Unidad 4

- Regresión Lineal y Correlación

11.1 Introducción
11.2 Un modelo probabilístico lineal simple
11.3 El método de mínimos cuadrados
11.4 Cálculo de s^2, un estimador de o^2
11.5 Inferencias relativas a la pendiente ß1 de una recta
11.6 Estimación de E(y/x), el valor esperado de “y” para un valor dado de “x”
11.7 Predicción de un valor particular de “y” para un valor dado de “x”
11.8 Coeficiente de correlación
11.9 La aditividad de sumas de cuadrados
11.10 Resumen

11 Regresión lineal y correlación

11.1 Introducción

Un problema de estimación que es particularmente importante en casi cualquier campo de estudio es el de **pronosticar** o **predecir** el valor de una variable de algún proceso, a partir de los valores conocidos de otras variables que estén relacionadas. Existe un gran número de ejemplos prácticos de predicción en negocios, industrias y en la ciencia. El corredor de bolsa busca predecir el valor futuro de las acciones en función de algunos índices clave que sean observables. El gerente de ventas de una cadena de comercios quiere conocer las ventas mensuales futuras de cada comercio en función del número de clientes a los cuales les ha sido otorgado crédito y quizás en función de los gastos que se han hecho en publicidad. El gerente de producción de una planta se interesa en conocer la relación entre el rendimiento en la obtención de un cierto producto químico y una serie de variables asociadas a su proceso de elaboración. Al haber obtenido lo anterior, emplearía en lo futuro los valores de las variables controlables asociadas a los rendimientos más altos en la elaboración del producto químico. El director de personal de una empresa, al igual que el encargado de las admisiones a una universidad, se interesa en medir algunas características individuales del candidato que le permitan saber si es la persona adecuada para el tipo de trabajo. El analista político desea relacionar el éxito finalmente obtenido de una campaña de un candidato con sus características personales, con el tipo de candidatos de la oposición y con las tácticas publicitarias seguidas. En muchos aspectos, todos los problemas de predicción son iguales.

En cierto sentido, el proceder de tipo estadístico que se haga en cada problema es simplemente una formalización del procedimiento que se haría

siguiendo la intuición. Suponga que un analista de la Bolsa de Valores está tratando de predecir el precio de las acciones de una compañía determinada a partir del índice industrial "Dow-Jones," las tasas bancarias de interés y el volumen de ventas de la misma compañía habidas el mes pasado. Al analista le parecerá lógico que el valor de las acciones aumente al aumentar el índice Dow-Jones o al aumentar el volumen de ventas; y al mismo tiempo, le parecerá también lógico que el valor de las acciones baje si aumentan las tasas bancarias de interés, ya que esto haría que disminuyeran las inversiones en la bolsa y aumentaran en los bancos. Cada una de las variables mencionadas parece tener su influencia individual en el precio de las acciones pero cuando se observan en conjunto, puede ser que las variables tengan un efecto de tipo interactivo en el precio de las acciones que es distinto de los efectos individuales. Por ejemplo, ¿cómo se afectará el precio si aumenta el volumen de ventas pero también aumentan las tasas bancarias de interés? Posiblemente para poder contestar lo anterior, habría que incluir otra variable que relacione al volumen de ventas y a las tasas de interés, que sea, por ejemplo, financiamiento bancario total de la compañía el mes pasado. Siguiendo este tipo de razonamiento, se llegaría a un punto extremo e idealizado en el cual el valor en el mercado de las acciones de la compañía se establecería como una **función** matemática de un cierto número de variables como las mencionadas y otras más que pudieran influenciar dicho valor y que pudieran ser observables o cuantificadas fácilmente. Esta función idealizada estaría dada por una ecuación matemática que relaciona al precio de las acciones con todas las variables relevantes y se usaría para predecir.

Note que el problema al que se ha referido es de naturaleza muy general. El interés se centra en una variable aleatoria y que está relacionada con una serie de **variables predictoras independientes** $x_1, x_2, x_3, \ldots$. En el ejemplo visto, la variable y es el precio de las acciones de una determinada compañía en algún momento también determinado y las variables independientes podrían ser:

x_1 = índice industrial Dow-Jones

x_2 = tasa preferencial* de interés

x_3 = volumen de ventas del mes pasado

y algunas otras más. El objetivo es el de medir $x_1, x_2, x_3, \ldots$, substituir los valores obtenidos en la ecuación de predicción y con ella predecir el precio de las acciones. Para poder hacer todo lo anterior, primero se tienen que identificar las variables $x_1, x_2, x_3, \ldots$ y obtener una medida de la fuerza con la que se asocian con y. Después se tiene que construir una ecuación de predicción satisfactoria que exprese a y como una función de las variables independientes seleccionadas.

En este capítulo se restringe el estudio al problema de predecir y cuando es una función lineal de una sola variable independiente. La solución al problema múltiple — como por ejemplo el de predecir el precio de las acciones en función de más de una variable predictora — será tratado en el capítulo 12.

11.2 Un modelo probabilístico lineal simple

Para efectos de ilustración se introduce el tema considerando el problema de predecir las ventas mensuales y de una compañía en la cual sus productos no experimentan una variación estacionaria en sus ventas. Como la variable predictora x se utiliza la cantidad gastada en publicidad por la compañía en el mes bajo estudio. Es de interés ver si en efecto hay una relación entre lo gastado en publicidad y lo vendido y además, si se puede predecir lo que se venderá, y, como una función de lo que se esté dispuesto a gastar en publicidad, x. La evidencia que se presenta en la tabla 11.1 es una lista de gastos publicitarios y volúmenes de ventas de 10 meses que fueron seleccionados al azar de los archivos. Se supondrá que los gastos publicitarios y ventas de estos 10 meses constituyen una muestra de mediciones de las operaciones pasadas y presentes de la compañía.

Tabla 11.1 *Gastos publicitarios y volúmenes de venta de una compañía durante 10 meses elegidos al azar*

MES	GASTOS PUBLICITARIOS x(× \$10,000)	VOLUMEN DE VENTAS y(× \$10,000)
1	1.2	101
2	.8	92
3	1.0	110
4	1.3	120
5	.7	90
6	.8	82
7	1.0	93
8	.6	75
9	.9	91
10	1.1	105

Lo primero que se hace para analizar los datos de la tabla 11.1 es el graficar los datos como puntos en una gráfica, representando el volumen mensual de ventas y en el eje vertical y los gastos publicitarios correspondientes x en el eje horizontal.

La gráfica, mostrada en la figura 11.1, es referida como **diagrama de dispersión.** Se observa en ella que aparentemente y crece cuando x crece. (¿Podría haber ocurrido un tal diagrama por casualidad si x y y no estuvieren relacionadas?)

