

Capítulo 11

Regresión lineal simple y correlación

11.1 Introducción a la regresión lineal

En la práctica a menudo se requiere resolver problemas que implican conjuntos de variables de las cuales se sabe que tienen alguna relación inherente entre sí. Por ejemplo, en una situación industrial quizá se sepa que el contenido de alquitrán en el flujo de salida de un proceso químico está relacionado con la temperatura en la entrada. Podría ser de interés desarrollar un método de pronóstico, es decir, un procedimiento que permita estimar el contenido de alquitrán para varios niveles de temperatura de entrada a partir de información experimental. Desde luego, es muy probable que para muchos ejemplos concretos en los que la temperatura de entrada sea la misma, por ejemplo 130°C, el contenido de alquitrán de salida no sea el mismo. Esto es muy similar a lo que ocurre cuando se estudian varios automóviles con un motor del mismo volumen; no todos tienen el mismo rendimiento de combustible. No todas las casas ubicadas en la misma zona del país, con la misma superficie de construcción, se venden al mismo precio. El contenido de alquitrán, el rendimiento del combustible (en millas por galón) y el precio de las casas (en miles de dólares) son **variables dependientes** naturales o respuestas en los tres escenarios. La temperatura en la entrada, el volumen del motor (pies cúbicos) y los metros cuadrados de superficie de construcción son, respectivamente, **variables independientes** naturales o **regresores**. Una forma razonable de relación entre la **respuesta** Y y el regresor x es la relación lineal,

$$Y = \beta_0 + \beta_1 x,$$

en la que, por supuesto, β_0 es la **intersección** y β_1 es la **pendiente**. Esta relación se ilustra en la figura 11.1.

Si la relación es exacta y no contiene ningún componente aleatorio o probabilístico, entonces se trata de una relación **determinista** entre dos variables científicas. Sin embargo, en los ejemplos que se mencionaron, así como en muchos otros fenómenos científicos y de ingeniería, la relación no es determinista, es decir, una x dada no siempre produce el mismo valor de Y . Como resultado, los problemas importantes en este caso son de naturaleza probabilística, toda vez que la relación anterior no puede considerarse exacta. El concepto de **análisis de regresión** se refiere a encontrar la mejor relación entre Y y x

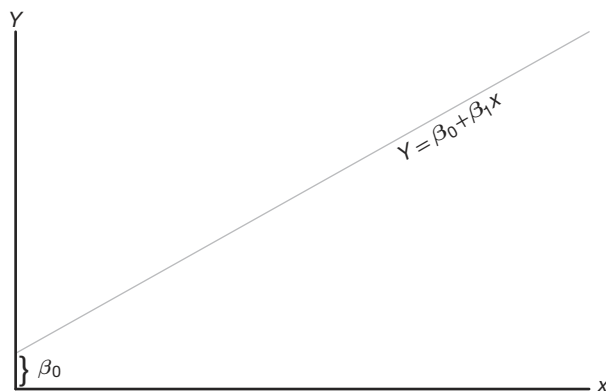


Figura 11.1: Una relación lineal; β_0 : intersección; β_1 : pendiente.

cuantificando la fuerza de esa relación, y empleando métodos que permitan predecir los valores de la respuesta dados los valores del regresor x .

En muchas aplicaciones habrá más de un regresor, es decir, más de una variable independiente **que ayude a explicar a Y** . Por ejemplo, si se tratara de explicar las razones para el precio de una casa, se esperaría que una de ellas fuera su antigüedad, en cuyo caso la estructura múltiple de la regresión se podría escribir como

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

donde Y es el precio, x_1 son los metros cuadrados y x_2 es la antigüedad de la casa en años. En el capítulo siguiente se estudiarán problemas con regresores múltiples. El análisis resultante se denomina **regresión múltiple**; en tanto que el análisis del caso con un solo regresor recibe el nombre de **regresión simple**. En un segundo ejemplo de la regresión múltiple, un ingeniero químico podría estar interesado en la cantidad de hidrógeno que se ha perdido en las muestras de un metal específico que se tiene almacenado. En este caso habría dos entradas, x_1 , el tiempo de almacenamiento en horas, y x_2 , la temperatura de almacenamiento en grados centígrados. De modo que la respuesta sería Y , la pérdida de hidrógeno en partes por millón.

En este capítulo estudiaremos el tema de la **regresión lineal simple**, que trata el caso de una sola variable regresora, en el que la relación entre x y y es lineal. Para el caso en el que hay más de una variable regresora el lector debe consultar el capítulo 12. Denotemos una muestra aleatoria de tamaño n mediante el conjunto $\{(x_i, y_i); i = 1, 2, \dots, n\}$. Si se tomaran muestras adicionales utilizando exactamente los mismos valores de x , se esperaría que los valores de y variaran. Así, el valor y_i en el par ordenado (x_i, y_i) es el valor de cierta variable aleatoria Y_i .

11.2 El modelo de regresión lineal simple (RLS)

Hemos limitado el uso del término *análisis de regresión* a los casos en los que las relaciones entre las variables no son deterministas, es decir, no son exactas. En otras palabras, debe existir un **componente aleatorio** en la ecuación que relaciona las variables. Este componente aleatorio toma en cuenta consideraciones que no son medibles o, de

hecho, que los científicos o los ingenieros no comprenden. En realidad, en la mayoría de aplicaciones de la regresión, la ecuación lineal, digamos, $Y = \beta_0 + \beta_1 x$ es una aproximación que representa de manera simplificada algo desconocido y mucho más complicado. Por ejemplo, en el caso que implica la respuesta $Y =$ contenido de alquitrán y $x =$ temperatura de entrada es probable que $Y = \beta_0 + \beta_1 x$ sea una aproximación razonable que podría funcionar dentro de un rango limitado de x . La mayoría de las veces los modelos que son simplificaciones de estructuras más complicadas y desconocidas son de naturaleza lineal, es decir, lineales en los **parámetros** β_0 y β_1 o, en el caso del modelo que implica el precio, el tamaño y la antigüedad de la casa, lineal en los **parámetros** β_0 , β_1 y β_2 . Estas estructuras lineales son sencillas y de naturaleza empírica, por lo que se denominan **modelos empíricos**.

Un análisis de la relación entre x y Y requiere el planteamiento de un **modelo estadístico**. Con frecuencia un estadístico utiliza un modelo como representación de un **ideal** que, en esencia, define cómo percibimos que el sistema en cuestión generó los datos. El modelo debe incluir al conjunto $\{(x_i, y_i); i = 1, 2, \dots, n\}$ de datos que implica n pares de valores (x, y) . No debemos olvidar que el valor de y_i depende de x_i por medio de una estructura lineal que también incluye el componente aleatorio. La base para el uso de un modelo estadístico se relaciona con la manera en que la variable aleatoria Y cambia con x y el componente aleatorio. El modelo también incluye lo que se asume acerca de las propiedades estadísticas del componente aleatorio. A continuación se presenta el modelo estadístico para la regresión lineal simple. La respuesta Y se relaciona con la variable independiente x a través de la ecuación

Modelo de
regresión lineal
simple

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

en la cual β_0 y β_1 son los parámetros desconocidos de la intersección y la pendiente, respectivamente, y ϵ es una variable aleatoria que se supone está distribuida con $E(\epsilon) = 0$ y $\text{Var}(\epsilon) = \sigma^2$. Es frecuente que a la cantidad σ^2 se le denomine varianza del error o varianza residual.

En el modelo anterior hay varias cuestiones evidentes. La cantidad Y es una variable aleatoria, ya que ϵ es aleatoria. El valor x de la variable regresora no es aleatorio y, de hecho, se mide con un error despreciable. La cantidad ϵ , que a menudo recibe el nombre de **error aleatorio** o **alteración aleatoria**, tiene varianza constante. Es común que a esta parte se le denomine **suposición de varianza homogénea**. La presencia de este error aleatorio ϵ evita que el modelo se convierta tan sólo en una ecuación determinista. Ahora, el hecho de que $E(\epsilon) = 0$ implica que para una x específica, los valores de y se distribuyen alrededor de la **recta verdadera** o **recta de regresión** de la población $y = \beta_0 + \beta_1 x$. Si se elige bien el modelo, es decir, si no hay otros regresores de importancia y la aproximación lineal es buena dentro de los rangos de los datos, entonces son razonables los errores positivos y negativos que rodean a la regresión verdadera. Debe recordarse que en la práctica β_0 y β_1 se desconocen y que deben estimarse a partir de los datos. Además, el modelo que se acaba de describir es de naturaleza conceptual. Como resultado, en la práctica nunca se observan los valores ϵ reales, por lo que nunca se puede trazar la verdadera recta de regresión, aunque suponemos que ahí está. Sólo es posible dibujar una recta estimada. En la figura 11.2 se ilustra la naturaleza de los datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para un caso en que sólo se dispone de $n = 5$ observaciones. Debemos destacar que lo que observamos en la figura 11.2 no es la recta que utilizan el científico o ingeniero. En vez de esa recta, ¡lo

que describe la ilustración es el significado de las suposiciones! Ahora describiremos la regresión que el usuario tiene a su disposición.

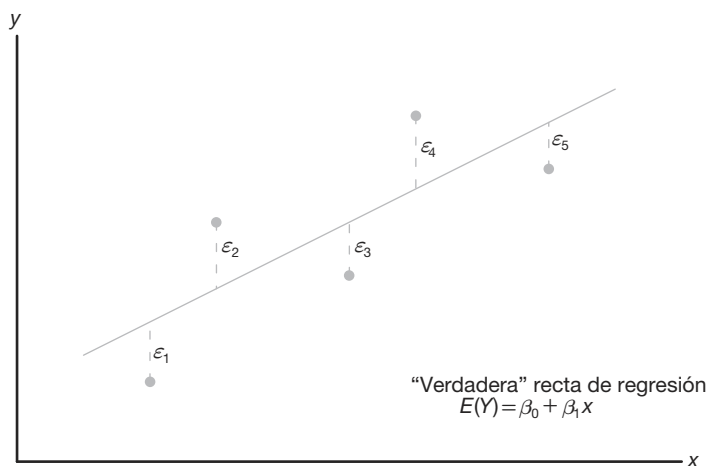


Figura 11.2: Datos (x, y) hipotéticos dispersos alrededor de la verdadera recta de regresión para $n = 5$.

La recta de regresión ajustada

Un aspecto importante del análisis de regresión es, en términos sencillos, estimar los parámetros β_0 y β_1 , es decir, estimar los llamados **coeficientes de regresión**. En la sección siguiente se estudiará el método para estimarlos. Suponga que denotamos los estimados b_0 para β_0 y b_1 para β_1 . Entonces, la recta de **regresión ajustada**, o estimada, es dada por

$$\hat{y} = b_0 + b_1 x,$$

donde \hat{y} es el valor pronosticado o ajustado. Es evidente que la recta ajustada es un estimado de la verdadera recta de regresión. Se espera que la recta ajustada esté más cerca de la verdadera línea de regresión cuando se dispone de una gran cantidad de datos. En el ejemplo siguiente se ilustra la recta ajustada para un estudio sobre contaminación en la vida real.

Uno de los problemas más desafiantes que enfrenta el campo del control de la contaminación del agua lo representa la industria de la peletería, ya que sus desechos son químicamente complejos; se caracterizan por valores elevados de la demanda de oxígeno químico, sólidos volátiles y otras medidas de contaminación. Considere los datos experimentales de la tabla 11.1, que se obtuvieron de 33 muestras de desechos tratados químicamente en un estudio realizado en Virginia Tech. Se registraron los valores de x , la reducción porcentual de los sólidos totales, y de y , el porcentaje de disminución de la demanda de oxígeno químico.

Los datos de la tabla 11.1 aparecen graficados en un **diagrama de dispersión** en la figura 11.3. Al inspeccionar dicho diagrama se observa que los puntos se acercan mucho a una línea recta, lo cual indica que la suposición de linealidad entre las dos variables parece ser razonable.

Tabla 11.1: Medidas de la reducción de los sólidos y de la demanda de oxígeno químico

Reducción de sólidos, x (%)	Reducción de la demanda de oxígeno, y (%)	Reducción de sólidos, x (%)	Reducción de la demanda de oxígeno, y (%)
3	5	36	34
7	11	37	36
11	21	38	38
15	16	39	37
18	16	39	36
27	28	39	45
29	27	40	39
30	25	41	41
30	35	42	40
31	30	42	44
31	40	43	37
32	32	44	44
33	34	45	46
33	32	46	46
34	34	47	49
36	37	50	51
36	38		

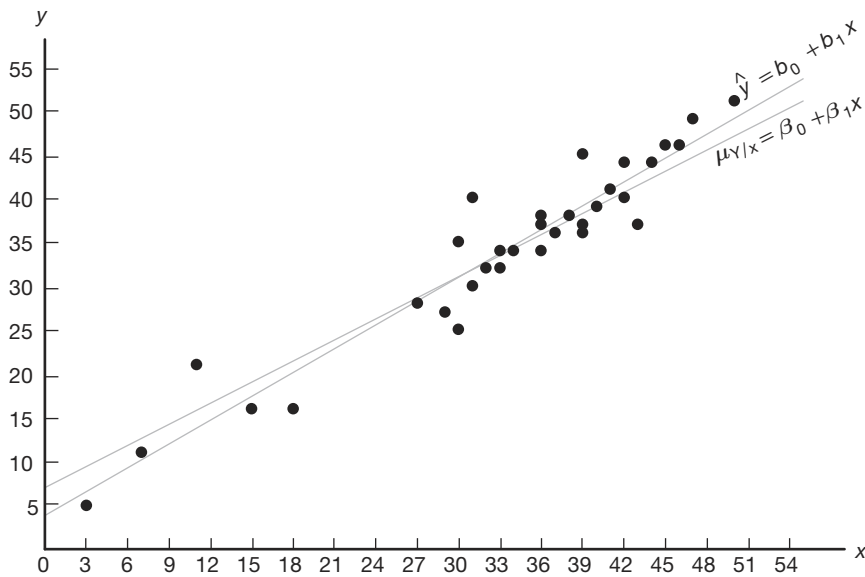


Figura 11.3: Diagrama de dispersión con rectas de regresión.

En el diagrama de dispersión de la figura 11.3 se ilustra la recta de regresión ajustada y una recta hipotética de regresión verdadera. Más adelante, en la sección 11.3, en la cual estudiaremos el método de estimación, revisaremos este ejemplo.

Otra mirada a las suposiciones del modelo

Resulta aleccionador repasar el modelo de regresión lineal simple que se presentó con anterioridad y analizar de forma gráfica la manera en que se relaciona con la denominada regresión verdadera. Daremos más detalles en la figura 11.2, cuando ilustremos no sólo el lugar en que los ϵ_i se localizan en la gráfica, sino también lo que implica la suposición de normalidad para los ϵ_i .

Suponga que tenemos una regresión lineal simple con $n = 6$, valores de x equidistantes y un valor único de y para cada x . Considere la gráfica de la figura 11.4, la cual debería proporcionar al lector una representación clara del modelo y de las suposiciones implicadas. La recta que aparece en la gráfica es la recta de regresión verdadera. Los puntos graficados (y, x) son puntos reales dispersos alrededor de la recta. Cada punto se ubica en su propia distribución normal, donde el centro de la distribución, es decir, la media de y , cae sobre la recta. Ciertamente esto es lo esperado, ya que $E(Y) = \beta_0 + \beta_1 x$. Como resultado, la verdadera recta de regresión **pasa a través de las medias de la respuesta** y las observaciones reales se encuentran sobre la distribución, alrededor de las medias. Observe también que todas las distribuciones tienen la misma varianza, que se denota con σ^2 . Desde luego, la desviación entre una y individual y el punto sobre la recta será su valor individual ϵ . Esto queda claro porque

$$y_i - E(Y_i) = y_i - (\beta_0 + \beta_1 x_i) = \epsilon_i.$$

Así, con una x dada, tanto Y como el ϵ correspondiente tienen varianza σ^2 .

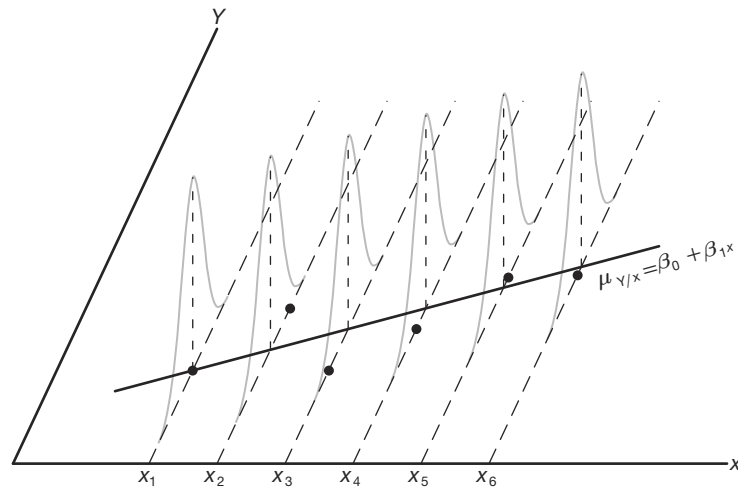


Figura 11.4: Observaciones individuales alrededor de la verdadera recta de regresión.

Note también que aquí escribimos la verdadera recta de regresión como $\mu_{y|x} = \beta_0 + \beta_1 x$ con el fin de reafirmar que la recta pasa a través de la media de la variable aleatoria Y .

11.3 Mínimos cuadrados y el modelo ajustado

En esta sección se estudia el método para ajustar una recta de regresión estimada a los datos, lo cual equivale a determinar los estimados b_0 para β_0 y b_1 para β_1 . Por supuesto,

esto permite el cálculo de los valores pronosticados a partir de la recta ajustada $\hat{y} = b_0 + b_1x$, y otros tipos de análisis y de información diagnóstica que determinarán la fuerza de la relación, así como la adecuación y el ajuste del modelo. Antes de analizar el método de estimación de los mínimos cuadrados es importante presentar el concepto de **residual**. En esencia, un residual es un error en el ajuste del modelo $\hat{y} = b_0 + b_1x$.

Residual: Error en el ajuste Dado un conjunto de datos de regresión $\{(x_i, y_i); i = 1, 2, \dots, n\}$ y un modelo ajustado $\hat{y}_i = b_0 + b_1x$, el i -ésimo residual e_i es dado por

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \dots, n.$$

Es evidente que si un conjunto de n residuales es grande, entonces el ajuste del modelo no es bueno. Los residuales pequeños son indicadores de un ajuste adecuado. Otra relación interesante, y que a veces es útil, es la siguiente:

$$y_i = b_0 + b_1x_i + e_i.$$

El uso de la ecuación anterior debería aclarar la diferencia entre los residuales e_i y los errores del modelo conceptual ϵ_i . No debemos olvidar que, mientras que los ϵ_i no se observan, los e_i no sólo se observan sino que desempeñan un papel importante en el análisis total.

La figura 11.5 ilustra el ajuste de la recta a este conjunto de datos: a saber $\hat{y} = b_0 + b_1x$, y la recta que refleja el modelo $\mu_{y|x} = \beta_0 + \beta_1x$. Desde luego, β_0 y β_1 son parámetros desconocidos. La recta ajustada es un estimado de la recta que genera el modelo estadístico. Hay que tener presente que la recta $\mu_{y|x} = \beta_0 + \beta_1x$ es desconocida.

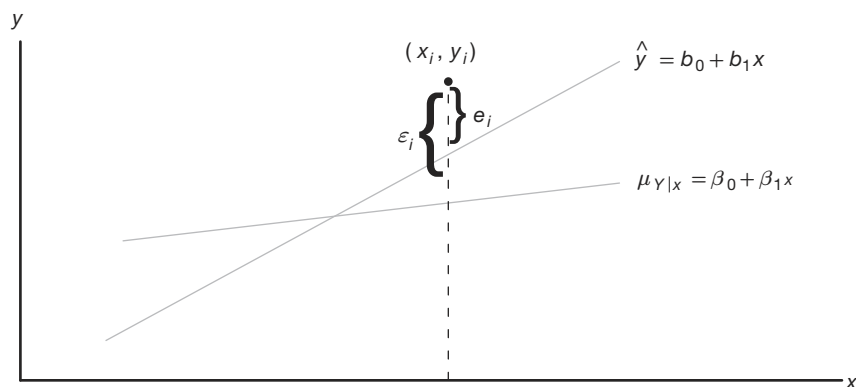


Figura 11.5: Comparación de ϵ_i con el residual e_i .

Método de mínimos cuadrados

Debemos calcular b_0 y b_1 , los estimados de β_0 y β_1 , de manera que la suma de los cuadrados de los residuales sea mínima. La suma residual de los cuadrados con frecuencia se denomina suma de los cuadrados del error respecto de la recta de regresión y se denota como *SCE*. Este procedimiento de minimización para estimar los parámetros

se denomina **método de mínimos cuadrados**. Por lo tanto, debemos calcular a y b para minimizar

$$SCE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2.$$

Al diferenciar la SCE con respecto a b_0 y b_1 , se obtiene

$$\frac{\partial(SCE)}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i), \quad \frac{\partial(SCE)}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i.$$

Al igualar a cero las derivadas parciales y reacomodar los términos, obtenemos las ecuaciones siguientes (llamadas **ecuaciones normales**)

$$nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \quad b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i,$$

que se resuelven simultáneamente para obtener fórmulas de cálculo para b_0 y b_1 .

