectivo para determinar el subconjunto de observaciones influyentes. Hay cierto "exceso", esto es identificar demasiadas observaciones como influyentes, pero el empleo de la distancia de Cook elimina con eficiencia las observaciones no influyentes. Al estudiar nueve conjuntos de datos publicados, los autores informan que no hubo incidentes de "encubrimiento", esto es, no poder determinar el subconjunto correcto de puntos influyentes. También informan buenos resultados con un extenso estudio de desempeño, hecho con simulación Monte Carlo.

6.7 TRATAMIENTO DE LAS OBSERVACIONES INFLUYENTES

Los diagnósticos de balanceo e influencia son parte importante del arsenal de herramientas de quien construye modelos. Pretenden ofrecer al analista mejor comprensión de los datos y señalar cuáles observaciones merecen más escrutinio. ¿Cuánto esfuerzo se debe dedicar al estudio de esos puntos? Es probable que dependa de la cantidad de puntos influyentes identificados, de su impacto real sobre el modelo y de la importancia del problema de construcción del modelo. Si el lector dedicó un año a reunir 30 observaciones, es probable que se justifique mucho análisis de seguimiento para los puntos dudosos. Esto es válido en especial si se obtiene un resultado inesperado debido a una sola observación influyente.

¿Se deben desechar las observaciones influyentes? Esta pregunta se parece a la cuestión de descartar los valores atípicos. Por regla general, si hay un error al anotar un valor medido, o si el punto de la muestra realmente es inválido o no es parte de la población que se pretendía muestrear, será adecuado descartar la observación. Sin embargo, si el análisis indica que un punto influyente es una observación válida, no hay justificación para su eliminación.

Un "compromiso" entre eliminar una observación y retenerla es considerar una técnica de estimación que no sea tan sensible a los puntos influyentes como lo son los mínimos cuadrados. Esas técnicas **robustas** de estimación en esencia **aligeran** las observaciones en proporción con la magnitud o influencia residual, de tal modo que una observación muy influyente recibirá menos peso que en el ajuste de mínimos cuadrados. En el capítulo 11 se describirán métodos robustos de regresión.

PROBLEMAS

6.1 Hacer un análisis minucioso de influencia con los datos de pruebas de energía térmica que aparecen en la tabla B.2 del apéndice. Comentar los resultados.

segmentos del polinomio como las restricciones de continuidad que no degraden en forma sustancial el ajuste, usando los métodos estándar de prueba de hipótesis para regresión múltiple.

Como ilustración, se tiene una función local cúbica con un solo nudo en t, y sin restricciones de continuidad, como la siguiente:

$$E(y) = S(x) = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{03}x^3 + \beta_{10}(x - t)_+^0 + \beta_{11}(x - t)_+^1 + \beta_{12}(x - t)_+^2 + \beta_{13}(x - t)_+^3$$

Nótese que ni $S(x)$, $S'(x)$ ni $S''(x)$ son necesariamente continuas en t, por la presencia de los términos con β_{10}, β_{11} y β_{12} en el modelo. Para determinar si al imponer restricciones de continuidad se reduce la calidad del ajuste, se prueban las hipótesis H_0: $\beta_{10} = 0$ [continuidad de $S(x)$], H_0: $\beta_{10} = \beta_{11} = 0$ [continuidad de $S(x)$ y $S'(x)$] y H_0: $\beta_{10} = \beta_{11} = \beta_{12} = 0$ [continuidad de $S(x)$, $S'(x)$ y $S''(x)$]. Para determinar si la spline cúbica se ajusta a los datos mejor que un solo polinomio cúbico dentro del rango de x, sólo se prueba H_0: $\beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = 0$.

En Smith [1979] se presenta una excelente descripción de este método para ajustar funciones spline. Una desventaja potencial de este método es que la matriz $\mathbf{X}'\mathbf{X}$ se vuelve mal acondicionada si hay una gran cantidad de nudos. Este problema se puede resolver usando una representación distinta de la función spline, llamada **spline *B* cúbica**. Las spline B cúbicas se definen en función de las diferencias divididas

$$B_i(x) = \sum_{j=i-4}^{i} \left[\frac{(x - t_j)_+^3}{\prod_{\substack{m=i-4 \\ m \neq j}}^{i} (t_j - t_m)} \right], \quad i = 1, 2, \ldots, h + 4 \tag{7.5}$$

y

$$E(y) = S(x) = \sum_{i=1}^{h+4} \gamma_i B_i(x) \tag{7.6}$$

en donde γ_i, $i = 1, 2, \ldots, h + 4$ son parámetros por estimar. En la ecuación (7.5) hay ocho nudos adicionales, $t_{-3} < t_{-2} < t_{-1} < t_0$, y $t_{h+1} < t_{h+2} < t_{h+3} < t_{h+4}$. Se suelen igualar $t_0 = x_{\text{mín}}$ y $t_{h+1} = x_{\text{máx}}$; los demás nudos son arbitrarios. Para mayor documentación acerca de funciones spline véanse Buse y Lim [1977], Curry y Schoenberg [1966], Eubank [1988], Gallant y Fuller [1973], Hayes [1970, 1974], Poirer [1973, 1975] y Wold [1974].

Ejemplo 7.2 *Datos de caída de voltaje*

La caída de voltaje en la batería del motor de un misil guiado, que se observa durante el tiempo de vuelo del misil, se muestra en la tabla 7.3. El diagrama de dispersión de la figura 7.6 parece indicar que la caída de voltaje se comporta en forma distinta en diferentes intervalos de tiempo, por lo que se modelarán los datos con una spline cúbica usando dos nudos en $t_1 = 6.5$ y $t_2 = 13$ segundos después del lanzamiento, respectivamente. La colocación de los nudos concuerda en forma aproximada con los cambios de curso del proyectil

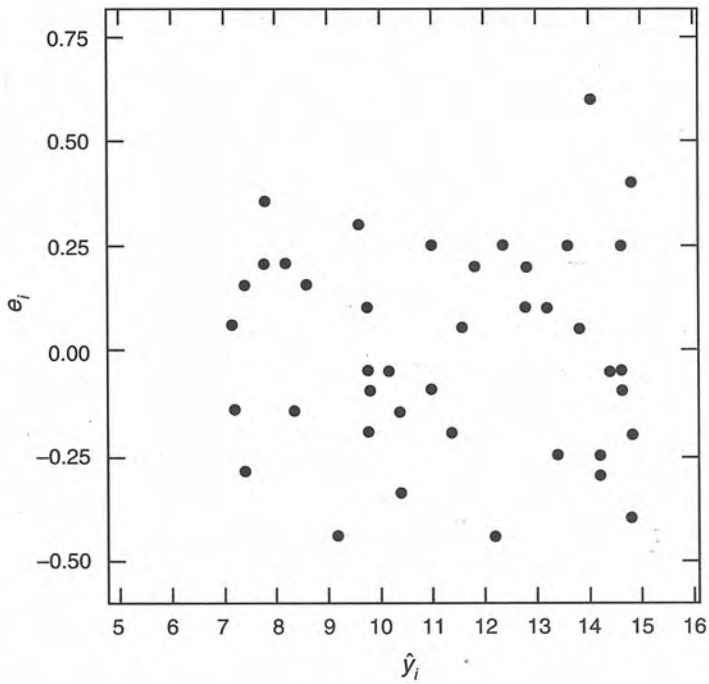


Figura 7.7 Gráfica de residuales e_i en función de valores ajustados $\hat{y}_i$, para el modelo con spline cúbica.

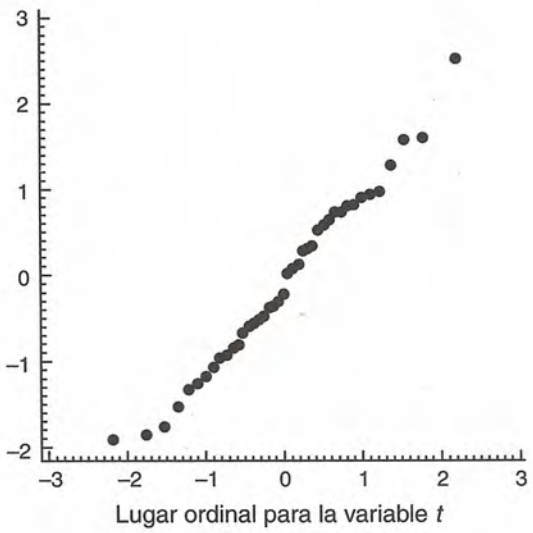


Figura 7.8 Gráfica de probabilidad normal de los residuales R de Student para el modelo con spline cúbica.

También se puede investigar si el modelo de spline cúbica mejora el ajuste, probando la hipótesis H_0: $\beta_1 = \beta_2 = 0$, con el método de la suma extra de cuadrados. La suma de cuadrados de la regresión para el polinomio cúbico es

$$SS_R(\beta_{01}, \beta_{02}, \beta_{03} | \beta_{00}) = 230.4444$$

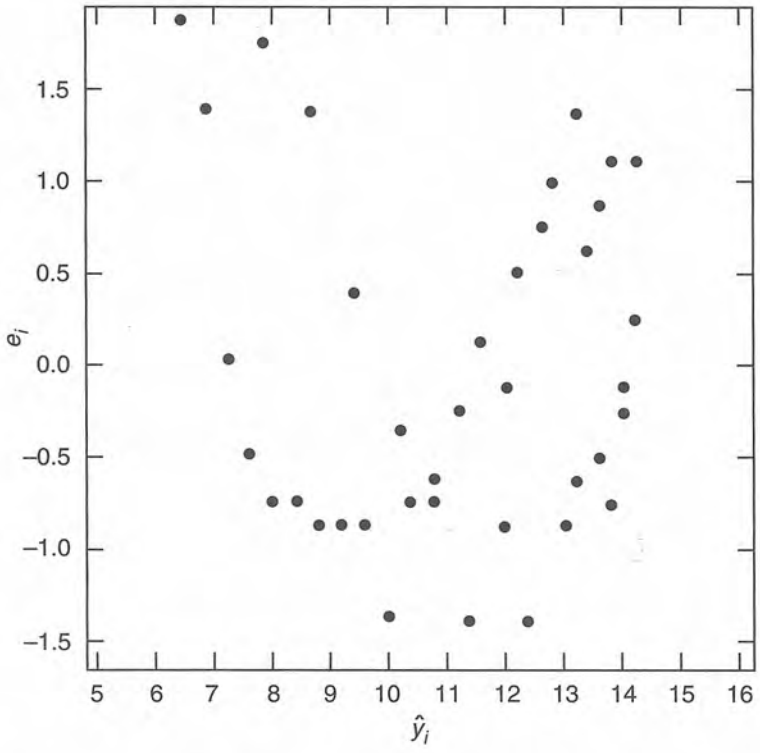


Figura 7.9 Gráfica de residuales e_i en función de valores ajustados $\hat{y}_i$, para el modelo de polinomio cúbico.

con tres grados de libertad. La suma extra de cuadrados para probar H_0: $\beta_1 = \beta_2 = 0$ es

$$\begin{aligned} SS_R(\beta_1, \beta_2|\beta_{00}, \beta_{01}, \beta_{02}, \beta_{03}) &= SS_R(\beta_{01}, \beta_{02}, \beta_{03}, \beta_2, \beta_2|\beta_{00}) \\ &\quad - SS_R(\beta_{01}, \beta_{02}, \beta_{03}|\beta_{00}) \\ &= 260.1784 - 230.4444 \\ &= 29.7340 \end{aligned}$$

con dos grados de libertad. Ya que

$$F_0 = \frac{SS_R(\beta_1, \beta_2|\beta_{00}, \beta_{01}, \beta_{02}, \beta_{03})/2}{MS_{Res}} = \frac{29.7340/2}{0.0717} = 207.35$$

que se compararía con la distribución $F_{2,35}$, se rechaza la hipótesis que H_0: $\beta_1 = \beta_2 = 0$. La conclusión es que el modelo de la spline cúbica proporciona mejor ajuste.