Un método para obtener una ecuación de predicción que relacione a y con x consiste en poner una regla de dibujo sobre la gráfica y moverla hasta que dé la aparencia de que pasa a través de los puntos. La línea recta que resulta se considera el "mejor ajuste" a los datos (véase figura 11.2). Se puede utilizar de-

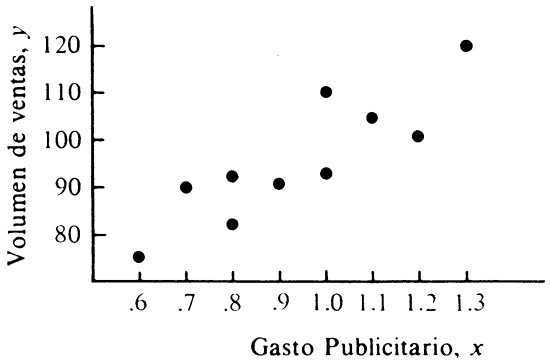

Figura 11.1 *Gráfica de los datos de la tabla 11.1*

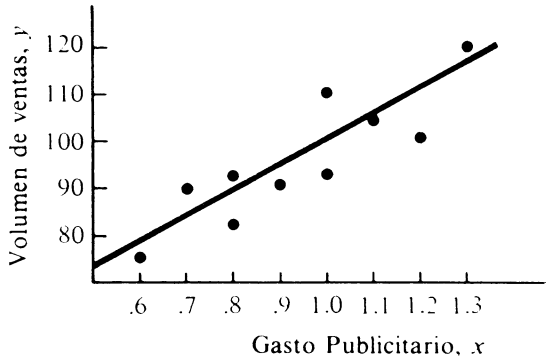

Figura 11.2 *Ajuste visual de una línea*

A continuación se revisan algunos aspectos relacionados con la graficación de funciones matemáticas. Primero, la ecuación matemática de una línea recta es

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

donde β_0 se conoce como la **ordenada al orígen** y β_1 se conoce como la **pendiente** de la recta. En segundo lugar, la línea recta que se puede graficar para cada ecuación lineal es única. Cada ecuación corresponde a una sola línea y viceversa. Así, cuando se dibuja una línea a través de los puntos del diagrama, se está escogiendo de manera automática un modelo para la respuesta* y, dado por

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

en donde β_0 y β_1 son valores numéricos únicos.

El modelo lineal $y = \beta_0 + \beta_1 x$ se dice ser un modelo matemático determinístico ya que una vez que un valor fijo de x sea sustituido en la ecuación, el valor de y queda determinado sin tomar en cuenta la posibilidad de error

*N. del T. Respuesta se utiliza en el sentido de que se busca o deduce de la x. También y se llama frecuentemente **variable dependiente.**

alguno. El ajustar visualmente una línea recta al conjunto de puntos del diagrama produce un modelo determinístico. Existen muchos ejemplos de modelos determinísticos que se pueden ver al hojear libros de texto elementales de química, física, economía o ingeniería.

Los modelos determinísticos resultan apropiados para explicar y predecir algunos fenómenos en los cuales el error de predicción resulta despreciable para efectos prácticos. Así, la ley de Newton que expresa la relación entre la fuerza F de un cuerpo en movimiento de masa m con aceleración a, está dada por

$$F = ma$$

y predice la fuerza con un error muy pequeño para efectos prácticos. El "muy pequeño" es desde luego un concepto relativo. Un error de 0.1 cm. en la construcción de la viga de un puente es muy pequeño, pero ese mismo error en la construcción de una pieza de relojería sería absurdamente grande. En todas las situaciones, el error de predicción no puede ser ignorado y consistentes con lo ya establecido, no confiaremos mucho en predicciones que no vengan acompañadas de una medida de su bondad. Por esta razón, el ajuste visual de una línea recta para relacionar gastos publicitarios con volúmenes de venta puede tener un valor bastante limitado.

Regresando al ejemplo, ¿qué tiene de malo el usar un modelo determinístico para representar la relación entre el volumen de ventas y lo gastado en publicidad? La respuesta es que aun cuando los modelos determinísticos sí permiten pronosticar (substituya el valor de x en la ecuación de predicción), no proporcionan formas para evaluar el error de predicción. Por ejemplo si se usa la recta de la figura 11.2 para predecir el volumen de ventas si se gasta $x = 1.2$ en publicidad, se obtiene un valor pronosticado de $y = 111$ pero, ¿dentro de qué tanto de más y de menos?

Puede observarse que los puntos gráficados se desvían considerablemente de la línea recta y aparentemente se desvían de la recta en una forma aleatoria. ¿Cuál debe ser una cota para el error de predicción? Un ejecutivo necesitaría de esa información para poder decidir si sería costeable realizar un determinado gasta en publicidad. Se requiere, en consecuencia, de un método que permita ajustar un modelo a los datos, un método que también permita exhibir cotas para los errores en la estimación de β_0 y β_1 y cotas en el error de predecir y para un valor dado de x. Más aún, se requiere de un método que pueda utilizarse para ajustar modelos en casos en los que más de una variable predictora x sea usada.

El modelo probabilístico que se usa para relacionar el volumen de ventas y y los gastos publicitarios x es una variante sencilla del modelo determinístico. En lugar de que se diga que y y x se relacionan a través del modelo determinístico

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

se dice que la media o el valor esperado de y para un valor dado de x, que se denota $E(y/x)$, tiene una gráfica que es una línea recta. Esto es, se escribe,

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

Para cualquier x, los valores y varían de manera aleatoria alrededor de la media $E(y/x)$. Por ejemplo, si la compañía gasta \$10,000 ($x = 1.0$) en publicidad cada mes, el volumen de ventas mensual tomará algún valor dentro de la población de valores posibles; pero el volumen *medio* de ventas mensuales se obtiene precisamente sustituyendo $x = 1.0$ en la ecuación

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

Para resumir, se escribirá el modelo probabilístico para cualquier valor observable de y como,

y = (valor medio para y dado el valor de x) + (error aleatorio)

o sea

$$= \overbrace{\beta_0 + \beta_1 x}^{E(y|x)} + \epsilon$$

en donde ϵ es un error aleatorio que representa a la diferencia entre el valor observado de y y su valor medio para ese valor de x. Se está suponiendo entonces, que para un valor dado de x, el valor observado de y varía de manera aleatoria y tiene una distribución de probabilidad cuya media es $E(y/x)$. Como un auxilio en la presentación de esta idea, las distribuciones de probabilidad de y alrededor de la línea recta de medias $E(y/x)$ se muestran en la figura 11.3 para tres valores hipotéticos de x: x_1, x_2, y x_3.

¿Qué se supondrá para la distribución de y dado un valor de x? Las suposiciones que aparentemente se satisfacen en un gran número de situaciones prácticas y que sirven de base para la metodología de las secciones 11.4, 11.5, 11.6 y 11.7, se presentan en el cuadro.

Suposiciones para el modelo probabilístico

Para cualquier valor de x, y tiene una distribución normal con media dada por la ecuación

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

y con varianza σ^2. Más aún, los valores de y son independientes entre sí.