Estimación de los coeficientes de regresión Dada la muestra $\{(x_i, y_i); i = 1, 2, \dots, n\}$, los estimados b_0 y b_1 de los mínimos cuadrados de los coeficientes de regresión β_0 y β_1 se calculan mediante las fórmulas

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} y$$

$$b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} = \bar{y} - b_1 \bar{x}.$$

En el ejemplo siguiente se ilustra el cálculo de b_0 y b_1 usando los datos de la tabla 11.1.

Ejemplo 11.1: Estime la recta de regresión para los datos de contaminación de la tabla 11.1.

Solución:

$$\sum_{i=1}^{33} x_i = 1104, \quad \sum_{i=1}^{33} y_i = 1124, \quad \sum_{i=1}^{33} x_i y_i = 41,355, \quad \sum_{i=1}^{33} x_i^2 = 41,086$$

Por lo tanto,

$$b_1 = \frac{(33)(41,355) - (1104)(1124)}{(33)(41,086) - (1104)^2} = 0.903643 \text{ y}$$

$$b_0 = \frac{1124 - (0.903643)(1104)}{33} = 3.829633.$$

Por consiguiente, la recta de regresión estimada es dada por

$$\hat{y} = 3.8296 + 0.9036x.$$



Si utilizáramos la recta de regresión del ejemplo 11.1, podríamos pronosticar una reducción de 31% en la demanda de oxígeno químico si los sólidos totales se redujeran

un 30%. La reducción de 31% en la demanda de oxígeno químico se puede interpretar como un estimado de la media de la población $\mu_{y|30}$, o como un estimado de una observación nueva si la reducción de sólidos totales es de 30%. Sin embargo, dichas estimaciones están sujetas a error. Incluso si el experimento estuviera controlado para que la reducción de los sólidos totales fuera de 30%, es improbable que la reducción en la demanda de oxígeno químico que se midiera fuera exactamente igual a 31%. De hecho, los datos originales registrados en la tabla 11.1 indican que se registraron medidas de 25% y de 35% en la reducción de la demanda de oxígeno, cuando la disminución de los sólidos totales se mantuvo en 30%.

¿Qué es lo bueno de los mínimos cuadrados?

Debemos señalar que el criterio de los mínimos cuadrados está diseñado para brindar una recta ajustada que resulte en la “cercanía” entre la recta y los puntos graficados. Existen muchas formas de medir dicha cercanía. Por ejemplo, quizá desearíamos determinar los valores de b_0 y b_1 para los que se minimiza $\sum_{i=1}^n |y_i - \hat{y}_i|$ o para los que se minimiza $\sum_{i=1}^n |y_i - \hat{y}_i|^{1.5}$. Ambos métodos son viables y razonables. Observe que los dos, así como el procedimiento de mínimos cuadrados, obligan a que los residuales sean “pequeños” en cierto sentido. Debemos recordar que los residuales son el equivalente empírico de los valores de ϵ . La figura 11.6 ilustra un conjunto de residuales. Observe que la línea ajustada tiene valores predichos como puntos sobre la recta y, en consecuencia, los residuales son desviaciones verticales desde los puntos hasta la recta. Como resultado, el procedimiento de mínimos cuadrados genera una recta que **minimiza la suma de los cuadrados de las desviaciones verticales** desde los puntos hasta la recta.

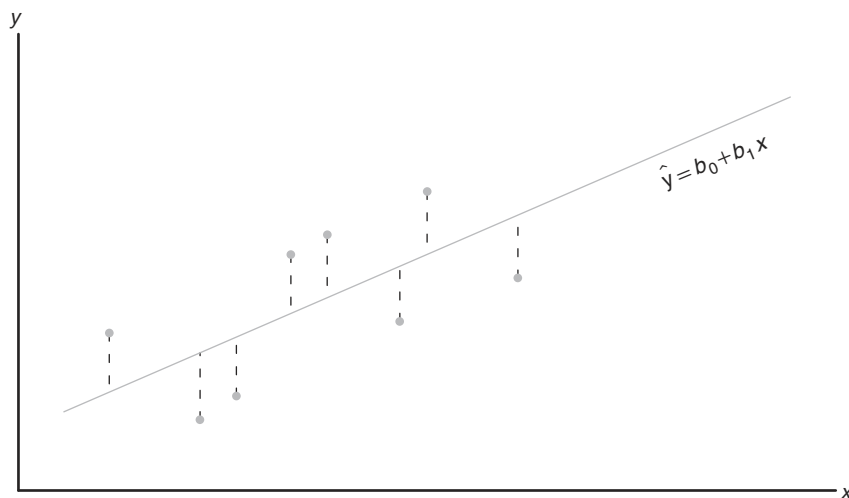


Figura 11.6: Los residuales como desviaciones verticales.

Ejercicios

11.1 Se realizó un estudio en Virginia Tech para determinar si ciertas medidas de la fuerza estática del brazo influyen en las características de “levantamiento dinámico” de un individuo. Veinticinco individuos se sometieron a pruebas de fuerza y luego se les pidió que hicieran una prueba de levantamiento de peso, en el que el peso se elevaba en forma dinámica por encima de la cabeza. A continuación se presentan los datos.

Individual	Fuerza del brazo, x	Levantamiento dinámico, y
1	17.3	71.7
2	19.3	48.3
3	19.5	88.3
4	19.7	75.0
5	22.9	91.7
6	23.1	100.0
7	26.4	73.3
8	26.8	65.0
9	27.6	75.0
10	28.1	88.3
11	28.2	68.3
12	28.7	96.7
13	29.0	76.7
14	29.6	78.3
15	29.9	60.0
16	29.9	71.7
17	30.3	85.0
18	31.3	85.0
19	36.0	88.3
20	39.5	100.0
21	40.4	100.0
22	44.3	100.0
23	44.6	91.7
24	50.4	100.0
25	55.9	71.7

- a) Estime los valores de β_0 y β_1 para la curva de regresión lineal $\mu_{y|x} = \beta_0 + \beta_1 x$.
- b) Calcule un estimado puntual de $\mu_{y|30}$.
- c) Grafique los residuales en comparación con las x (fuerza del brazo). Comente los resultados.

11.2 Las siguientes son las calificaciones de un grupo de 9 estudiantes en un informe de medio semestre (x) y en el examen final (y):

x	77	50	71	72	81	94	96	99	67
y	82	66	78	34	47	85	99	99	68

- a) Estime la recta de regresión lineal.
- b) Calcule la calificación final de un estudiante que obtuvo 85 de calificación en el informe de medio semestre.

11.3 Se registraron las cantidades de un compuesto químico y que se disuelve en 100 gramos de agua a distintas temperaturas x :

x ($^{\circ}\text{C}$)	y (gramos)		
0	8	6	8
15	12	10	14
30	25	21	24
45	31	33	28
60	44	39	42
75	48	51	44

- a) Calcule la ecuación de la recta de regresión.
- b) Grafique la recta en un diagrama de dispersión.
- c) Estime la cantidad de producto químico que se disolverá en 100 gramos de agua a 50°C .

11.4 Para fines de calibración se recabaron los siguientes datos, los cuales permitirían determinar la relación entre la presión y la lectura correspondiente en la escala.

Presión, x (lb/pulg ²)	Lectura en la escala, y
10	13
10	18
10	16
10	15
10	20
50	86
50	90
50	88
50	88
50	92

- a) Calcule la ecuación de la recta de regresión.
- b) En esta aplicación el propósito de la calibración es estimar la presión a partir de una lectura observada en la escala. Estime la presión para una lectura en la escala de 54, usando $\hat{x} = (54 - b_0)/b_1$.

11.5 Se realizó un estudio sobre la cantidad de azúcar convertida en cierto proceso a distintas temperaturas. Los datos se codificaron y registraron como sigue:

Temperatura, x	Azúcar convertida, y
1.0	8.1
1.1	7.8
1.2	8.5
1.3	9.8
1.4	9.5
1.5	8.9
1.6	8.6
1.7	10.2
1.8	9.3
1.9	9.2
2.0	10.5

- a) Estime la recta de regresión lineal.
- b) Calcule la cantidad media de azúcar convertida que se produce cuando se registra una temperatura codificada de 1.75.
- c) Grafique los residuales en comparación con la temperatura. Comente sus resultados.

11.6 En cierto tipo de espécimen de prueba metálico se sabe que la tensión normal sobre un espécimen se relaciona funcionalmente con la resistencia al corte. El siguiente es un conjunto de datos experimentales codificados para las dos variables:

Tensión normal, x	Resistencia al corte, y
26.8	26.5
25.4	27.3
28.9	24.2
23.6	27.1
27.7	23.6
23.9	25.9
24.7	26.3
28.1	22.5
26.9	21.7
27.4	21.4
22.6	25.8
25.6	24.9

- Estime la recta de regresión $\mu_{y|x} = \beta_0 + \beta_1 x$.
- Estime la resistencia al corte para una tensión normal de 24.5.

11.7 Los siguientes son algunos de los datos contenidos en un conjunto clásico denominado “datos piloto de graficación” que aparecen en *Fitting Equations to Data*, de Daniel y Wood, publicado en 1971. La respuesta y es el contenido de ácido del material determinado por análisis volumétrico; mientras que el regresor x es el contenido de ácido orgánico determinado por extracción y ponderación.

y	x	y	x
76	123	70	109
62	55	37	48
66	100	82	138
58	75	88	164
88	159	43	28

- Grafique los datos; ¿la regresión lineal simple parece un modelo adecuado?
- Haga un ajuste de regresión lineal simple; calcule la pendiente y la intersección.
- Grafique la recta de regresión en la gráfica del inciso a .

11.8 Se aplica un examen de colocación de matemáticas a todos los estudiantes de nuevo ingreso en una universidad pequeña. Se negará la inscripción al curso regular de matemáticas a los estudiantes que obtengan menos de 35 puntos y se les enviará a clases de regularización. Se registraron los resultados del examen de colocación y las calificaciones finales de 20 estudiantes que tomaron el curso regular:

- Elabore un diagrama de dispersión.
- Calcule la ecuación de la recta de regresión para predecir las calificaciones en el curso a partir de las del examen de colocación.
- Grafique la recta en el diagrama de dispersión.

- Si la calificación aprobatoria mínima fuera 60 puntos, ¿qué calificación en el examen de colocación se debería usar en el futuro como criterio para negar a los estudiantes el derecho de admisión a ese curso?

Examen de colocación	Calificación en el curso
50	53
35	41
35	61
40	56
55	68
65	36
35	11
60	70
90	79
35	59
90	54
80	91
60	48
60	71
60	71
40	47
55	53
50	68
65	57
50	79

11.9 Un comerciante minorista realizó un estudio para determinar la relación que hay entre los gastos semanales de publicidad y las ventas.

Costos de publicidad (\$)	Ventas (\$)
40	385
20	400
25	395
20	365
30	475
50	440
40	490
20	420
50	560
40	525
25	480
50	510

- Elabore un diagrama de dispersión.
- Calcule la ecuación de la recta de regresión para pronosticar las ventas semanales a partir de los gastos de publicidad.
- Estime las ventas semanales si los costos de publicidad son de \$35.
- Grafique los residuales en comparación con los costos de publicidad. Comente sus resultados.

11.10 Los siguientes datos son los precios de venta z de cierta marca y modelo de automóvil usado con w años de antigüedad. Ajuste una curva de la forma $\mu_{z|w} = \gamma \delta^w$ mediante la ecuación de regresión muestral no lineal $\hat{z} = cd^w$ [Sugerencia: Escriba $\ln \hat{z} = \ln c + (\ln d)w = b_0 + b_1 w$].

<i>w</i> (años)	<i>z</i> (dólares)	<i>w</i> (años)	<i>z</i> (dólares)
1	6350	3	5395
2	5695	5	4985
2	5750	5	4895

11.11 La fuerza de impulso de un motor (*y*) es una función de la temperatura de escape (*x*) en °F cuando otras variables de importancia se mantienen constantes. Considere los siguientes datos.

<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>
4300	1760	4010	1665
4650	1652	3810	1550
3200	1485	4500	1700
3150	1390	3008	1270
4950	1820		

- a) Grafique los datos.
- b) Ajuste una recta de regresión simple a los datos y grafíquela a través de ellos.

11.12 Se realizó un estudio para analizar el efecto de la temperatura ambiente *x* sobre la energía eléctrica consumida por una planta química *y*. Otros factores se mantuvieron constantes y se recabaron los datos de una planta piloto experimental.

<i>y</i> (BTU)	<i>x</i> (°F)	<i>y</i> (BTU)	<i>x</i> (°F)
250	27	265	31
285	45	298	60
320	72	267	34
295	58	321	74

- a) Grafique los datos.
- b) Estime la pendiente y la intersección en un modelo de regresión lineal simple.
- c) Pronostique el consumo de energía para una temperatura ambiente de 65°F.

11.13 Un estudio sobre la cantidad de lluvia y la de contaminación del aire eliminada produjo los siguientes datos:

Cantidad de lluvia diaria, <i>x</i> (0.01 cm)	Partículas eliminadas, <i>y</i> (μg/m ³)
4.3	126
4.5	121
5.9	116
5.6	118
6.1	114
5.2	118
3.8	132
2.1	141
7.5	108

- a) Calcule la ecuación de la recta de regresión para predecir las partículas eliminadas de la cantidad de precipitación diaria.
- b) Estime la cantidad de partículas eliminadas si la precipitación diaria es *x* = 4.8 unidades.

11.14 Un profesor de la Escuela de Negocios de una universidad encuestó a una docena de colegas acerca del número de reuniones profesionales a que acudieron en los últimos cinco años (*x*) y el número de trabajos que enviaron a revistas especializadas (*y*) durante el mismo periodo. A continuación se presenta el resumen de los datos:

$$n = 12, \quad \bar{x} = 4, \quad \bar{y} = 12, \\ \sum_{i=1}^n x_i^2 = 232, \quad \sum_{i=1}^n x_i y_i = 318.$$

Ajuste un modelo de regresión lineal simple entre *x* y *y* calculando los estimados de la intersección y la pendiente. Comente si la asistencia a más reuniones profesionales da como resultado más publicaciones de artículos.

11.4 Propiedades de los estimadores de mínimos cuadrados

Además de los supuestos de que el término del error en el modelo

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

es una variable aleatoria con media igual a cero y varianza σ^2 constante, suponga que además damos por hecho que $\epsilon_1, \epsilon_2, \dots, \epsilon_n$ son independientes de una corrida a otra del experimento, lo cual proporciona la base para calcular las medias y varianzas de los estimadores de β_0 y β_1 .

Es importante recordar que nuestros valores de b_0 y b_1 , basados en una muestra dada de *n* observaciones, sólo son estimaciones de los parámetros verdaderos β_0 y β_1 . Si el experimento se repitiera una y otra vez, usando en cada ocasión los mismos valores fijos de *x*, los estimados resultantes de β_0 y β_1 muy probablemente diferirían de un experimento a otro. Estos estimados distintos podrían ser considerados como valores adoptados por las variables aleatorias B_0 y B_1 ; en tanto que b_0 y b_1 son ejecuciones específicas.

Como los valores de *x* permanecen fijos, los valores de B_0 y B_1 dependen de las variaciones en los valores de *y* o, con más precisión, en los valores de las variables aleatorias

Y_1, Y_2, \dots, Y_n . Las suposiciones sobre la distribución implican que las Y_i , $i = 1, 2, \dots, n$ también están distribuidas de manera independiente, con media $\mu_{Y|x_i} = \beta_0 + \beta_1 x_i$ y varianzas σ^2 iguales, es decir,

$$\sigma_{Y|x_i}^2 = \sigma^2 \quad \text{para } i = 1, 2, \dots, n.$$

Media y varianza de los estimadores

En la exposición que sigue mostramos que el estimador B_1 es insesgado para β_1 , y se demuestran tanto las varianzas de B_0 como las de B_1 . Esto inicia una serie de procedimientos que conducen a la prueba de hipótesis y a la estimación de intervalos de confianza para la intersección y la pendiente.

Como el estimador

$$B_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

es de la forma $\sum_{i=1}^n c_i Y_i$,

$$c_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad i = 1, 2, \dots, n,$$

podemos concluir a partir del teorema 7.11 que B_1 tiene una distribución $n(\mu_{B_1}, \sigma_{B_1})$ con

$$\mu_{B_1} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} = \beta_1 \quad \text{y} \quad \sigma_{B_1}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma_{Y_i}^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

También se puede demostrar (véase el ejercicio de repaso 11.60 de la página 438) que la variable aleatoria B_0 se distribuye normalmente con

$$\text{media } \mu_{B_0} = \beta_0 \quad \text{y varianza } \sigma_{B_0}^2 = \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2.$$

A partir de estos resultados es evidente que los **estimadores de mínimos cuadrados tanto para β_0 como para β_1 son insesgados**.

Partición de la variabilidad total y estimación de σ^2

Para hacer inferencias sobre β_0 y β_1 es necesario llegar a una estimación del parámetro σ^2 que aparece en las dos fórmulas anteriores de la varianza de B_0 y B_1 . El parámetro σ^2 , el modelo de la varianza del error, refleja una variación aleatoria o una variación del

error experimental alrededor de la recta de regresión. En gran parte de lo que sigue se recomienda emplear la notación

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

De manera que la suma de los cuadrados del error se puede escribir como sigue:

$$\begin{aligned} SCE &= \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^n [(y_i - \bar{y}) - b_1 (x_i - \bar{x})]^2 \\ &= \sum_{i=1}^n (y_i - \bar{y})^2 - 2b_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) + b_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= S_{yy} - 2b_1 S_{xy} + b_1^2 S_{xx} = S_{yy} - b_1 S_{xy}, \end{aligned}$$

que es el paso final que surge del hecho de que $b_1 = S_{xy} / S_{xx}$.

Teorema 11.1: Un estimador insesgado de σ^2 es

$$s^2 = \frac{SCE}{n-2} = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{n-2} = \frac{S_{yy} - b_1 S_{xy}}{n-2}.$$

La prueba del teorema 11.1 se deja como ejercicio (véase el ejercicio de repaso 11.59).

El estimador de σ^2 como error cuadrado medio

Para darnos una idea del estimador de σ^2 deberíamos observar el resultado del teorema 11.1. El parámetro σ^2 mide la varianza o las desviaciones cuadradas entre los valores de Y y su media, dada por $\mu_{Y|x}$, es decir, las desviaciones cuadradas entre Y y $\beta_0 + \beta_1 x$. Por supuesto, $\beta_0 + \beta_1 x$ se estima por medio de $\hat{y} = b_0 + b_1 x$. Por consiguiente, tendría sentido que la varianza σ^2 se describa mejor como una desviación cuadrada de la observación típica y_i con respecto a la media estimada \hat{y}_i , que es el punto correspondiente sobre la recta ajustada. Entonces, los valores $(y_i - \hat{y}_i)$ revelan la varianza apropiada, de manera muy similar a como los valores $(y_i - \bar{y})^2$ miden la varianza cuando se realiza un muestreo en un escenario no relacionado con la regresión. En otras palabras, \bar{y} estima la media en la última situación sencilla, mientras que \hat{y}_i estima la media de y_i en una estructura de regresión. Ahora, ¿qué significa el divisor $n-2$? En las secciones que siguen observaremos que éstos son los grados de libertad asociados con el estimador s^2 de σ^2 . En tanto que en el escenario i.i.d. (independiente e idénticamente distribuidas), la normal estándar se resta un grado de libertad de n en el denominador, para lo cual una explicación razonable es que se estima un parámetro, que es la media μ por medio de, digamos, \bar{y} , pero en el problema de la regresión **se estiman dos parámetros**, que son β_0 y β_1 , por medio de b_0 y b_1 . Así, el parámetro importante σ^2 , que se estima mediante

$$s^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 / (n-2),$$

se denomina **error cuadrado medio**, que describe un tipo de media (división entre $n-2$) de los residuales cuadrados.

11.5 Inferencias sobre los coeficientes de regresión

Además de tan sólo estimar la relación lineal entre x y Y para fines de predicción, el experimentador podría estar interesado en hacer ciertas inferencias acerca de la pendiente y la intersección. Para dar ocasión a la prueba de hipótesis y a la construcción de intervalos de confianza para β_0 y β_1 , debemos estar dispuestos a hacer la suposición adicional de que cada ϵ_i , $i = 1, 2, \dots, n$, se distribuye de forma normal. Esta suposición implica que Y_1, Y_2, \dots, Y_n también están distribuidas normalmente, cada una con una distribución de probabilidad $n(y_i; \beta_0 + \beta_1 x_i, \sigma)$.

A partir de la sección 11.4 sabemos que B_1 tiene una distribución normal, y suponiendo normalidad, un resultado muy parecido al que se plantea en el teorema 8.4 nos permite concluir que $(n-2)S^2/\sigma^2$ es una variable chi cuadrada con $n-2$ grados de libertad, independiente de la variable aleatoria B_1 . Entonces, el teorema 8.5 garantiza que el estadístico

$$T = \frac{(B_1 - \beta_1)/(\sigma/\sqrt{S_{xx}})}{S/\sigma} = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}}$$

tenga una distribución t con $n-2$ grados de libertad. Podemos utilizar el estadístico T para construir un intervalo de confianza del $100(1-\alpha)\%$ para el coeficiente β_1 .