Ejemplo 7.3 *Regresión lineal por segmentos*

Un caso especial, de interés práctico, implica ajustar modelos de regresión lineal por segmentos. Esto se puede manejar con facilidad usando **spline lineales**, por ejemplo, supóngase que hay un solo nudo en t, y que en él podría haber un cambio de pendiente y una discontinuidad. El modelo de spline lineal resultante es

$$E(y) = S(x) = \beta_{00} + \beta_{01}x + \beta_{10}(x - t)_+^0 + \beta_{11}(x - t)_+^1$$

Ahora bien, si $x \leq t$, el modelo de línea recta es

$$E(y) = \beta_{00} + \beta_{01}x$$

y si $x > t$, el modelo es

$$\begin{aligned} E(y) &= \beta_{00} + \beta_{01}x + \beta_{10}(1) + \beta_{11}(x - t) \\ &= (\beta_{00} + \beta_{10} - \beta_{11}t) + (\beta_{01} + \beta_{11})x \end{aligned}$$

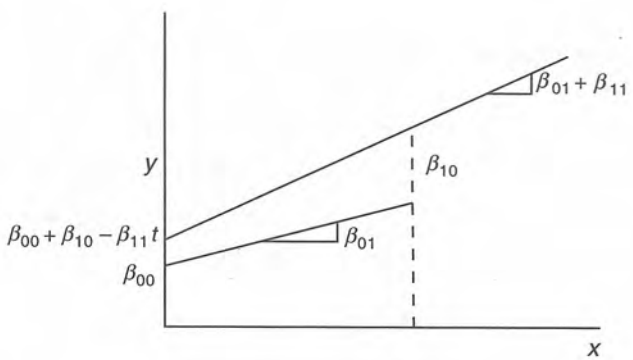
Esto es, si $x \leq t$, el modelo tiene ordenada al origen β_{00} y pendiente β_{01}, mientras que si $x > t$, la ordenada al origen es $\beta_{00} + \beta_{10} - \beta_{11}t$ y la pendiente es $\beta_{01} + \beta_{11}$. La función de regresión se muestra en la figura 7.10a. Nótese que el parámetro β_{10} representa la diferencia en la respuesta promedio en el nudo t.

Si se pidiera que la función de regresión sea continua en el nudo se obtendría una función más lisa. Eso se hace con facilidad eliminando el término $\beta_{10}(x - t)_+^0$ del modelo original, y se obtiene

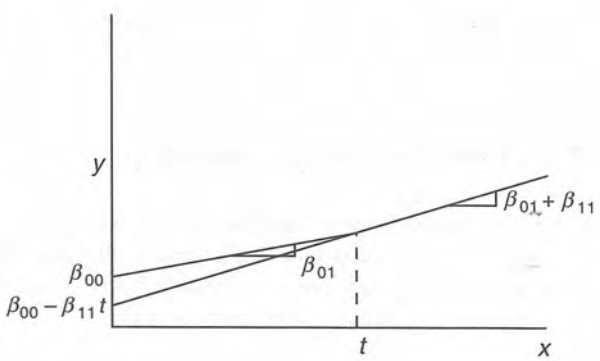
$$E(y) = S(x) = \beta_{00} + \beta_{01}x + \beta_{11}(x - t)_+^1$$

Ahora bien, si $x \leq t$, el modelo es

$$E(y) = \beta_{00} + \beta_{01}x$$



a)



b)

Figura 7.10 Regresión lineal por intervalos: a) discontinuidad en el nudo; b) modelo de regresión lineal continua por intervalos.

TABLA 7.12 Coeficientes de los polinomios ortogonales para el ejemplo 7.5

i	$P_0(x_i)$	$P_1(x_i)$	$P_2(x_i)$
1	1	−9	6
2	1	−7	2
3	1	−5	−1
4	1	−3	−3
5	1	−1	−4
6	1	1	−4
7	1	3	−3
8	1	5	−1
9	1	7	2
10	1	9	6
	$\sum_{i=1}^{10} P_0^2(x_i) = 10$	$\sum_{i=1}^{10} P_1^2(x_i) = 330$	$\sum_{i=1}^{10} P_2^2(x_i) = 132$
		$\lambda_1 = 2$	$\lambda_2 = 1/2$

Entonces,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} \sum_{i=1}^{10} P_0^2(x_i) & 0 & 0 \\ 0 & \sum_{i=1}^{10} P_1^2(x_i) & 0 \\ 0 & 0 & \sum_{i=1}^{10} P_2^2(x_i) \end{bmatrix} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 330 & 0 \\ 0 & 0 & 132 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} \sum_{i=1}^{10} P_0(x_i)y_i \\ \sum_{i=1}^{10} P_1(x_i)y_i \\ \sum_{i=1}^{10} P_2(x_i)y_i \end{bmatrix} = \begin{bmatrix} 3243 \\ 245 \\ 369 \end{bmatrix}$$

y

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{1}{10} & 0 & 0 \\ 0 & \frac{1}{330} & 0 \\ 0 & 0 & \frac{1}{132} \end{bmatrix} \begin{bmatrix} 3243 \\ 245 \\ 369 \end{bmatrix} = \begin{bmatrix} 324.3000 \\ 0.7424 \\ 2.7955 \end{bmatrix}$$

El modelo ajustado es

$$\hat{y} = 324.30 + 0.7424P_1(x) + 2.7955P_2(x)$$

TABLA 7.13 Análisis de varianza para el modelo cuadrático del ejemplo 7.5

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor P
Regresión	1213.43	2	606.72	159.24	< 0.0001
Lineal, α_1	(181.89)	1	181.89	47.74	< 0.0002
Cuadrática, α_2	(1031.54)	1	1031.54	270.75	< 0.0001
Residual	26.67	7	3.81		
Total	1240.10	9			

La suma de cuadrados de regresión es

$$SS_R(\alpha_1, \alpha_2) = \sum_{j=1}^{2} \hat{\alpha}_j \left[\sum_{i=1}^{10} P_j(x_i) y_i \right]$$
$$= 0.7424(245) + 2.7955(369)$$
$$= 181.89 + 1\,031.54 = 1\,213.43$$

El análisis de varianza se ve en la tabla 7.13. Tanto los términos lineales como los cuadráticos contribuyen en forma significativa al modelo. Como esos términos explican la mayor parte de la variación en los datos, se adoptará el modelo cuadrático en forma tentativa, a reserva de hacerle un análisis satisfactorio de residuales.

Se puede obtener una ecuación ajustada en términos del regresor original, sustituyendo $P_j(x_j)$ como sigue:

$$\hat{y} = 324.30 + 0.7424P_1(x) + 2.7955P_2(x)$$
$$= 324.30 + 0.7424(2)\left(\frac{x - 162.5}{25}\right)$$
$$+ 2.7955\frac{1}{2}\left[\left(\frac{x - 162.5}{25}\right)^2 - \frac{(10)^2 - 1}{12}\right]$$
$$= 312.7686 + 0.0594(x - 162.5) + 0.0022(x - 162.5)^2$$

Ésta es la forma de la ecuación que se debe proporcionar al usuario.

PROBLEMAS

7.1 Para los valores de x mostrados a continuación:

$$x = 1.00, 1.70, 1.25, 1.20, 1.45, 1.85, 1.60, 1.50, 1.95, 2.00$$

Supóngase que se desea ajustar un modelo de segundo orden, utilizando estos valores de la variable regresora x. Calcular la correlación entre x y x^2. ¿Se ven algunas dificultades potenciales para ajustar el modelo?

7.2 Un combustible sólido para cohetes pierde peso después de haber sido producido. Se dispone de los siguientes datos:

Meses después de producido, x	Pérdida de peso, y (kg)
0.25	1.42
0.50	1.39
0.75	1.55
1.00	1.89
1.25	2.43
1.50	3.15
1.75	4.05
2.00	5.15
2.25	6.43
2.50	7.89

a. Ajustar un polinomio de segundo orden que exprese la pérdida de peso en función de la cantidad de meses después de haber sido producido.
b. Pruebe la significancia de la regresión.
c. Pruebe la hipótesis H_0: $\beta_2 = 0$. Comente la necesidad del término cuadrático en este modelo.
d. ¿Hay riesgos potenciales al extrapolar con este modelo?

7.3 Acerca del problema 7.2, calcular los residuales del modelo de segundo orden. Analizar los residuales y comentar la adecuación del modelo.

7.4 Se tienen los datos siguientes:

x	y
4.00	24.60
4.00	24.71
4.00	23.90
5.00	39.50
5.00	39.60
6.00	57.12
6.50	67.11
6.50	67.24
6.75	67.15
7.00	77.87
7.10	80.11
7.30	84.67

a. Ajustar un modelo de polinomio de segundo orden a esos datos.
b. Probar la significancia de la regresión.
c. Probar la falta de ajuste y comentar sobre la adecuación del modelo de segundo orden.
d. Probar la hipótesis H_0: $\beta_2 = 0$. ¿Se puede eliminar el término cuadrático de esta ecuación?

7.5 Para el problema 7.4 calcular los residuales del modelo de segundo orden. Analizar los residuales y llegar a conclusiones acerca de la adecuación del modelo.

7.6 El grado de carbonatación de una bebida gaseosa se afecta por la temperatura del producto y por la presión de funcionamiento de la llenadora. Se obtuvieron 12 observaciones, y los datos resultantes se presentan a continuación.

Carbonatación, y	Temperatura, x_1	Presión, x_2
2.60	31.0	21.0
2.40	31.0	21.0
17.32	31.5	24.0
15.60	31.5	24.0
16.12	31.5	24.0
5.36	30.5	22.0
6.19	31.5	22.0
10.17	30.5	23.0
2.62	31.0	21.5
2.98	30.5	21.5
6.92	31.0	22.5
7.06	30.5	22.5

a. Ajustar un polinomio de segundo orden.
b. Probar la significancia de la regresión.
c. Probar la falta de ajuste y llegar a conclusiones.
d. ¿Contribuye al modelo el término de interacción, en forma significativa?
e. ¿Contribuyen al modelo los términos de segundo orden, en forma significativa?

7.7 Para el problema 7.6, calcular los residuales del modelo de segundo orden. Analizar esos residuales y comentar la adecuación del modelo.

7.8 Examínense los datos del problema 7.2.
a. Ajustar un modelo de segundo orden a esos datos, usando polinomios ortogonales.
b. Suponer que se desea investigar la adición de un término de tercer orden a este modelo. Comentar la necesidad de este término adicional. Respaldar las conclusiones con un análisis estadístico adecuado.

7.9 Suponer que se trata de ajustar el polinomio cuadrático por tramos, con un nudo en $x = t$:

$$E(y) = S(x) = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{10}(x - t)_+^0 + \beta_{11}(x - t)_+^1 + \beta_{12}(x - t)_+^2$$

a. Demostrar cómo se debe probar la hipótesis de que este modelo de spline cuadrática se ajusta a los datos, en forma significativamente mejor que un polinomio cuadrático ordinario.
b. El modelo de spline cuadrática no es continuo en el nudo t. ¿Se puede modificar el modelo para obtener continuidad en $x = t$?
c. Indicar cómo se puede modificar el modelo para que tanto $E(y)$ como $dE(y)/dx$ sean continuas en $x = t$.
d Comentar el significado de las restricciones de continuidad para el modelo, en las partes b y c. En la práctica, ¿cómo se seleccionaría el tipo de restricciones de continuidad que se van a imponer?

7.10 Se trata de ajustar un modelo polinomial por segmentos con tres segmentos: si $x < t_1$, el polinomio es lineal; si $t_1 \le x < t_2$, el polinomio es cuadrático, y si $x > t_2$, el polinomio es lineal. Considérese el modelo

$$\begin{aligned} E(y) = S(x) = \beta_{00} + \beta_{01}x + \beta_{02}x^2 + \beta_{10}(x - t_1)_+^0 + \beta_{11}(x - t_1)_+^1 \\ + \beta_{12}(x - t_1)_+^2 + \beta_{20}(x - t_2)_+^0 + \beta_{21}(x - t_2)_+^1 \\ + \beta_{22}(x - t_2)_+^2 \end{aligned}$$

a. ¿Satisface los requisitos este polinomio segmentado? Si no es así, indicar cómo se puede modificar para que los satisfaga.

b. Indicar cómo se modificaría el modelo segmentado para asegurar que $E(y)$ sea continua en los nudos t_1 y t_2.

c. Indicar cómo se modificaría el modelo segmentado para asegurar que tanto $E(y)$ como $dE(y)/dx$ sean continuas en los nudos t_1 y t_2.

7.11 Un analista de investigación de operaciones estudia la relación entre el tamaño del lote de producción, x, y el costo unitario promedio de producción, y. Un estudio de las operaciones recientes produjo los siguientes datos:

x	100	120	140	160	180	200	220	240	260	280	300
y	\$9.73	9.61	8.15	6.98	5.87	4.98	5.09	4.79	4.02	4.46	3.82

El analista cree que un modelo de regresión lineal por segmentos debe ajustar a esos datos. Estimar los parámetros de ese modelo, suponiendo que la pendiente de la recta cambia en $x = 200$ unidades. ¿Respaldan los datos el uso de ese modelo?