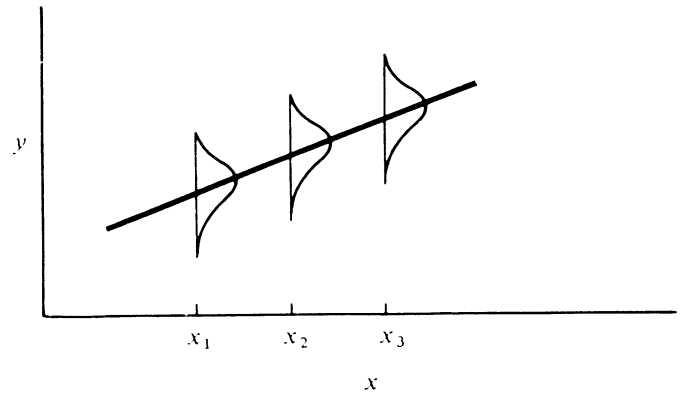

Figura 11.3 *Modelo probabilístico lineal*

Estas suposiciones permiten construir pruebas de hipótesis y permiten construir intervalos de confianza para β_0, β_1 y $E(y/x)$. Además se puede construir un intervalo de predicción para y cuando se usa el modelo ajustado (la línea recta encontrada usando un conjunto de datos) para predecir el valor que tendrá y, para valores particulares de x. La cantidad de problemas prácticos que pueden resolverse utilizando los procedimientos anteriores se hará evidente después de haber hecho algunos ejemplos.

Ejercicios

11.1. Para la ecuación lineal $y = 5 - 4x$:
a. Obtenga la ordenada al origen y la pendiente de la recta.
b. Grafique la recta dada por la ecuación.

11.2. Grafique la recta correspondiente a la ecuación $2y = 3x + 1$.

11.3. ¿Cuál es la diferencia entre modelos matemáticos determinísticos y probabilísticos?

11.3 El método de mínimos cuadrados

El procedimiento estadístico para encontrar la línea recta de "mejor ajuste" a un conjunto de puntos es en cierto modo, una formalización del procedimiento empleado al hacer un ajuste visual de una recta. Por ejemplo, cuando se ajusta visualmente una recta, se mueve la regla hasta que se piensa que se han minimizado las **desviaciones** que presentan los puntos en relación a la recta que se propone. Si se denota el valor ajustado o de pronóstico para y por $\hat{y}$, entonces la ecuación de pronóstico o predicción es,

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

en donde $\hat{\beta}_0$ y $\hat{\beta}_1$ representan estimaciones hechas de las verdaderas β_0 y β_1. La línea para los datos de la tabla 11.1 se muestra en la figura 11.4.

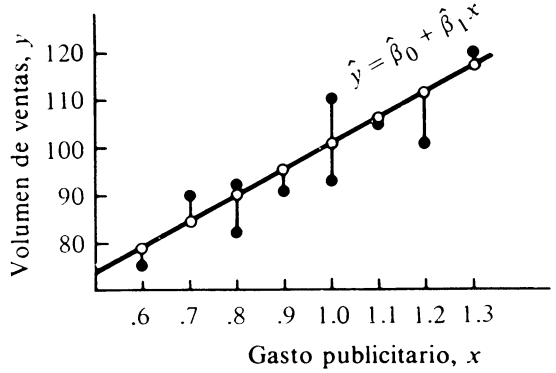

Figura 11.4 *Ecuación lineal de predicción*

Las líneas verticales dibujadas de la recta de predicción a cada uno de los puntos representan las desviaciones entre cada punto y su valor pronosticado (ajustado) para y. La desviación del i-ésimo punto es

$$y_i - \hat{y}_i$$

en donde

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Habiendo decidido que de un modo o de otro, lo que se busca es minimizar las desviaciones escogiendo la recta que se ajuste "mejor," se tiene ahora que definir lo que se entiende por "mejor." Esto es, se requiere de un criterio para calificar "mejor ajuste," que resulte tanto razonable como objetivo y que bajo ciertas suposiciones dé buenas predicciones de y dado un valor de x.

Se empleará un criterio de "bondad" que se conoce como el **principio de mínimos cuadrados,** y que se establece como sigue. **Escoja como la recta de mejor ajuste a la que minimice la suma de los cuadrados de las desviaciones entre los valores observados y los pronosticados.** En expresión matemática, el criterio se lee: minimice

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

El término SCE representa a la suma de los cuadrados de las desviaciones y es referido comúnmente como, **suma de cuadrados del error.**

Ya que el valor pronosticado $\hat{y}_i$ correspondiente a $x = x_i$ es $\hat{\beta}_0 + \hat{\beta}_1 x_i$, se puede sustituir esta expresión en lugar de $\hat{y}_i$ en SCE y se obtiene la siguiente expresión:

$$\text{SCE} = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

El método para encontrar los valores numéricos $\hat{\beta}_0$, $\hat{\beta}_1$ que minimicen SCE utilizan el cálculo diferencial y por ello quedan fuera del alcance de este libro. Se presentan sin embargo las fórmulas con las cuales se obtienen los valores numéricos.

Estimadores de mínimos cuadrados de β_0 y β_1

$$\hat{\beta}_1 = \frac{\text{SC}_{xy}}{\text{SC}_x} \quad \text{y} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

en donde

$$\text{SC}_x = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

y

$$\text{SC}_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Nótese que SC_x (suma de cuadrados para x) se calcula usando una fórmula similar a la usada en el cómputo de la suma de cuadrados de desviaciones s^2 del capítulo 3. SC_{xy} se calcula de manera muy similar. Una vez que $\hat{\beta}_0$ y $\hat{\beta}_1$ se hayan calculado, se sustituyen sus valores en la ecuación de la recta y se obtiene la ecuación de predicción de mínimos cuadrados:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Cabe hacer una aclaración. Los errores de redondeo en el cálculo de SC_x y SC_{xy} pueden afectar enormemente la respuesta. Si se tiene que recurrir al redondeo de ciertos números, se recomienda trabajar con por lo menos seis cifras significativas. (En los ejercicios numéricos del texto, las respuestas del texto y las que usted obtenga podrán tener pequeñas discrepencias, debido a los errores de redondeo.)

El uso de las fórmulas para encontrar $\hat{\beta}_0$ y $\hat{\beta}_1$ y la recta de mínimos cuadrados se ilustra con el siguiente ejemplo.

Ejemplo 11.1 Obtenga la recta de predicción de mínimos cuadrados para el problema cuyos datos aparecen en la tabla 11.1.