Intervalo de confianza para β_1	Un intervalo de confianza de $100(1-\alpha)\%$ para el parámetro β_1 en la recta de regresión $\mu_{Y x} = \beta_0 + \beta_1 x$ es
---------------------------------------	--

$$b_1 - t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}} < \beta_1 < b_1 + t_{\alpha/2} \frac{s}{\sqrt{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n-2$ grados de libertad.

Ejemplo 11.2: Calcule un intervalo de confianza de 95% para β_1 en la recta de regresión $\mu_{Y|x} = \beta_0 + \beta_1 x$, con base en los datos de contaminación de la tabla 11.1.

Solución: A partir de los resultados dados en el ejemplo 11.1, se determina que $S_{xx} = 4152.18$ y $S_{xy} = 3752.09$. Además, se observa que $S_{yy} = 3713.88$. Recuerde que $b_1 = 0.903643$. En consecuencia,

$$s^2 = \frac{S_{yy} - b_1 S_{xy}}{n-2} = \frac{3713.88 - (0.903643)(3752.09)}{31} = 10.4299.$$

Por lo tanto, al sacar la raíz cuadrada obtenemos $s = 3.2295$. Si usamos la tabla A.4 encontramos que $t_{0.025} \approx 2.045$ para 31 grados de libertad. Así, un intervalo de confianza de 95% para β_1 es

$$0.903643 - \frac{(2.045)(3.2295)}{\sqrt{4152.18}} < \beta_1 < 0.903643 + \frac{(2.045)(3.2295)}{\sqrt{4152.18}},$$

que se simplifica a

$$0.8012 < \beta_1 < 1.0061.$$



Prueba de hipótesis sobre la pendiente

Para probar la hipótesis nula H_0 de que $\beta_1 = \beta_{10}$, en comparación con una alternativa posible, utilizamos de nuevo la distribución t con $n - 2$ grados de libertad con el fin de establecer una región crítica y después basar nuestra decisión en el valor de

$$t = \frac{b_1 - \beta_{10}}{s/\sqrt{S_{xx}}}.$$


El método se ilustra con el ejemplo siguiente.

Ejemplo 11.3: Utilice el valor estimado $b_1 = 0.903643$ del ejemplo 11.1 y pruebe la hipótesis de que $\beta_1 = 1.0$ en comparación con la alternativa de que $\beta_1 < 1.0$.

Solución: Las hipótesis son $H_0: \beta_1 = 1.0$ y $H_1: \beta_1 < 1.0$. Por lo tanto,

$$t = \frac{0.903643 - 1.0}{3.2295/\sqrt{4152.18}} = -1.92,$$

con $n - 2 = 31$ grados de libertad ($P \approx 0.03$).

Decisión: El valor t es significativo al nivel 0.03, lo cual sugiere evidencia sólida de que $\beta_1 < 1.0$. 

Una prueba t importante sobre la pendiente es la prueba de la hipótesis

$$H_0: \beta_1 = 0 \text{ en comparación con } H_1: \beta_1 \neq 0.$$

Cuando no se rechaza la hipótesis nula la conclusión es que no hay relación lineal significativa entre $E(y)$ y la variable independiente x . La gráfica de los datos del ejemplo 11.1 sugeriría que existe una relación lineal. Sin embargo, en ciertas aplicaciones en las que σ^2 es grande y, por ende, hay “ruido” considerable en los datos, una gráfica, aunque útil, quizá no produzca información clara para el investigador. El rechazo anterior de H_0 implica que hay una relación lineal significativa.

La figura 11.7 muestra la salida de resultados de MINITAB que presenta la prueba t para

$$H_0: \beta_1 = 0 \text{ en comparación con } H_1: \beta_1 \neq 0,$$

para los datos del ejemplo 11.1. Observe el coeficiente de regresión (Coef), el error estándar (EE Coef), el valor t (T) y el valor P (P). Se rechaza la hipótesis nula. Es claro que existe una relación lineal significativa entre la reducción de la demanda media del oxígeno químico y la reducción de los sólidos. Observe que el estadístico t se calcula como

$$t = \frac{\text{coeficiente}}{\text{error estándar}} = \frac{b_1}{s/\sqrt{S_{xx}}}.$$

El no rechazo de $H_0: \beta_1 = 0$ sugiere que no hay una relación lineal entre Y y x . La figura 11.8 es una ilustración de la implicación de este resultado; podría significar que los cambios de x tienen poco efecto sobre los cambios de Y , como se ve en el inciso a . Sin embargo, también puede indicar que la relación verdadera es no lineal, como se aprecia en b .

Cuando se rechaza $H_0: \beta_1 = 0$ existe la implicación de que el término lineal en x que reside en el modelo explica una parte significativa de la variabilidad de Y . Las dos gráfi-

Regression Analysis: COD versus Per_Red					
The regression equation is COD = 3.83 + 0.904 Per_Red					
Predictor	Coef	SE Coef	T	P	
Constant	3.830	1.768	2.17	0.038	
Per_Red	0.90364	0.05012	18.03	0.000	
S = 3.22954 R-Sq = 91.3% R-Sq(adj) = 91.0%					
Analysis of Variance					
Source	DF	SS	MS	F	P
Regression	1	3390.6	3390.6	325.08	0.000
Residual Error	31	323.3	10.4		
Total	32	3713.9			

Figura 11.7: Salida de resultados de MINITAB para la prueba t de los datos del ejemplo 11.1.

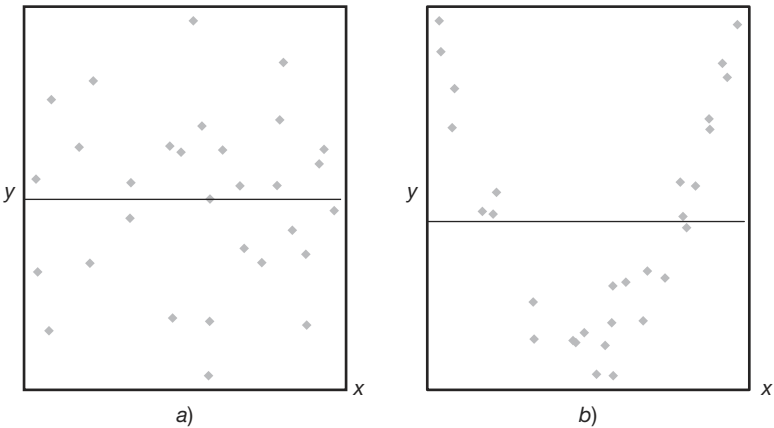


Figura 11.8: No se rechaza la hipótesis $H_0: \beta_1 = 0$.

cas que aparecen en la figura 11.9 ilustran los escenarios posibles. Como se muestra en el inciso a de la figura, el rechazo de H_0 sugiere que la relación en efecto es lineal. En el caso del inciso b , lo que se observa sugiere que, aunque el modelo contenga un efecto lineal, se podría obtener una mejor representación si se incluyera un término polinomial (tal vez cuadrático), es decir, términos que complementen el término lineal.

Inferencia estadística sobre la intersección

Los intervalos de confianza y la prueba de hipótesis del coeficiente β_0 se podrían establecer a partir del hecho de que B_0 también se distribuye de forma normal. No es difícil demostrar que

$$T = \frac{B_0 - \beta_0}{S \sqrt{\sum_{i=1}^n x_i^2 / (nS_{xx})}}$$

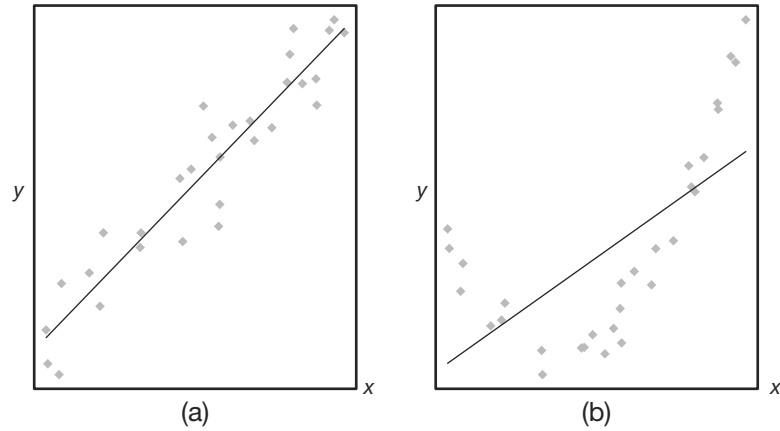


Figura 11.9: Se rechaza la hipótesis de que $H_0: \beta_1 = 0$.

tiene una distribución t con $n - 2$ grados de libertad, de manera que podemos construir un intervalo de confianza de $100(1 - \alpha)\%$ para α .

Intervalo de confianza para β_0 Un intervalo de confianza de $100(1 - \alpha)\%$ para el parámetro β_0 en la recta de regresión $\mu_{y|x} = \beta_0 + \beta_1 x$ es

$$b_0 - t_{\alpha/2} \frac{s}{\sqrt{nS_{xx}}} \sqrt{\sum_{i=1}^n x_i^2} < \beta_0 < b_0 + t_{\alpha/2} \frac{s}{\sqrt{nS_{xx}}} \sqrt{\sum_{i=1}^n x_i^2},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Ejemplo 11.4: Calcule un intervalo de confianza de 95% para β_0 en la recta de regresión $\mu_{y|x} = \beta_0 + \beta_1 x$ con base en los datos de la tabla 11.1.

Solución: En los ejemplos 11.1 y 11.2 se encontró que

$$S_{xx} = 4152.18 \quad \text{y} \quad s = 3.2295.$$

Del ejemplo 11.1 se tiene que

$$\sum_{i=1}^n x_i^2 = 41,086 \quad \text{y} \quad b_0 = 3.829633.$$

Si usamos la tabla A.4, encontramos que $t_{0.025} \approx 2.045$ para 31 grados de libertad. Por lo tanto, un intervalo de confianza de 95% para β_0 es

$$3.829633 - \frac{(2.045)(3.2295) \sqrt{41,086}}{\sqrt{(33)(4152.18)}} < \beta_0 < 3.829633 + \frac{(2.045)(3.2295) \sqrt{41,086}}{\sqrt{(33)(4152.18)}},$$

que se simplifica a $0.2132 < \beta_0 < 7.4461$. ▀

Para probar la hipótesis nula H_0 de que $\beta_0 = \beta_{00}$ en comparación con una alternativa posible utilizamos la distribución t con $n - 2$ grados de libertad para establecer una región crítica y, luego, basar nuestra decisión en el valor de

$$t = \frac{b_0 - \beta_{00}}{s \sqrt{\sum_{i=1}^n x_i^2 / (nS_{xx})}}.$$

Ejemplo 11.5: Utilice el valor estimado de $b_0 = 3.829633$ del ejemplo 11.1 y, a un nivel de significancia de 0.05, pruebe la hipótesis de que $\beta_0 = 0$ en comparación con la alternativa de que $\beta_0 \neq 0$. Entonces

Solución: Las hipótesis son $H_0: \beta_0 = 0$ y $H_1: \beta_0 \neq 0$. Así que,

$$t = \frac{3.829633 - 0}{3.2295 \sqrt{41,086 / [(33)(4152.18)]}} = 2.17,$$

con 31 grados de libertad. Por lo tanto, $P = \text{valor } P \approx 0.038$ y concluimos que $\beta_0 \neq 0$. Observe que esto tan sólo es Coef/desviación estándar, como se aprecia en la salida de resultados de MINITAB en la figura 11.7. El SE Coef es el error estándar de la intersección estimada. ─

Una medida de la calidad del ajuste: el coeficiente de determinación

Observe en la figura 11.7 que aparece un elemento denotado con R-Sq, cuyo valor es 91.3%. Esta cantidad, R^2 , se denomina **coeficiente de determinación** y es una medida de la **proporción de la variabilidad explicada por el modelo ajustado**. En la sección 11.8 se presentará el concepto del método del análisis de varianza para la prueba de hipótesis en la regresión. El enfoque del análisis de varianza utiliza la suma de los cuadrados del error $SCE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ y la **suma total de los cuadrados corregida STCC** $= \sum_{i=1}^n (y_i - \bar{y})^2$. Esta última representa la variación en los valores de respuesta que *idealmente* serían explicados con el modelo. El valor de la SCE es la variación debida al error, o la **variación no explicada**. Resulta claro que si la $SCE = 0$, toda variación queda explicada. La cantidad que representa la variación explicada es $STCC - SCE$. R^2 es el

$$\text{Coeficiente de determinación: } R^2 = 1 - \frac{SCE}{STCC}.$$

Advierta que si el ajuste es perfecto, *todos los residuales son cero*, y así $R^2 = 1.0$. Pero si la SCE es tan sólo un poco menor que la $STCC$, $R^2 \approx 0$. Observe en la salida de resultados de la figura 11.7 que el coeficiente de determinación sugiere que el modelo ajustado a los datos explica el 91.3% de la variabilidad observada en la respuesta, la reducción en la demanda de oxígeno químico.

La figura 11.10 ofrece ejemplos de una gráfica con un buen ajuste ($R^2 \approx 1.0$) en a) y una gráfica con un ajuste deficiente ($R^2 \approx 0$) en b).

Errores en el uso de R^2

Los analistas citan con mucha frecuencia los valores de R^2 , quizá debido a su simplicidad. Sin embargo, hay errores en su interpretación. La confiabilidad de R^2 depende del

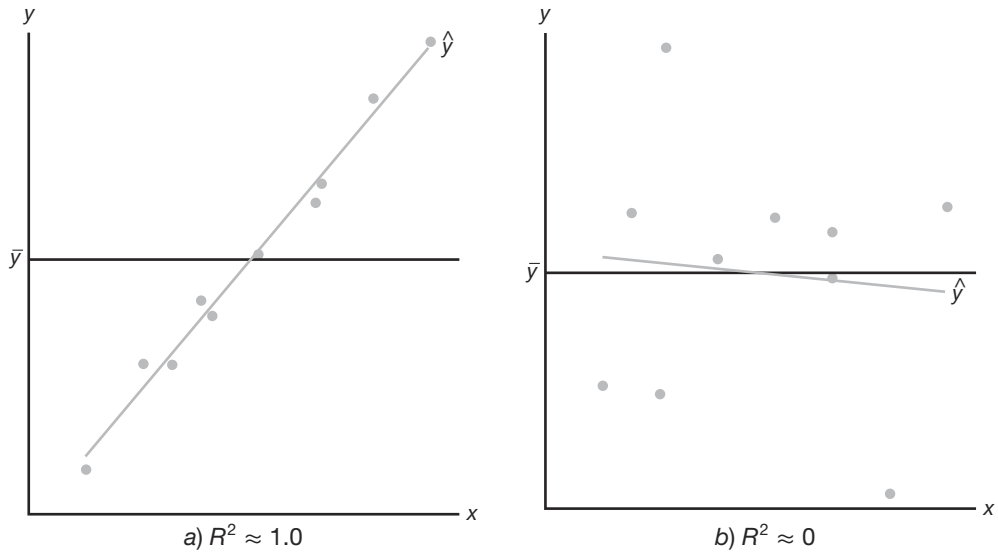


Figura 11.10: Gráficas que ilustran un ajuste muy bueno y otro deficiente.

tamaño del conjunto de los datos de la regresión y del tipo de aplicación. Resulta claro que $0 \leq R^2 \leq 1$, y el límite superior se logra cuando el ajuste a los datos es perfecto, es decir, cuando todos los residuales son cero. ¿Cuál es un valor aceptable de R^2 ? Se trata de una pregunta difícil de responder. Un químico encargado de establecer una calibración lineal de una pieza de equipo de alta precisión seguramente esperaría obtener un valor muy alto de R^2 (quizá superior a 0.99); mientras que un científico del comportamiento, que trabaja con datos en los que influye la variabilidad de la conducta humana, quizá se sentiría afortunado si obtuviera un valor de R^2 de hasta 0.70. Un individuo con experiencia en el ajuste de modelos tiene la sensibilidad para saber cuándo un valor es suficientemente grande dada la situación que está enfrentando. Es evidente que algunos fenómenos científicos se prestan más a un modelamiento más preciso que otros.

Es peligroso usar el criterio de R^2 para comparar *modelos en competencia* para el mismo conjunto de datos. Cuando se agregan términos adicionales al modelo, por ejemplo un regresor más, disminuye la *SCE*, lo que provoca que R^2 aumente (o al menos no disminuya). Esto implica que R^2 se puede volver artificialmente elevado por medio de la práctica inapropiada de **sobreajustar**, es decir, de incluir demasiados términos en el modelo. Por consiguiente, el incremento inevitable de R^2 que se logra al agregar términos adicionales no implica que éstos se necesitaban. En realidad, el modelo simple puede ser mejor para predecir los valores de la respuesta. En el capítulo 12, cuando se presente el concepto de los modelos que implican **más de un solo regresor**, se estudiará con detalle el papel del sobreajuste y su influencia sobre la capacidad de predicción. En este momento baste decir que *para seleccionar un modelo no se debe adoptar un proceso de selección que sólo incluya la consideración de R^2 .*

11.6 Predicción

Hay varias razones para construir un modelo de regresión lineal. Una de ellas es, desde luego, predecir valores de respuesta para uno o más valores de la variable independiente. En esta sección se centra el enfoque en los errores asociados con la predicción.

La ecuación $\hat{y} = b_0 + b_1x$ se puede utilizar para predecir o estimar la **respuesta media** $\mu_{Y|x_0}$ en $x = x_0$, donde x_0 no necesariamente es uno de los valores preestablecidos, o cuando $x = x_0$, se podría emplear para pronosticar un solo valor y_0 de la variable Y_0 . Se esperaría que el error de predicción fuera mayor para el caso de un solo valor pronosticado que para aquel en que se predice una media. Entonces, esto afectaría la anchura de los intervalos para los valores que se predicen.

Suponga que el experimentador desea construir un intervalo de confianza para $\mu_{Y|x_0}$. En tal caso debe usar el estimador puntual $\hat{Y}_0 = B_0 + B_1x_0$ para estimar $\mu_{Y|x_0} = \beta_0 + \beta_1x$. Se puede demostrar que la distribución muestral de \hat{Y}_0 es normal con media

$$\mu_{Y|x_0} = E(\hat{Y}_0) = E(B_0 + B_1x_0) = \beta_0 + \beta_1x_0 = \mu_{Y|x_0}$$

y varianza

$$\sigma_{\hat{Y}_0}^2 = \sigma_{B_0 + B_1x_0}^2 = \sigma_{\hat{Y} + B_1(x_0 - \bar{x})}^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right],$$

esta última surge del hecho de que $\text{Cov}(\bar{Y}_0, B_1) = 0$ (véase el ejercicio de repaso 11.61 de la página 438). Por consiguiente, ahora podemos construir un intervalo de confianza de $100(1 - \alpha)\%$ sobre la respuesta media $\mu_{Y|x_0}$ a partir del estadístico

$$T = \frac{\hat{Y}_0 - \mu_{Y|x_0}}{S \sqrt{1/n + (x_0 - \bar{x})^2/S_{xx}}},$$

que tiene una distribución t con $n - 2$ grados de libertad.

Intervalo de confianza para $\mu_{Y|x_0}$ Un intervalo de confianza de $100(1 - \alpha)\%$ para la respuesta media $\mu_{Y|x_0}$ es

$$\hat{y}_0 - t_{\alpha/2}s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < \mu_{Y|x_0} < \hat{y}_0 + t_{\alpha/2}s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

$t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Ejemplo 11.6: Con los datos de la tabla 11.1 construya límites de confianza de 95% para la respuesta media $\mu_{Y|x_0}$.

Solución: A partir de la ecuación de regresión encontramos que, para $x_0 = 20\%$ de reducción de sólidos, digamos,

$$\hat{y}_0 = 3.829633 + (0.903643)(20) = 21.9025.$$

Además, $\bar{x} = 33.4545$, $S_{xx} = 4152.18$, $s = 3.2295$ y $t_{0.025} \approx 2.045$ para 31 grados de libertad. Por lo tanto, un intervalo de confianza de 95% para $\mu_{Y|20}$ es

$$\begin{aligned} 21.9025 - (2.045)(3.2295) \sqrt{\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} &< \mu_{Y|20} \\ &< 21.9025 + (2.045)(3.2295) \sqrt{\frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}}, \end{aligned}$$

o simplemente, $20.1071 < \mu_{Y|20} < 23.6979$. ▀

Si repetimos los cálculos anteriores para cada uno de los diferentes valores de x_0 , obtenemos los límites de confianza correspondientes para cada $\mu_{Y|x_0}$. En la figura 11.11 se presentan los datos de los puntos, la recta de regresión estimada y los límites de confianza superior e inferior sobre la media de $Y|x$.