7.12 Modificar el modelo del problema 7.11 para investigar la posibilidad de que exista una discontinuidad en la función de regresión en $x = 200$ unidades. Estimar los parámetros en este modelo. Probar las hipótesis adecuadas para determinar si la función de regresión cambia, tanto de pendiente como de ordenada, en $x = 200$ unidades.

7.13 Para el modelo polinomial del problema 7.11, calcular los factores de inflación de varianza, y comentar sobre la multicolinealidad en este modelo.

7.14 Se tienen los datos del problema 7.2.

a. Ajustar un modelo de segundo orden $y = \beta_0 + \beta_1 x + \beta_{11}x^2 + \varepsilon$ a esos datos. Evaluar los factores de inflación de la varianza.

b. Ajustar un modelo de segundo orden $y = \beta_0 + \beta_1(x - \bar{x}) + \beta_{11}(x - \bar{x})^2 + \varepsilon$ a los datos. Evaluar los factores de inflación de la varianza.

c. ¿Cuál sería la conclusión acerca del impacto, sobre la multicolinealidad, de centrar las x en un modelo polinomial?

7.15 Con frecuencia, en ingeniería química y mecánica se necesita conocer la presión de vapor de agua a diversas temperaturas; para ello se pueden usar las "infames" tablas de vapor. A continuación se presentan datos de la presión de vapor y del agua a diversas temperaturas.

y = presión de vapor (mm Hg)	x = Temperatura (°C)
9.2	10
17.5	20
31.8	30
55.3	40
92.5	50
149.4	60

a. Ajustar un modelo de primer orden a los datos. Sobreponer el modelo ajustado al diagrama de dispersión de y en función de x. Comentar el ajuste aparente del modelo.

b. Preparar un diagrama de dispersión de valores ajustados de y en función de los valores de y observada. ¿Qué parece indicar ese diagrama acerca del ajuste del modelo?

c. Graficar los residuales en función de la y ajustada o predicha. Comentar la adecuación del modelo.

d. Ajustar un modelo de segundo orden a los datos. ¿Hay evidencia de que el término cuadrático sea estadísticamente importante?

e. Repetir las partes de a a c, usando el modelo de segundo orden. ¿Hay pruebas de que el modelo de segundo orden produce un mejor ajuste para los datos de presión de vapor?

7.16 Un artículo en la revista *Journal of Pharmaceutical Sciences* (**80**, 971-977, 1991) presenta datos sobre la solubilidad observada, en fracción molar, de un soluto a temperatura

Observación número	y	x_1	x_2	x_3
1	0.22200	7.3	0.0	0.0
2	0.39500	8.7	0.0	0.3
3	0.42200	8.8	0.7	1.0
4	0.43700	8.1	4.0	0.2
5	0.42800	9.0	0.5	1.0
6	0.46700	8.7	1.5	2.8
7	0.44400	9.3	2.1	1.0
8	0.37800	7.6	5.1	3.4
9	0.49400	10.0	0.0	0.3
10	0.45600	8.4	3.7	4.1
11	0.45200	9.3	3.6	2.0
12	0.11200	7.7	2.8	7.1
13	0.43200	9.8	4.2	2.0
14	0.10100	7.3	2.5	6.8
15	0.23200	8.5	2.0	6.6
16	0.30600	9.5	2.5	5.0
17	0.09230	7.4	2.8	7.8
18	0.11600	7.8	2.8	7.7
19	0.07640	7.7	3.0	8.0
20	0.43900	10.3	1.7	4.2
21	0.09440	7.8	3.3	8.5
22	0.11700	7.1	3.9	6.6
23	0.07260	7.7	4.3	9.5
24	0.04120	7.4	6.0	10.9
25	0.25100	7.3	2.0	5.2
26	0.00002	7.6	7.8	20.7

constante, junto con x_1 = solubilidad parcial de la dispersión, x_2 = solubilidad parcial dipolar y x_3 = solubilidad parcial de Hansen por puentes de hidrógeno. La respuesta y es el logaritmo negativo de la solubilidad en fracción mol.

a. Ajustar un modelo cuadrático completo a los datos.
b. Probar el significado de la regresión y formar las estadísticas t para cada parámetro del modelo. Interpretar esos resultados.
c. Graficar los residuales y comentar la adecuación del modelo.
d. Usar el método de suma extra de cuadrados para probar la contribución de todos los términos de segundo orden al modelo.

7.17 Véase el modelo cuadrático de regresión del problema 7.16. Determinar los factores de inflación de varianza, y comentar la multicolinealidad en ese modelo.

7.18 Para los datos de solubilidad del problema 7.16, suponer que un punto de interés es $x_1 = 8.0$, $x_2 = 3.0$ y $x_3 = 5.0$.

a. Para el modelo cuadrático del problema 7.16, pronosticar la respuesta en el punto de interés y determinar un intervalo de confianza de 95% para la respuesta media en ese punto.
b. Ajustar un modelo que sólo incluya los efectos principales, e interacciones de dos factores, de los datos de solubilidad. Usar este modelo para predecir la respuesta en el punto de interés. Determinar un intervalo de confianza de 95% para la respuesta media en ese punto.
c. Comparar las longitudes de los intervalos de confianza en las partes a y b. ¿Puede el lector llegar a conclusiones acerca del óptimo modelo de acuerdo con esta comparación?

7.19 A continuación se presentan datos de y = concentración de licor verde (g/l) y x = velocidad de la máquina de papel (pies/min) en una fábrica de papel kraft. Los datos se tomaron de una gráfica, de un artículo en la revista *Tappi Journal*, marzo de 1986.

y	16.0	15.8	15.6	15.5	14.8
x	1700	1720	1730	1740	1750
y	14.0	13.5	13.0	12.0	11.0
x	1760	1770	1780	1790	1795

a. Ajustar el modelo $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ a los datos.
b. Probar el significado de la regresión con $\alpha = 0.05$. ¿Cuáles son sus conclusiones?
c. Probar la contribución del término cuadrático al modelo, y la del término lineal, usando una dócima F. Si $\alpha = 0.05$, ¿qué conclusión se puede sacar?
d. Traficar los residuales del modelo. ¿Parece satisfactorio el ajuste del modelo?

7.20 Reconsiderar los datos del problema 7.19. Suponer que es importante predecir la respuesta en los puntos $x = 1\,750$ y $x = 1\,775$.

a. Calcular la respuesta predicha en esos puntos, y los intervalos de predicción de 95% para la respuesta observada en el futuro, en esos puntos.
b. Suponer que también se considera un modelo de primer orden. Ajustar ese modelo y determinar la respuesta predicha en esos puntos. Calcular los intervalos de predicción de 95% para la respuesta futura observada en esos puntos. ¿Permite lo anterior decidir cuál modelo se debe preferir?

CAPÍTULO 8

VARIABLES INDICADORAS

8.1 EL CONCEPTO GENERAL DE LAS VARIABLES INDICADORAS

Las variables que se emplean en el análisis de regresión se suelen llamar **variables cuantitativas**, lo que significa que las variables tienen una escala bien definida de medición. Las variables como temperatura, distancia, presión e ingreso son variables cuantitativas, en ocasiones es necesario usar variables **cualitativas** o **categóricas** como variables predictoras en la regresión. Como ejemplos de variables cualitativas o categóricas están los operadores, estado de empleo (empleado o desempleado), los turnos (diurno, mixto o nocturno) y el sexo (masculino o femenino). En general, una variable cualitativa no tiene escala natural de medida. Se debe asignar un conjunto de niveles a una variable cualitativa para tener en cuenta el efecto que pueda tener la variable sobre la respuesta, esto se hace usando **variables indicadoras;** a veces, a las variables indicadoras se les llama **variables ficticias**.

Supóngase que un ingeniero mecánico desea relacionar la vida útil y de una herramienta (un *buril*) en un torno, con la velocidad del torno, en revoluciones por minuto (x_1), y con la clase de buril que se usa. La segunda variable regresora, la clase de buril, es cualitativa y puede tener dos niveles (por ejemplo, los tipos A y B de buril). Se usará una variable indicadora que asuma los valores 0 y 1 para identificar las clases de la variable regresora "tipo de buril", sea

$$x_2 = \begin{cases} 0 \text{ si la observación procede de la herramienta tipo A} \\ 1 \text{ si la observación procede de la herramienta tipo B} \end{cases}$$

La elección de 0 y 1 para identificar los niveles de una variable cualitativa es arbitraria. Dos valores cualesquiera distintos para x_2 serían satisfactorios, aunque por lo general los mejores son 0 y 1.

Suponiendo que es adecuado un modelo de primer orden, se tiene

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{8.1}$$

Para interpretar los parámetros en este modelo, se examinará el primer tipo de buril, el A, para el cual $x_2 = 0$. El modelo de regresión se transforma en

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \varepsilon \end{aligned} \tag{8.2}$$

Así, la relación entre la vida útil y la velocidad del torno para el buril tipo A es una recta con ordenada al origen β_0 y pendiente β_1. Para el tipo de herramienta B, $x_2 = 1$ y

$$y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \varepsilon$$
$$= (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon \tag{8.3}$$

Esto es, para el buril de tipo B, la relación entre la vida de la herramienta y la velocidad del torno también es una recta β_1, pero con ordenada al origen $\beta_0 + \beta_2$.

Las dos funciones de respuesta se ven en la figura 8.1. Los modelos (8.2) y (8.3) describen dos líneas de regresión **paralelas**, esto es, dos rectas con una pendiente común β_1 y distintas ordenadas al origen. También, se supone que la varianza de los errores ε es igual para ambos tipos de buril, A y B. El parámetro β_2 expresa la diferencia de alturas entre las dos líneas de regresión, ya que, β_2 es una medida de la diferencia de vida media de herramienta que resulta de cambiar del tipo A al tipo B.

Se puede generalizar este método a factores cualitativos con cualquier cantidad de niveles, por ejemplo, supóngase que interesan tres tipos de herramienta: A, B y C, se requieren dos variables indicadoras, como x_2 y x_3, para manejar los tres niveles de tipo de herramienta en el modelo. Los niveles de las variables indicadoras son

x_2	x_3	
0	0	si la observación procede del buril tipo A
1	0	si la observación procede del buril tipo B
0	1	si la observación procede del buril tipo C

y el modelo de regresión es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

En general, una variable cualitativa con a niveles se representa con $a - 1$ variables indicadoras, y cada una asume los valores 0 y 1.

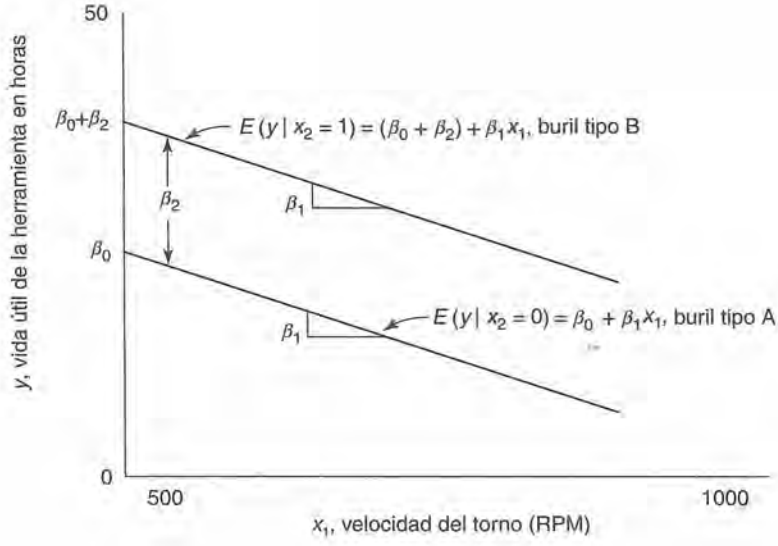


Figura 8.1 Funciones de respuesta para el ejemplo de la vida útil del buril.