Solución El cálculo de $\hat{\beta}_0$ y $\hat{\beta}_1$ a partir de los datos de la tabla 11.1 se simplifica usando la tabla 11.2

Tabla 11.2 *Cálculos para los datos de la tabla 11.1*

	y_i	x_i	x_i^2	$x_i y_i$	y_i^2
	101	1.2	1.44	121.2	10,201
	92	.8	.64	73.6	8,464
	110	1.0	1.00	110.0	12,100
	120	1.3	1.69	156.0	14,400
	90	.7	.49	63.0	8,100
	82	.8	.64	65.6	6,724
	93	1.0	1.00	93.0	8,649
	75	.6	.36	45.0	5,625
	91	.9	.81	81.9	8,281
	105	1.1	1.21	115.5	11,025
SUMAS	959	9.4	9.28	924.8	93,569

Al sustituir las sumas correspondientes de la tabla 11.2 en las ecuaciones de mínimos cuadrados, se obtiene lo siguiente:

$$SC_x = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} = 9.28 - \frac{(9.4)^2}{10} = .444$$

$$SC_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} = 924.8 - \frac{(9.4)(959)}{10} = 23.34$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{959}{10} = 95.9 \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{9.4}{10} = .94$$

De ahí que,

$$\hat{\beta}_1 = \frac{SC_{xy}}{SC_x} = \frac{23.34}{.444} = 52.5676 \approx 52.57$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 95.9 - (52.5676)(.94) \approx 46.49$$

De acuerdo al principio de mínimos cuadrados, la línea recta de mejor ajuste (comúnmente referida como **recta de regresión),** que relaciona los gastos en publicidad con el volumen de ventas es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad \text{o sea} \quad \hat{y} = 46.49 + 52.57x$$

La gráfica de dicha recta se encuentra en la figura 11.4. Se puede ahora predecir un y para un valor dado de x simplemente usando la gráfica de la figura 11.4 o sustituyendo ese valor de x en la ecuación de predicción. Por ejemplo si se han presupuestado \$10,000 para publicidad este mes, el volumen de ventas pronosticado para este mes es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 46.49 + (52.57)(1.0) = 99.06$$

o expresando en dólares, el volumen de venta pronosticado es \$990,600.

No olvide que la línea recta de mejor ajuste, esto es, la ecuación de predicción

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

también estima a la recta de medias

$$E(y|x) = \beta_0 + \beta_1 x$$

En el ejemplo 11.1, la mejor estimación para el volumen medio de ventas mensuales para un gasto publicitario $x = 1.0$ (\$10,000) es también \$990,600. Esto es, la ecuación de predicción se puede usar tanto para predecir un valor futuro de y como para estimar el valor medio de y para un determinado gasto publicitario x.

El siguiente paso será el de encontrar cotas para el error de estimación o predicción. Este problema y otros relacionados se verán en las siguientes secciones.

Ejercicios

11.4. Suponga que le son dados cinco puntos cuyas coordenadas son las de la tabla.

a. Encuentre la recta de mínimos cuadrados para los datos.

b. Como una verificación de sus cálculos en a, grafique los puntos y la recta de mínimos cuadrados.

x	−3	−1	1	1	2
y	6	4	3	1	1

11.5. Por presupuesto flexible, se entiende la relación entre ingresos y costos. Suponga que un ejecutivo de una empresa quiere establecer un presupuesto flexible para estimar sus costos para un cierto rango de producción. Los costos y predicciones pasadas se encuentran en la tabla.

a. Encuentre la recta de mínimos cuadrados que le permita estimar costos a partir de la producción.

b. Como verificación de sus cálculos, grafique los 7 puntos y la recta de mínimos cuadrados.

Producción (× $10,000)	3	4	5	6	7	8	9
Costos fijos (× $1,000)	12	10.5	13	12	13	13.3	16.5

11.4 Cálculo de s^2, un estimador de σ^2

Recuerde que el modelo probabilístico para y en la sección 11.2 supone que y está relacionada con x a través de la ecuación

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

Para un valor de x dado, los valores de y son independientes y tienen distribución normal de media $E(y|x)$ y varianza σ^2. Así, cada valor observado de y está sujeto a un error ϵ que de algún modo entra en el cálculo de $\hat{\beta}_0$ y $\hat{\beta}_1$ e introduce un error en estos estimadores. Más aún, si se usa la recta de mínimos cuadrados.

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

para predecir un valor futuro de y, los errores aleatorios afectarán el error de predicción. En consecuencia, la variabilidad de los errores aleatorios, medida por σ^2, juega un papel muy importante cuando se estima o se pronostica con la recta de mínimos cuadrados.

El primer paso si se quieren encontrar cotas para el error de predicción es estimar σ^2, la varianza de y para un valor dado de x. Para esto parece razonable el usar SCE, la suma de cuadrados de los errores respecto a la recta de predicción. Se puede mostrar que ése es el caso y en el cuadro aparece la fórmula del estimador de σ^2 que resulta ser un buen estimador, insesgado y basado en $(n - 2)$ grados de libertad.

Un estimador para σ^2

$$\hat{\sigma}^2 = s^2 = \frac{\text{SCE}}{n - 2}$$

La suma de cuadrados de las desviaciones SCE se puede calcular directamente por medio del uso de la ecuación de predicción para calcular $\hat{y}$ para cada punto y de ahí, calculando las desviaciones $(y_i - \hat{y}_i)$, y finalmente

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

Lo anterior es un procedimiento tedioso y un tanto malo desde el punto de vista computacional debido a la cantidad de operaciones que requiere que solamente introducen errores de redondeo. Un método más simple y con menos desventajas computacionales es el que se da en el cuadro.

Fórmula para calcular SCE

$$\text{SCE} = \text{SC}_y - \hat{\beta}_1 \text{SC}_{xy}$$

en donde

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

$$\text{SC}_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Nótese que SC_{xy} ya se había usado en el cálculo de $\hat{\beta}_1$ por lo que ya se tenía calculada. Si tuviera que redondear un número se recomienda trabajar con no menos de seis cifras significativas en los cálculos para evitar problemas serios de falta de aproximación en la respuesta final.

Ejemplo 11.2 Calcule un estimador de σ^2 para los datos de la tabla 11.1.

Solución Primero se calcula

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

$$= 93{,}569 - \frac{(959)^2}{10} = 1{,}600.9$$

Se sustituye este valor de SC_y y los de $\hat{\beta}_1$ y SC_{xy} que ya se obtuvieron en el ejemplo 11.1 en la fórmula para SCE y se obtiene

$$\text{SCE} = \text{SC}_y - \hat{\beta}_1 \text{SC}_{xy} = 1{,}600.9 - (52.5676)(23.34) = 373.97$$

Esto es,

$$s^2 = \frac{\text{SCE}}{n-2} = \frac{373.97}{8} = 46.75$$

Ejercicios

11.6. Calcule SCE y s^2 para los datos del ejercicio 11.4.

11.7. Calcule SCE y s^2 para los datos del ejercicio 11.5.

11.5 Inferencias relativas a la pendiente β_1 de una recta

La primera inferencia que se debe hacer cuando se estudia la relación entre y y x concierne a la existencia misma de dicha relación. En otras palabras se quiere contestar a las siguientes preguntas. ¿Muestran los datos suficiente evidencia como para pensar que el conocimiento de x contribuye para predecir y, en alguna región de observación? ¿Al contrario, se puede considerar que es bastante probable que, aun sin estar y y x relacionadas, los puntos observados formen un diagrama como el de la figura 11.1?