En el ejemplo 11.6 tenemos 95% de confianza en que la reducción media poblacio-

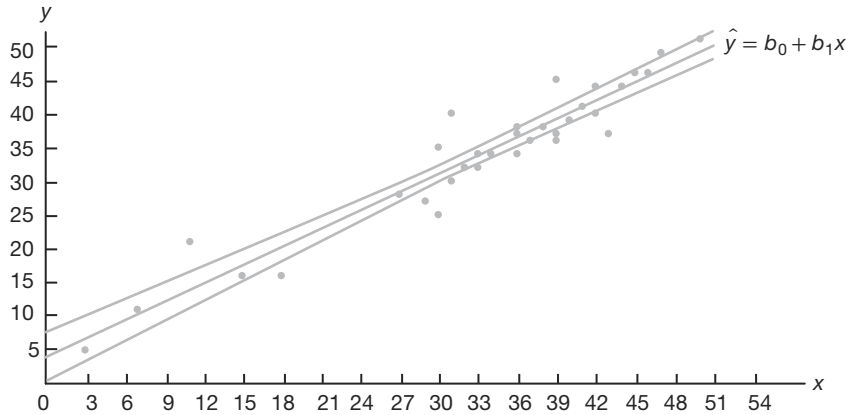


Figura 11.11: Límites de confianza para el valor medio de $Y|x$.

nal en la demanda de oxígeno químico estará entre el 20.1071% y 23.6979%, cuando la reducción de sólidos sea de 20%.

Predicción del intervalo

Otro tipo de intervalo que con frecuencia se malinterpreta y se confunde con aquel dado para $\mu_{Y|x}$ es el intervalo de la predicción para una respuesta futura observada. En realidad, en muchos casos el intervalo de la predicción es más relevante para el científico o el ingeniero que el intervalo de confianza sobre la media. En el ejemplo del contenido de alquitrán y la temperatura de entrada, mencionado en la sección 11.1, seguramente sería interesante no sólo estimar la media del contenido de alquitrán a una temperatura específica, sino también construir un intervalo que refleje el error en la predicción de una cantidad futura observada del contenido de alquitrán a la temperatura dada.

Para obtener un **intervalo de predicción** para cualquier valor único y_0 de la variable Y_0 es necesario estimar la varianza de las diferencias entre las ordenadas \hat{y}_0 , obtenidas de las rectas de regresión calculadas en el muestreo repetido cuando $x = x_0$, y la ordenada verdadera correspondiente y_0 . Podríamos considerar la diferencia $\hat{y}_0 - y_0$ como un valor de la variable aleatoria $\hat{Y}_0 - Y_0$, cuya distribución muestral se podría demostrar que es normal con media

$$\mu_{\hat{Y}_0 - Y_0} = E(\hat{Y}_0 - Y_0) = E[B_0 + B_1x_0 - (\beta_0 + \beta_1x_0 + \epsilon_0)] = 0$$

y varianza

$$\sigma_{\hat{Y}_0 - Y_0}^2 = \sigma_{B_0 + B_1x_0 - \epsilon_0}^2 = \sigma_{\hat{Y} + B_1(x_0 - \bar{x}) - \epsilon_0}^2 = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right].$$

Así, un intervalo de predicción de $100(1 - \alpha)\%$ para un solo valor pronosticado y_0 se puede construir a partir del estadístico

$$T = \frac{\hat{Y}_0 - Y_0}{S \sqrt{1 + 1/n + (x_0 - \bar{x})^2 / S_{xx}}},$$

que tiene una distribución t con $n - 2$ grados de libertad.

Intervalo de predicción para y_0 Un intervalo de predicción de $100(1 - \alpha)\%$ para una sola respuesta y_0 es dado por

$$\hat{y}_0 - t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} < y_0 < \hat{y}_0 + t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}},$$

donde $t_{\alpha/2}$ es un valor de la distribución t con $n - 2$ grados de libertad.

Es claro que hay una diferencia entre el concepto de un intervalo de confianza y el del intervalo de predicción antes descrito. La interpretación del intervalo de confianza es idéntica a la que se describió para todos los intervalos de confianza sobre los parámetros de la población estudiados en el libro. De hecho, $\mu_{Y|x_0}$ es un parámetro de la población. Sin embargo, el intervalo de la predicción calculado representa un intervalo que tiene una probabilidad igual a $1 - \alpha$ de contener no un parámetro sino un valor futuro de y_0 de la variable aleatoria Y_0 .

Ejemplo 11.7: Con los datos de la tabla 11.1 construya un intervalo de predicción de 95% para y_0 cuando $x_0 = 20\%$.

Solución: Tenemos que $n = 33$, $x_0 = 20$, $\bar{x} = 33.4545$, $\hat{y}_0 = 21.9025$, $S_{xx} = 4152.18$, $s = 3.2295$, y $t_{0.025} \approx 2.045$ para 31 grados de libertad. Por lo tanto, un intervalo de predicción de 95% para y_0 es

$$\begin{aligned} 21.9025 - (2.045)(3.2295) \sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} &< y_0 \\ &< 21.9025 + (2.045)(3.2295) \sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}}, \end{aligned}$$

que se simplifica como $15.0585 < y_0 < 28.7464$. ▀

En la figura 11.12 se presenta otra gráfica de los datos de reducción de la demanda de oxígeno químico, tanto con los intervalos de confianza de la respuesta media como con el intervalo de predicción sobre una respuesta individual. En el caso de la respuesta media la gráfica refleja un intervalo mucho más angosto alrededor de la recta de regresión.

Ejercicios

11.15 Remítase al ejercicio 11.1 de la página 398,

- evalúe s^2 ;
- pruebe la hipótesis de que $\beta_1 = 0$ en comparación con la alternativa de que $\beta_1 \neq 0$ a un nivel de significancia de 0.05, e interprete la decisión resultante.

11.16 Remítase al ejercicio 11.2 de la página 398,

- evalúe s^2 ;
- construya un intervalo de confianza de 95% para β_0 ;
- construya un intervalo de confianza de 95% para β_1 .

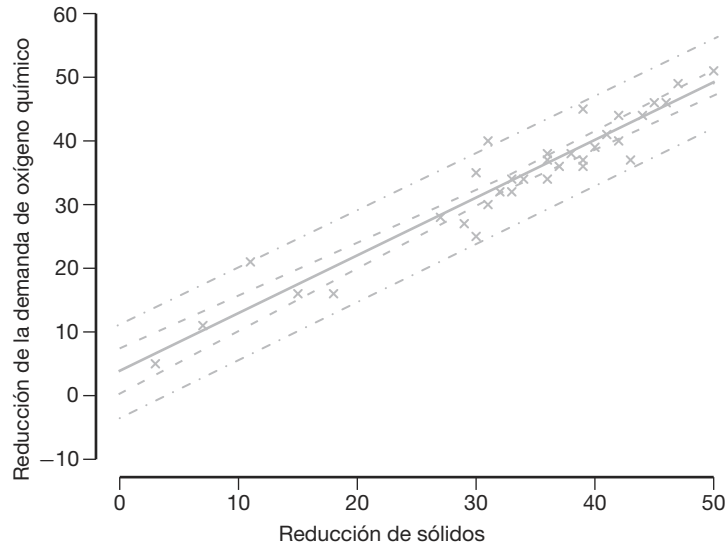


Figura 11.12: Intervalos de confianza y predicción para los datos de la reducción de la demanda de oxígeno químico; las bandas internas indican los límites de confianza para las respuestas medias y las externas señalan los límites de predicción para las respuestas futuras.

11.17 Remítase al ejercicio 11.5 de la página 398,

- evalúe s^2 ;
- construya un intervalo de confianza de 95% para β_0 ;
- construya un intervalo de confianza de 95% para β_1 .

11.18 Remítase al ejercicio 11.6 de la página 399,

- evalúe s^2 ;
- construya un intervalo de confianza de 99% para β_0 ;
- construya un intervalo de confianza de 99% para β_1 .

11.19 Remítase al ejercicio 11.3 de la página 398,

- evalúe s^2 ;
- construya un intervalo de confianza de 99% para β_0 ;
- construya un intervalo de confianza de 99% para β_1 .

11.20 Pruebe la hipótesis de que $\beta_0 = 10$ en el ejercicio 11.8 de la página 399, en comparación con la alternativa de que $\beta_0 < 10$. Utilice un nivel de significancia de 0.05.

11.21 Pruebe la hipótesis de que $\beta_1 = 6$ en el ejercicio 11.9 de la página 399, en comparación con la alternativa de que $\beta_1 < 6$. Utilice un nivel de significancia de 0.025.

11.22 Utilice el valor de s^2 que se obtuvo en el ejercicio 11.16a para construir un intervalo de confianza de 95% para $\mu_{Y|85}$ en el ejercicio 11.2 de la página 398.

11.23 Remítase al ejercicio 11.6 de la página 399 y utilice el valor de s^2 que se obtuvo en el ejercicio 11.18a para calcular

- un intervalo de confianza de 95% para la resistencia media al corte cuando $x = 24.5$;
- un intervalo de predicción de 95% para un solo valor pronosticado de la resistencia al corte cuando $x = 24.5$.

11.24 Utilice el valor de s^2 que se obtuvo en el ejercicio 11.17a) y grafique la regresión lineal y las bandas de confianza de 95% para la respuesta media $\mu_{Y|x}$ en el caso de los datos del ejercicio 11.5 de la página 398.

11.25 Utilice el valor de s^2 que se obtuvo en el ejercicio 11.17a) y construya un intervalo de confianza de 95% para la cantidad de azúcar convertida correspondiente a $x = 1.6$ en el ejercicio 11.5 de la página 398.

11.26 Remítase al ejercicio 11.3 de la página 398, y utilice el valor de s^2 que se obtuvo en el ejercicio 11.19a para calcular

- un intervalo de confianza de 99% para la cantidad promedio del producto químico que se disolverá en 100 gramos de agua a 50°C;

- b) un intervalo de predicción de 99% para la cantidad de producto químico que se disolverá en 100 gramos de agua a 50°C.

11.27 Considere la regresión de la distancia recorrida para ciertos automóviles, en millas por galón (mpg) y su peso en libras (wt). Los datos son de la revista *Consumer Reports* (abril de 1997). En la figura 11.13 se presenta una parte de la salida del SAS con los resultados del procedimiento.

- a) Estime la distancia recorrida para un vehículo que pesa 4000 libras.
 b) Suponga que los ingenieros de Honda afirman que, en promedio, el Civic (o cualquier otro modelo que pese 2440 libras) recorre más de 30 millas por galón (mpg). Con base en los resultados del análisis de regresión, ¿creería usted dicha afirmación? Explique su respuesta.
 c) Los ingenieros de diseño del Lexus ES300 consideraron que un rendimiento de 18 mpg sería el objetivo ideal para dicho modelo (o cualquier otro modelo que pese 3390 libras), aunque se espera que haya cierta variación. ¿Es probable que ese objetivo sea realista? Comente al respecto.

11.28 Existen aplicaciones importantes en las que, debido a restricciones científicas conocidas, la recta de regresión **debe atravesar el origen**, es decir, la intersección debe estar en el cero. En otras palabras, el modelo debe ser

$$Y_i = \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n,$$

y tan sólo se requiere estimar un parámetro sencillo. Con frecuencia a este modelo se le denomina **modelo de regresión por el origen**.

- a) Demuestre que el estimador de mínimos cuadrados para la pendiente es

$$b_1 = \left(\sum_{i=1}^n x_i y_i \right) / \left(\sum_{i=1}^n x_i^2 \right).$$

- b) Demuestre que $\sigma_{B_1}^2 = \sigma^2 / \left(\sum_{i=1}^n x_i^2 \right)$.

- c) Demuestre que b_1 del inciso a es un estimador insesgado para β_1 . Es decir, demuestre que $E(B_1) = \beta_1$.

11.29 Dado el conjunto de datos

y	x
7	2
50	15
100	30
40	10
70	20

- a) Grafique los datos.
 b) Ajuste una recta de regresión por el origen.
 c) Grafique la recta de regresión sobre la gráfica de los datos.
 d) Calcule una fórmula general (en términos de y_i y la pendiente b_1) para el estimador de σ^2 .
 e) Calcule una fórmula para $\text{Var}(\hat{y}_i)$; $i = 1, 2, \dots, n$, aplicable a este caso.
 f) Grafique límites de confianza de 95% para la respuesta media alrededor de la recta de regresión.

11.30 Para los datos del ejercicio 11.29 calcule un intervalo de predicción de 95% en $x = 25$.

		Root MSE		1.48794	R-Square	0.9509
		Dependent Mean		21.50000	Adj R-Sq	0.9447
		Parameter Estimates				
		Variable	DF	Estimate	Error	t Value
		Intercept	1	44.78018	1.92919	23.21
		WT	1	-0.00686	0.00055133	-12.44
						Pr > t
						<.0001
						<.0001
MODEL	WT	MPG	Predict	LMean	UMean	Lpred
GMC	4520	15	13.7720	11.9752	15.5688	9.8988
Geo	2065	29	30.6138	28.6063	32.6213	26.6385
Honda	2440	31	28.0412	26.4143	29.6681	24.2439
Hyundai	2290	28	29.0703	27.2967	30.8438	25.2078
Infiniti	3195	23	22.8618	21.7478	23.9758	19.2543
Isuzu	3480	21	20.9066	19.8160	21.9972	17.3062
Jeep	4090	15	16.7219	15.3213	18.1224	13.0158
Land	4535	13	13.6691	11.8570	15.4811	9.7888
Lexus	3390	22	21.5240	20.4390	22.6091	17.9253
Lincoln	3930	18	17.8195	16.5379	19.1011	14.1568
						21.4822
						0.18051

Figura 11.13: Salida de resultados del SAS para el ejercicio 11.27.

11.7 Selección de un modelo de regresión

Gran parte de lo que se ha presentado hasta ahora acerca de la regresión que involucra una sola variable independiente depende de la suposición de que el modelo elegido es correcto, la suposición de que $\mu_{Y|x}$ se relaciona con x linealmente en los parámetros. Es cierto que no se esperaría que la predicción de la respuesta fuera buena si hubiera diversas variables independientes que no se tomaran en cuenta en el modelo, que afectarían la respuesta y variarían en el sistema. Además, la predicción seguramente sería inadecuada si la estructura verdadera que relaciona $\mu_{Y|x}$ con x fuera extremadamente no lineal en el rango de las variables consideradas.

Es frecuente que se utilice el modelo de regresión lineal simple aun cuando se sepa que el modelo no es lineal o que se desconozca la estructura verdadera. Este método suele ser acertado, en particular cuando el rango de las x es estrecho. De esta manera, el modelo que se utiliza se vuelve una función de aproximación que se espera sea una representación adecuada del panorama verdadero en la región de interés. Sin embargo, hay que señalar el efecto que tendría un modelo inadecuado sobre los resultados presentados hasta este momento. Por ejemplo, si el modelo verdadero, desconocido para el experimentador, es lineal en más de una x , digamos,

$$\mu_{Y|x_1, x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

entonces el estimado $b_1 = S_{xy}/S_{xx}$ de los mínimos cuadrados ordinarios que se calcula considerando tan sólo x_1 en el experimento es, en circunstancias generales, un estimado sesgado del coeficiente β_1 , donde el sesgo es una función del coeficiente adicional β_2 (véase el ejercicio de repaso 11.65 en la página 438). Asimismo, el estimado s^2 para σ^2 es sesgado debido a la variable adicional.

11.8 El método del análisis de varianza

Con frecuencia el problema de analizar la calidad de la recta de regresión estimada se maneja por medio del método del **análisis de varianza** (ANOVA), que es un procedimiento mediante el cual la variación total de la variable dependiente se subdivide en componentes significativos, que luego se observan y se tratan en forma sistemática. El análisis de varianza, que se estudia en el capítulo 13, es un recurso poderoso que se emplea en muchas situaciones.

Suponga que tenemos n puntos de datos experimentales en la forma usual (x_i, y_i) y que se estima la recta de regresión. En la sección 11.4 para la estimación de σ^2 se estableció la identidad

$$S_{yy} = b_1 S_{xy} + SCE.$$

Una formulación alternativa y quizá más informativa es la siguiente:

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

Logramos hacer una partición de la **suma total de los cuadrados corregida de y** en dos componentes que deberían proporcionar un significado particular para el experimentador. Esta partición se debería indicar en forma simbólica como

$$STCC = SCR + SCE.$$

El primer componente de la derecha, *SCR*, se denomina **suma de cuadrados de la regresión** y refleja la cantidad de variación de los valores *y* que se **explica con el modelo**, que en este caso es la línea recta postulada. El segundo componente es la ya conocida suma de cuadrados del error, que refleja la variación alrededor de la recta de regresión.

Suponga que nos interesa probar la hipótesis

$$H_0: \beta_1 = 0 \text{ en comparación con } H_1: \beta_1 \neq 0,$$

donde la hipótesis nula en esencia dice que el modelo es $\mu_{Y|x} = \beta_0$; es decir, la variación en los resultados *Y* debida a las fluctuaciones de probabilidad o aleatorias que son independientes de los valores de *x*. Esta condición se refleja en la figura 11.10*b*). En las condiciones de esta hipótesis nula se puede demostrar que SCR/σ^2 , y SCE/σ^2 son valores de variables chicuadradas independientes con 1 y $n - 2$ grados de libertad, respectivamente y, usando el teorema 7.12, se sigue que $STCC/\sigma^2$ también es un valor de una variable chi cuadrada con $n - 1$ grados de libertad. Para probar la hipótesis anterior calculamos

$$f = \frac{SCR / 1}{SCE / (n-2)} = \frac{SCR}{s^2}$$

y rechazamos H_0 al nivel de significancia α cuando $f > f_\alpha(1, n - 2)$.

Por lo general los cálculos se resumen mediante las medias de una **tabla de análisis de varianza**, como se indica en la tabla 11.2. Es costumbre referirse a las distintas sumas de los cuadrados divididos entre sus respectivos grados de libertad como **cuadrados medios**.

Tabla 11.2: Análisis de varianza para la prueba de $\beta_1 = 0$

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio	<i>f</i> calculada
Regresión	<i>SCR</i>	1	$\frac{SCR}{s^2}$	$\frac{SCR}{s^2}$
Error	<i>SCE</i>	$n - 2$	$s^2 = \frac{SCE}{n-2}$	
Total	<i>STCC</i>	$n - 1$		

Cuando se rechaza la hipótesis nula, es decir, cuando el estadístico *F* calculado excede al valor crítico $f_\alpha(1, n - 2)$, concluimos que **hay una cantidad significativa de variación en la respuesta justificada por el modelo postulado, que es la función de la línea recta**. Si el estadístico *F* está en la región de no rechazo, se concluye que los datos no reflejan evidencia suficiente para apoyar el modelo que se postula.

En la sección 11.5 se presentó un procedimiento donde se usa el estadístico

$$T = \frac{B_1 - \beta_{10}}{S/\sqrt{S_{xx}}}$$

para probar la hipótesis

$$H_0: \beta_1 = \beta_{10} \text{ contra } H_1: \beta_1 \neq \beta_{10},$$

donde *T* sigue la distribución *t* con $n - 2$ grados de libertad. La hipótesis se rechaza si $|t| > t_{\alpha/2}$ para un nivel de significancia α . Es interesante observar que en el caso especial en que probamos

$$H_0: \beta_1 = 0 \text{ en comparación con } H_1: \beta_1 \neq 0,$$

el valor del estadístico T se convierte en

$$t = \frac{b_1}{s/\sqrt{S_{xx}}},$$

y la hipótesis a considerar es idéntica a la que se prueba en la tabla 11.2. En otras palabras, la hipótesis nula establece que la variación en la respuesta se debe tan sólo al azar. El análisis de varianza utiliza la distribución F en vez de la distribución t . Para la alternativa bilateral ambos enfoques son idénticos. Esto se observa si se escribe

$$t^2 = \frac{b_1^2 S_{xx}}{s^2} = \frac{b_1 S_{xy}}{s^2} = \frac{SCR}{s^2},$$

que da como resultado un valor idéntico al valor f utilizado en el análisis de varianza. La relación fundamental entre la distribución t con ν grados de libertad y la distribución F con 1 y ν grados de libertad es

$$t^2 = f(1, \nu).$$

Desde luego, la prueba t permite probar en comparación con una alternativa unilateral, en tanto que la prueba F está restringida a una prueba en comparación con una alternativa bilateral.

Salida de resultados por computadora comentados para la regresión lineal simple

Considere nuevamente los datos de la tabla 11.1 sobre la reducción de la demanda de oxígeno químico. En las figuras 11.14 y 11.15 se presentan salidas de los resultados por computadora más completos. De nuevo se ilustran con el software *MINITAB*. La columna de la razón t indica pruebas para la hipótesis nula de valores de cero en el parámetro. El término “Fit” denota los valores \hat{y} , que con frecuencia se denominan **valores ajustados**. El término “SE Fit” se emplea para calcular los intervalos de confianza sobre la respuesta media. El elemento R^2 se calcula como $(SCR/STCC) \times 100$, y significa la proporción de variación en y explicada por la regresión de la línea recta. Asimismo, se incluyen los intervalos de confianza sobre la respuesta media y los intervalos de predicción sobre una observación nueva.

11.9 Prueba para la linealidad de la regresión: datos con observaciones repetidas

En ciertos tipos de situaciones experimentales el investigador tiene la capacidad de efectuar observaciones repetidas de la respuesta para cada valor de x . Aunque no es necesario tener dichas repeticiones para estimar β_0 y β_1 , las repeticiones permiten al experimentador obtener información cuantitativa acerca de lo apropiado que resulta el modelo. De hecho, si se generan observaciones repetidas, el investigador puede efectuar una prueba de significancia para determinar si el modelo es o no adecuado.