Ejemplo 8.1 *Datos de vida de herramienta*

En la tabla 8.1 se presentan 20 observaciones de duración de herramienta y velocidad del torno, y el diagrama de dispersión se ve en la figura 8.2. Al revisar este diagrama de dispersión se ve que se requieren dos líneas de regresión para modelar bien estos datos, y que la ordenada al origen depende del tipo de herramienta que se usa; en consecuencia, se ajustará el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

TABLA 8.1 Datos, valores ajustados y residuales para el ejemplo 8.1

i	y_i(Horas)	x_{i1} (rpm)	Tipo de herramienta	$\hat{y}_i$	e_i
1	18.73	610	A	20.7552	−2.0252
2	14.52	950	A	11.7087	2.8113
3	17.43	720	A	17.8284	−0.3984
4	14.54	840	A	14.6355	−0.0955
5	13.44	980	A	10.9105	2.5295
6	24.39	530	A	22.8838	1.5062
7	13.34	680	A	18.8927	−5.5527
8	22.71	540	A	22.6177	0.0923
9	12.68	890	A	13.3052	−0.6252
10	19.32	730	A	17.5623	1.7577
11	30.16	670	B	34.1630	−4.0030
12	27.09	770	B	31.5023	−4.4123
13	25.40	880	B	28.5755	−3.1755
14	26.05	1000	B	25.3826	0.6674
15	33.49	760	B	31.7684	1.7216
16	35.62	590	B	36.2916	−0.6716
17	26.07	910	B	27.7773	−1.7073
18	36.78	650	B	34.6952	2.0848
19	34.95	810	B	30.4380	4.5120
20	43.67	500	B	38.6862	4.9838

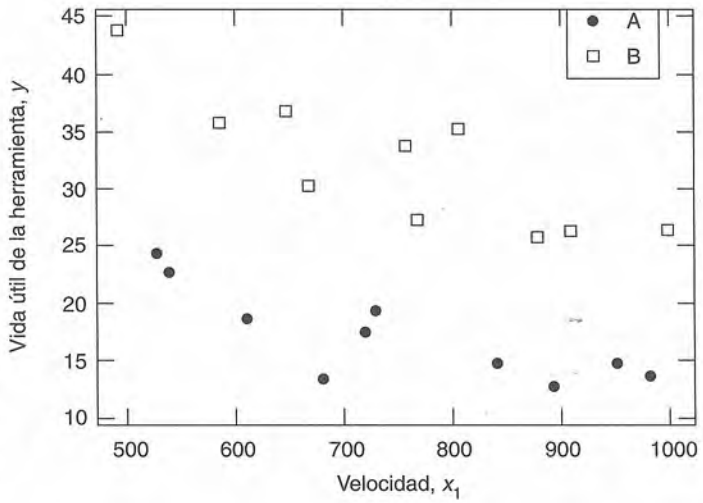


Figura 8.2 Gráfica de la vida útil de la herramienta y en función de la velocidad del torno x_1, para los tipos de buril A y B.

en donde la variable indicadora $x_2 = 0$ si la observación procede de la herramienta tipo A, y $x_2 = 1$ si procede de la herramienta tipo B. La matriz **X** y el vector **y** para ajustar este modelo son

$$\mathbf{X} = \begin{bmatrix} 1 & 610 & 0 \\ 1 & 950 & 0 \\ 1 & 720 & 0 \\ 1 & 840 & 0 \\ 1 & 980 & 0 \\ 1 & 530 & 0 \\ 1 & 680 & 0 \\ 1 & 540 & 0 \\ 1 & 890 & 0 \\ 1 & 730 & 0 \\ 1 & 670 & 1 \\ 1 & 770 & 1 \\ 1 & 880 & 1 \\ 1 & 1000 & 1 \\ 1 & 760 & 1 \\ 1 & 590 & 1 \\ 1 & 910 & 1 \\ 1 & 650 & 1 \\ 1 & 810 & 1 \\ 1 & 500 & 1 \end{bmatrix} \quad \text{y} \quad \mathbf{y} = \begin{bmatrix} 18.73 \\ 14.52 \\ 17.43 \\ 14.54 \\ 13.44 \\ 24.39 \\ 13.34 \\ 22.71 \\ 12.68 \\ 19.32 \\ 30.16 \\ 27.09 \\ 25.40 \\ 26.05 \\ 33.49 \\ 35.62 \\ 26.07 \\ 36.78 \\ 34.95 \\ 43.67 \end{bmatrix}$$

El ajuste por mínimos cuadrados es

$$\hat{y} = 36.986 - 0.027x_1 + 15.004x_2$$

En la tabla 8.2 se presenta el análisis de varianza y otras estadísticas de resumen para este modelo. Como el valor observado de F_0 tiene un valor P muy pequeño, se rechaza la hipótesis de significancia de la regresión, y en vista de que las estadísticas t para β_1 y β_2 tienen valores P pequeños, se llega a la conclusión de que los dos regresores, x_1 (rpm) y x_2 (tipo de herramienta) contribuyen al modelo. El parámetro β_2 es el cambio en la duración

TABLA 8.2 Estadísticas de resumen para el modelo de regresión del ejemplo 8.1

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor P
Regresión	1418.034	2	709.017	76.75	3.12×10^{-9}
Residual	157.055	17	9.239		
Total	1575.089	19			

Coeficiente	Estimado	Error estándar	t_0	Valor P
β_0	36.986			
β_1	−0.027	0.005	−5.887	8.97×10^{-6}
β_2	15.004	1.360	11.035	1.79×10^{-9}
	$R^2 = 0.9003$			

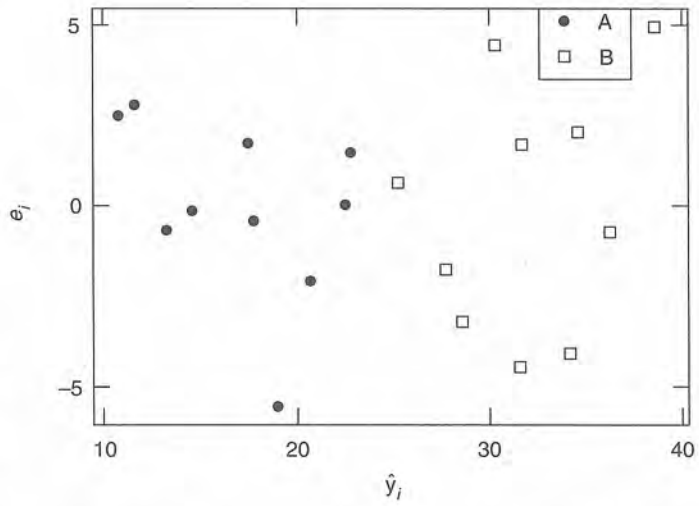


Figura 8.3 Gráfica de residuales e_i en función de valores ajustados $\hat{y}_i$, ejemplo 8.1.

promedio de la herramienta, debido a un cambio del tipo A al tipo B de herramienta. El intervalo de confianza de 95% para β_2 es

$$\hat{\beta}_2 - t_{0.025,\,17}\,\mathrm{se}\left(\hat{\beta}_2\right) \le \beta_2 \le \hat{\beta}_2 + t_{0.025,\,17}\,\mathrm{se}\left(\hat{\beta}_2\right)$$
$$= 15.004 - 2.110(1.360) \le \beta_2 \le 15.004 + 2.110(1.360)$$

o sea que

$$12.135 \le \beta_2 \le 17.873$$

En vista de lo anterior, se tiene el 95% de confianza que al cambiar del tipo de buril A al B aumenta la duración de la herramienta entre 12.135 horas y 17.873 horas.

Los valores ajustados $\hat{y}_i$ y los residuales e_i de este modelo se ven en las dos últimas columnas de la tabla 8.1. En la figura 8.3 se muestra la gráfica de los residuales en función de $\hat{y}_i$; en esta gráfica, se identifican por tipo de herramienta (A o B); si la varianza de los errores no fuera la misma para los dos tipos, se vería en esta gráfica. Nótese que los residuales "B" en la figura 8.3 muestran un poco más dispersión que los "A", y eso implica que puede haber un problema moderado de desigualdad de varianza. La figura 8.4 es la gráfica de probabilidad normal de los residuales. No hay indicios de inadecuación grave del modelo.

En vista de que se emplean dos líneas de regresión distintas para modelar la relación entre la vida útil del buril y la velocidad del torno, desde el principio se podrían haber ajustado dos modelos separados rectilíneos, en lugar de uno solo con una variable indicadora. Sin embargo, se prefiere el método con un solo modelo, porque el analista sólo tiene una ecuación final con la que trabajar, y no dos; es un resultado práctico mucho más simple, además, como se supone que las dos rectas tienen la misma pendiente, tiene sentido combinar los datos de ambos tipos para producir un solo estimado de este parámetro común; este método también proporciona una estimación de la varianza común del error σ^2, y se tienen más grados de libertad que los que resultarían de ajustar dos líneas separadas de regresión.

Ahora supóngase que se espera que las rectas de regresión que relacionan la vida útil del buril con la velocidad del torno difieren tanto en la ordenada al origen como en

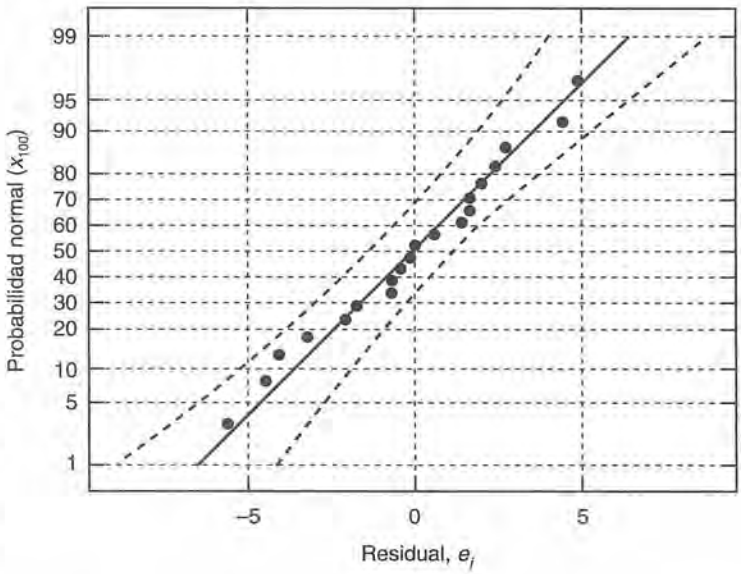


Figura 8.4 Gráfica de probabilidad normal de los residuales, ejemplo 8.1.

la pendiente. Es posible modelar este caso con una sola ecuación de regresión, usando variables indicadoras. El modelo es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon \tag{8.4}$$

Al comparar la ecuación (8.4) con la (8.1) se observa que se agregó al modelo un producto cruzado entre x_1, la velocidad del torno y la variable indicadora que representa el tipo de buril, x_2. Para interpretar los parámetros en este modelo, se examinará primero la herramienta tipo A, para la cual $x_2 = 0$. El modelo (8.4) se transforma en

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(0) + \beta_3 x_1(0) + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \varepsilon \end{aligned} \tag{8.5}$$

que es una recta con ordenada al origen β_0 y pendiente β_1. Para el buril tipo B, $x_2 = 1$ y

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(1) + \beta_3 x_1(1) + \varepsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \varepsilon \end{aligned} \tag{8.6}$$

Es un modelo rectilíneo con ordenada al origen $\beta_0 + \beta_2$ y pendiente $\beta_1 + \beta_3$. Las dos funciones de regresión se grafican en la figura 8.5. Nótese que la ecuación (8.4) define dos rectas de regresión con distintas pendientes y ordenadas al origen. En consecuencia, el parámetro β_2 refleja el cambio de ordenada al origen asociado con el cambio de buril tipo A a buril tipo B (las clases 0 y 1 de la variable indicadora x_2), y β_3 indica el cambio de pendiente asociado con el cambio de tipos de herramienta, de A a B.

El ajuste del modelo (8.4) equivale a ajustar dos ecuaciones de regresión separadas. Una ventaja del uso de variables indicadoras es que las pruebas de hipótesis se pueden hacer en forma directa, con el método de la suma extra de cuadrados. Por ejemplo, para probar si los dos modelos de regresión son idénticos, las hipótesis serían

$$H_0: \beta_2 = \beta_3 = 0$$
$$H_1: \beta_2 \neq 0 \text{ y/o } \beta_3 \neq 0.$$

Si no se rechaza H_0: $\beta_2 = \beta_3 = 0$, entonces un solo modelo de regresión puede explicar la relación entre la vida útil de la herramienta y la velocidad del torno. Para probar si las dos rectas de regresión tienen la misma pendiente pero quizá distintas ordenadas al origen, las hipótesis son

$$H_0: \beta_3 = 0$$
$$H_1: \beta_3 \neq 0$$

Si se usa el modelo (8.4), las dos rectas de regresión se pueden ajustar, y se pueden hacer esas pruebas en una sola corrida de computadora, siempre que el programa produzca las sumas de cuadrados $SS_R(\beta_1|\beta_0)$, $SS_R(\beta_2|\beta_0, \beta_1)$ y $SS_R(\beta_3|\beta_0, \beta_1, \beta_2)$.

