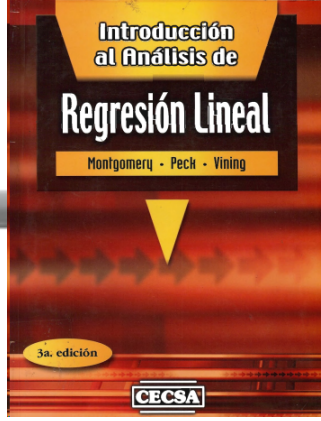

CAPÍTULO 2

REGRESIÓN LINEAL SIMPLE

2.1 MODELO DE REGRESIÓN LINEAL SIMPLE

Este capítulo explica el **modelo de regresión lineal simple**, un modelo con un solo regresor x que tiene una relación con una respuesta y, donde la relación es una línea recta. Este modelo de regresión lineal simple es

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.1}$$

donde la ordenada al origen β_0 y la pendiente β_1 son constantes desconocidas, y ε es un componente aleatorio de error. Se supone que los errores tienen promedio cero y varianza σ^2 desconocida. Además, se suele suponer que los errores no están correlacionados. Esto quiere decir que el valor de un error no depende del valor de cualquier otro error.

Conviene considerar que el regresor x está controlado por el analista de datos, y se puede medir con error despreciable, mientras que la respuesta y es una variable aleatoria. Con lo que hay una distribución de probabilidades de y para cada valor posible de x. La media de esta distribución es

$$E(y|x) = \beta_0 + \beta_1 x \tag{2.2a}$$

y la varianza es

$$\text{Var}(y|x) = \text{Var}(\beta_0 + \beta_1 x + \varepsilon) = \sigma^2 \tag{2.2b}$$

Así, la media de y es una función lineal de x, aunque la varianza de y no depende del valor de x. Además, ya que los errores no están correlacionados, las respuestas tampoco lo están.

A los parámetros β_0 y β_1 se les suele llamar **coeficientes de regresión**. Éstos tienen una interpretación simple y, frecuentemente, útil. La pendiente β_1 es el cambio de la media de la distribución de y producido por un cambio unitario en x. Si el intervalo de los datos incluye a $x = 0$, entonces la ordenada al origen, β_0, es la media de la distribución de la respuesta y cuando $x = 0$. Si no incluye al cero, β_0 no tiene interpretación práctica.

2.2 ESTIMACIÓN DE LOS PARÁMETROS POR MÍNIMOS CUADRADOS

Los parámetros β_0 y β_1 son desconocidos, y se deben estimar con los datos de la muestra. Supongamos que hay n pares de datos: (y_1, x_1), (y_2, x_2), . . . , (y_n, x_n). Como se indicó en el capítulo 1, esos datos pueden obtenerse en un experimento controlado, diseñado en forma específica para recolectarlos, o en un estudio observacional, o a partir de registros históricos existentes (lo que se llama un estudio retrospectivo).

2.2.1 Estimación de β_0 y β_1

Para estimar β_0 y β_1 se usa el **método de mínimos cuadrados**. Esto es, se estiman β_0 y β_1 tales que la suma de los cuadrados de las diferencias entre las observaciones y_i y la línea recta sea mínima. Según la ecuación (2.1), se puede escribir

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n \tag{2.3}$$

Se puede considerar que la ecuación (2.1) es un **modelo poblacional de regresión**, mientras que la ecuación (2.3) es un **modelo muestral de regresión**, escritos en términos de los n pares de datos (y_i, x_i) $(i = 1, 2, \ldots, n)$. Así, el criterio de mínimos cuadrados es

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{2.4}$$

Los estimadores, por mínimos cuadrados, de β_0 y β_1, que se designarán por $\hat{\beta}_0$ y $\hat{\beta}_1$, deben satisfacer

$$\left.\frac{\partial S}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0$$

y

$$\left.\frac{\partial S}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0$$