The regression equation is COD = 3.83 + 0.904 Per_Red						
Predictor	Coef	SE Coef	T	P		
Constant	3.830	1.768	2.17	0.038		
Per_Red	0.90364	0.05012	18.03	0.000		
S = 3.22954	R-Sq = 91.3%	R-Sq(adj) = 91.0%				
Analysis of Variance						
Source	DF	SS	MS	F	P	
Regression	1	3390.6	3390.6	325.08	0.000	
Residual Error	31	323.3	10.4			
Total	32	3713.9				

Obs	Per_Red	COD	Fit	SE Fit	Residual	St Resid
1	3.0	5.000	6.541	1.627	-1.541	-0.55
2	36.0	34.000	36.361	0.576	-2.361	-0.74
3	7.0	11.000	10.155	1.440	0.845	0.29
4	37.0	36.000	37.264	0.590	-1.264	-0.40
5	11.0	21.000	13.770	1.258	7.230	2.43
6	38.0	38.000	38.168	0.607	-0.168	-0.05
7	15.0	16.000	17.384	1.082	-1.384	-0.45
8	39.0	37.000	39.072	0.627	-2.072	-0.65
9	18.0	16.000	20.095	0.957	-4.095	-1.33
10	39.0	36.000	39.072	0.627	-3.072	-0.97
11	27.0	28.000	28.228	0.649	-0.228	-0.07
12	39.0	45.000	39.072	0.627	5.928	1.87
13	29.0	27.000	30.035	0.605	-3.035	-0.96
14	40.0	39.000	39.975	0.651	-0.975	-0.31
15	30.0	25.000	30.939	0.588	-5.939	-1.87
16	41.0	41.000	40.879	0.678	0.121	0.04
17	30.0	35.000	30.939	0.588	4.061	1.28
18	42.0	40.000	41.783	0.707	-1.783	-0.57
19	31.0	30.000	31.843	0.575	-1.843	-0.58
20	42.0	44.000	41.783	0.707	2.217	0.70
21	31.0	40.000	31.843	0.575	8.157	2.57
22	43.0	37.000	42.686	0.738	-5.686	-1.81
23	32.0	32.000	32.746	0.567	-0.746	-0.23
24	44.0	44.000	43.590	0.772	0.410	0.13
25	33.0	34.000	33.650	0.563	0.350	0.11
26	45.0	46.000	44.494	0.807	1.506	0.48
27	33.0	32.000	33.650	0.563	-1.650	-0.52
28	46.0	46.000	45.397	0.843	0.603	0.19
29	34.0	34.000	34.554	0.563	-0.554	-0.17
30	47.0	49.000	46.301	0.881	2.699	0.87
31	36.0	37.000	36.361	0.576	0.639	0.20
32	50.0	51.000	49.012	1.002	1.988	0.65
33	36.0	38.000	36.361	0.576	1.639	0.52

Figura 11.14: Salida de resultados de *MINITAB* de la regresión lineal simple para los datos de reducción de la demanda de oxígeno químico; parte I.

Seleccionemos una muestra aleatoria de n observaciones utilizando k valores distintos de x , por ejemplo, x_1, x_2, \dots, x_k , tales que la muestra contenga n_1 valores observados de la variable aleatoria Y_1 correspondientes a los valores x_1 , con n_2 valores observados de Y_2 correspondientes a x_2, \dots, n_k valores observados de Y_k correspondientes a x_k . Necesariamente, $n = \sum_{i=1}^k n_i$.

Obs	Fit	SE Fit	95% CI	95% PI
1	6.541	1.627	(3.223, 9.858)	(-0.834, 13.916)
2	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)
3	10.155	1.440	(7.218, 13.092)	(2.943, 17.367)
4	37.264	0.590	(36.062, 38.467)	(30.569, 43.960)
5	13.770	1.258	(11.204, 16.335)	(6.701, 20.838)
6	38.168	0.607	(36.931, 39.405)	(31.466, 44.870)
7	17.384	1.082	(15.177, 19.592)	(10.438, 24.331)
8	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
9	20.095	0.957	(18.143, 22.047)	(13.225, 26.965)
10	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
11	28.228	0.649	(26.905, 29.551)	(21.510, 34.946)
12	39.072	0.627	(37.793, 40.351)	(32.362, 45.781)
13	30.035	0.605	(28.802, 31.269)	(23.334, 36.737)
14	39.975	0.651	(38.648, 41.303)	(33.256, 46.694)
15	30.939	0.588	(29.739, 32.139)	(24.244, 37.634)
16	40.879	0.678	(39.497, 42.261)	(34.149, 47.609)
17	30.939	0.588	(29.739, 32.139)	(24.244, 37.634)
18	41.783	0.707	(40.341, 43.224)	(35.040, 48.525)
19	31.843	0.575	(30.669, 33.016)	(25.152, 38.533)
20	41.783	0.707	(40.341, 43.224)	(35.040, 48.525)
21	31.843	0.575	(30.669, 33.016)	(25.152, 38.533)
22	42.686	0.738	(41.181, 44.192)	(35.930, 49.443)
23	32.746	0.567	(31.590, 33.902)	(26.059, 39.434)
24	43.590	0.772	(42.016, 45.164)	(36.818, 50.362)
25	33.650	0.563	(32.502, 34.797)	(26.964, 40.336)
26	44.494	0.807	(42.848, 46.139)	(37.704, 51.283)
27	33.650	0.563	(32.502, 34.797)	(26.964, 40.336)
28	45.397	0.843	(43.677, 47.117)	(38.590, 52.205)
29	34.554	0.563	(33.406, 35.701)	(27.868, 41.239)
30	46.301	0.881	(44.503, 48.099)	(39.473, 53.128)
31	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)
32	49.012	1.002	(46.969, 51.055)	(42.115, 55.908)
33	36.361	0.576	(35.185, 37.537)	(29.670, 43.052)

Figura 11.15: Salida de resultados de *MINITAB* de la regresión lineal simple para los datos de reducción de la demanda de oxígeno químico; parte II.

Definimos

y_{ij} = el j -ésimo valor de la variable aleatoria Y_i ,

$$y_i = T_i = \sum_{j=1}^{n_i} y_{ij},$$

$$\bar{y}_i = \frac{T_i}{n_i}.$$

Entonces, si se realizaron $n_4 = 3$ mediciones de Y que corresponden a $x = x_4$, estas observaciones se indicarían por medio de y_{41} , y_{42} y y_{43} . Por lo tanto,

$$T_i = y_{41} + y_{42} + y_{43}.$$

El concepto de la falta de ajuste

La suma de cuadrados del error consta de dos partes: la cantidad debida a la variación entre los valores de Y dentro de valores dados de x , y un componente que normalmente

se denomina contribución a la **falta de ajuste**. El primer componente refleja tan sólo la variación aleatoria, o **error experimental puro**, en tanto que el segundo es una medida de la variación sistemática introducida por los términos de orden superior. En nuestro caso éstos son términos de x distintos de la contribución lineal o de primer orden. Observe que al elegir un modelo lineal en esencia asumimos que este segundo componente no existe y que, en consecuencia, la suma de cuadrados del error se debe por completo a errores aleatorios. Si éste fuera el caso, entonces $s^2 = SCE/(n-2)$ es un estimado insesgado de σ^2 . Sin embargo, si el modelo no se ajusta a los datos en forma apropiada, entonces la suma de cuadrados del error estará inflada y producirá un estimador sesgado de σ^2 . Ya sea que el modelo se ajuste o no a los datos, siempre que se tienen observaciones repetidas es posible obtener un estimador insesgado de σ^2 calculando

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n_i - 1}, \quad i = 1, 2, \dots, k,$$

para cada uno de los k valores distintos de x y, después, agrupando estas varianzas, tenemos

$$s^2 = \frac{\sum_{i=1}^k (n_i - 1)s_i^2}{n - k} = \frac{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}{n - k}.$$

El numerador de s^2 es una **medida del error experimental puro**. A continuación se presenta un procedimiento de cálculo para separar la suma de los cuadrados del error en los dos componentes que representan el error puro y la falta de ajuste:

Cálculo de la suma de los cuadrados de la falta de ajuste	1. Calcular la suma de los cuadrados del error puro
---	--

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2.$$

Esta suma de cuadrados tiene $n - k$ grados de libertad asociados con ella, y el cuadrado medio resultante es el estimador insesgado s^2 de σ^2 .

2. Restar la suma de los cuadrados del error puro de la suma de los cuadrados del error, SCE , con lo que se obtiene la suma de los cuadrados debida a la falta de ajuste. Los grados de libertad de la falta de ajuste también se obtienen simplemente restando $(n-2) - (n-k) = k-2$.

Los cálculos necesarios para probar hipótesis en un problema de regresión con mediciones repetidas de la respuesta se pueden resumir como se muestra en la tabla 11.3.

Las figuras 11.16 y 11.17 ilustran los puntos muestrales para las situaciones del “modelo correcto” y del “modelo incorrecto”. En la figura 11.16, donde $\mu_{Y|x}$ cae sobre una línea recta, no hay falta de ajuste cuando se asume un modelo lineal, por lo que la variación muestral alrededor de la recta de regresión es un error puro que resulta de la variación que ocurre entre observaciones repetidas. En la figura 11.17, donde es evidente que $\mu_{Y|x}$ no cae sobre una línea recta, la responsable de la mayor parte de la variación alrededor de la recta de regresión, además del error puro, es la falta de ajuste que resulta de seleccionar por error un modelo lineal.

Tabla 11.3: Análisis de varianza para la prueba de linealidad de la regresión

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	f calculada
Regresión	SCR	1	SCR	$\frac{SCR}{s^2}$
Error	SCE	$n - 2$		
Falta de ajuste	$\left\{ \begin{array}{l} SCE - SCE \text{ (puro)} \\ SCE \text{ (puro)} \end{array} \right.$	$\left\{ \begin{array}{l} k - 2 \\ n - k \end{array} \right.$	$\frac{SCE - SCE \text{ (puro)}}{k - 2}$ $s^2 = \frac{SCE \text{ (puro)}}{n - k}$	$\frac{SCE - SCE \text{ (puro)}}{s^2 (k - 2)}$
Error puro				
Total	$STCC$	$n - 1$		

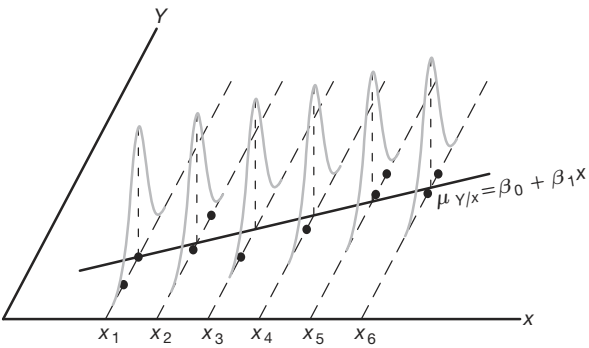


Figura 11.16: Modelo lineal correcto con componente sin falta de ajuste.

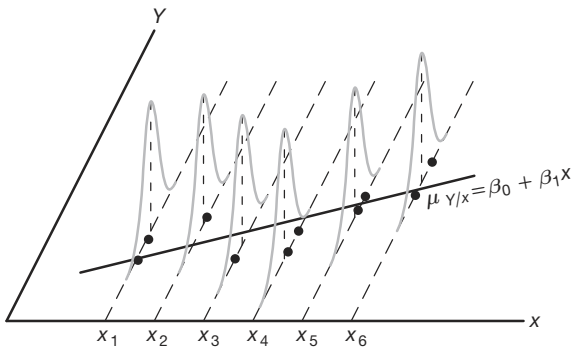


Figura 11.17: Modelo lineal incorrecto con componente de falta de ajuste.

¿Por qué es importante detectar la falta de ajuste?

El concepto de falta de ajuste es muy importante en las aplicaciones del análisis de regresión. De hecho, la necesidad de construir o diseñar un experimento que tome en cuenta la falta de ajuste se vuelve más crítica a medida que el problema y el mecanismo subyacente implicados se vuelven más complicados. Es cierto que no siempre se puede tener la certeza de que la estructura que se postula, en este caso el modelo de regresión lineal, sea una representación correcta o incluso adecuada. El ejemplo siguiente muestra la manera en que se parte la suma de cuadrados del error en los dos componentes que representan el error puro y la falta de ajuste. Lo adecuado del modelo se prueba al nivel de significancia α , comparando el cuadrado medio de la falta de ajuste dividido entre s^2 con $f_{\alpha}(k - 2, n - k)$.

Ejemplo 11.8: En la tabla 11.4 se presenta el registro de las observaciones del producto de una reacción química tomadas a distintas temperaturas. Calcule el modelo lineal $\mu_{Y|x} = \beta_0 + \beta_1x$ y pruebe la falta de ajuste.

Solución: Los resultados de los cálculos se presentan en la tabla 11.5.

Conclusión: La partición de la variación total de esta manera revela una variación significativa debida al modelo lineal y una cantidad insignificante de variación debida a la falta de ajuste. Por consiguiente, los datos experimentales no parecen sugerir la necesidad de considerar en el modelo términos superiores a los de primer orden y no se rechaza la hipótesis nula.

Tabla 11.4: Datos para el ejemplo 11.8

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

Tabla 11.5: Análisis de varianza de los datos de producto-temperatura

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios	f calculada	Valores P
Regresión	509.2507	1	509.2507	1531.58	< 0.0001
Error	3.8660	10			
Falta de ajuste	1.2060	2	0.6030	1.81	0.2241
Error puro	2.6600	8	0.3325		
Total	513.1167	11			

Salida de resultados por computadora comentados para la prueba de falta de ajuste

En la figura 11.18 se presenta una salida de resultados por computadora para el análisis de los datos del ejemplo 11.8 con el programa SAS. Observe la “LOF” con 2 grados de libertad, que representa las contribuciones cuadrática y cúbica al modelo, y el valor P de 0.22, que sugiere que el modelo lineal (de primer orden) es adecuado.

Dependent Variable: yield

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	510.4566667	170.1522222	511.74	<.0001
Error	8	2.6600000	0.3325000		
Corrected Total	11	513.1166667			
R-Square					
0.994816		0.666751	Root MSE	yield Mean	
			0.576628	86.48333	
Source	DF	Type I SS	Mean Square	F Value	Pr > F
temperature	1	509.2506667	509.2506667	1531.58	<.0001
LOF	2	1.2060000	0.6030000	1.81	0.2241

Figura 11.18: Salida de resultados del SAS que incluye el análisis de los datos del ejemplo 11.8.

Ejercicios

11.31 En el ejercicio 11.3 de la página 398 pruebe la linealidad de la regresión. Use un nivel de significancia de 0.05. Haga comentarios al respecto.

11.32 En el ejercicio 11.8 de la página 399 pruebe la linealidad de la regresión. Haga comentarios al respecto.

11.33 Suponga que tenemos una ecuación lineal que pasa por el origen $\mu_{Y|x} = \beta x$ (ejercicio 11.28).

a) Estime la regresión lineal que pasa por el origen para los siguientes datos:

x	0.5	1.5	3.2	4.2	5.1	6.5
y	1.3	3.4	6.7	8.0	10.0	13.2

- b) Suponga que se desconoce si la regresión verdadera debería pasar por el origen. Estime el modelo lineal $\mu_{Y|x} = \beta_0 + \beta_1 x$ y pruebe la hipótesis de que $\beta_0 = 0$ a un nivel de significancia de 0.10, en comparación con la alternativa de que $\beta_0 \neq 0$.

11.34 En el ejercicio 11.5 de la página 398 utilice el método del análisis de varianza para probar la hipótesis de que $\beta_1 = 0$, en comparación con la hipótesis alternativa de que $\beta_1 \neq 0$, a un nivel de significancia de 0.05.

11.35 Los siguientes datos son el resultado de una investigación sobre el efecto de la temperatura de reacción x sobre la conversión porcentual de un proceso químico y . (Véase Myers, Montgomery y Anderson-Cook, 2009). Ajuste una regresión lineal simple y utilice pruebas de falta de ajuste para determinar si el modelo es adecuado. Analice los resultados.

Observación	Temperatura (°C), x	Conversión (%), y
1	200	43
2	250	78
3	200	69
4	250	73
5	189.65	48
6	260.35	78
7	225	65
8	225	74
9	225	76
10	225	79
11	225	83
12	225	81

11.36 La ganancia de un transistor en un dispositivo de circuito integrado, entre el emisor y el colector (hFE), se relaciona con dos variables (Myers, Montgomery y Anderson-Cook, 2009) que se controlan en el proceso de deposición, controlado por el emisor en el tiempo (x_1 , en minutos) y la dosis del emisor (x_2 , en iones $\times 10^{14}$). Se observaron 14 muestras después de la deposición y los datos resultantes se presentan en la tabla siguiente. Consideraremos modelos de regresión lineal usando la ganancia como respuesta y el control del emisor en el tiempo o la dosis del emisor como la variable regresora.

Obs.	x_1 (tiempo de control, min)	x_2 (dosis, iones $\times 10^{14}$)	y (ganancia o hFE)
1	195	4.00	1004
2	255	4.00	1636
3	195	4.60	852
4	255	4.60	1506
5	255	4.20	1272
6	255	4.10	1270
7	255	4.60	1269
8	195	4.30	903
9	255	4.30	1555

10	255	4.00	1260
11	255	4.70	1146
12	255	4.30	1276
13	255	4.72	1225
14	340	4.30	1321

- a) Determine si el tiempo de control del emisor influye en la ganancia en una relación lineal. Es decir, pruebe $H_0: \beta_1 = 0$, donde β_1 es la pendiente de la variable regresora.
- b) Efectúe una prueba de falta de ajuste para determinar si la relación lineal es adecuada. Saque sus conclusiones.
- c) Determine si la dosis del emisor influye en la ganancia en una relación lineal. ¿Cuál variable regresora es el mejor predictor de la ganancia?

11.37 En los pesticidas se utilizan compuestos de organofosfatos (OF). Sin embargo, es importante estudiar el efecto que tienen sobre las especies expuestas a ellos. Como parte del estudio de laboratorio *Some Effects of Organophosphate Pesticides on Wildlife Species*, elaborado por el Departamento de Pesca y Vida Silvestre de Virginia Tech, se realizó un experimento en el cual se suministraron distintas dosis de un pesticida de OF específico a 5 grupos de 5 ratones (*peromysius leucopus*). Los 25 ratones eran hembras de edad y condiciones similares. Un grupo no recibió el producto. La respuesta básica y consistió en medir la actividad cerebral. Se postuló que dicha actividad disminuiría con un incremento en la dosis de OF. A continuación se presentan los datos:

Animal	Dosis, x (mg/kg de peso corporal)	Actividad, y (moles/litro/min)
1	0.0	10.9
2	0.0	10.6
3	0.0	10.8
4	0.0	9.8
5	0.0	9.0
6	2.3	11.0
7	2.3	11.3
8	2.3	9.9
9	2.3	9.2
10	2.3	10.1
11	4.6	10.6
12	4.6	10.4
13	4.6	8.8
14	4.6	11.1
15	4.6	8.4
16	9.2	9.7
17	9.2	7.8
18	9.2	9.0
19	9.2	8.2
20	9.2	2.3
21	18.4	2.9
22	18.4	2.2
23	18.4	3.4
24	18.4	5.4
25	18.4	8.2

- a) Con el modelo

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, 25,$
 calcule los estimados de los mínimos cuadrados de β_0 y β_1 .

- b) Construya una tabla de análisis de varianza en la cual aparezcan por separado el error puro y el error por falta de ajuste. Determine si la falta de ajuste es significativa al nivel de 0.05. Interprete los resultados.

11.38 Es frecuente que se utilice el tratamiento con calor para carburar partes metálicas como los engranes. El espesor de la capa carburada se considera una característica importante del engrane que contribuye a la confiabilidad general de la parte. Debido a la naturaleza crítica de esta característica, se realiza una prueba de laboratorio para cada lote del horno. La prueba es destructiva, ya que una parte real se corta en forma transversal y se sumerge en un producto químico durante cierto tiempo. Esta prueba requiere que se efectúe un análisis del carbono sobre la superficie, tanto de la parte superior del engrane (arriba de los dientes) como de su raíz (entre los dientes). Los datos siguientes son los resultados de la prueba de análisis de carbono en 19 partes.

Tiempo de inmersión		Tiempo de inmersión	
Grado		Grado	
0.58	0.013	1.17	0.021
0.66	0.016	1.17	0.019
0.66	0.015	1.17	0.021
0.66	0.016	1.20	0.025
0.66	0.015	2.00	0.025
0.66	0.016	2.00	0.026
1.00	0.014	2.20	0.024
1.17	0.021	2.20	0.025
1.17	0.018	2.20	0.024
1.17	0.019		

- a) Ajuste una regresión lineal simple que relacione el grado del análisis de carbono y en comparación con el tiempo de inmersión. Pruebe $H_0: \beta_1 = 0$.
- b) Si se rechaza la hipótesis del inciso a, determine si el modelo lineal es adecuado.

11.39 Se desea obtener un modelo de regresión que relacione la temperatura con la proporción de impurezas de una sustancia que pasa a través de helio sólido. Se lista la temperatura en grados centígrados. A continuación se presentan los datos.