Ejemplo 8.2 *Datos de duración de herramienta*

Se ajustará el modelo de regresión

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

a los datos de vida útil de herramienta de la tabla 8.1. La matriz **X** y el vector **y** para e modelo son:

$$\mathbf{X} = \begin{array}{c} \\ \end{array} \begin{array}{c} \begin{matrix} 1 \end{matrix} \\ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{array} \begin{array}{c} \begin{matrix} x_1 \end{matrix} \\ \begin{bmatrix} 610 \\ 950 \\ 720 \\ 840 \\ 980 \\ 530 \\ 680 \\ 540 \\ 890 \\ 730 \\ 670 \\ 770 \\ 880 \\ 1000 \\ 760 \\ 590 \\ 910 \\ 650 \\ 810 \\ 500 \end{bmatrix} \end{array} \begin{array}{c} \begin{matrix} x_2 \end{matrix} \\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \end{array} \begin{array}{c} \begin{matrix} x_1x_2 \end{matrix} \\ \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 670 \\ 770 \\ 880 \\ 1000 \\ 760 \\ 590 \\ 910 \\ 650 \\ 810 \\ 500 \end{bmatrix} \end{array}$$

y $$\mathbf{y} = \begin{bmatrix} 18.73 \\ 14.52 \\ 17.43 \\ 14.54 \\ 13.44 \\ 24.39 \\ 13.34 \\ 22.71 \\ 12.68 \\ 19.32 \\ 30.16 \\ 27.09 \\ 25.40 \\ 26.05 \\ 33.49 \\ 35.62 \\ 26.07 \\ 36.78 \\ 34.95 \\ 43.67 \end{bmatrix}$$

El modelo de regresión ajustado es

$$\hat{y} = 32.775 - 0.021\, x_1 + 23.971 x_2 - 0.012 x_1 x_2$$

En la tabla 8.3 se ve el resumen de análisis para este modelo. Para probar la hipótesis que las dos rectas de regresión son idénticas (H_0: $\beta_2 = \beta_3 = 0$), se usa la estadística

$$F_0 = \frac{SS_R(\beta_2, \beta_3|\beta_1, \beta_0)/2}{MS_{Res}}$$

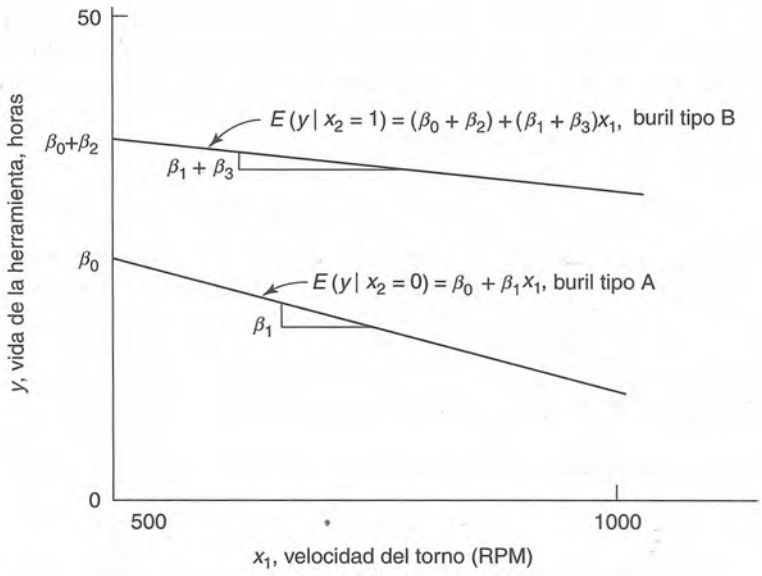


Figura 8.5 Funciones de respuesta para la ecuación (8.4).

En vista de que

$$\begin{aligned} SS_R(\beta_2, \beta_3 | \beta_1, \beta_0) &= SS_R(\beta_1, \beta_2, \beta_3 | \beta_0) - SS_R(\beta_1 | \beta_0) \\ &= 1434.112 - 293.005 \\ &= 1141.107 \end{aligned}$$

la estadística de prueba es

$$F_0 = \frac{SS_R(\beta_2, \beta_3 | \beta_1, \beta_0)/2}{MS_{Res}} = \frac{1141.107/2}{8.811} = 64.75$$

TABLA 8.3 Estadísticas de resumen para el modelo de regresión de la vida útil de herramientas, en el ejemplo 8.2

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0	Valor F
Tratamientos	1434.112	3	478.037	54.25	1.32×10^{-9}
Error	140.976	16	8.811		
Total	1575.088	19			

Coeficiente	Estimado	Error estándar	t_0	Suma de cuadrados
β_0	32.775			
β_1	–0.021	0.0061	–3.45	$SS_R(\beta_1 \mid \beta_0) = 293.005$
β_2	23.971	6.7690	3.54	$SS_R(\beta_2 \mid \beta_1, \beta_0) = 1125.029$
β_3	–0.012	0.0088	–1.35	$SS_R(\beta_3 \mid \beta_2, \beta_1, \beta_0) = 16.078$
		$R^2 = 0.9105$		

y como el valor P para esta estadística es $P = 2.14 \times 10^{-8}$, se concluye que los dos modelos de regresión no son idénticos. Para probar la hipótesis que las dos rectas tienen distintas ordenadas al origen y una pendiente común (H_0: $\beta_3 = 0$) se usa la estadística

$$F_0 = \frac{SS_R(\beta_3|\beta_2, \beta_1, \beta_0)/1}{MS_{Res}} = \frac{16.078}{8.811} = 1.82$$

y como el valor P para esta estadística es $P = 0.20$, se llega a la conclusión de que las pendientes de las dos rectas son iguales. Esto también se puede determinar con la estadística t para β_2 y β_3 en la tabla 8.3.

Las variables indicadoras son útiles en diversos casos de regresión. Ahora se presentarán más aplicaciones características de ellas.

Ejemplo 8.3 Una variable indicadora con más de dos niveles

Una empresa eléctrica está investigando el efecto que tiene el tamaño de la vivienda familiar y el tipo del acondicionamiento de aire que se usa en ella, sobre el consumo total de electricidad durante los meses calurosos. Sea y el consumo eléctrico total (en kilowatt-horas), durante el periodo de junio a septiembre, y x_1 el tamaño de la casa (pies cuadrados de construcción). Hay cuatro tipos de sistemas de acondicionamiento de aire: 1) sin acondicionamiento, 2) unidades de ventana, 3) bomba térmica, y 4) acondicionamiento central. Los cuatro niveles de ese factor se pueden modelar con tres variables indicadoras, x_2, x_3 y x_4, que se definen como sigue:

Tipo de acondicionamiento de aire	x_2	x_3	x_4
Sin acondicionamiento de aire	0	0	0
Unidades de ventana	1	0	0
Bomba térmica	0	1	0
Acondicionamiento central de aire	0	0	1

El modelo de regresión es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon \tag{8.7}$$

Si la casa no tiene acondicionamiento de aire, la ecuación (8.7) se transforma en

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Si la casa tiene unidades de ventanas, entonces

$$y = (\beta_0 + \beta_2) + \beta_1 x_1 + \varepsilon$$

Si la casa tiene bomba térmica, el modelo de regresión es

$$y = (\beta_0 + \beta_3) + \beta_1 x_1 + \varepsilon$$

y si la casa tiene acondicionamiento central, entonces

$$y = (\beta_0 + \beta_4) + \beta_1 x_1 + \varepsilon$$

Así, en el modelo (8.7) se supone que la relación entre el consumo eléctrico en tiempo caluroso, y el tamaño de la casa es lineal, y que la pendiente no depende del tipo de sistema de acondicionamiento de aire que se emplee. Los parámetros β_2, β_3 y β_4 modifican la altura (u ordenada al origen) del modelo de regresión para los distintos sistemas de acondicionamiento de aire. Esto es, β_2, β_3 y β_4 miden el efecto de las unidades de ventana, de bomba térmica y de acondicionamiento central, respectivamente, en comparación con la falta de acondicionamiento de aire; además, se pueden determinar otros efectos comparando en forma directa los coeficientes adecuados de regresión. Por ejemplo, $\beta_3 - \beta_4$ refleja la eficiencia relativa de una bomba térmica respecto al acondicionamiento central de aire, también nótese la hipótesis que la varianza del consumo de energía no depende del tipo de sistema de acondicionamiento usado; esta hipótesis puede ser inadecuada.

En este problema parece irreal suponer que la pendiente de la función de regresión que relaciona el consumo eléctrico medio con el tamaño de la vivienda no depende del tipo de sistema de acondicionamiento de aire. Por ejemplo, cabría esperar que el consumo eléctrico medio aumentará al aumentar el tamaño de la casa, pero la tasa de aumento debería ser distinta para un sistema central de acondicionamiento de aire que para las unidades de ventana, porque el primero debería ser más eficiente que las unidades de ventana para las casas más grandes. Esto es, debería haber una **interacción** entre el tamaño de la casa y la clase de sistema de acondicionamiento. Esto se puede incorporar al modelo ampliando la ecuación (8.7) para incluir términos de interacción. El modelo resultante es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \varepsilon \quad (8.8)$$

Los cuatro modelos de regresión, que corresponden a las cuatro clases de sistema de acondicionamiento de aire son:

$$\begin{aligned}
y &= \beta_0 + \beta_1 x_1 + \varepsilon_1 && \text{(sin acondicionamiento de aire)} \\
y &= (\beta_0 + \beta_2) + (\beta_1 + \beta_5) x_1 + \varepsilon && \text{(unidades de ventana)} \\
y &= (\beta_0 + \beta_3) + (\beta_1 + \beta_6) x_1 + \varepsilon && \text{(bomba térmica)} \\
y &= (\beta_0 + \beta_4) + (\beta_1 + \beta_7) x_1 + \varepsilon && \text{(acondicionamiento central de aire)}
\end{aligned}$$

Nótese que el modelo (8.8) implica que cada clase de sistema de acondicionamiento de aire puede tener una recta separada de regresión, con su pendiente y ordenada al origen correspondientes.

Ejemplo 8.4 Más de una variable indicadora

Con frecuencia hay varias variables cualitativas diferentes que se deben incorporar al modelo. Para ilustrarlo, supóngase que en el ejemplo 8.1 se debe considerar un segundo factor cualitativo, el tipo de lubricante de corte que se usa, suponiendo que este factor tiene dos niveles, se puede definir una segunda variable indicadora, x_3, como sigue:

$$x_3 = \begin{cases} 0 \text{ si se usa aceite de baja viscosidad} \\ 1 \text{ si se usa aceite de viscosidad intermedia} \end{cases}$$

Un modelo de regresión que relacione la vida útil de la herramienta (y) con la velocidad de corte (x_1), el tipo de herramienta (x_2) y el tipo de lubricante de corte (x_3) es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{8.9}$$

Es claro que la pendiente β_1 del modelo de regresión que relaciona la vida de la herramienta con la velocidad de corte no depende ni del tipo de herramienta ni del tipo de lubricante de corte. La ordenada al origen de la recta de regresión sí depende de esos factores, en una forma aditiva.

Se pueden agregar diversos tipos de efecto de interacción al modelo. Por ejemplo, supóngase que se consideran interacciones entre la velocidad de corte y los dos factores cualitativos, por lo que el modelo (8.9) se transforma en

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \varepsilon \tag{8.10}$$

Lo anterior conlleva la siguiente situación:

Tipo de buril	Lubricante de corte	Modelo de regresión
A	Baja viscosidad	$y = \beta_0 + \beta_1 x_1 + \varepsilon$
B	Baja viscosidad	$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \varepsilon$
A	Viscosidad intermedia	$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \varepsilon$
B	Viscosidad intermedia	$y = (\beta_0 + \beta_2 + \beta_3) + (\beta_1 + \beta_4 + \beta_5) x_1 + \varepsilon$

Nótese que cada combinación de tipo de buril y de lubricante de corte da como resultado una línea de regresión aparte, con distintas pendientes y ordenadas al origen, sin embargo, el modelo sigue siendo aditivo con respecto a los niveles de las variables de regresión. Esto es, al cambiar el lubricante de corte de viscosidad baja a intermedia, la ordenada al origen cambia en β_3, y la pendiente en β_5, independientemente de la clase de buril que se use.