Se simplifican estas dos ecuaciones y se obtiene

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned} \tag{2.5}$$

Las ecuaciones (2.5) son llamadas **ecuaciones normales de mínimos cuadrados**. Su solución es la siguiente:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.6}$$

y

$$\hat{\beta}_1 = \frac{\displaystyle\sum_{i=1}^{n} y_i x_i - \frac{\left(\displaystyle\sum_{i=1}^{n} y_i\right)\left(\displaystyle\sum_{i=1}^{n} x_i\right)}{n}}{\displaystyle\sum_{i=1}^{n} x_i^2 - \frac{\left(\displaystyle\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{2.7}$$

en donde

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \text{y} \qquad \bar{x} \quad \frac{1}{n}\sum_{i=1}^{n} x_i$$

son los promedios de y_i y x_i, respectivamente. Por consiguiente, $\hat{\beta}_0$ y $\hat{\beta}_1$ en las ecuaciones (2.6) y (2.7) son los **estimadores por mínimos cuadrados** de la ordenada al origen y la pendiente, respectivamente. El modelo ajustado de regresión lineal simple es, entonces,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{2.8}$$

La ecuación (2.8) produce un estimado puntual, de la media de y para una determinada x.

Como el denominador de la ecuación (2.7) es la suma corregida de cuadrados de las x_i y el numerador es la suma corregida de los productos cruzados de x_i y y_i, estas ecuaciones se pueden escribir en una forma más compacta como sigue:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2.9}$$

y

$$S_{xy} = \sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n} = \sum_{i=1}^{n} y_i(x_i - \bar{x}) \tag{2.10}$$

Entonces, una forma cómoda de escribir la ecuación (2.7) es

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{2.11}$$

La diferencia entre el valor observado y_i y el valor ajustado correspondiente $\hat{y}_i$ se llama **residual**. Matemáticamente, el i-ésimo residual es

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1, 2, \ldots, n \tag{2.12}$$

Los residuales tienen un papel importante para investigar la **adecuación** del modelo de regresión ajustado, y para detectar diferencias respecto a las hipótesis básicas. Este tema se describirá en capítulos posteriores.

Ejemplo 2.1 *Datos del propelente*

Un motor cohete se forma pegando entre sí un propelente de ignición y un propelente de sostenimiento dentro de una caja metálica. La resistencia al corte de la pegadura entre los dos propelentes es una característica importante de la calidad. Se cree que la resistencia al corte se relaciona con la edad, en semanas, del lote del propelente de sostenimiento. Se

hicieron 20 observaciones de resistencia al corte y la edad del lote correspondiente de propelente, y se ven en la tabla 2.1. El diagrama de dispersión que se ve en la figura 2.1 parece indicar que hay una fuerte relación estadística entre la resistencia al cortante y la edad del propelente, y que parece razonable la hipótesis tentativa del modelo de línea recta, $y = \beta_0 + \beta_1 x + \varepsilon$.

TABLA 2.1 Datos para el ejemplo 2.1

Observación i	Resistencia al corte (psi) y_i	Edad del propelente (semanas) x_i
1	2158.70	15.50
2	1678.15	23.75
3	2316.00	8.00
4	2061.30	17.00
5	2207.50	5.50
6	1708.30	19.00
7	1784.70	24.00
8	2575.00	2.50
9	2357.90	7.50
10	2256.70	11.00
11	2165.20	13.00
12	2399.55	3.75
13	1779.80	25.00
14	2336.75	9.75
15	1765.30	22.00
16	2053.50	18.00
17	2414.40	6.00
18	2200.50	12.50
19	2654.20	2.00
20	1753.70	21.50

Para estimar los parámetros del modelo se calcula primero:

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 4\,677.69 - \frac{71\,422.56}{20} = 1\,106.56$$

y

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} = 528\,492.64 - \frac{(267.25)\,(42\,627.15)}{20}$$

$$= -41\,112.65$$

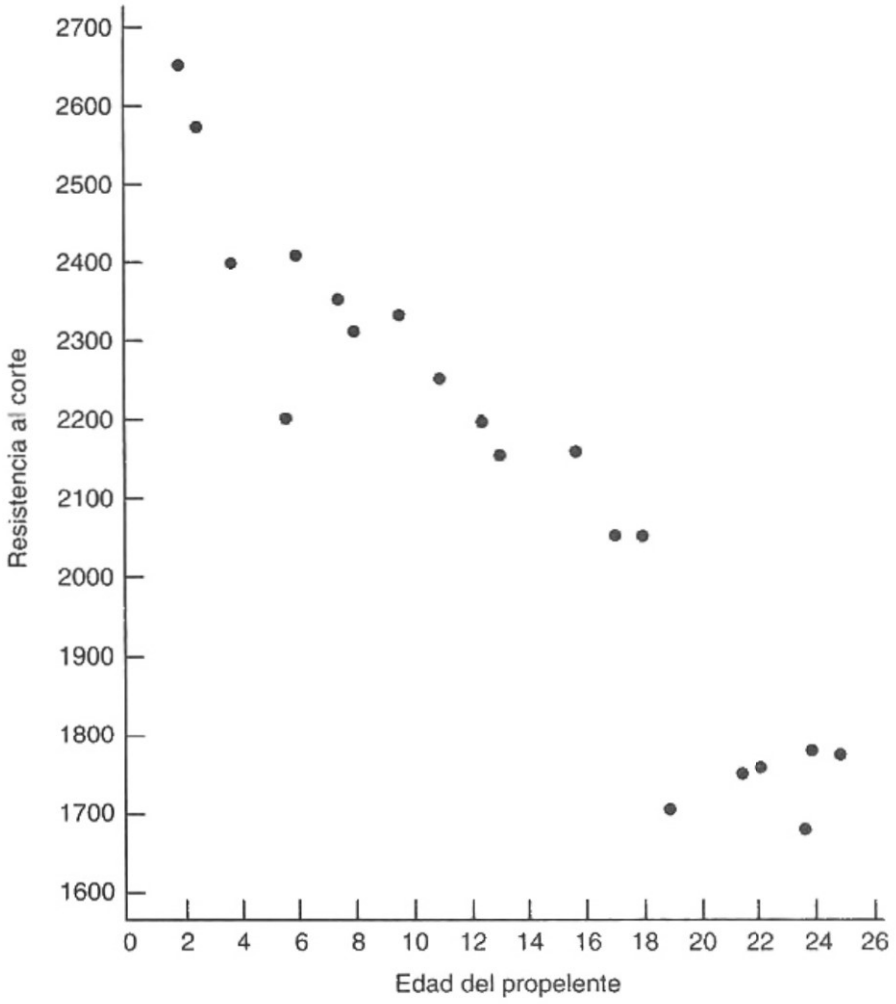

Figura 2.1 Diagrama de dispersión de la resistencia cortante en función de la edad del propelente. Ejemplo 2.1.

Por consiguiente, según las ecuaciones (2.11) y (2.6), se ve que

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-41\,112.65}{1\,106.56} = -37.15$$

y

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 2\,131.3575 - (-37.15)13.3625 = 2\,627.82$$

El ajuste de mínimos cuadrados es

$$\hat{y} = 2\,627.82 - 37.15x$$

Se puede interpretar que la pendiente de –37.15 es la disminución semanal promedio de resistencia del propelente al corte, debido a su edad. Como el límite inferior de las x está cerca del origen, la ordenada al origen de 2 627.82 representa la resistencia al corte de un lote de propelente inmediatamente después de ser fabricado. La tabla 2.2 muestra los valores observados y_i, los valores calculados o "ajustados" $\hat{y}_i$ y los residuales.