Las preguntas desde el punto de vista práctico se refieren al valor de β_1; el cambio medio que se experimenta en y por unidad de cambio en x. El decir que x no proporciona información para predecir y es equivalente a decir $\beta_1 = 0$. (Si $\beta_1 = 0$, se predice siempre el mismo valor para y sin importar cuál sea el valor de x.) Así que lo primero que hay que hacer es probar la hipótesis de que $\beta_1 = 0$ contra la alternativa de que $\beta_1 \neq 0$. Como ya se habrá sospechado, el estimador $\hat{\beta}_1$ es extremadamente útil en la construcción de un procedimiento para probar esta hipótesis. De ahí que sea necesario examinar la distribución de probabilidad que se obtiene para los valores estimados $\hat{\beta}_1$ cuando se muestrea repetidas veces de la población de interés.

Con base en las suposiciones acerca de la distribución de y para un valor de x, se puede mostrar que tanto $\hat{\beta}_1$ como $\hat{\beta}_0$ tienen una distribución normal y que el valor esperado y varianza para $\hat{\beta}_1$ son

$$E(\hat{\beta}_1) = \beta_1 \qquad \text{y} \qquad \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\mathrm{SC}_x}$$

Esto es, $\hat{\beta}_1$ es un estimador insesgado para β_1; y como se sabe cual es su desviación estándar, se puede construir una estadística como la descrita en la sección 8.14, esto es

$$z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} = \frac{\hat{\beta}_1 - \beta_1}{\sigma / \sqrt{\mathrm{SC}_x}}$$

que tiene una distribución normal estandarizada para muestreo repetido. Ahora bien, como el valor σ^2 es desconocido, se debe obtener un estimador de él que permita estimar finalmente la desviación estándar de $\hat{\beta}_1$, la cual queda estimada por

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{\mathrm{SC}_x}}$$

Sustituyendo a σ por s en la expresión para z, se obtiene, al igual que en el capítulo 9, una estadística

$$t = \frac{\hat{\beta}_1 - \beta_1}{s / \sqrt{\mathrm{SC}_x}} = \frac{\hat{\beta}_1 - \beta_1}{s} \sqrt{\mathrm{SC}_x}$$

la cual puede mostrarse que se distribuye como una t de Student con $(n - 2)$ grados de libertad en muestreo repetido. Note que el número de grados de

libertad asociados a s^2 es el mismo que el número de grados de libertad asociados a t.

De lo anterior, para probar la hipótesis de que β_1 es igual a algún valor numérico, por ejemplo $\beta_1 = 0$, se utiliza el procedimiento de la t que se vió en el capítulo 9. Denote por β_{10} el valor hipotetizado para β_1. El procedimiento de prueba se dá en el cuadro.

Un procedimiento para la prueba de hipótesis relativa a la pendiente de la recta

Hipótesis nula $H_0: \beta_1 = \beta_{10}$

Hipótesis alternativa: La da el experimentador y depende de los valores de β_1 que desee detectar.

Estadística: $t = \dfrac{\hat{\beta}_1 - \beta_{10}}{s}\sqrt{SC_x}$

Región de rechazo: Vea los valores críticos para t, tabla 4 del apéndice para $(n-2)$ grados de libertad.

Ejemplo 11.3 Utilice los datos de la tabla 11.2 para determinar si existe evidencia que indique que β_1 difiere de 0 al utilizar una relación lineal entre el gasto publicitario x y el volumen mensual medio, y, de ventas.

Solución Se desea probar la hipótesis nula

$$H_0: \beta_1 = 0$$

contra la hipótesis alternativa

$$H_a: \beta_1 \neq 0$$

a partir de los datos para gastos publicitarios y volumen mensual de ventas de la tabla 11.2. La estadística para probar la hipótesis es

$$t = \frac{\hat{\beta}_1 - 0}{s}\sqrt{SC_x}$$

y si se escoge $\alpha = .05$, se rechazaría H_0 si $t > 2.306$ o $t < -2.306$, el valor crítico leído de la tabla cuando $(n-2) = 8$ grados de libertad. Sustituyendo estos valores, que ya han sido calculados en los ejemplos 11.1 y 11.2, se obtiene

$$t = \frac{\hat{\beta}_1}{s}\sqrt{SC_x} = \frac{52.57}{6.84}\sqrt{.444} = 5.12$$

El valor de la estadística resultó ser mayor que el valor crítico para la t, 2.306, de donde se rechaza la hipótesis $\beta_1 = 0$ y se concluye que hay evidencia que indica que los gastos publicitarios proporcionan información para la predicción de los volúmenes mensuales de ventas.

Una vez que se ha decidido que β_1 es diferente de 0, resulta de interés el examinar con mayor detalle la relación entre y y x. ¿Si x aumenta una unidad cuál será el cambio estimado para y y qué confianza se puede tener en dicha estimación? En otras palabras, se requiere un estimador para la pendiente β_1. Es probable que hasta este momento no se haya sorprendido usted por la similitud entre los procedimientos de los capítulos 9 y 11, así que tampoco le sorprenderá que el intervalo de confianza del $(1 - \alpha)100\%$ para β_1 sea el que aparece en el cuadro, cosa que puede demostrarse.

Un intervalo de confianza para β_1

$$\hat{\beta}_1 \pm \frac{t_{\alpha/2}\, s}{\sqrt{SC_x}}$$

Ejemplo 11.4 Encuentre el intervalo de confianza del 95% para β_1 basándose en los datos de la tabla 11.1.

Solución El intervalo de confianza del 95% para β_1, basado en los datos de la tabla 11.1 es

$$\hat{\beta}_1 \pm \frac{t_{.025}\, s}{\sqrt{SC_x}}$$

Sustituyendo, se obtiene,

$$52.57 \pm \frac{(2.306)(6.84)}{\sqrt{.444}} = 52.57 \pm 23.67$$

Lo anterior quiere decir que si se aumenta en una unidad la x, esto es si se aumenta en \$10,000 el gasto publictario, se estima que el aumento en los volúmenes mensuales de venta correspondientes será entre 28.90 y 76.24 o en las unidades originales para y, será entre \$289,000 y \$762,400.