Supóngase que se agrega un término de producto cruzado, que implica las dos variables indicadoras x_2 y x_3 al modelo, y se obtiene

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \varepsilon \tag{8.11}$$

En este caso se obtiene lo siguiente:

Tipo de buril	Lubricante de corte	Modelo de regresión
A	Baja viscosidad	$y = \beta_0 + \beta_1 x_1 + \varepsilon$
B	Baja viscosidad	$y = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) x_1 + \varepsilon$
A	Viscosidad intermedia	$y = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) x_1 + \varepsilon$
B	Viscosidad intermedia	$y = (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5) x_1 + \varepsilon$

La adición del término de producto cruzado $\beta_6 x_2 x_3$ en la ecuación (8.11) da como resultado que el efecto de una variable indicadora sobre la ordenada al origen dependa del nivel de la otra variable indicadora, lo que significa que al cambiar el lubricante de corte de viscosidad baja a intermedia, cambia la ordenada al origen en β_3, si se usa el buril del tipo A, pero ese mismo cambio de lubricante de corte cambia la ordenada al origen en $\beta_3 + \beta_6$

si se usa el buril del tipo B. Si se agregara un término de interacción $\beta_7 x_1 x_2 x_3$ al modelo (8.11), entonces al cambiar el lubricante de corte de viscosidad baja a viscosidad intermedia tendría un efecto **tanto** en la ordenada al origen **como** en la pendiente, y ese efecto dependería de la clase de buril empleado.

A menos que se disponga de información anterior sobre el efecto esperado del tipo de buril y de la viscosidad del lubricante de corte sobre la duración de la herramienta, habrá que dejar que los datos guíen a uno para seleccionar la forma correcta del modelo. Esto se puede hacer, en general, probando hipótesis sobre los coeficientes individuales de regresión, aplicando la prueba F parcial. Por ejemplo, la prueba de H_0: $\beta_6 = 0$ para el modelo (8.11) permitiría discriminar entre los dos modelos candidato, el (8.11) y el (8.10).

Ejemplo 8.5 *Comparación de modelos de regresión*

Se examinará el caso de la regresión lineal simple, en el que las n observaciones se pueden dividir en M grupos, y el m-ésimo grupo tiene n_m observaciones. El modelo más general consiste en M ecuaciones separadas, como por ejemplo

$$y = \beta_{0m} + \beta_{1mx} + \varepsilon, \qquad m = 1, 2, \ldots, M \tag{8.12}$$

Con frecuencia interesa comparar este modelo general con uno más restrictivo; las variables indicadoras son útiles en este respecto. Se considerarán los siguientes casos:

a. Líneas paralelas En este caso todas las M pendientes son idénticas, $\beta_{11} = \beta_{12} = \cdots = \beta_{1M}$, pero las ordenadas al origen pueden ser distintas, nótese que ésta es la clase de problema que se vio en el ejemplo 8.1 (en donde $M = 2$); condujo al uso de una variable indicadora. En forma más general se puede aplicar el método de la suma extra de cuadrados para probar la hipótesis H_0: $\beta_{11} = \beta_{12} = \cdots = \beta_{1M}$. Recuérdese que este procedimiento implica ajustar un **modelo completo** (FM) y un **modelo reducido** (RM) restringido a la hipótesis nula, y calcular la estadística F:

$$F_0 = \frac{\left[SS_{\text{Res}}(RM) - SS_{\text{Res}}(FM) \right] / (df_{RM} - df_{FM})}{SS_{\text{Res}}(FM) / df_{FM}} \tag{8.13}$$

Si el modelo reducido es tan satisfactorio como el modelo completo, entonces F_0 será pequeña en comparación con $F_{\alpha, df_{RM} - df_{FM}, df_{FM}}$. Los valores grandes de F_0 implican que el modelo reducido es inadecuado.

Para ajustar el modelo completo (8.12) tan sólo se ajustan M ecuaciones separadas de regresión, a continuación se calcula $SS_{\text{Res}}(FM)$ sumando las sumas de cuadrados de residuales obtenidas en cada regresión separada. Los grados de libertad para $SS_{\text{Res}}(FM)$ son $df_{FM} = \sum_{m=1}^{M} (n_m - 2) = n - 2M$. Para ajustar el modelo reducido se definen $M - 1$ variables indicadoras, $D_1, D_2, \ldots, D_{M-1}$ que corresponden a los M grupos, y entonces se ajusta

$$y = \beta_0 + \beta_1 x + \beta_2 D_1 + \beta_3 D_2 + \cdots + \beta_M D_{M-1} + \varepsilon$$

La suma de cuadrados de residuales de este modelo es $SS_{\text{Res}}(RM)$ con $df_{RM} = n - (M + 1)$ grados de libertad.

Si la prueba F, ecuación (8.13) indica que los M modelos de regresión tienen una pendiente común, entonces $\hat{\beta}_1$ obtenida en el modelo reducido es un estimado de este parámetro, que se determina agrupando o combinando todos los datos, esto se ilustró en el

ejemplo 8.1. En forma más general, el **análisis de covarianza** se usa para agrupar los datos para estimar la pendiente común. En consecuencia, el análisis de covarianza es un tipo especial de modelo lineal, que es una combinación de un modelo de regresión (con factores cuantitativos) con un modelo de análisis de varianza (con factores cualitativos). Para conocer una introducción al análisis de covarianza, véase Montgomery [1997].

b. Líneas concurrentes. En esta sección, las M ordenadas al origen son iguales, $\beta_{01} = \beta_{02} = \cdots = \beta_{0M}$, pero las pendientes pueden ser distintas. El modelo reducido es

$$y = \beta_0 + \beta_1 x + \beta_2 Z_1 + \beta_2 Z_2 + \cdots + \beta_M Z_{M-1} + \varepsilon$$

en donde $Z_k = xD_k$, $k = 1, 2, \ldots, M - 1$. La suma de cuadrados de residuales de este modelo es $SS_{\text{Res}}(RM)$ y $df_{RM} = n - (M + 1)$ grados de libertad, nótese que se está suponiendo la concurrencia en el origen. El caso más general, de concurrencia en cualquier punto x_0, se describe en Graybill [1976] y en Seber [1977].

c. Líneas coincidentes En este caso las M pendientes y las M ordenadas al origen son iguales, es decir, $\beta_{01} = \beta_{02} = \cdots = \beta_{0M}$, y $\beta_{11} = \beta_{12} = \cdots \beta_{1M}$. El modelo reducido es tan sólo

$$y = \beta_0 + \beta_1 x + \varepsilon$$

y la suma de cuadrados de residuales $SS_{\text{Res}}(RM)$ tiene $df_{RM} = n - 2$ grados de libertad. No son necesarias variables indicadoras en la prueba de coincidencia, pero se incluye este caso para completar la explicación.

8.2 COMENTARIOS SOBRE EL USO DE VARIABLES INDICADORAS

8.2.1 Variables indicadoras en función de la regresión con códigos asignados

Otro método para manejar una variable cualitativa en regresión es medir los niveles de la variable mediante un código asignado. Recuérdese que en el ejemplo 8.3, donde se está investigando el efecto del tamaño de la vivienda y la clase de sistema de acondicionamiento de aire sobre el consumo residencial, por parte de una empresa eléctrica. En lugar de usar tres variables indicadoras para representar los cuatro niveles del factor cualitativo clase de sistema de acondicionamiento de aire, se podría usar un factor cuantitativo, x_2, con el siguiente código asignado:

Tipo de acondicionamiento de aire	x_2
Sin acondicionamiento de aire	1
Unidades de ventana	2
Bombas térmicas	3
Acondicionamiento central de aire	4

Ahora se puede ajustar el modelo de regresión

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \tag{8.14}$$

en donde x_1 es el tamaño de la vivienda. Este modelo implica que

$$\begin{aligned} E(y|x_1, \text{sin acondicionamiento de aire}) &= \beta_0 + \beta_1 x_1 + \beta_2 \\ E(y|x_1, \text{unidades de ventana}) &= \beta_0 + \beta_1 x_1 + 2\beta_2 \\ E(y|x_1, \text{bomba térmica}) &= \beta_0 + \beta_1 x_1 + 3\beta_2 \\ E(y|x_1, \text{acondicionamiento de aire central}) &= \beta_0 + \beta_1 x_1 + 4\beta_2 \end{aligned}$$

Una consecuencia directa de lo anterior es que

$$\begin{aligned} &E(y|x_1, \text{acondicionamiento de aire central}) - E(y|x_1, \text{bomba térmica}) \\ &\quad = E(y|x_1, \text{bomba térmica}) - E(y|x_1, \text{unidades de ventana}) \\ &\quad = E(y|x_1, \text{unidades de ventana}) - E(y|x_1, \text{sin acondicionamiento de aire}) \\ &\quad = \beta_2 \end{aligned}$$

lo cual puede ser muy poco realista. Los códigos asignados imponen determinada métrica a los niveles del factor cualitativo. Otras opciones del código asignado implicarían diversas distancias entre los niveles del factor cualitativo, pero nadie garantiza que determinado código asignado lleve a un espaciamiento que sea adecuado.

Las variables indicadoras son más informativas para este problema, porque no forzan a determinada métrica sobre los niveles del factor cualitativo, además, la regresión que usa variables indicadoras siempre produce R^2 mayor que la regresión sobre los códigos asignados (por ejemplo, véase Searle y Udell [1970]).

8.2.2 Variables indicadoras como sustituto de un regresor cuantitativo

También se pueden representar regresores cuantitativos por medio de variables indicadoras. A veces esto se hace necesario, porque es difícil reunir información exacta del regresor cuantitativo. Considere el estudio sobre uso de energía eléctrica en el ejemplo 8.3, y supóngase que en el análisis se incluye un segundo regresor cuantitativo, el ingreso familiar. Como es difícil obtener esta información con precisión, se puede recopilar el regresor cuantitativo ingreso, agrupándolo en clases como las siguientes:

0 a 4 999 dólares
5 000 a 9 999 dólares
10 000 a 14 999 dólares
15 000 a 19 999 dólares
20 000 o más dólares

Ahora se puede representar el factor "ingreso" en el modelo usando cuatro variables regresoras.

Una desventaja de este método es que se requieren más parámetros para representar el contenido de información del factor cuantitativo. En general, si el regresor cuantitativo

se agrupa en a clases, se requerirán $a - 1$ parámetros, mientras que si se usara el regresor cuantitativo original sólo se requeriría un parámetro. Así, el manejar un factor cuantitativo como se hace con uno cualitativo, aumenta la complejidad del modelo. Este método también reduce los grados de libertad del error, aunque si los datos son numerosos, eso no es problema serio. Una ventaja del método de la variable indicadora es que no requiere que el analista haga hipótesis anteriores acerca de la forma funcional de la relación entre la respuesta y la variable regresora.

8.3 MÉTODO DE REGRESIÓN PARA ANÁLISIS DE VARIANZA

El **análisis de varianza** es una técnica que se usa con frecuencia para analizar los datos de **experimentos planeados** o **diseñados**. Aunque se suelen usar procedimientos especiales de cómputo para el análisis de varianza, cualquier problema de análisis de varianza también se puede manejar como un problema de regresión lineal. De ordinario no se recomienda usar métodos de regresión para el análisis de varianza, porque las técnicas de cómputo especializadas suelen ser muy eficientes, sin embargo, hay ciertos casos de análisis de varianza, en especial los que implican diseños no balanceados, donde puede ayudar el método de regresión, además, muchos analistas no se percatan de la estrecha conexión entre los dos procedimientos. En esencia, cualquier problema de análisis de varianza se puede manejar como un problema de regresión en el que todos los regresores son variables indicadoras.

En esta sección ilustraremos la alternativa de regresión para la clasificación unilateral, o análisis de varianza de un solo factor. Para conocer más ejemplos de la relación entre la regresión y el análisis de varianza, véanse Draper y Smith [1998], Montgomery [2001], Schilling [1974a, b] y Seber [1977].