TABLA 2.2 Datos, valores ajustados y residuales para el ejemplo 2.1

Valor observado, y_i	Valor ajustado, $\hat{y}_i$	Residual, e_i
2158.70	2051.94	106.76
1678.15	1745.42	−67.27
2316.00	2330.59	−14.59
2061.30	1996.21	65.09
2207.50	2423.48	−215.98
1708.30	1921.90	−213.98
1784.70	1736.14	48.56
2575.00	2534.94	40.06
2357.90	2349.17	8.73
2256.70	2219.13	37.57
2165.20	2144.83	20.37
2399.55	2488.50	−88.95
1779.80	1698.98	80.82
2336.75	2265.58	71.17
1765.30	1810.44	−45.14
2053.50	1959.06	94.44
2414.40	2404.90	9.50
2200.50	2163.40	37.10
2654.20	2553.52	100.68
1753.70	1829.02	−75.32
$\Sigma y_i = 42627.15$	$\Sigma \hat{y}_i = 42627.15$	$\Sigma e_i = 0.00$

Después de obtener el ajuste por mínimos cuadrados, surgen varias preguntas interesantes:

1. ¿Qué tan bien se ajusta esta ecuación a los datos?
2. ¿Es probable que el modelo sea útil como predictor?
3. ¿Se viola alguna de las hipótesis básicas (como la de varianza constante y la de errores no correlacionados)? y en caso afirmativo, ¿qué tan grave es eso?

Se deben investigar todos estos asuntos antes de adoptar al modelo en forma definitiva y usarlo. Como se dijo anteriormente, los residuales juegan un papel clave para evaluar la adecuación del modelo. Se puede considerar que los residuales son realizaciones de los errores ε_i del modelo. Así, para comprobar la constancia de la varianza y la hipótesis de errores no correlacionados, uno se debe preguntar si los residuales parecen ser realmente una muestra aleatoria de una distribución con esas propiedades. Regresaremos a estas cuestiones en el capítulo 4, al investigar el uso de los residuales en la comprobación de la adecuación del modelo.

Resultado en computadora

Los paquetes de programas de cómputo se usan mucho para ajustar modelos de regresión. Las rutinas de regresión se encuentran tanto en programas estadísticos para computadora central, como para PC, y también en muchos de los paquetes más usados de hojas de cálculo. La tabla 2.3 representa el resultado obtenido con Minitab, un paquete estadístico para PC de uso

muy frecuente, con los datos del propelente de cohetes del ejemplo 2.1. La parte superior de la tabla contiene el modelo ajustado de regresión. Obsérvese que antes de redondear, los coeficientes de regresión concuerdan con los que se calcularon manualmente. También, la tabla 2.3 contiene más información sobre el modelo de regresión. Regresaremos a este resultado y explicaremos esas cantidades más adelante.

TABLA 2.3 Resultados de regresión para el ejemplo 2.1 con Minitab

Regression Analysis

The regression equation is
Strength = 2628 – 37.2 Age

Predictor	Coef	StDev	T	P
Constant	2627.82	44.18	59.47	0.000
Age	–37.154	2.889	–12.86	0.000
S = 96.11	R-Sq = 90.2%		R-Sq(adj) = 89.6%	

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	1527483	1527483	165.38	0.000
Error	18	166255	9236		
Total	19	1693738			

2.2.2 Propiedades de los estimadores por mínimos cuadrados y el modelo ajustado de regresión

Los estimadores por cuadrados mínimos $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen algunas propiedades importantes. Primero, obsérvese que, según las ecuaciones (2.6) y (2.7), $\hat{\beta}_0$ y $\hat{\beta}_1$ son **combinaciones lineales** de las observaciones y_i. Por ejemplo

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} c_i y_i$$

donde $c_i = (x_i - \bar{x})/S_{xx}$, para $i = 1, 2, \ldots, n$.