Algunos aspectos relativos a la interpretación de los resultados requieren de comentario. Como ya se dijo, β_1 es la pendiente de la recta supuesta sobre una región de observación, que indica el cambio en $E(y/x)$ por unidad de cambio en x. Si no se rechaza la hipótesis $\beta_1 = 0$, eso no quiere decir que y y x no estén relacionadas. En primer lugar se debe tener presente la probabilidad de cometer un error tipo II, esto es, el aceptar que la pendiente es 0 cuando esta hipótesis es falsa. En segundo lugar, puede ser que y y x estén perfectamente relacionadas pero en un modo curvilíneo, no lineal. Como ejemplo, en la figura 11.5 se bosqueja una relación curvilínea entre y y x para x en el dominio: $a \leq x \leq f$. Nótese de la figura 11.5, que una línea recta proporcionaría un buen predictor para y si ésta se ajustara sobre un intervalo pequeño del dominio de x, por ejemplo $b \leq x \leq c$. La recta que hubiera resultado sería la recta 1. Por otro lado si se ajusta una recta sobre la región $c \leq x \leq d$, β_1 resultaría cero y la recta de mejor ajuste sería la recta 2. Lo anterior ocurriría aun en el caso en el que todos los puntos cayesen perfectamente en la curva y y y x tuvieran una

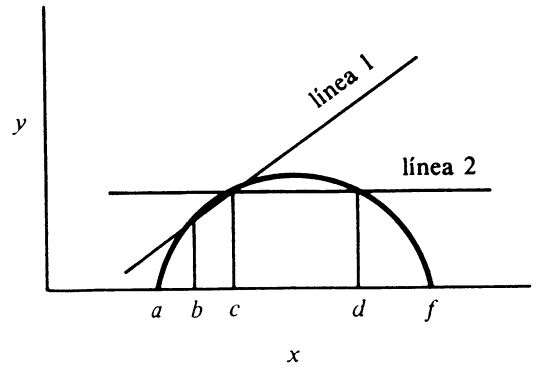

Figura 11.5 *Relación curvilínea*

relación funcional como la definida en la sección 11.2. Se debe tener cuidado al sacar conclusiones si es que no se encuentra evidencia que indique que β_1 difiere de cero, ya que posiblemente la razón se deba a haber escogido el modelo probabilístico equivocado para la situación particular que se estudia.

Los comentarios anteriores contienen una segunda implicación. **Si los datos provienen de x en el intervalo $b \leq x \leq c$, entonces la ecuación de predicción resulta apropiada pero sólo sobre esa región.** El extrapolar para predecir valores de y para x fuera de la región $b \leq x \leq c$ en la situación ilustrada en la figura 11.5 traería como consecuencia un error de predicción considerable.

Si los datos proporcionan evidencia que indique que β_1 difiere de cero, no se puede concluir que la relación entre y y x es lineal. Sin lugar a dudas que y debe ser función de un buen número de variables, cuya existencia queda mostrada en mayor o menor grado por la presencia de un error aleatorio ϵ en el modelo. Lo anterior es la razón por la cual se ha tenido que recurrir a un modelo probabilístico. Los errores grandes en la predicción se deben entonces a curvaturas en la relación entre y y x y a la presencia de otras variables a las cuales no se les toma en cuenta en el modelo y que influyen en los valores de y. Todo lo que se puede decir, sin embargo, es que se tiene evidencia de que el valor de y cambia si el valor de x cambia y de que se puede obtener una mejor predicción para y si se usa x y la línea recta de mejor ajuste que la predicción que se obtendría al usar simplemente $\bar{y}$ e ignorar x. **Lo anterior no implica una relación causal entre *x* y *y*.** Posiblemente hubo una tercera variable que causó tanto el cambio en x como en y, produciendo la relación que se observó.

La desviación estándar del estimador de β_1.

$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SC_x}} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

arroja información sobre el cómo seleccionar los valores de x, esto es, el modo de diseñar un experimento para obtener el mejor estimador de la pendiente β_1 y de la media de los valores de y para un x dado, $E(y/x)$. Para ilustrar lo anterior note que

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

aparece en la fórmula para $\sigma_{\hat{\beta}_1}$. Esa cantidad mide la dispersión (variación) de los valores de x. Mientras más grande sea la dispersión de los valores de x, más grande será

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

Al examinar la fórmula para $\sigma_{\hat{\beta}_1}$, se ve que mientras más grande sea

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

más pequeño será $\sigma_{\hat{\beta}_1}$. Esto proporciona una sugerencia para encontrar un buen diseño de experimento para hacer el mejor ajuste lineal a un conjunto de datos. Fije la gran mayoría de los valores de x en los extremos de la región experimental, la mitad en cada extremo y sólo unos cuantos al centro que permitan detectar presencia de curvatura, si es que existe.

Ejercicios

11.8. En los datos del ejercicio 11.4, ¿hay suficiente evidencia de que y y x estén relacionadas linealmente? (Pruebe la hipótesis $\beta_1 = 0$; use $\alpha = .05$.)

11.9. En los datos del ejercicio 11.5, ¿hay suficiente evidencia de que los costos y la producción estén relacionadas linealmente? Use $\alpha = .05$.

11.10. En un estudio de distintos fondos para inversión se desarrolló un procedimiento consistente en construir la llamada "recta característica" para cada posible fondo. Dicha recta no es otra cosa más que la recta de regresión de la redituabilidad del fondo considerado sobre la redituabilidad promedio del mercado bursátil. Si para un fondo de inversión la pendiente de su recta característica es significativamente distinta de cero, se dice que ese fondo es muy sensible a las fluctuaciones de la bolsa de valores y por ende es una inversión riesgosa. Si el fondo tiene una recta característica con pendiente cercana a cero se dice que es una inversión estable y de poco riesgo. La redituabilidad tanto del fondo "Penn Square Mutual" como la promedio en el mercado bursátil se observó en el período 1964 a 1973 y se dan en la tabla siguiente.

Año	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973
"Penn Square"	18.4	29.7	−12.3	10.8	23.6	−16.2	5.8	7.2	7.7	−8.8
Promedio en mercado	12.9	9.1	−13.1	20.1	7.7	−11.4	.1	10.8	15.6	−17.4

Fuente: Wiesenberger Financial Services; *Investment Companies, Mutual Funds and Other Types*, 1974.

a. Encuentre la "recta característica" del fondo "Penn Square Mutual" (esto es, la recta de regresión de la redituabilidad del fondo sobre la redituabilidad promedio).

b. Grafique los puntos y la recta de regresión para verificar sus cálculos.

c. Describa el tipo de riesgo asociado a invertir en el "Penn Square Mutual" (esto es, pruebe la hipótesis $\beta_1 = 0$; use $\alpha = .05$).

11.11. Encuentre un intervalo confidencial del 95% para la pendiente de la recta característica del fondo "Penn Square Mutual" del ejercicio 11.10.

11.12. Encuentre un intervalo confidencial del 90% para la pendiente de la recta que relaciona costos con producción del ejercicio 11.5.

11.13. Un experimento de mercados se realizó para estudiar la relación entre el tiempo que requiere un comprador para decidirse en su compra y el número de presentaciones distintas del producto exhibidas. Las marcas se eliminaron de los productos para reducir el efecto de las preferencias a determinadas marcas. Los compradores seleccionaron los artículos basados exclusivamente en las descripciones y diseños de las presentaciones de cada producto. El tiempo utilizado hasta llegar a una selección fue registrado para los 15 participantes en el estudio.