El modelo para el análisis de varianza de clasificación de una vía es

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \; i = 1, 2, \ldots, k, \quad j = 1, 2, \ldots, n \tag{8.15}$$

en donde y_{ij} es la j-ésima observación del i-ésimo **tratamiento** o **nivel de factor**, μ es un parámetro común a los k tratamientos (que se suele llamar **media general**), τ_i es un parámetro que representa el efecto del i-ésimo tratamiento y ε_{ij} son los componentes del error, distribuidos normalmente con media cero y varianza σ^2. Se acostumbra definir los efectos del tratamiento en el caso balanceado (es decir, una cantidad igual de observaciones por tratamiento) como sigue:

$$\tau_1 + \tau_2 + \cdots + \tau_k = 0$$

Además, la media del i-ésimo tratamiento es $\mu_i = \mu + \tau_i$, $i = 1, 2, \ldots, k$. En el caso de efectos fijos (o modelo I), el análisis de varianza se usa para probar la hipótesis que todos las k medias poblacionales son iguales, o lo que es lo mismo,

$$H_0: \tau_1 = \tau_2 = \cdots = \tau_k = 0$$
$$H_1: \tau_i \neq 0 \text{ para al menos una } i \tag{8.16}$$

La tabla 8.4 muestra el análisis estándar de varianza con un solo factor, en este caso se tiene un término de error verdadero, y no un término residual, porque la replicación permite hacer una estimación del error independiente del modelo. La estadística de prueba

TABLA 8.4 Análisis unilateral de varianza

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	F_0
Tratamientos	$n\sum_{i=1}^{k}(\bar{y}_{i.}-\bar{y}_{..})^2$	$k-1$	$\frac{SS_{\text{Tratamientos}}}{k-1}$	$\frac{MS_{\text{Tratamientos}}}{MS_{\text{Res}}}$
Error	$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{i.})^2$	$k(n-1)$	$\frac{SS_{\text{Res}}}{k(n-1)}$	
Total	$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_{..})^2$	$kn-1$		

F_0 se compara con $F_{\alpha,\,k-1,\,k(n-1)}$. Si F_0 es mayor que este valor crítico, se rechaza la hipótesis nula H_0 de la ecuación (8.16); esto es, se llega a la conclusión que los k promedios de tratamiento no son iguales. Nótese que en la tabla 8.4 se ha empleado la notación estándar de "subíndice de punto", asociada al análisis de varianza, lo que indica que el promedio de las n observaciones en el i-ésimo tratamiento es

$$\bar{y}_{i.}=\frac{1}{n}\sum_{j=1}^{n}y_{ij},\quad i=1,2,\ldots,k$$

y el promedio general es

$$\bar{y}_{..}=\frac{1}{kn}\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij}$$

Para ilustrar la relación entre el análisis de varianza con efectos fijos y un solo factor, y la regresión, supóngase que se tienen $k=3$ tratamientos, de modo que la ecuación (8.15) se transforma en

$$y_{ij}=\mu+\tau_i+\varepsilon_{ij},\qquad i=1,2,3;\qquad j=1,3,\ldots,n \tag{8.17}$$

Esos tres tratamientos se pueden considerar como tres niveles de un **factor cualitativo** y se pueden manejar con variables indicadoras. En forma específica, un factor cualitativo con tres niveles necesitaría dos variables indicadoras, definidas como sigue:

$$x_1=\begin{cases}1 & \text{si la observación procede del tratamiento 1}\\ 0 & \text{en cualquier caso}\end{cases}$$

$$x_2=\begin{cases}1 & \text{si la observación procede del tratamiento 2}\\ 0 & \text{en cualquier caso}\end{cases}$$

Por consiguiente, el modelo de regresión viene a ser

$$y_{ij}=\beta_0+\beta_1x_{1j}+\beta_2x_{2j}+\varepsilon_{ij},\quad i=1,2,3;\quad j=1,2,\ldots,n \tag{8.18}$$

en donde x_{1j} es el valor de la variable indicadora x_1 para la observación j en el tratamiento i, y x_{2j} es el valor de x_2 para la observación j en el tratamiento i.

La relación entre el parámetro β_u ($u = 0, 1, 2$) en el modelo de regresión, y los parámetros μ y τ_i ($i = 1, 2, \ldots, k$) en el modelo de análisis de varianza se determina con facilidad.

Considérense las observaciones del tratamiento 1, para las cuales

$$x_{1j} = 1 \text{ y } x_{2j} = 0$$

El modelo de regresión (8.18) se transforma en

$$\begin{aligned} y_{1j} &= \beta_0 + \beta_1(1) + \beta_2(0) + \varepsilon_{1j} \\ &= \beta_0 + \beta_1 + \varepsilon_{1j} \end{aligned}$$

Como en el modelo de análisis de varianza una observación del tratamiento 1 se representa con $y_{1j} = \mu + \tau_1 + \varepsilon_{1j} = \mu_1 + \varepsilon_{1j}$, esto implica que

$$\beta_0 + \beta_1 = \mu_1$$

De igual modo, si las observaciones proceden del tratamiento 2, entonces $x_{1j} = 0$, $x_{2j} = 1$ y

$$\begin{aligned} y_{2j} &= \beta_0 + \beta_1(0) + \beta_2(1) + \varepsilon_{2j} \\ &= \beta_0 + \beta_2 + \varepsilon_{2j} \end{aligned}$$

En cuanto al modelo de análisis de varianza, $y_{2j} = \mu + \tau_2 + \varepsilon_{2j} = \mu_2 + \varepsilon_{2j}$, así que

$$\beta_0 + \beta_2 = \mu_2$$

Por último, en cuanto a las observaciones del tratamiento 3, ya que $x_{1j} = x_{2j} = 0$, el modelo de regresión se transforma en

$$\begin{aligned} y_{3j} &= \beta_0 + \beta_1(0) + \beta_2(0) + \varepsilon_{3j} \\ &= \beta_0 + \varepsilon_{3j} \end{aligned}$$

El modelo de análisis de varianza correspondiente es $y_{3j} = \mu + \tau_3 + \varepsilon_{3j} = \mu_3 + \varepsilon_{3j}$, así que

$$\beta_0 = \mu_3$$

Por lo anterior, en la formulación del modelo de regresión del análisis unifactorial de varianza, los coeficientes de regresión describen comparaciones de los dos primeros tratamientos μ_1 y μ_2 con la media μ_3 del tercer tratamiento. Esto es,

$$\begin{aligned} \beta_0 &= \mu_3 \\ \beta_1 &= \mu_1 - \mu_2 \\ \beta_2 &= \mu_2 - \mu_3 \end{aligned}$$

En general, si hay k tratamientos, el modelo de regresión para el análisis de varianza con un solo factor necesitará $k - 1$ variables indicadoras, por ejemplo,

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_{k-1} x_{k-1,j} + \varepsilon_{ij}$$
$$i = 1, 2, \ldots, k, \; j = 1, 2, \ldots, n \tag{8.19}$$

en donde

$$x_{ij} = \begin{cases} 1 \text{ si la observación } j \text{ procede del tratamiento } i \\ 0 \text{ en cualquier otro caso} \end{cases}$$

La relación entre los parámetros del modelo de regresión y los modelos de análisis de varianza es

$$\beta_0 = \mu_k$$
$$\beta_i = \mu_i - \mu_k, \quad i = 1, 2, \ldots, k - 1$$

Así, β_0 siempre estima la media del k-ésimo tratamiento, y β_i estima las diferencias en las medias del tratamiento i y del tratamiento k.

Ahora se verá el ajuste del modelo de regresión para el análisis de varianza de una vía. De nuevo, supóngase que se tienen $k = 3$ tratamientos, pero ahora sean $n = 3$ observaciones por tratamiento. La matriz $\mathbf{X}$ y el vector $\mathbf{y}$ son los siguientes:

$$\mathbf{y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \\ y_{31} \\ y_{32} \\ y_{33} \end{bmatrix} \qquad \mathbf{X} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \quad x_1 \quad x_2 \\ \end{array} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Nótese que la matriz $\mathbf{X}$ consiste sólo de 0 y 1. Ésta es una característica de la formulación de cualquier modelo de análisis de varianza por regresión. Las ecuaciones normales de mínimos cuadrados son

$$(\mathbf{X}'\mathbf{X})\,\hat{\beta} = \mathbf{X}'\mathbf{y}$$

o sea

$$\begin{bmatrix} 9 & 3 & 3 \\ 3 & 3 & 0 \\ 3 & 0 & 3 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{bmatrix}$$

en donde y_i es el total de todas las observaciones en el tratamiento i, y $y_{..}$ es el gran total de las nueve observaciones (es decir, $y_{..} = y_{1.} + y_{2.} + y_{3.}$). La solución de las ecuaciones normales es

$$\hat{\beta}_0 = \bar{y}_{..} - \bar{y}_{1.} - \bar{y}_{2.} = \bar{y}_{3.}$$

$$\hat{\beta}_1 = \bar{y}_{1.} - \bar{y}_{3.}$$

$$\hat{\beta}_2 = \bar{y}_{2.} - \bar{y}_{3.}$$

Se puede usar el método de la suma extra de cuadrados para probar si hay diferencias en las medias de los tratamientos. Para el modelo completo, la suma de cuadrados de regresión es

$$SS_R(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = \hat{\boldsymbol{\beta}}' \mathbf{X}'\mathbf{y} = \begin{bmatrix} \bar{y}_{3.}, \bar{y}_{1.} - \bar{y}_{3.}, \bar{y}_{2.} - \bar{y}_{3.} \end{bmatrix} \begin{bmatrix} y_{..} \\ y_{1.} \\ y_{2.} \end{bmatrix}$$

$$= y_{..}\bar{y}_{3.} + y_{1.}(\bar{y}_{1.} - \bar{y}_{3.}) + y_{2.}(\bar{y}_{2.} - \bar{y}_{3.})$$

$$= (y_{1.} + y_{2.} + y_{3.})\bar{y}_{3.} + y_{1.}(\bar{y}_{1.} - \bar{y}_{3.}) + y_{2.}(\bar{y}_{2.} - \bar{y}_{3.})$$

$$= \bar{y}_{1.}y_{1.} + \bar{y}_{2.}y_{2.} + \bar{y}_{3.}y_{3.}$$

$$= \sum_{i=1}^{3} \frac{y_{i.}^2}{3}$$

con tres grados de libertad. La suma de cuadrados residuales de error, para el modelo completo, es

$$SS_{Res} = \sum_{i=1}^{3}\sum_{j=1}^{3} y_{ij}^2 - SS_R(\beta_0, \beta_1, \beta_2)$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{3} y_{ij}^2 - \sum_{i=1}^{3} \frac{y_{i.}^2}{3}$$

$$= \sum_{i=1}^{3}\sum_{j=1}^{3} (y_{ij} - \bar{y}_{i.})^2 \qquad (8.20)$$

con $9 - 3 = 6$ grados de libertad. Nótese que la ecuación (8.20) es la suma de cuadrados de error en la tabla de análisis de varianza (Tabla 8.4) para $k = n = 3$.

La prueba de diferencias en las medias de tratamientos equivale a probar

$$H_0: \tau_1 = \tau_2 = \tau_3 = 0$$

$$H_1: \text{al menos una } \tau_i \neq 0$$

Si H_0 es cierta, los parámetros en el modelo de regresión vienen a ser

$$\beta_0 = \mu, \quad \beta_1 = 0, \quad \beta_2 = 0$$

Por consiguiente, el modelo reducido sólo contiene un parámetro, esto es

$$y_{ij} = \beta_0 + \varepsilon_{ij}$$

El estimado de β_0 en el modelo reducido es $\hat{\beta}_0 = \bar{y}_{..}$, y la suma de cuadrados de regresión con un grado de libertad para este modelo es

$$SS_R(\beta_0) = \frac{y_{..}^2}{9}$$

La suma de cuadrados para probar la igualdad de medias de los tratamientos es la diferencia en las sumas de cuadrados de regresión, entre los modelos completo y reducido, o sea

$$\begin{aligned} SS_R(\beta_1, \beta_2|\beta_0) &= SS_R(\beta_0, \beta_1, \beta_2) - SS_R(\beta_0) \\ &= \sum_{i=1}^{3} \frac{y_{i.}^2}{3} - \frac{y_{..}^2}{9} \\ &= 3\sum_{j=1}^{3} (\bar{y}_{i.} - \bar{y}_{..})^2 \end{aligned} \tag{8.21}$$

Esta suma tiene 3 – 1 = 2 grados de libertad. Nótese que la ecuación (8.21) es la suma de cuadrados de tratamiento, en la tabla 8.4, suponiendo que $k = n = 3$. La estadística de prueba adecuada es

$$\begin{aligned} F_0 &= \frac{SS_R(\beta_1, \beta_2|\beta_0)/2}{SS_{Res}/6} \\ &= \frac{3\sum_{i=1}^{3} (\bar{y}_{i.} - \bar{y}_{..})^2/2}{\sum_{i=1}^{3}\sum_{j=1}^{3} (y_{ij} - \bar{y}_{i.})^2/6} \\ &= \frac{MS_{Tratamientos}}{MS_{Res}} \end{aligned}$$

Si es cierta H_0: $\tau_1 = \tau_2 = \tau_3 = 0$, entonces F_0 sigue la distribución $F_{2,6}$. Es la misma estadística de prueba que aparece en la tabla de análisis de varianza (Tabla 8.4). Por consiguiente, el método de regresión es idéntico al procedimiento de análisis de una vía de varianza que se describió en la tabla 8.4.