Los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ por mínimos cuadrados son **estimadores insesgados** de los parámetros β_0 y β_1 del modelo. Para demostrarlo con $\hat{\beta}_1$, considérese

$$E(\hat{\beta}_1) = E\left(\sum_{i=1}^{n} c_i y_i\right) = \sum_{i=1}^{n} c_i E(y_i)$$

$$= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i$$

ya que, se supuso que, $E(\varepsilon_i) = 0$. Ahora se puede demostrar en forma directa que $\sum_{i=1}^{n} c_i = 0$ y que $\sum_{i=1}^{n} c_i x_i = 1$, y entonces

$$E(\hat{\beta}_1) = \beta_1$$

Esto es, si se supone que el modelo es correcto [que $E(y_i) = \beta_0 + \beta_1 x_1$], entonces $\hat{\beta}_1$ es un estimador insesgado de β_1. De igual manera se puede demostrar que $\hat{\beta}_0$ es un estimador insesgado de β_0, es decir,

$$E\left(\hat{\beta}_0\right) = \beta_0$$

La varianza de $\hat{\beta}_1$ se calcula como sigue:

$$\begin{aligned} \operatorname{Var}\left(\hat{\beta}_1\right) &= \operatorname{Var}\left(\sum_{i=1}^{n} c_i y_i\right) \\ &= \sum_{i=1}^{n} c_i^2 \operatorname{Var}(y_i) \end{aligned} \tag{2.13}$$

ya que las observaciones y_i son no correlacionadas, por lo que la varianza de la suma es igual a la suma de las varianzas. La varianza de cada término en la suma es $c_i^2 \operatorname{Var}(y_i)$ y hemos supuesto que $\operatorname{Var}(y_i) = \sigma^2$; en consecuencia,

$$\begin{aligned} \operatorname{Var}\left(\hat{\beta}_1\right) &= \sigma^2 \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{S_{xx}^2} \\ &= \frac{\sigma^2}{S_{xx}} \end{aligned} \tag{2.14}$$

La varianza de $\hat{\beta}_0$ es

$$\begin{aligned} \operatorname{Var}\left(\hat{\beta}_1\right) &= \operatorname{Var}\left(\bar{y} - \hat{\beta}_1 \bar{x}\right) \\ &= \operatorname{Var}(\bar{y}) + \bar{x}^2 \operatorname{Var}\left(\hat{\beta}_1\right) - 2\bar{x}\operatorname{Cov}\left(\bar{y}, \hat{\beta}_1\right) \end{aligned}$$

Ahora bien, la varianza de $\bar{y}$ no es más que $\operatorname{Var}(\bar{y}) = \sigma^2/n$, y se puede demostrar que la covarianza entre $\bar{y}$ y $\hat{\beta}_1$ es cero (véase el Prob. 2.19). Así,

$$\begin{aligned} \operatorname{Var}\left(\hat{\beta}_0\right) &= \operatorname{Var}(\bar{y}) + \bar{x}^2 \operatorname{Var}\left(\hat{\beta}_1\right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \end{aligned} \tag{2.15}$$

Otro resultado importante acerca de la calidad de los estimadores por mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ es el **teorema de Gauss-Markov**, que establece que para el modelo de regresión (2.1) con las hipótesis $E(\varepsilon) = 0$, $\operatorname{Var}(\varepsilon) = \sigma^2$ y con errores no correlacionados, los estimadores por mínimos cuadrados son insesgados y tienen varianza mínima en comparación con todos los demás estimadores insesgados que sean combinaciones lineales de las y_i. Con frecuencia se dice que los estimadores por mínimos cuadrados son los **estimadores lineales insesgados óptimos**, donde "óptimos" implica que son de varianza mínima.

En el apéndice C.4 se demuestra el teorema de Gauss-Markov para el caso más general de regresión lineal múltiple, del cual la regresión lineal simple es un caso especial.

Hay varias otras propiedades útiles del ajuste por mínimos cuadrados:

1. La suma de los residuales en cualquier modelo de regresión que contenga una ordenada al origen β_0 siempre es igual a cero, esto es,

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

   Esta propiedad es consecuencia directa de la primera de las ecuaciones normales (2.5), y se demuestra en la tabla 2.2 para los residuales del ejemplo 2.1. Redondear los errores puede afectar la suma.