Temperatura (°C)	Proporción de impurezas
-260.5	0.425
-255.7	0.224
-264.6	0.453
-265.0	0.475
-270.0	0.705

-272.0	0.860
-272.5	0.935
-272.6	0.961
-272.8	0.979
-272.9	0.990

- a) Ajuste un modelo de regresión lineal.
- b) ¿Parece que la proporción de impurezas que pasan a través del helio aumenta a medida que la temperatura se acerca a -273 grados centígrados?
- c) Calcule R^2 .
- d) Con base en la información anterior, ¿parece adecuado el modelo lineal? ¿Qué información adicional necesitaría usted para responder mejor a la pregunta?

11.40 Existe interés por estudiar el efecto que tiene el tamaño de la población de varias ciudades de Estados Unidos sobre las concentraciones de ozono. Los datos consisten en la población de 1999 en millones de habitantes y en la cantidad de ozono presente por hora en partes por mil millones (ppmm). Los datos son los siguientes:

Ozono (ppmm/hora), y	Población, x
126	0.6
135	4.9
124	0.2
128	0.5
130	1.1
128	0.1
126	1.1
128	2.3
128	0.6
129	2.3

- a) Ajuste un modelo de regresión lineal que relacione la concentración de ozono con la población. Pruebe $H_0: \beta_1 = 0$ usando el método ANOVA.
- b) Haga una prueba para la falta de ajuste. Con base en los resultados de la prueba, ¿es apropiado el modelo lineal?
- c) Pruebe la hipótesis del inciso a) utilizando el cuadrado medio del error puro en la prueba F . ¿Cambian los resultados? Comente las ventajas de cada prueba.

11.41 Evaluar la deposición del nitrógeno de la atmósfera es una tarea importante del National Atmospheric Deposition Program (NADP), que está asociado con muchas instituciones. Este programa está estudiando la deposición atmosférica y su efecto sobre los cultivos agrícolas, las aguas superficiales de los bosques y otros recursos. Los óxidos del nitrógeno pueden tener efectos sobre el ozono atmosférico y la cantidad de nitrógeno puro que se encuentra en el aire que respiramos. Los datos son los siguientes:

Año	Óxido de nitrógeno
1978	0.73
1979	2.55
1980	2.90
1981	3.83
1982	2.53
1983	2.77
1984	3.93
1985	2.03
1986	4.39
1987	3.04
1988	3.41
1989	5.07
1990	3.95
1991	3.14
1992	3.44
1993	3.63
1994	4.50
1995	3.95
1996	5.24
1997	3.30
1998	4.36
1999	3.33

- Grafique los datos.
- Ajuste un modelo de regresión lineal y calcule R^2 .
- ¿Qué puede decir acerca de la tendencia del óxido de nitrógeno con el paso del tiempo?

11.42 Para una variedad particular de planta los investigadores desean desarrollar una fórmula para predecir la cantidad de semillas (en gramos) como una función de la densidad de las plantas. Efectuaron un estudio con cuatro niveles del factor x , el número de plantas por parcela. Se utilizaron cuatro réplicas para cada nivel de x . A continuación se muestran los datos:

Plantas por parcela, x	Cantidad de semillas, y (gramos)			
10	12.6	11.0	12.1	10.9
20	15.3	16.1	14.9	15.6
30	17.9	18.3	18.6	17.8
40	19.2	19.6	18.9	20.0

¿Es adecuado un modelo de regresión lineal simple para analizar este conjunto de datos?

11.10 Gráficas de datos y transformaciones

En este capítulo se estudia la construcción de modelos de regresión en los que hay una variable independiente o regresora. Además, se supone que durante la construcción del modelo tanto x como y entran en el modelo en *forma lineal*. Con frecuencia es aconsejable trabajar con un modelo alternativo en el que x o y (o ambas) intervengan en una forma no lineal. Se podría recomendar una **transformación** de los datos debido a consideraciones teóricas inherentes al estudio científico, o bien, una simple graficación de los datos podría sugerir la necesidad de *reexpresar* las variables en el modelo. La necesidad de llevar a cabo una transformación es muy fácil de diagnosticar en el caso de la regresión lineal simple, ya que las gráficas en dos dimensiones brindan un panorama verdadero de la manera en que las variables se comportan en el modelo.

Un modelo en el que x o y se transforman no debería considerarse como un *modelo de regresión no lineal*. Por lo general denominamos a un modelo de regresión como lineal cuando es **lineal en los parámetros**. En otras palabras, suponga que el aspecto de los datos u otra información científica sugiere que debe hacerse la **regresión de y^* en comparación con la de x^*** , donde cada una de ellas es una transformación de las variables naturales x y y . Entonces, el modelo de la forma

$$y_i^* = \beta_0 + \beta_1 x_i^* + \epsilon_i$$

es lineal porque lo es en los parámetros β_0 y β_1 . El material que se estudió en las secciones 11.2 a 11.9 permanece sin cambio, donde y_i^* y x_i^* reemplazan a y_i y x_i . Un ejemplo sencillo y útil es el modelo log-log:

$$\log y_i = \beta_0 + \beta_1 \log x_i + \epsilon_i.$$

Aunque este modelo es no lineal en x y y , sí lo es en los parámetros y por ello recibe el tratamiento de un modelo lineal. Por otro lado, un ejemplo de modelo verdaderamente no lineal es:

$$y_i = \beta_0 + \beta_1 x^{\beta_2} + \epsilon_i,$$

donde se debe estimar el parámetro β_2 , así como β_0 y β_1 . El modelo es no lineal en β_2 . Las transformaciones susceptibles de mejorar el ajuste y la capacidad de predicción de un modelo son muy numerosas. Para un análisis completo de las transformaciones el lector podría consultar a Myers (1990, véase la bibliografía). Decidimos incluir aquí algunas de ellas y mostrar la apariencia de las gráficas que sirven como herramientas diagnósticas. Considere la tabla 11.6, donde se presentan varias funciones que describen relaciones entre y y x que pueden producir una *regresión lineal* por medio de la transformación indicada. Además, en aras de que el análisis sea más exhaustivo, se presentan al lector las variables dependiente e independiente que se utilizan en la *regresión lineal simple* resultante. La figura 11.19 ilustra las funciones que se listan en la tabla 11.6, las cuales sirven como guía para el analista en la elección de una transformación a partir de la observación de la gráfica de y contra x .

Tabla 11.6: Algunas transformaciones útiles para linealizar

Forma funcional que relaciona y con x	Transformación propia	Forma de la regresión lineal simple
Exponencial: $y = \beta_0 e^{\beta_1 x}$	$y^* = \ln y$	Hacer la regresión de y^* contra x
Potencia: $y = \beta_0 x^{\beta_1}$	$y^* = \log y$; $x^* = \log x$	Hacer la regresión de y^* contra x^*
Recíproca: $y = \beta_0 + \beta_1 \left(\frac{1}{x}\right)$	$x^* = \frac{1}{x}$	Hacer la regresión de y contra x^*
Hiperbólica: $y = \frac{x}{\beta_0 + \beta_1 x}$	$y^* = \frac{1}{y}$; $x^* = \frac{1}{x}$	Hacer la regresión de y^* contra x^*

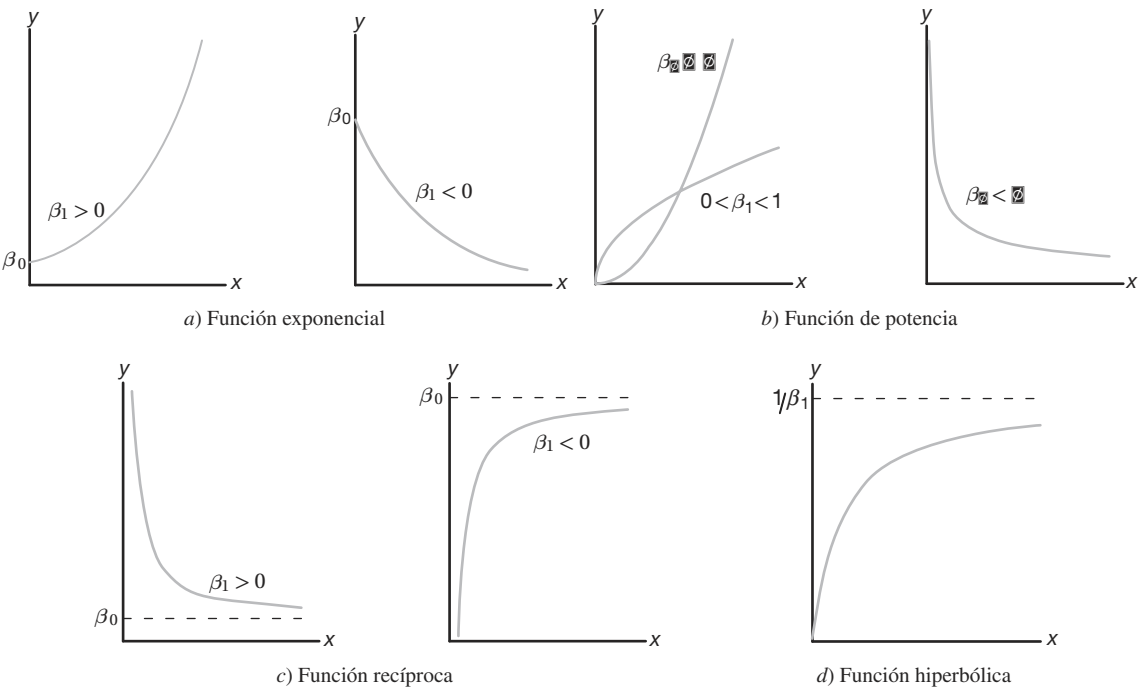


Figura 11.19: Diagramas que ilustran las funciones listadas en la tabla 11.6.

¿Cuáles son las implicaciones de un modelo transformado?

Lo que sigue intenta ser una ayuda para el analista cuando es evidente que una transformación producirá una mejoría. Sin embargo, antes de dar un ejemplo hay que mencionar dos puntos importantes. El primero tiene que ver con la escritura formal del modelo una vez que se hayan transformado los datos. Con mucha frecuencia el analista no piensa en esto y simplemente lleva a cabo la transformación sin preocuparse por la forma del modelo *antes ni después* de la transformación. El modelo exponencial sirve como una buena ilustración de esto. El modelo en las variables naturales (no transformadas) que produce un *modelo de error aditivo* en las variables transformadas es dado por

$$y_i = \beta_0 e^{\beta_1 x_i} \cdot \epsilon_i,$$

que es un *modelo de error multiplicativo*. Al aplicar logaritmos es claro que se obtiene

$$\ln y_i = \ln \beta_0 + \beta_1 x_i + \ln \epsilon_i.$$

Como resultado, las suposiciones básicas se efectúan sobre $\ln \epsilon_i$. El propósito de esta presentación sólo es recordar al lector que no debemos considerar una transformación tan sólo como una manipulación algebraica a la cual se suma un error. Con frecuencia, un modelo en las variables transformadas que tiene una adecuada *estructura de error aditivo* es resultado de un modelo en las variables naturales con un tipo de estructura de error diferente.

El segundo aspecto importante se refiere a la noción de las medidas de mejoría. Las medidas evidentes de comparación son, por supuesto, el valor de R^2 y el cuadrado medio de los residuales s^2 . (En el capítulo 12 se estudian otras medidas de rendimiento que se usan para comparar modelos que compiten). Ahora, si la respuesta y no se transforma, entonces es claro que s^2 y R^2 se pueden usar para medir la utilidad de la transformación. Los residuales estarán en las mismas unidades para los dos modelos, el transformado y el que no se transformó. No obstante, cuando se transforma y los criterios de rendimiento para el modelo transformado deberían basarse en los valores de los residuales en las unidades de medida de la respuesta no transformada. De esta manera las comparaciones son más apropiadas. El siguiente ejemplo proporciona una ilustración de lo anterior.

Ejemplo 11.9: Se registra la presión P de un gas que corresponde a distintos volúmenes V y los datos se presentan en la tabla 11.7.

Tabla 11.7: Datos para el ejemplo 11.9

V (cm ³)	50	60	70	90	100
P (kg/cm ²)	64.7	51.3	40.5	25.9	7.8

La ley del gas ideal es dada por la forma funcional $PV^\gamma = C$, donde γ y C son constantes. Estime las constantes C y γ .

Solución: Se toman logaritmos naturales en ambos lados del modelo

$$P_i V_i^\gamma = C \cdot \epsilon_i, \quad i = 1, 2, 3, 4, 5.$$

Como resultado, es posible escribir el modelo lineal

$$\ln P_i = \ln C - \gamma \ln V_i + \epsilon_i^*, \quad i = 1, 2, 3, 4, 5,$$

Donde $\epsilon_i^* = \ln \epsilon_i$. Los siguientes son los resultados de la regresión lineal simple:

Intersección $\widehat{\ln C} = 14.7589$, $\widehat{C} = 2,568,862.88$, Pendiente: $\hat{\gamma} = 2.65347221$.

Varianza no homogénea

Una suposición importante que se hace en el análisis de regresión es la varianza homogénea. A menudo las transgresiones se detectan mediante la apariencia de la gráfica de residuales. Es común que en los datos científicos se incremente la varianza del error con el aumento de la variable regresora. Una varianza grande del error produce residuales grandes y, por ende, una gráfica de residuales como la que se presenta en la figura 11.22 es una señal de varianza no homogénea. En el capítulo 12, en el cual se expone la regresión lineal múltiple, se presenta un análisis más amplio acerca de las gráficas de los residuales e información acerca de los diferentes tipos de residuales.

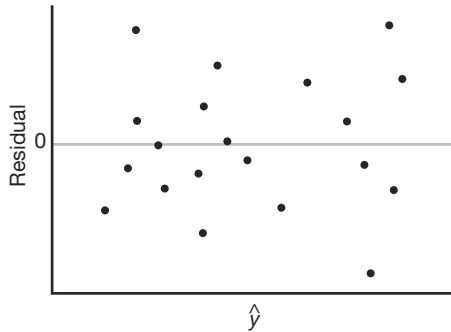


Figura 11.21: Gráfica ideal de los residuales.

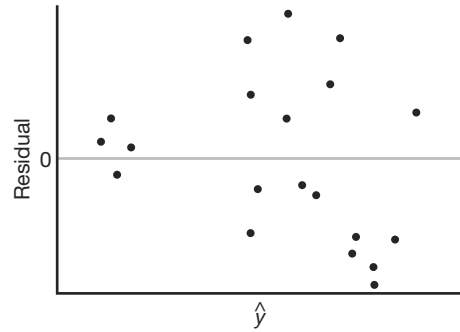


Figura 11.22: Gráfica de los residuales que ilustra una varianza heterogénea del error.

Gráfica de la probabilidad normal

La suposición de que los errores del modelo son normales se hace cuando el analista de los datos se ocupa de las pruebas de hipótesis o de la estimación de intervalos de confianza. De nuevo, los equivalentes numéricos de los ϵ_i , es decir, los residuales, son sujetos de diagnóstico mediante la graficación para detectar cualesquiera transgresiones extremas. En el capítulo 8 se presentaron las gráficas normales cuantil-cuantil y se analizaron en forma breve las de probabilidad normal. En el estudio de caso que se presenta en la siguiente sección se ilustran estas gráficas de residuales.

11.11 Estudio de caso de regresión lineal simple

En la fabricación de productos comerciales de madera es importante estimar la relación que hay entre la densidad de un producto de madera y su rigidez. Se está considerando un tipo relativamente nuevo de aglomerado que se puede formar con mucha mayor facilidad que el producto comercial ya aceptado. Es necesario saber a qué densidad su rigidez es comparable con la del producto comercial bien conocido y documentado. Terrance E. Connors realizó un estudio titulado *Investigation of Certain Mechanical Properties of a Wood-Foam Composite* (Tesis para el doctorado, Departamento de Bosques y Vida Silvestre, University of Massachusetts). Se produjeron 30 tableros de aglomerado con densidades que variaban aproximadamente de 8 a 26 libras por pie cúbico y se midió su rigidez en libras por pulgada cuadrada. En la tabla 11.8 se presentan los datos.

Es necesario que el analista de datos se concentre en un ajuste apropiado para los datos y que utilice los métodos de inferencia que se estudian en este capítulo. Tal vez lo más apropiado sea una prueba de hipótesis sobre la pendiente de la regresión, así como

la estimación de los intervalos de confianza o de predicción. Se comenzará presentando un simple diagrama de dispersión de los datos brutos con una regresión lineal simple sobrepuesta. En la figura 11.23 se observa dicha gráfica.

El ajuste de regresión lineal simple a los datos produce el modelo ajustado

$$\hat{y} = -25,433.739 + 3884.976x \quad (R^2 = 0.7975),$$

Tabla 11.8: Densidad y rigidez de 30 tableros de aglomerado

Densidad, x	Rigidez, y	Densidad, x	Rigidez, y
9.50	14,814.00	8.40	17,502.00
9.80	14,007.00	11.00	19,443.00
8.30	7573.00	9.90	14,191.00
8.60	9714.00	6.40	8076.00
7.00	5304.00	8.20	10,728.00
17.40	43,243.00	15.00	25,319.00
15.20	28,028.00	16.40	41,792.00
16.70	49,499.00	15.40	25,312.00
15.00	26,222.00	14.50	22,148.00
14.80	26,751.00	13.60	18,036.00
25.60	96,305.00	23.40	104,170.00
24.40	72,594.00	23.30	49,512.00
19.50	32,207.00	21.20	48,218.00
22.80	70,453.00	21.70	47,661.00
19.80	38,138.00	21.30	53,045.00

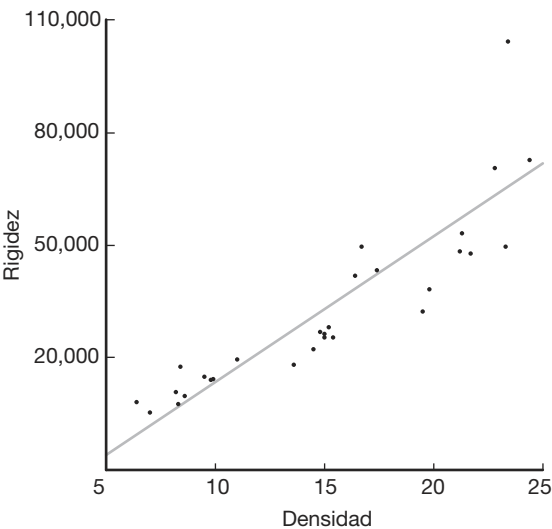


Figura 11.23: Diagrama de dispersión de los datos de densidad de la madera.

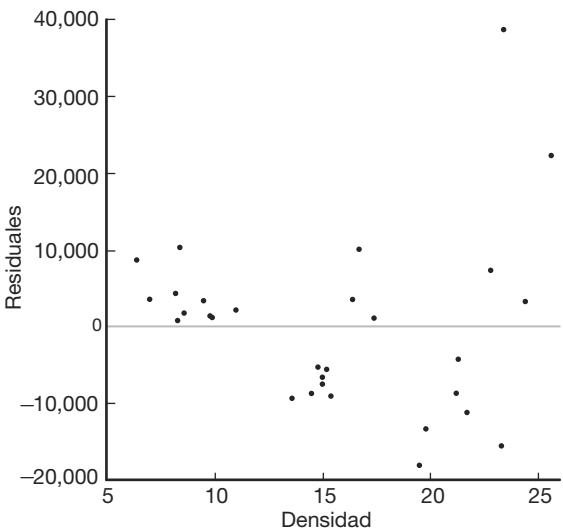


Figura 11.24: Gráfica de los residuales para los datos de densidad de la madera.

y se calcularon los residuales. En la figura 11.24 se presentan los residuales graficados contra las mediciones de la densidad. Difícilmente se trata de un conjunto de residuales ideal o satisfactorio, pues no muestran una distribución aleatoria alrededor del valor de cero. En realidad, los agrupamientos de valores positivos y negativos sugerirían que se debe investigar una tendencia curvilínea en los datos.

Para darnos una idea respecto a la suposición de error normal se dibujó una gráfica de probabilidad normal de los residuales. Es el tipo de gráfica que estudiamos en la sección 8.8, donde el eje horizontal representa la función de distribución normal empírica en una escala que produce una gráfica con línea recta cuando se grafica contra los residuales. En la figura 11.25 se presenta la gráfica de probabilidad normal de los residuales. Esta gráfica no refleja la apariencia de recta que a uno le gustaría ver, lo cual es otro síntoma de una selección errónea, quizá sobresimplificada, de un modelo de regresión.

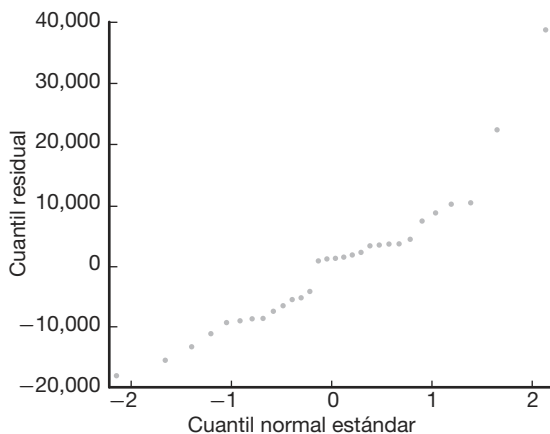


Figura 11.25: Gráfica de probabilidad normal de los residuales para los datos de densidad de la madera.