Tiempo requerido (en seg.)	5,8,8,7,9	7,9,8,9,10	10,11,10,12,9
Número de alternativas (presentaciones)	2	3	4

a. Encuentre la recta de mínimos cuadrados para estos datos.
b. Grafique los puntos y la recta para verificar sus cálculos.
c. Calcule s^2.
d. ¿Presentan los datos suficiente evidencia que indique que el tiempo requerido para decidir está linealmente relacionado al número de presentaciones? (Pruebe al nivel $\alpha = .05$.)

11.6 Estimación de $E(y/x)$, el valor esperado de y para un valor dado de x

En los capítulos 8 y 9 se estudiaron métodos para estimar la media de una población μ y se vieron muchas situaciones prácticas en las que se aplicaban estos métodos, en los ejemplos y ejercicios. Se considera ahora una generalización de este problema.

El estimar el valor medio de y para un valor de x dado (esto es, estimar $E(y/x)$) puede ser de mucha aplicación. El encargado de seguridad industrial en una empresa puede estar interesado en estimar el número medio de algún tipo de accidentes dado el número de horas que cada empleado ha estado sujeto a entrenamiento especial para seguridad. O quizá el director de personal desea estimar el número medio de años que un empleado permanecerá en la compañía dado un cierto puntaje obtenido en un examen diseñado para medir su compatibilidad con el tipo de trabajo. Si en una empresa, la ganancia y se encuentra linealmente relacionada a los gastos publicitarios x, el gerente de ventas querrá estimar la ganancia media para un cierto nivel de publicidad x. Por ejemplo si la compañía invierte \$10,000 en publicidad, ¿cuánto debe esperar que sea $E(y/x)$? El encontrar un intervalo de confianza para $E(y/x)$ será el tema de esta sección.

Suponga que y y x están linealmente relacionadas de acuerdo al modelo probabilístico definido en la sección 11.2 y por ende que $E(y/x) = \beta_0 + \beta_1 x$ representa al valor esperado de y para un valor dado de x. La recta ajustada

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

pretende estimar la recta de medias $E(y/x)$ (estimando β_0 y β_1). De ahí que y pueda emplearse para estimar el valor esperado de y al igual que para estimar algún valor futuro de y. Parece razonable suponer que los errores en la predic-

ción serán distintos entre sí para estos dos problemas de estimación. Los dos procedimientos de estimación son, de hecho, diferentes. En esta sección se considera la estimación del valor medio de y para un valor dado de x.

Observe las dos rectas de la figura 11.6. Una representa a la recta de medias

$$E(y|x) = \beta_0 + \beta_1 x$$

y la otra es la recta ajustada o ecuación de predicción,

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

Se puede observar en la figura que el error al estimar el valor esperado de y cuando $x = x_p$ es la desviación entre las dos rectas sobre el punto x_p. También se observa que ese error aumenta a medida que x_p se aleja hacia los extremos del intervalo sobre el cual x toma valores.

Se puede mostrar que el valor pronosticado (predicción para y)

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

es un estimador insesgado de $E(y/x)$, esto es, $E(\hat{y}) = \beta_0 + \beta_1 x$, y que $\hat{y}$ tiene distribución normal, con varianza

$$\sigma^2_{\hat{y}} = \sigma^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x} \right]$$

La estimación correspondiente para $\sigma^2_{\hat{y}}$ usa s^2 en lugar de σ^2 en la expresión anterior.

Los resultados anteriores pueden ser usados para probar una hipótesis acerca del valor medio o esperado de y para un valor dado de x, por ejemplo x_p.* (El problema de probar hipótesis sobre la ordenada al orígen β_0 quedaría incluído en el anterior al hacer $x_p = 0$.) La hipótesis nula es

$$H_0 : E(y|x = x_p) = E_0$$

en donde E_0 es un valor numérico hipotético para $E(y)$ cuando $x = x_p$. En esta ocasión también puede mostrarse que la expresión

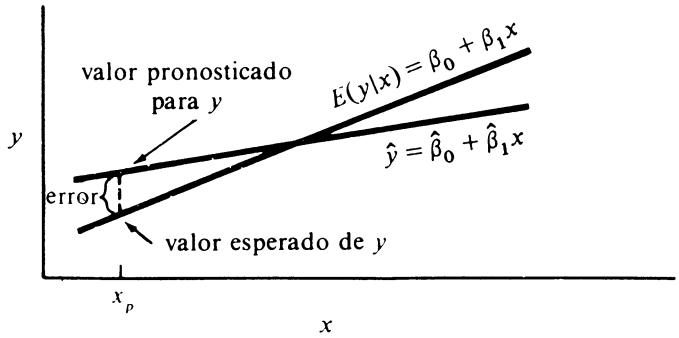

Figura 11.6 *Valores esperados y pronosticados para y*

*Vea los comentarios de la sección 11.5 en relación a la conveniencia de que x_p esté dentro de la región de valores observados de x.

$$t = \frac{\hat{y} - E_0}{\text{estimador de } o_{\hat{y}}} = \frac{\hat{y} - E_0}{s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x}}}$$

tiene una distribución t de Student con $(n - 2)$ grados de libertad en muestreos repetidos. El procedimiento estadístico para probar la hipótesis se desarrolla exactamente igual que para la otra prueba t vista antes.

Un procedimiento para la prueba de hipótesis relativa al valor esperado de y

Hipótesis nula $H_0: E(y/x = x_p) = E_0$.

Hipótesis alternativa: La da el experimentador y depende de los valores de $E(y/x)$ que desea detectar

Estadística de prueba: $t = \dfrac{\hat{y} - E_0}{s\sqrt{\dfrac{1}{n} + \dfrac{(x_p - \bar{x})^2}{SC_x}}}$

Región crítica o de rechazo: Vea los valores críticos para t, tabla 4 del apéndice, para $(n - 2)$ grados de libertad.

El intervalo de confianza correspondiente, con coeficiente de confianza $(1 - \alpha)$, para el valor esperado de y cuando $x = x_p$ se da en el cuadro.

Un intervalo de confianza para $E(y/x)$

$$\hat{y} \pm t_{\alpha/2} s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x}}$$

Ejemplo 11.5 Utilice los datos de la tabla 11.1 y encuentre un intervalo de confianza del 95% para el volumen mensual esperado de ventas cuando los gastos en publicidad son $x_p = 1.0$ ($10,000).

Solución En la estimación del volumen mensual esperado de ventas para un gasto publicitario de $x_p = 1.0$, se utiliza

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_p$$

Con ello se calcula $\hat{y}$, y se estima $E(y/x = 1.0)$. De los valores calculados en ejemplos anteriores, se tiene

$$\hat{y} = 46.49 + (52.57)(1.0) = 99.06 \qquad \text{ó} \qquad \$990.600$$

La expresión para el intervalo de confianza del 95% es

$$\hat{y} \pm t_{.025} s\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x}}$$

Sustituyendo en esta expresión, se encuentra que el intervalo de confianza del 95% para el volumen mensual medio (esperado) asociado a un gasto en publicidad de 1.0 ($\times$ \$10,000) es

$$99.06 \pm (2.306)(6.84)\sqrt{\frac{1}{10} + \frac{(1.0 - .94)^2}{.444}}$$

Efectuando estas operaciones se llega a

$$99.06 \pm 5.19 \quad \text{o lo que es lo mismo, entre} \quad 93.87 \text{ y } 104.25$$

Como cada unidad representa \$10,000, en unidades monetarias se estima que las ventas mensuales esperadas sobre la población de los meses en los que la compañía gasta \$10,000 están entre \$938,700 y \$1,042,500.