2. La suma de los valores observados y_i es igual a la suma de los valores ajustados $\hat{y}_i$:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

   La tabla 2.2 demuestra este resultado para el ejemplo 2.1.

3. La línea de regresión de mínimos cuadrados siempre pasa por el **centroide** de los datos, que es el punto $(\bar{y}, \bar{x})$.

4. La suma de los residuales, ponderados por el valor correspondiente de la variable regresora, siempre es igual a cero:

$$\sum_{i=1}^{n} x_i e_i = 0$$

5. La suma de los residuales, ponderados por el valor ajustado correspondiente, siempre es igual a cero:

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0$$

2.2.3 Estimación de σ^2

Además de estimar β_0 y β_1, se requiere un estimado de σ^2 para probar hipótesis y formar estimados de intervalo pertinentes al modelo de regresión. En el caso ideal este estimado no debería depender de la adecuación del modelo ajustado. Eso sólo es posible cuando hay varias observaciones de y para cuando menos un valor de x (véase la Sec. 3.4) o cuando se dispone de información anterior acerca de σ^2. Cuando no se puede usar este método, el estimado de σ^2 se obtiene de la suma de cuadrados de residuales, o suma de cuadrados de error:

$$SS_{\text{Res}} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.16}$$

Se puede deducir una fórmula cómoda para calcular SS_{Res} sustituyendo $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ en la ecuación (2.16), y simplificando.

Así se llega a

$$SS_{\text{Res}} = \sum_{i=1}^{n} y_i^2 = n\bar{y}^2 - \hat{\beta}_1 S_{xy} \tag{2.17}$$

Pero

$$\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \equiv SS_{\text{T}}$$

es justo la suma de cuadrados corregida, de las observaciones de la respuesta, por lo que

$$SS_{\text{Res}} = SS_{\text{T}} - \hat{\beta}_1 S_{xy} \tag{2.18}$$

La suma de cuadrados de residuales tiene $n - 2$ grados de libertad, porque dos grados de libertad se asocian con los estimados $\hat{\beta}_0$ y $\hat{\beta}_1$ que se usan para obtener $\hat{y}_i$. El apéndice C.3 demuestra que el valor esperado de SS_{Res} es $E(SS_{\text{Res}}) = (n - 2)\sigma^2$, por lo que un **estimador insesgado de** σ^2 es

$$\hat{\sigma}^2 = \frac{SS_{\text{Res}}}{n - 2} = MS_{\text{Res}} \tag{2.19}$$

La cantidad MS_{Res} se llama **cuadrado medio residual**. La raíz cuadrada de $\hat{\sigma}^2$ se llama, a veces, el **error estándar de la regresión** y tiene las mismas unidades que la variable de respuesta y.

Ya que $\hat{\sigma}^2$ depende de la suma de cuadrados de residuales, cualquier violación de las hipótesis sobre los errores del modelo, o cualquier especificación equivocada de la forma del modelo pueden dañar gravemente la utilidad de $\hat{\sigma}^2$ como estimado de σ^2. Como σ^2 se calcula con los residuales del modelo de regresión, se dice que es un estimado de σ^2 **dependiente del modelo.**

Ejemplo 2.2 *Datos del propelente de reacción*

Para estimar σ^2 de los datos del propelente de cohetes en el ejemplo 2.1, primero se calcula

$$SS_{\text{T}} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$= 92\,547{,}433.45 - \frac{(42\,627.15)^2}{20} = 1\,693{,}737.60$$

De acuerdo con la ecuación (2.19), la suma de cuadrados de residuales es

$$SS_{\text{Res}} = SS_{\text{T}} - \hat{\beta}_1 S_{xy}$$

$$= 1\,693\,737.60 - (-37.15)(-41\,112.65) = 166\,402.65$$

Por consiguiente el estimado de σ^2 se calcula con la ecuación (2.19) como sigue:

$$\hat{\sigma}^2 = \frac{SS_{\text{Res}}}{n-2} = \frac{166\,402.65}{18} = 9\,244{,}59$$