Los dos tipos de gráficas de residuales y, de hecho, el propio diagrama de dispersión, sugieren que sería adecuado un modelo algo más complicado. Una posibilidad es usar un modelo con transformación de logaritmos naturales. En otras palabras, hay que elegir hacer la regresión de $\ln y$ contra x . Esto produce la regresión

$$\widehat{\ln y} = 8.257 + 0.125x \quad (R^2 = 0.9016).$$

Para darse una idea de si el modelo transformado es más apropiado considere las figuras 11.26 y 11.27, que muestran las gráficas de los residuales de la rigidez [es decir, y_i -antilog $(\widehat{\ln y})$] en comparación con las de la densidad. La figura 11.26 parece más cercana a un patrón aleatorio alrededor del cero, en tanto que la figura 11.27 con seguridad se acerca más a una línea recta. Esto, además de un valor de R^2 más elevado, sugeriría que el modelo transformado es más apropiado.

11.12 Correlación

Hasta este momento se ha supuesto que la variable regresora independiente x es una variable científica o física en lugar de una variable aleatoria. De hecho, en este contexto es frecuente que x se denomine **variable matemática**, la cual, en el proceso de muestreo, se mide con un error despreciable. En muchas aplicaciones de las técnicas de regresión es más realista suponer que tanto X como Y son variables aleatorias y que las mediciones $\{(x_i, y_i); i = 1, 2, \dots, n\}$ son observaciones de una población que tiene la función de

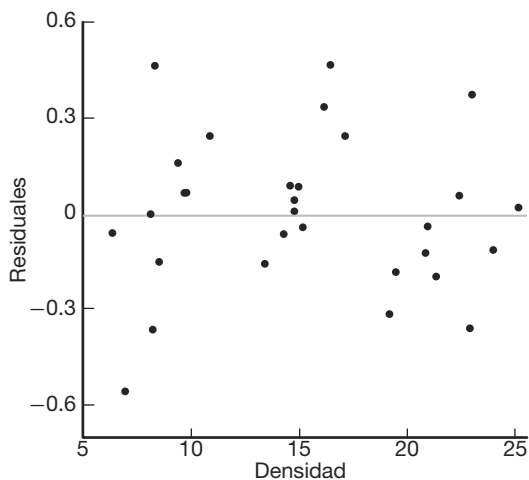


Figura 11.26: Gráfica de residuales donde se utiliza una transformación logarítmica para los datos de densidad de la madera.

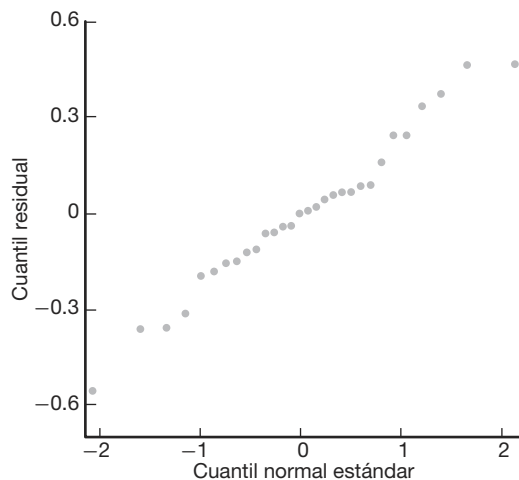


Figura 11.27: Gráfica de probabilidad normal de residuales en la cual se utiliza una transformación logarítmica para los datos de densidad de la madera.

densidad conjunta $f(x, y)$. Debemos considerar el problema de medir la relación entre las dos variables X y Y . Por ejemplo, si X y Y representaran la longitud y la circunferencia de una clase particular de hueso en el cuerpo de un adulto, podríamos realizar un estudio antropológico para determinar si los valores grandes de X se asocian con valores grandes de Y , y viceversa.

Por otro lado, si X representa la antigüedad de un automóvil usado y Y representa su precio de lista al menudeo, se esperaría que los valores grandes de X correspondan a valores pequeños de Y y que los valores pequeños de X correspondan a valores grandes de Y . El **análisis de correlación** intenta medir la fuerza de tales relaciones entre dos variables por medio de un solo número denominado **coeficiente de correlación**.

En teoría, con frecuencia se supone que la distribución condicional $f(y|x)$ de Y , para valores fijos de X , es normal con media $\mu_{y|x} = \beta_0 + \beta_1 x$ y varianza $\sigma_{y|x}^2 = \sigma^2$, y que, de igual manera, X se distribuye de forma normal con media μ y varianza σ_x^2 . Entonces, la densidad conjunta de X y Y es

$$\begin{aligned} f(x, y) &= n(y|x; \beta_0 + \beta_1 x, \sigma) n(x; \mu, \sigma_x) \\ &= \frac{1}{2\pi\sigma_x\sigma} \exp \left\{ -\frac{1}{2} \left[\left(\frac{y - \beta_0 - \beta_1 x}{\sigma} \right)^2 + \left(\frac{x - \mu_x}{\sigma_x} \right)^2 \right] \right\}, \\ &\text{para } -\infty < x < \infty \text{ y } -\infty < y < \infty. \end{aligned}$$

Escribamos la variable aleatoria Y en la forma

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

donde ahora X es una variable aleatoria independiente del error aleatorio ϵ . Como la media del error aleatorio ϵ es cero, se deduce que

$$\mu_Y = \beta_0 + \beta_1 \mu_X \quad \text{y} \quad \sigma_Y^2 = \sigma^2 + \beta_1^2 \sigma_X^2.$$

Al sustituir para α y σ^2 en la expresión anterior para $f(x, y)$, se obtiene la **distribución normal bivariada**

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\},$$

para $-\infty < x < \infty$ y $-\infty < y < \infty$, donde

$$\rho^2 = 1 - \frac{\sigma^2}{\sigma_Y^2} = \beta_1^2 \frac{\sigma_X^2}{\sigma_Y^2}.$$

La constante ρ (ro) se denomina **coeficiente de correlación de la población** y desempeña un papel importante en muchos problemas de análisis de datos bivariados. Es importante que el lector entienda la interpretación física de este coeficiente de correlación, así como la diferencia entre correlación y regresión. El término *regresión* aún tiene algún significado aquí. De hecho, la línea recta dada por $\mu_{Y|x} = \beta_0 + \beta_1 x$ se sigue llamando recta de regresión, igual que antes, y los estimadores de β_0 y β_1 son idénticos a los que se presentaron en la sección 11.3. El valor de ρ es 0 cuando $\beta_1 = 0$, que resulta cuando en esencia no existe regresión lineal; es decir, cuando la recta de regresión es horizontal y cualquier conocimiento de X es inútil para predecir Y . Como $\sigma_Y^2 \geq \sigma^2$, se debe tener $\rho^2 \leq 1$ y, por lo tanto, $-1 \leq \rho \leq 1$. Los valores de $\rho \pm 1$ sólo ocurren cuando $\sigma^2 = 0$, en cuyo caso se tiene una relación lineal perfecta entre las dos variables. Así, un valor de ρ igual a $+1$ implica una relación lineal perfecta con pendiente positiva, en tanto que un valor de ρ igual a -1 resulta de una relación lineal perfecta con pendiente negativa. Entonces, se podría decir que los estimadores muestrales de ρ con magnitud cercana a la unidad implican una buena correlación o **asociación lineal** entre X y Y , mientras que valores cercanos a cero indican poca o ninguna correlación.

Para obtener un estimador muestral de ρ recordemos que en la sección 11.4 aprendimos que la suma de los cuadrados del error es

$$SCE = S_{yy} - b_1 S_{xy}.$$

Al dividir ambos lados de esta ecuación entre S_{yy} y reemplazar S_{xy} con $b_1 S_{xx}$, se obtiene la relación

$$b_1^2 \frac{S_{xx}}{S_{yy}} = 1 - \frac{SCE}{S_{yy}}.$$

El valor de $b_1^2 S_{xx} / S_{yy}$ es igual a cero cuando $b_1 = 0$, lo que ocurrirá cuando los puntos muestrales no tengan relación lineal. Como $S_{yy} \geq SCE$, se concluye que $b_1^2 S_{xx} / S_{yy}$ debe estar entre 0 y 1. En consecuencia, $b_1 \sqrt{S_{xx} / S_{yy}}$ debe variar entre -1 y $+1$, y los valores negativos corresponden a rectas con pendientes negativas, mientras que los valores positivos corresponden a rectas con pendientes positivas. Un valor de -1 o $+1$ sucederá cuando $SCE = 0$, pero éste es el caso en el que todos los puntos muestrales caen sobre una línea recta. Por lo tanto, una relación lineal perfecta se da en los datos muestrales cuando $b_1 \sqrt{S_{xx} / S_{yy}} = \pm 1$. Es claro que la cantidad $b_1 \sqrt{S_{xx} / S_{yy}}$, la cual se designará de aquí en adelante como r , se puede usar como un estimado del coeficiente de correlación ρ de la población. Se acostumbra hacer referencia al estimado r como **coeficiente de correlación producto-momento de Pearson**, o sólo como **coeficiente de correlación muestral**.

Coeficiente de correlación La medida ρ de la asociación lineal entre dos variables X y Y se estima por medio del **coeficiente de correlación muestral** r , donde

$$r = b_1 \sqrt{\frac{S_{xx}}{S_{yy}}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}.$$

Hay que tener cuidado en la interpretación de valores de r entre -1 y $+1$. Por ejemplo, valores de r iguales a 0.3 y 0.6 significan sólo que hay dos correlaciones positivas, una un poco más fuerte que la otra. Sería un error concluir que $r = 0.6$ indica una relación lineal dos veces mejor que la del valor $r = 0.3$. Por otro lado, si escribimos

$$r^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{SCR}{S_{yy}},$$

entonces r^2 , que por lo general se denomina **coeficiente muestral de determinación**, representa la proporción de la variación de S_{yy} explicada por la regresión de Y sobre x , a saber, la SCR . Es decir, r^2 expresa la proporción de la variación total de los valores de la variable Y que son ocasionados o explicados por una relación lineal con los valores de la variable aleatoria X . Así, una correlación de 0.6 significa que 0.36 , o 36% , de la variación total de los valores de Y en la muestra se explica mediante la relación lineal con los valores de X .

Ejemplo 11.10: Es importante que los investigadores científicos del área de productos forestales sean capaces de estudiar la correlación entre la anatomía y las propiedades mecánicas de los árboles. Para el estudio *Quantitative Anatomical Characteristics of Plantation Grown Loblolly Pine (Pinus Taeda L.) and Cottonwood (Populus deltoides Bart. Ex Marsh.) and Their Relationships to Mechanical Properties*, realizado por el Departamento de Bosques y Productos Forestales de Virginia Tech, se seleccionaron al azar 29 pinos de Arkansas para investigarlos. En la tabla 11.9 se presentan los datos resultantes sobre la gravedad específica en gramos/cm³ y el módulo de ruptura en kilopascales (kPa). Calcule e interprete el coeficiente de correlación muestral.

Tabla 11.9: Datos de 29 pinos de Arkansas para el ejemplo 11.10

Gravedad específica, x (g/cm ³)	Módulo de ruptura, y (kPa)	Gravedad específica, x (g/cm ³)	Módulo de ruptura, y (kPa)
0.414	29,186	0.581	85,156
0.383	29,266	0.557	69,571
0.399	26,215	0.550	84,160
0.402	30,162	0.531	73,466
0.442	38,867	0.550	78,610
0.422	37,831	0.556	67,657
0.466	44,576	0.523	74,017
0.500	46,097	0.602	87,291
0.514	59,698	0.569	86,836
0.530	67,705	0.544	82,540
0.569	66,088	0.557	81,699
0.558	78,486	0.530	82,096
0.577	89,869	0.547	75,657
0.572	77,369	0.585	80,490
0.548	67,095		

Solución: A partir de los datos se encuentra que

$$S_{xx} = 0.11273, \quad S_{yy} = 11,807,324,805, \quad S_{xy} = 34,422.27572.$$

Por lo tanto,

$$r = \frac{34,422.27572}{\sqrt{(0.11273)(11,807,324,805)}} = 0.9435.$$

Un coeficiente de correlación de 0.9435 indica una buena relación lineal entre X y Y . Como $r^2 = 0.8902$, se puede decir que aproximadamente 89% de la variación de los valores de Y es ocasionada por una relación lineal con X . ─

Una prueba de la hipótesis especial $\rho = 0$ en comparación con una alternativa apropiada es equivalente a probar $\beta_1 = 0$ para el modelo de regresión lineal simple y, por lo tanto, son aplicables los procedimientos de la sección 11.8, donde se usaba la distribución t con $n - 2$ grados de libertad o la distribución F con 1 y $n - 2$ grados de libertad. Sin embargo, si se desea evitar el procedimiento del análisis de varianza y tan sólo calcular el coeficiente de correlación muestral, se podría verificar (véase el ejercicio de repaso 11.66 en la página 438) que el valor t

$$t = \frac{b_1}{s/\sqrt{S_{xx}}}$$

también se puede escribir como

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

que, como antes, es un valor del estadístico T que tiene una distribución t con $n - 2$ grados de libertad.

Ejemplo 11.11: Para los datos del ejemplo 11.10 pruebe la hipótesis de que no existe asociación lineal entre las variables.

Solución: 1. $H_0: \rho = 0$.

2. $H_1: \rho \neq 0$.

3. $\alpha = 0.05$.

4. Región crítica: $t < -2.052$ o $t > 2.052$.

5. Cálculos: $t = \frac{0.9435 \sqrt{27}}{\sqrt{1-0.9435^2}} = 4.79, P \approx 0.0001$.

6. Decisión: Rechazar la hipótesis de que no existe asociación lineal. ─

A partir de la información muestral es fácil efectuar una prueba de la hipótesis más general de que $\rho = \rho_0$ en comparación con una hipótesis alternativa adecuada. Si X y Y siguen una distribución normal bivariada, la cantidad

$$\frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

es el valor de una variable aleatoria que sigue aproximadamente la distribución normal con media $\frac{1}{2} \ln \frac{1+\rho}{1-\rho}$ y varianza $1/(n-3)$. Entonces, el procedimiento de prueba consiste en calcular

$$z = \frac{\sqrt{n-3}}{2} \left[\ln \left(\frac{1+r}{1-r} \right) - \ln \left(\frac{1+\rho_0}{1-\rho_0} \right) \right] = \frac{\sqrt{n-3}}{2} \ln \left[\frac{(1-r)(1-\rho_0)}{(1-r)(1-\rho_0)} \right]$$

y compararlo con los puntos críticos de la distribución normal estándar.

Ejemplo 11.12: Para los datos del ejemplo 11.10 pruebe la hipótesis nula de que $\rho = 0.9$ en comparación con la alternativa de que $\rho > 0.9$. Utilice un nivel de significancia de 0.05.

Solución: 1. $H_0: \rho = 0.9$.

2. $H_1: \rho > 0.9$.

3. $\alpha = 0.05$.

4. Región crítica: $z > 1.645$.

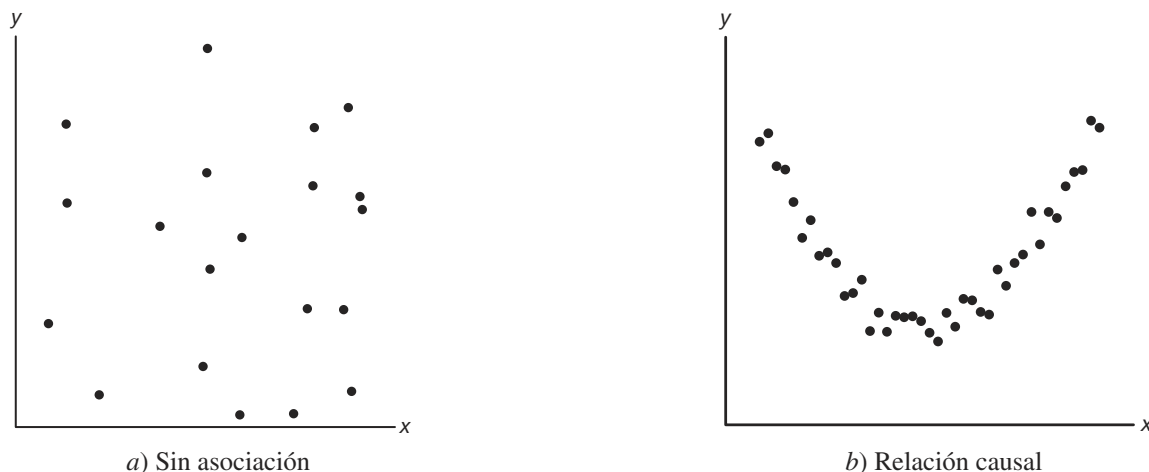


Figura 11.28: Diagrama de dispersión que muestra correlación de cero.

5. Cálculos:

$$z = \frac{\sqrt{26}}{2} \ln \left[\frac{(1 - 0.9435)(0.1)}{(1 - 0.9435)(1.9)} \right] = -1.51, \quad P = 0.0655.$$

6. Decisión: Existe con certeza alguna evidencia de que el coeficiente de correlación no excede a 0.9. ■

Debe precisarse que en los estudios de correlación, como en los problemas de regresión lineal, los resultados obtenidos sólo son tan buenos como el modelo que se adopte. En las técnicas de correlación estudiadas aquí se supone que las variables X y Y tienen una densidad normal bivariada, con el valor medio de Y para cada valor de x relacionado en forma lineal con x . Con frecuencia es útil elaborar una gráfica preliminar de los datos experimentales para observar qué tan adecuada es la suposición de linealidad. Un valor del coeficiente de correlación muestral cercano a cero resultará de datos que muestren un efecto estrictamente aleatorio, como los de la figura 11.28a, lo que implica que hay poca o ninguna relación causal. Es importante recordar que el coeficiente de correlación entre dos variables es una medida de su relación lineal, y que un valor de $r = 0$ implica *falta de linealidad* y *no falta de asociación*. Por lo tanto, si existiera una relación cuadrática fuerte entre X y Y , como la que se observa en la figura 11.28b, aún se podría obtener una correlación de cero que indicaría una relación no lineal.

Ejercicios

11.43 Calcule e interprete el coeficiente de correlación para las siguientes calificaciones de 6 estudiantes seleccionados al azar:

Calificación en matemáticas	70	92	80	74	65	83
Calificación en inglés	74	84	63	87	78	90

11.44 Remítase al ejercicio 11.1 de la página 398 y suponga que x y y son variables aleatorias con una distribución normal bivariada:

- Calcule r .
- Pruebe la hipótesis de que $\rho = 0$ en comparación con la alternativa de que $\rho \neq 0$ a un nivel de significancia de 0.05.

11.45 Remítase al ejercicio 11.13 de la página 400, suponga una distribución normal bivariada para x y y .

- Calcule r .
- Pruebe la hipótesis nula de que $\rho = -0.5$, en comparación con la alternativa de que $\rho < -0.5$, a un nivel de significancia de 0.025.
- Determine el porcentaje de la variación en la cantidad de partículas eliminadas que se debe a cambios en la cantidad de lluvia diaria.

11.46 En el ejercicio 11.43 pruebe la hipótesis de que $\rho = 0$ en comparación con la alternativa de que $\rho \neq 0$. Utilice un nivel de significancia de 0.05.

11.47 Los datos siguientes se obtuvieron en un estudio de la relación entre el peso y el tamaño del pecho de niños al momento de nacer.

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

- Calcule r .
- Pruebe la hipótesis nula de que $\rho = 0$ en comparación con la alternativa de que $\rho > 0$ a un nivel de significancia de 0.01.
- ¿Qué porcentaje de la variación del tamaño del pecho de los niños es explicado por la diferencia de peso?

Ejercicios de repaso

11.48 Remítase al ejercicio 11.8 de la página 399 y construya

- un intervalo de confianza de 95% para la calificación promedio en el curso de los estudiantes que obtuvieron 35 puntos en el examen de colocación;
- un intervalo de predicción de 95% para la calificación del curso de un estudiante que obtuvo 35 puntos en el examen de colocación.

11.49 El Centro de Consulta Estadística de Virginia Tech analizó datos sobre las marmotas normales para el Departamento de Veterinaria. Las variables de interés fueron el peso corporal en gramos y el peso del corazón en gramos. Se deseaba desarrollar una ecuación de regresión lineal con el fin de determinar si había una relación lineal significativa entre el peso del corazón y el peso total del cuerpo.

Peso corporal (gramos)	Peso del corazón (gramos)
4050	11.2
2465	12.4
3120	10.5
5700	13.2
2595	9.8
3640	11.0
2050	10.8
4235	10.4
2935	12.2
4975	11.2
3690	10.8
2800	14.2
2775	12.2
2170	10.0
2370	12.3
2055	12.5
2025	11.8
2645	16.0
2675	13.8

Utilice el peso del corazón como la variable independiente, el peso del cuerpo como la dependiente y haga un ajuste de regresión lineal simple con los siguientes datos. Además, pruebe la hipótesis de que $H_0: \beta_1 = 0$ en comparación con $H_1: \beta_1 \neq 0$. Saque conclusiones.