PROBLEMAS

8.1 Para el modelo de regresión (8.8) descrito en el ejemplo 8.3, graficar la función de respuesta e indicar el papel que juegan los parámetros del modelo en la determinación de la forma de esta función.

8.2 Considérense los modelos de regresión descritos en el ejemplo 8.4.
- **a.** Graficar la función de respuesta asociada con la ecuación (8.10).
- **b.** Graficar la función de respuesta asociada con la ecuación (8.11).

8.3 Acerca de los datos del tiempo de entrega del ejemplo 3.1, en la sección 4.2.5 se hizo notar que esas observaciones se tomaron en cuatro ciudades: San Diego, Boston, Austin y Minneapolis.
- **a.** Desarrollar un modelo que relacione el tiempo de entrega y con las cajas x_1, la distancia x_2 y la ciudad donde se hizo la entrega. Estimar los parámetros del modelo.
- **b.** ¿Hay algún indicio de que el lugar de la entrega es una variable importante?
- **c.** Analizar los residuales de este modelo. ¿A qué conclusiones se puede llegar, acerca de la adecuación del modelo?

8.4 Para los datos de rendimiento de gasolina en coches, en la tabla B.3 del apéndice.
- **a.** Formar un modelo de regresión lineal que relacione el rendimiento de gasolina y con la cilindrada del motor x_1 y con el tipo de transmisión x_{11}. ¿Afecta en forma importante el tipo de transmisión al rendimiento de la gasolina?
- **b.** Modificar el modelo desarrollado en la parte a, para incluir una interacción entre la cilindrada del motor y el tipo de transmisión. ¿Qué conclusiones se pueden sacar acerca del efecto del tipo de transmisión sobre el rendimiento de la gasolina? Interpretar los parámetros en este modelo.

8.5 Para los datos de rendimiento de gasolina en coches, de la tabla B.3 del apéndice.
- **a.** Formar un modelo de regresión lineal que relacione el rendimiento de la gasolina y con el peso del vehículo x_{10}, y con el tipo de transmisión x_{11}. El tipo de transmisión ¿afecta el rendimiento de la gasolina en forma importante?
- **b.** Modificar el modelo desarrollado en la parte a, para incluir una interacción entre el peso del vehículo y el tipo de transmisión. ¿A qué conclusiones se puede llegar acerca del efecto del tipo de transmisión sobre el rendimiento de la gasolina? Interpretar los parámetros de este modelo.

8.6 Considérense los datos de la Liga Nacional de Futbol de la tabla B.1 del apéndice. Formular un modelo de regresión lineal que relacione la cantidad de juegos ganados con las yardas por tierra de los contrarios, x_8, el porcentaje de jugadas por tierra x_7 y una modificación del diferencial de pérdidas de balón, x_5. En forma específica, sea el diferencial de pérdidas de balón una variable indicadora cuyo valor se determina si ese diferencial real es positivo, negativo o cero. ¿Qué conclusiones se pueden sacar acerca del efecto de las pérdidas de balón sobre la cantidad de juegos ganados?

8.7 **Regresión lineal por segmentos.** En el ejemplo 7.3 se mostró cómo se puede ajustar un modelo de regresión lineal con un cambio de pendiente en algún punto t ($x_{\text{mín}} < t < x_{\text{máx}}$) usando funciones spline. Desarrollar una formulación del modelo de regresión lineal, por supuestos usando variables indicadoras. Suponer que la función es continua en el punto t.

8.8 **Continuación del problema 8.7.** Indicar cómo se pueden usar variables indicadoras para formar un modelo de regresión lineal por segmentos con una discontinuidad en el punto de unión t.

8.9 Suponer que un análisis unilateral de varianza implica cuatro tratamientos, pero que se ha tomado una cantidad distinta de observaciones (por ejemplo, n_i), para cada tratamiento. Suponer que $n_1 = 3$, $n_2 = 2$, $n_3 = 4$ y $n_4 = 3$, escribir el vector $\mathbf{y}$ y la matriz $\mathbf{X}$ para analizar esos datos en forma de un modelo de regresión múltiple. ¿Hay complicaciones introducidas por la naturaleza desbalanceada de esos datos?

8.10 Esquemas alternativos de codificación para el método de regresión de análisis de varianza. La ecuación (8.18) representa al modelo de regresión correspondiente a un análisis de varianza con tres tratamientos y n observaciones por tratamiento. Suponer que las variables indicadoras x_1 y x_2 se definen como sigue:

$$x_1 = \begin{cases} 1 & \text{si la observación procede del tratamiento 1} \\ -1 & \text{si la observación procede del tratamiento 2} \\ 0 & \text{en cualquier otro caso} \end{cases}$$

$$x_2 = \begin{cases} 1 & \text{si la observación procede del tratamiento 2} \\ -1 & \text{si la observación procede del tratamiento 3} \\ 0 & \text{en cualquier otro caso} \end{cases}$$

a. Demostrar que la relación entre los parámetros en los modelos de regresión y de análisis de varianza es

$$\beta_0 = \frac{\mu_1 + \mu_2 + \mu_3}{3} = \bar{\mu}$$

$$\beta_1 = \mu_1 - \bar{\mu}$$

$$\beta_2 = \mu_2 - \bar{\mu}$$

b. Escribir el vector **y** y la matriz **X**.

c. Desarrollar una suma adecuada de cuadrados para probar la hipótesis H_0: $\tau_1 = \tau_2 = \tau_3 = 0$. ¿Es ésta la suma de cuadrados que se maneja en forma acostumbrada en el análisis de varianza de una vía?

8.11 Montgomery [2001] presenta un experimento acerca de la resistencia a la tensión de las fibras sintéticas con que se fabrican telas para camisas de hombre. Se cree que la resistencia está influida por el porcentaje de algodón en la fibra. Los datos se ven a continuación.

Porcentaje de algodón	Resistencia a la tensión				
15	7	7	15	11	9
20	12	17	12	18	18
25	14	18	18	19	19
30	19	25	22	19	23
35	7	10	11	15	11

a. Escribir el factor **y** y la matriz **X** para el modelo de regresión correspondiente.

b. Determinar los estimados de los parámetros del modelo, por mínimos cuadrados.

c. Determinar un estimado de punto para la diferencia en resistencia media con 15% y 25% de algodón.

d. Probar la hipótesis que la resistencia media a la tensión es la misma para los cinco porcentajes de algodón.

8.12 Análisis de varianza de dos vías. Suponer que son de interés dos conjuntos distintos de tratamientos. Sea y_{ijk} la k-ésima observación en el nivel i del primer tipo de tratamiento y el nivel j del segundo tipo de tratamiento. El modelo de análisis de varianza de dos vías es

$$y_{ijk} = \mu + \tau_i + \gamma_j + (\tau\gamma)_{ij} + \varepsilon_{ijk}$$

$$i = 1, 2, \ldots, a, \quad j = 1, 2, \ldots, b, \quad k = 1, 2, \ldots, n,$$

en donde τ_i es el efecto del nivel i sobre el primer tipo de tratamiento, γ_j es el efecto del nivel j sobre el segundo tipo de tratamiento, $(\tau\gamma)_{ij}$ es un efecto de interacción entre los dos tipos de tratamiento y ε_{ijk} es un componente aleatorio de error NID(0, σ^2).

a. Para el caso $a = b = n = 2$, escribir un modelo de regresión que corresponda al análisis de varianza de dos vías.

b. ¿Cuáles son el vector **y** y la matriz **X** para este modelo de regresión?

c. Describir cómo se podría usar el modelo de regresión para probar la hipótesis H_0: $\tau_1 = \tau_2 = 0$ (las medias del tratamiento tipo 1 son iguales), H_0: $\gamma_1 = \gamma_2 = 0$ (las medias del tratamiento tipo 2 son iguales) y H_0: $(\tau\gamma)_{11} = (\tau\gamma)_{12} = (\tau\gamma)_{22} = 0$ (no hay interacción entre los tipos de tratamiento).

8.13 La tabla B.11 del apéndice presenta datos de la calidad del vino Pinot Noir.

a. Formular un modelo de regresión que relacione la calidad y con el aroma x_4, que incorpore la región de información que se ve en la última columna. ¿Tiene algún impacto esa región sobre la calidad del vino?

b. Hacer un análisis de residuales para este modelo, y comentar sobre la adecuación modelo.

c. ¿Hay algunos valores atípicos u observaciones influyentes en este conjunto de dat

d. Modificar el modelo de la parte a para incluir términos de interacción entre el aro y las variables de región. ¿Es mejor este modelo que el que se formó en la parte a?

8.14 Usar los datos de calidad de vino en la tabla B.11 del apéndice, para ajustar un modelo que relacione la calidad del vino y con el aroma x_4, usando la región como código de asignación, que toma los valores que se ven en la tabla: 1, 2 y 3. Describir la interpretación de los parámetros en este modelo. Comparar el modelo con uno que haya formado el lector usando variables indicadoras, en el problema 8.13.

CAPÍTULO 9

SELECCIÓN DE VARIABLE Y CONSTRUCCIÓN DEL MODELO

9.1 INTRODUCCIÓN

9.1.1 El problema de la construcción del modelo

En los capítulos anteriores se ha supuesto que se conoce que las variables regresoras incluidas en el modelo son importantes. Nuestro enfoque fue hacia las técnicas que aseguren que la forma funcional del modelo es la correcta y que no se violen las suposiciones básicas. En algunas aplicaciones, las consideraciones teóricas o la experiencia pueden ayudar a seleccionar los regresores que se van a usar en el modelo. Sin embargo, en la mayoría de los problemas prácticos el analista tiene un grupo de **regresores candidatos**, que deberían incluir a todos los factores influyentes, y debe determinar el subconjunto real de regresores que debe usarse en el modelo. La definición de un subconjunto adecuado de regresores para el modelo es lo que se llama **problema de selección de variable.**

La construcción de un modelo de regresión que sólo incluya un subconjunto de los regresores disponibles implica dos objetivos contrapuestos: 1) Se desea que el modelo incluya tantos regresores como sea posible, para que el contenido de información en ellos pueda influir sobre el valor predicho de y. 2) Se desea que el modelo incluya los menos regresores que sea posible, porque la varianza de la predicción $\hat{y}$ aumenta a medida que aumenta la cantidad de regresores. También, mientras más regresores haya en un modelo, los costos de recolección de datos y los de mantenimiento de modelo serán mayores. El proceso de encontrar un modelo que sea un término medio entre los dos objetivos se llama selección de la **"mejor" ecuación de regresión**. Desafortunadamente, como se verá en este capítulo, no hay definición única de "mejor", además hay varios algoritmos que se pueden usar para seleccionar variables, y esos procedimientos especifican como mejores, con frecuencia, subconjuntos distintos de los regresores candidatos.

El problema de selección de variables se suele describir en un entorno idealizado. Por lo general se supone conocida la especificación funcional correcta de los regresores (por ejemplo, $1/x_1$, $\ln x_2$), y que no hay valores atípicos ni observaciones influyentes. En la práctica rara vez se cumplen esas premisas. El **análisis de residuales**, como el que se describió en el capítulo 4, es útil para revelar formas funcionales de regresores, que se debieran investigar, o para señalar nuevos regresores candidatos, así como para identificar defectos en los datos, como son los valores atípicos. El efecto de las **observaciones influyentes** o de **gran balanceo** también se debe determinar. La investigación de la adecuación del modelo está ligada al problema de selección de variable; aunque en el caso ideal esos problemas se deben resolver al mismo tiempo, con frecuencia se usa un método iterativo, en el que 1) se emplea determinada estrategia de selección de variable, y 2) el modelo resultante para el