Recuérdese que este estimado de σ^2 **depende del modelo.**

2.2.4 Una forma alterna del modelo

Hay una forma alterna del modelo de regresión lineal simple, que a veces es útil. Supongamos que se redefine la variable regresora x_i como la desviación respecto a su propio promedio, esto es, $x_i - \bar{x}$. Entonces, el modelo de regresión se transforma en

$$\begin{aligned} y_i &= \beta_0 + \beta_1(x_i - \bar{x}) + \beta_1\bar{x} + \varepsilon_i \\ &= (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + \varepsilon_i \\ &= \beta_0' + \beta_1(x_i - \bar{x}) + \varepsilon_i \end{aligned} \tag{2.20}$$

Nótese que al redefinir la variable regresora en la ecuación (2.20), el origen de las x se desplazó de cero a $\bar{x}$. Para mantener iguales los valores ajustados tanto en el modelo original como en el transformado, es necesario modificar la ordenada al origen que se tenía. La relación entre la ordenada al origen original y la transformada es

$$\beta_0' = \beta_0 + \beta_1\bar{x} \tag{2.21}$$

Es fácil demostrar que el estimador de mínimos cuadrados de la ordenada al origen transformada es $\hat{\beta}_0' = \bar{y}$. El estimador de la pendiente no se afecta con la transformación. Esta forma alterna del modelo tiene ciertas ventajas. La primera es que los estimadores $\hat{\beta}_0' = \bar{y}$ y $\hat{\beta}_1 = S_{xy}/S_{xx}$, de mínimos cuadrados, son **no correlacionados**, esto es, $\text{Cov}(\hat{\beta}_0', \hat{\beta}_1) = 0$. Esto facilita algunas aplicaciones del modelo, como por ejemplo la determinación de los intervalos de confianza del promedio de y (véase la Sec. 2.4.2). Por último, el modelo ajustado es

$$\hat{y} = \bar{y} + \hat{\beta}_1(x - \bar{x}) \tag{2.22}$$

Aunque las ecuaciones (2.22) y (2.8) son equivalentes (las dos dan como resultado el mismo valor de $\hat{y}$ para el mismo valor de x), la ecuación 2.22 recuerda al analista, en forma directa, que el modelo de regresión sólo es válido dentro del intervalo de x en los **datos originales**. Esta región está centrada en $\bar{x}$.

2.3 PRUEBA DE HIPÓTESIS DE LA PENDIENTE Y DE LA ORDENADA AL ORIGEN

Con frecuencia interesa probar hipótesis y establecer intervalos de confianza de los parámetros del modelo. La prueba de hipótesis se explicará en esta sección, y en la sección 2.4 veremos los intervalos de confianza. Estos procedimientos requieren hacer la hipótesis adicional de que los errores ε_i del modelo estén distribuidos normalmente. Así, las hipótesis completas son: que los errores estén distribuidos en forma normal e independiente,

con media 0 y varianza σ^2, lo cual se abrevia "NID(0, σ^2)". NID viene de *normally and independently distributed* (distribuido normal e independientemente). En el capítulo 4 se describirá cómo se pueden comprobar esas hipótesis a través del **análisis de residuales**.