11.50 A continuación se presentan las cantidades de sólidos eliminados de cierto material cuando se expone a periodos de secado de diferentes duraciones.

x (horas)	y (gramos)
4.4	13.1 14.2
4.5	9.0 11.5
4.8	10.4 11.5
5.5	13.8 14.8
5.7	12.7 15.1
5.9	9.9 12.7
6.3	13.8 16.5
6.9	16.4 15.7
7.5	17.6 16.9
7.8	18.3 17.2

- Estime la recta de regresión lineal.
- Pruebe si es adecuado el modelo lineal a un nivel de significancia de 0.05.

11.51 Remítase al ejercicio 11.9 de la página 399 y construya

- un intervalo de confianza de 95% para las ventas semanales promedio cuando se gastan \$45 en publicidad.
- un intervalo de predicción de 95% para las ventas semanales cuando se gastan \$45 en publicidad.

11.52 Se diseñó un experimento para el Departamento de Ingeniería de Materiales de Virginia Tech con el fin de estudiar las propiedades de deterioro del nitrógeno con base en las mediciones de la presión de hidrógeno

electrolítico. Se utilizó una solución al 0.1 *N* NaOH y el material era cierto tipo de acero inoxidable. La densidad de corriente de carga catódica fue controlada y variada en cuatro niveles. Se observó la presión de hidrógeno efectiva como la respuesta. A continuación se presentan los datos.

Ensayo	Densidad de corriente de carga, x (mA/cm ²)	Presión de hidrógeno efectiva, y (atm)
1	0.5	86.1
2	0.5	92.1
3	0.5	64.7
4	0.5	74.7
5	1.5	223.6
6	1.5	202.1
7	1.5	132.9
8	2.5	413.5
9	2.5	231.5
10	2.5	466.7
11	2.5	365.3
12	3.5	493.7
13	3.5	382.3
14	3.5	447.2
15	3.5	563.8

- Efectúe un análisis de regresión lineal simple de y con x .
- Calcule la suma de cuadrados del error puro y haga una prueba para la falta de ajuste.
- ¿La información del inciso *b* indica la necesidad de un modelo en x más allá de una regresión de primer orden? Explique su respuesta.

11.53 Los datos siguientes representan la calificación en química de una muestra aleatoria de 12 estudiantes de nuevo ingreso a cierta universidad, así como sus calificaciones en una prueba de inteligencia aplicada mientras estudiaban el último año de preparatoria.

Estudiante	Calificación en la prueba, x	Calificación en química, y
1	65	85
2	50	74
3	55	76
4	65	90
5	55	85
6	70	87
7	65	94
8	70	98
9	55	81
10	70	91
11	50	76
12	55	74

- Calcule e interprete el coeficiente de correlación de la muestra.
- Establezca las suposiciones necesarias acerca de las variables aleatorias.

- Pruebe la hipótesis de que $\rho = 0.5$ en comparación con la alternativa de que $\rho > 0.5$. Use un valor P para las conclusiones.

11.54 La sección de negocios del *Washington Times* de marzo de 1997 listaba 21 diferentes computadoras e impresoras usadas, así como sus precios de lista. También se listaba la oferta promedio. En la figura 11.29 de la página 439 se presenta una parte de los resultados impresos por computadora del análisis de regresión usando el programa SAS.

- Explique la diferencia entre el intervalo de confianza sobre la media y el intervalo de predicción.
- Explique por qué los errores estándar de la predicción varían de una observación a otra.
- ¿Cuál observación tiene el menor error estándar de la predicción? Explique su respuesta.

11.55 Considere los datos de los vehículos de *Consumer Reports* que se incluyen en la figura 11.30 de la página 440. El peso se indica en toneladas, el rendimiento en millas por galón y también se incluye el cociente de manejo. Se ajustó un modelo de regresión que relaciona el peso x con el rendimiento y . En la figura 11.30 de la página 440 se observa una salida parcial del SAS con los resultados de dicho análisis de regresión, y en la figura 11.31 de la página 441 se incluye una gráfica de los residuales y el peso de cada vehículo.

- A partir del análisis y la gráfica de los residuales, ¿se podría concluir que cabría la posibilidad de encontrar un modelo mejorado si se usara una transformación? Explique su respuesta.
- Ajuste el modelo reemplazando el peso con el logaritmo del peso. Comente los resultados.
- Ajuste un modelo reemplazando mpg con los galones por cada 100 millas recorridas, como se reporta con frecuencia el rendimiento del combustible en otros países. ¿Cuál de los tres modelos es preferible? Explique su respuesta.

11.56 A continuación se presentan las observaciones registradas del producto de una reacción química tomadas a temperaturas diferentes:

x (°C)	y (%)	x (°C)	y (%)
150	75.4	150	77.7
150	81.2	200	84.4
200	85.5	200	85.7
250	89.0	250	89.4
250	90.5	300	94.8
300	96.7	300	95.3

- Grafique los datos.
- ¿La gráfica indica que la relación es lineal?
- Haga un análisis de regresión lineal simple y pruebe la falta de ajuste.

- d) Saque conclusiones con base en el resultado del inciso c.

11.57 La prueba de acondicionamiento físico es un aspecto importante del entrenamiento atlético. Una medida común para determinar la aptitud cardiovascular es el volumen máximo de oxígeno que se inhala al realizar un ejercicio extenuante. Se realizó un estudio con 24 hombres de mediana edad para analizar cómo el tiempo que les tomaba correr una distancia de dos millas influía en el oxígeno que consumían, el cual se midió con métodos estándar de laboratorio mientras los sujetos se ejercitaban en una banda sin fin. El trabajo fue publicado en el artículo “Maximal Oxygen Intake Prediction in Young and Middle Aged Males”, *Journal of Sports Medicine* 9, 1969, 17-22. A continuación se presentan los datos.

Sujeto	y, Volumen máximo de O ₂	x, Tiempo en segundos
1	42.33	918
2	53.10	805
3	42.08	892
4	50.06	962
5	42.45	968
6	42.46	907
7	47.82	770
8	49.92	743
9	36.23	1045
10	49.66	810
11	41.49	927
12	46.17	813
13	46.18	858
14	43.21	860
15	51.81	760
16	53.28	747
17	53.29	743
18	47.18	803
19	56.91	683
20	47.80	844
21	48.65	755
22	53.67	700
23	60.62	748
24	56.73	775

- a) Estime los parámetros en un modelo de regresión lineal simple.
- b) ¿El tiempo que toma correr dos millas influye de forma significativa en la cantidad máxima de oxígeno consumido? Utilice $H_0: \beta_0 = 0$ en comparación con $H_1: \beta_1 \neq 0$.
- c) Grafique los residuales en una gráfica en comparación con x y haga comentarios sobre qué tan apropiado es el modelo lineal simple.

11.58 Suponga que cierto científico postula el modelo

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n,$$

y β_0 es un **valor conocido** no necesariamente igual a cero.

- a) ¿Cuál es el estimador apropiado de mínimos cuadrados de β_1 ? Justifique su respuesta.
- b) ¿Cuál es la varianza del estimador de la pendiente?

11.59 Para el modelo de regresión lineal simple demuestre que $E(s^2) = \sigma^2$.

11.60 Suponga que las ϵ_i son independientes y que se distribuyen normalmente con medias de cero y varianza común σ^2 , y demuestre que B_0 , el estimador de mínimos cuadrados de β_0 en $\mu_{y|x} = \beta_0 + \beta_1 x$, se distribuye de manera normal con media β_0 y varianza

$$\sigma_{B_0}^2 = \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2.$$

11.61 Para un modelo de regresión lineal simple

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n,$$

donde las ϵ_i son independientes y se distribuyen normalmente con medias de cero y varianzas iguales σ^2 , demuestre que

$$B_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

y tienen covarianza de cero.

11.62 Demuestre, en el caso de un ajuste de mínimos cuadrados al modelo de regresión lineal simple

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \dots, n,$$

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

11.63 Considere la situación del ejercicio de repaso 11.62 pero suponga que $n = 2$, es decir, que sólo disponemos de dos puntos de datos. Argumente que la recta de regresión de mínimos cuadrados tendrá como resultado $(y_1 - \hat{y}_1) = (y_2 - \hat{y}_2) = 0$. También demuestre que para este caso $R^2 = 1.0$.

11.64 En el ejercicio de repaso 11.62 se pidió al estudiante que demostrara que $\sum_{i=1}^n (y_i - \hat{y}_i) = 0$ para un

modelo de regresión lineal simple estándar. ¿Se cumple también para un modelo con intersección en el origen? Demuestre su respuesta, ya sea afirmativa o negativa.

11.65 Suponga que un experimentador plantea un modelo como

$$Y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i, \quad i = 1, 2, \dots, n,$$

cuando en realidad una variable adicional, digamos x_2 , también contribuye linealmente a la respuesta. Entonces, el verdadero modelo es dado por

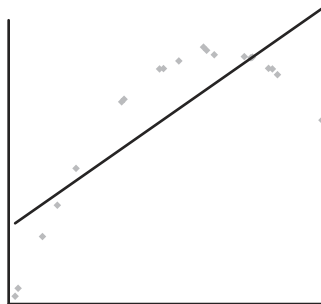
$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, 2, \dots, n.$$

Calcule el valor esperado del estimador

$$B_1 = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1) Y_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}.$$

11.66 Demuestre los pasos necesarios para convertir la ecuación $r = \frac{b_1}{s/\sqrt{S_{xx}}}$ a la forma equivalente $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

11.67 Considere el siguiente grupo ficticio de datos, donde la línea que los atraviesa representa la recta de regresión lineal simple ajustada. Grafique los residuales.



11.68 Proyecto: Este proyecto se puede realizar en grupos o de manera individual. Cada grupo o persona debe encontrar un grupo de datos, preferiblemente de su campo de estudios, aunque también pueden ser de otro campo. Los datos se deben ajustar al esquema de regresión, con una variable de regresión x y una variable de respuesta y . Determine con cuidado cuál variable es x y cuál es y . Tal vez necesite consultar una revista científica de su campo si no cuenta con otros datos experimentales.

- Grafique y contra x . Comente sobre la relación que se observa en la gráfica.
- Diseñe un modelo de regresión adecuado a partir de los datos. Utilice una regresión lineal simple o ajuste un modelo polinomial a los datos. Comente acerca de medidas de calidad.
- Grafique los residuales como se indica en el texto. Verifique posibles violaciones de los supuestos. Muestre de forma gráfica una representación de los intervalos de confianza de una respuesta media graficada en comparación con x . Haga comentarios al respecto.

R-Square	Coeff Var	Root MSE	PriceMean				
0.967472	7.923338	70.83841	894.0476				
		Standard					
Parameter	Estimate	Error	t Value	Pr > t			
Intercept	59.93749137	38.34195754	1.56	0.1345			
Buyer	1.04731316	0.04405635	23.77	<.0001			
PredictStd Err Lower 95% Upper 95% Lower 95% Upper 95%							
product	Buyer Price	Value Predict	Mean	Mean	Predict	Predict	
IBM PS/1 486/66420MB	325	375	400.3125.8906	346.12	454.50	242.46	558.17
IBM ThinkPad500	450	625	531.2321.7232	485.76	576.70	376.15	686.31
IBM Think-Dad755CX	1700	1850	1840.3742.7041	1750.99	1929.75	1667.25	2013.49
AST Pentium90 540MB	800	875	897.7915.4590	865.43	930.14	746.03	1049.54
Dell Pentium75 1GB	650	700	740.6916.7503	705.63	775.75	588.34	893.05
Gateway486/75320MB	700	750	793.0616.0314	759.50	826.61	641.04	945.07
Clone 586/1331GB	500	600	583.5920.2363	541.24	625.95	429.40	737.79
CompaqContura4/25 120MB	450	600	531.2321.7232	485.76	576.70	376.15	686.31
CompaqDeskproP90 1.2GB	800	850	897.7915.4590	865.43	930.14	746.03	1049.54
MicronP75 810MB	800	675	897.7915.4590	865.43	930.14	746.03	1049.54
MicronP100 1.2GB	900	975	1002.5216.1176	968.78	1036.25	850.46	1154.58
Mac Quadra840AV 500MB	450	575	531.2321.7232	485.76	576.70	376.15	686.31
Mac Performer6116 700MB	700	775	793.0616.0314	759.50	826.61	641.04	945.07
PowerBook540c 320MB	1400	1500	1526.1830.7579	1461.80	1590.55	1364.54	1687.82
PowerBook5300 500MB	1350	1575	1473.8128.8747	1413.37	1534.25	1313.70	1633.92
Power Mac 7500/1001GB	1150	1325	1264.3521.9454	1218.42	1310.28	1109.13	1419.57
NEC Versa486 340MB	800	900	897.7915.4590	865.43	930.14	746.03	1049.54
Toshiba1960CS320MB	700	825	793.0616.0314	759.50	826.61	641.04	945.07
Toshiba4800VCT500MB	1000	1150	1107.2517.8715	1069.85	1144.66	954.34	1260.16
HP Laser jet III	350	475	426.5025.0157	374.14	478.86	269.26	583.74
Apple Laser Writer Pro 63	750	800	845.4215.5930	812.79	878.06	693.61	997.24

Figura 11.29: Salida por computadora de los resultados del SAS que presenta el análisis parcial de datos del ejercicio de repaso 11.54.

Obs	Model	WT	MPG	DR_RATIO
1	Buick EstateWagon	4.360	16.9	2.73
2	Ford CountrySquireWagon	4.054	15.5	2.26
3	ChevyMa libu Wagon	3.605	19.2	2.56
4	ChryslerLeBaronWagon	3.940	18.5	2.45
5	Chevette	2.155	30.0	3.70
6	ToyotaCorona	2.560	27.5	3.05
7	Datsun510	2.300	27.2	3.54
8	Dodge Omni	2.230	30.9	3.37
9	Audi 5000	2.830	20.3	3.90
10	Volvo 240 CL	3.140	17.0	3.50
11	Saab 99 GLE	2.795	21.6	3.77
12	Peugeot694 SL	3.410	16.2	3.58
13	Buick CenturySpecial	3.380	20.6	2.73
14	MercuryZephyr	3.070	20.8	3.08
15	Dodge Aspen	3.620	18.6	2.71
16	AMC ConcordD/L	3.410	18.1	2.73
17	Chevy CapriceClassic	3.840	17.0	2.41
18	Ford LTP	3.725	17.6	2.26
19	MercuryGrandMarquis	3.955	16.5	2.26
20	Dodge St Regis	3.830	18.2	2.45
21	Ford Mustang4	2.585	26.5	3.08
22	Ford MustangGhia	2.910	21.9	3.08
23	Macda GLC	1.975	34.1	3.73
24	Dodge Colt	1.915	35.1	2.97
25	AMC Spirit	2.670	27.4	3.08
26	VW Scirocco	1.990	31.5	3.78
27	Honda AccordLX	2.135	29.5	3.05
28	Buick Skylark	2.570	28.4	2.53
29	Chevy Citation	2.595	28.8	2.69
30	Olds Omega	2.700	26.8	2.84
31	PontiacPhoenix	2.556	33.5	2.69
32	PlymouthHorizon	2.200	34.2	3.37
33	Datsun210	2.020	31.8	3.70
34	Fiat Strada	2.130	37.3	3.10
35	VW Dasher	2.190	30.5	3.70
36	Datsun810	2.815	22.0	3.70
37	BMW 320i	2.600	21.5	3.64
38	VW Rabbit	1.925	31.9	3.78
R-Square		Coeff Var	Root MSE	MPG Mean
0.817244		11.46010	2.837580	24.76053
		Standard		
Parameter	Estimate	Error	t Value	Pr > t
Intercept	48.67928080	1.94053995	25.09	<.0001
WT	-8.36243141	0.65908398	-12.69	<.0001

Figura 11.30: Salida de computadora de los resultados del SAS que muestra el análisis parcial de los datos del ejercicio de repaso 11.55.

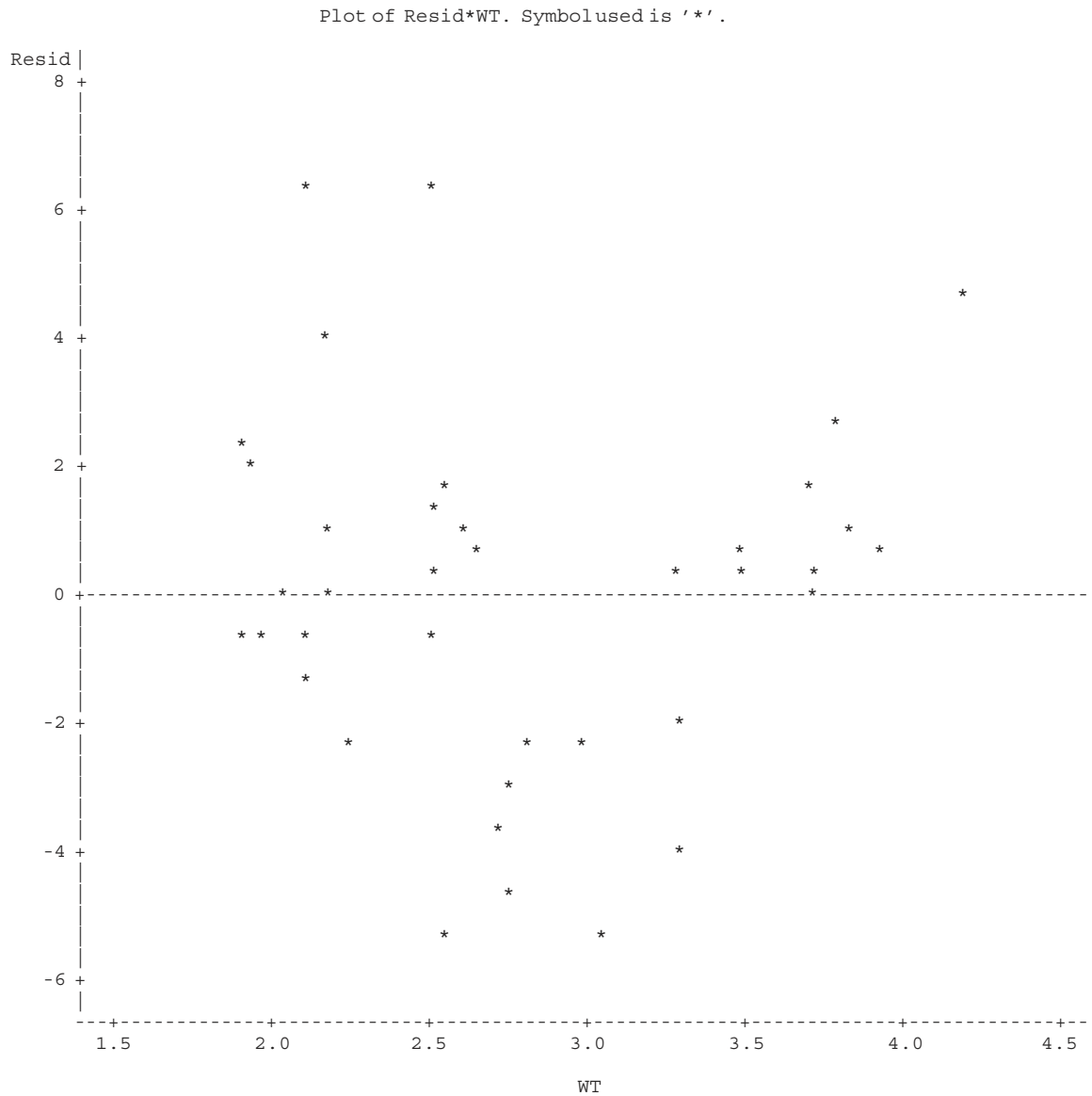


Figura 11.31: Salida de computadora de los resultados del SAS que muestra la gráfica de residuales del ejercicio de repaso 11.55.

11.13 Posibles riesgos y errores conceptuales; relación con el material de otros capítulos

Cada vez que se considere utilizar la regresión lineal simple no sólo es recomendable elaborar una gráfica de los datos, sino esencial. Siempre es edificante elaborar una gráfica de los residuales ordinarios y otra de la probabilidad normal de los mismos. Además, en el capítulo 12 se presentará e ilustrará un tipo adicional de residual en forma estandarizada. Todas esas gráficas están diseñadas para detectar la transgresión de las suposiciones.

El uso de los estadísticos t para las pruebas sobre los coeficientes de regresión es razonablemente robusto para la suposición de normalidad. La suposición de varianza homogénea es crucial y las gráficas de los residuales están diseñadas para detectar una violación.

El material de este capítulo se utiliza ampliamente en los capítulos 12 a 15. Toda la información acerca del método de los mínimos cuadrados para la elaboración de modelos de regresión se utilizará en el capítulo 12. La diferencia es que en ese capítulo se abordan las condiciones científicas en las que hay más de una sola variable x , es decir, más de una variable de regresión. Sin embargo, también utilizaremos el material de este capítulo en el que se exponen los diagnósticos de regresión, los tipos de gráficas residuales, las medidas de la calidad del modelo, etcétera. El estudiante notará que en el capítulo 12 habrá más complicaciones, lo cual se debe a que los problemas de los modelos de regresión múltiple suelen incluir el fundamento de las cuestiones respecto a cómo las diversas variables de regresión entran en el modelo, e incluso el tema de cuáles variables deben permanecer en el modelo. De hecho, el capítulo 15 incluye el uso constante de los modelos de regresión, pero en el resumen al final del capítulo 12 presentaremos una vista preliminar de la conexión.