2.3.1 Uso de las pruebas *t*

Supongamos que se desea probar la hipótesis que la pendiente es igual a una constante, por ejemplo, a β_{10}. Las hipótesis correspondientes son

$$H_0: \beta_1 = \beta_{10}$$
$$H_1: \beta_1 \neq \beta_{10} \tag{2.23}$$

en donde se ha especificado una alternativa bilateral. Como los errores ε_i son NID(0, σ^2), las observaciones y_i son NID($\beta_0 + \beta_1 x_i$, σ^2). Ahora, $\hat{\beta}_1$ es una combinación lineal de las observaciones, de modo que $\hat{\beta}_1$ está distribuido normalmente con promedio β_1 y varianza σ^2/S_{xx}, usando la media y la varianza de $\hat{\beta}_1$ que se determinó en la sección 2.2.2. Por consiguiente, el estadístico

$$Z_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\sigma^2/S_{xx}}}$$

está distribuido $N(0, 1)$ si es cierta la hipótesis nula H_0: $\beta_1 = \beta_{10}$. Si se conociera σ^2, se podría usar Z_0 para probar la hipótesis (2.23). Comúnmente se desconoce σ^2. Ya se ha visto que MS_{Res} es un estimador insesgado de σ^2. En el apéndice C.3 se establece que $(n - 2)\, MS_{\text{Res}}/\sigma^2$ tiene una distribución χ^2_{n-2} y que MS_{Res} y $\hat{\beta}_1$ son independientes. De acuerdo con la definición del estadístico t que se presenta en el apéndice C.1,

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{MS_{\text{Res}}/S_{xx}}} \tag{2.24}$$

sigue una distribución t_{n-2} si es cierta la hipótesis nula H_0: $\beta_1 = \beta_{10}$. La cantidad de grados de libertad asociados con t_0 es igual a la cantidad de grados de libertad asociados con MS_{Res}. Así, la razón t_0 es el estadístico con que se prueba H_0: $\beta_1 = \beta_{10}$. El procedimiento de prueba calcula t_0 y compara su valor observado de acuerdo con la ecuación (2.24) con el punto porcentual $\alpha/2$ superior de t_{n-2} de la distribución ($t_{\alpha/2, n-2}$). Este procedimiento rechaza la hipótesis nula si

$$|t_0| > t_{\alpha/2, n-2} \tag{2.25}$$

También se podría usar el método del valor P para tomar la decisión.

El denominador del estadístico t_0 en la ecuación (2.24) se llama con frecuencia el **error estándar estimado**, o más sencillamente el **error estándar** de la pendiente. Esto es,

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{MS_{\text{Res}}}{S_{xx}}} \tag{2.26}$$

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

$$H_0: \beta_0 = \beta_{00}$$
$$H_1: \beta_0 \neq \beta_{00} \tag{2.28}$$

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:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0 \tag{2.30}$$

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.**

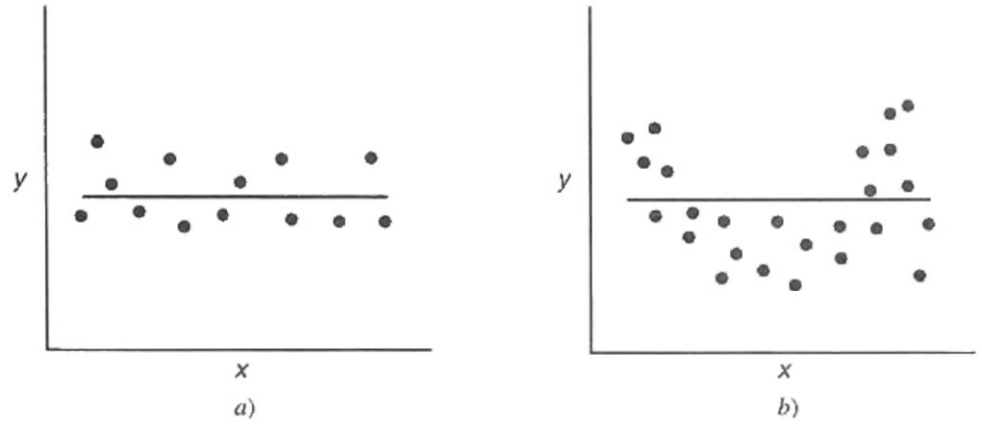

Figura 2.2 Casos en los que no se rechaza la hipótesis H_0: $\beta_1 = 0$.