Ejercicios

11.14. Refiérase al ejercicio 11.4. Estime el valor esperado de y dado $x = 1$, usando un intervalo de confianza del 90%.

11.15. Refiérase al ejercicio 11.5. Estime el costo medio asociado a una predicción de 55,000 unidades ($x = 5.5$). Utilice un intervalo de confianza del 95%.

11.16. El contínuo aumento en el precio del petróleo en los últimos años ha originado un aumento también contínuo en los costos para el industrial que tiene que transportar sus bienes terminados al mercado. Para abatir esos costos de transporte, el industrial ha sustituido los medios usuales de transporte por otros más baratos; por ejemplo flete ferroviario en lugar de carga aérea. En un estudio hecho en una compañía para estudiar los costos de transporte aéreo, se seleccionaron al azar 9 facturas de transporte aéreo utilizado para enviar mercancía, para estimar la relación entre el costo por unidad transportada y la distancia recorrida. Los resultados se encuentran en la tabla.

Distancia ($\times$ 100 km.)	6	13	27	15	9	11	21	14	12
Costo por unidad transportada	\$49	\$93	\$159	\$115	\$66	\$90	\$139	\$98	\$88

a. Encuentre la recta de mínimos cuadrados para estimar el costo de transporte y (por unidad transportada) a partir de la distancia recorrida (x).

b. Por lógica, si la distancia recorrida fuese 0 km., entonces el costo debiera ser \$0. Explique por qué la recta de mínimos cuadrados no pasa por el orígen. ¿Debería pasar por el orígen?

c. Estime el costo medio de transporte por unidad para una carga que se enviará 1700 km ($x = 17$), usando un intervalo confidencial del 90%.

11.7 Predicción de un valor particular de y para un valor dado de x

Suponga que la ecuación de predicción obtenida de las 10 observaciones de la tabla 11.1 se hubiera utilizado para predecir el volumen de ventas de un mes seleccionado al azar. Aun cuando resulta de interés considerar el valor esperado de y para un valor particular de x, el interés principal es el poder usar la ecuación de predicción $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ basada en los datos observados, para poder

predecir el volumen de ventas de un mes particular (en el cual se supone que la compañía ha estado o está en operación). Si los gastos en publicidad para este mes son x_p, se observa, intuitivamente, que el error en la predicción que se haga (la desviación entre $\hat{y}$ y el volumen de ventas reales y) puede descomponerse en dos aportaciones. La cantidad ($\hat{y}$ − y) es igual a la diferencia entre $\hat{y}$ y el valor esperado de y cuando $x = x_p$ (descrita en la sección 11.6 y mostrada en la figura 11.6) sumada a un error aleatorio ϵ que representa la diferencia entre el valor esperado de y y el valor real de y (véase la figura 11.7). De ahí que la variabilidad en el error en la predicción de un solo valor de y sea superior a la variabilidad del error en la estimación del valor esperado de y.

Se puede mostrar que la varianza del error $(y - \hat{y})$ en la predicción de un valor particular de y cuando $x = x_p$ es

$$\sigma^2_{\text{error}} = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x} \right]$$

Cuando n es muy grande, el segundo y tercer término dentro del paréntesis cuadrado, son muy pequeños y así la varianza del error de predicción se parece (y más mientras más grande sea n) a σ^2. Estos resultados pueden usarse en la construcción de un intervalo de predicción del $(1 - \alpha)$.

Un intervalo de predicción para y

$$\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x}}$$

Ejemplo 11.6 Encuentre un intervalo de predicción del 95% para el volumen de ventas que experimentará la compañía el próximo mes, si destina para publicidad $10,000. Suponga para lo anterior, que todas las condiciones de tipo económico que afectan el problema se mantendrán en el próximo mes aproximadamente igual que durante los meses considerados para la construcción de la tabla 11.1.

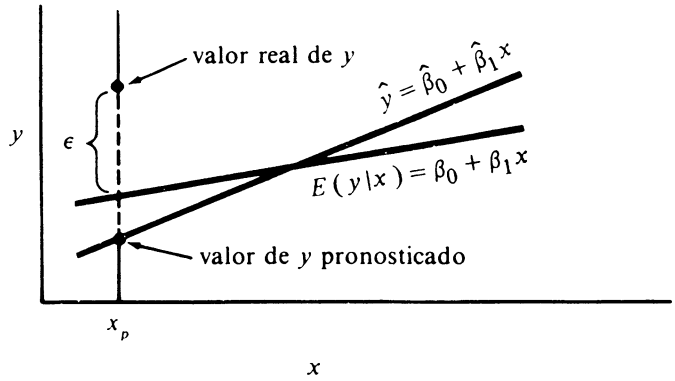

Figura 11.7 *Error en la predicción de un valor particular de y*

Solución Si en un mes particular el gasto en publicidad es de \$10,000 eso quiere decir $x_p = 1.0$ y se predice el volumen de ventas como

$$99.06 \pm (2.306)(6.84)\sqrt{1 + \frac{1}{10} + \frac{(1.0 - .94)^2}{.444}} = 99.06 \pm 16.60$$

o lo que es lo mismo entre 82.46 y 115.66. En las unidades monetarias originales, el intervalo de predicción del 95%, para las ventas del mes próximo, es \$990,600 ± \$166,000, o lo que es lo mismo,

entre \$824,600 y \$1,156,600

Note que en el ejemplo práctico, se tendrían registros de más de 10 meses sobre gastos publicitarios y ventas. El contar con más información, reduciría el ancho del intervalo de predicción, al reducir la cantidad que aparece dentro de la raíz cuadrada en la expresión del intervalo.

Observe la diferencia entre el intervalo confidencial para $E(y/x)$ discutido en la sección 11.6 y el intervalo de predicción presentado en esta sección. $E(y/x)$ es una media, un parámetro de la población de los valores de y y y es una variable que varía alrededor de $E(y/x)$ en forma aleatoria. El valor esperado de y cuando $x = 1.0$ es completamente diferente (conceptualmente) de algún valor de y escogido al azar de entre todos los valores de y asociados a $x = 1.0$. Para hacer una distinción clara al hacer inferencias *siempre se* **estima** *el valor de un parámetro mientras que siempre se* **predice** *(pronóstica) el valor de una variable aleatoria.* Como se indicó previamente y a través de las figuras 11.7 y 11.8, el error en la predicción de y es diferente del

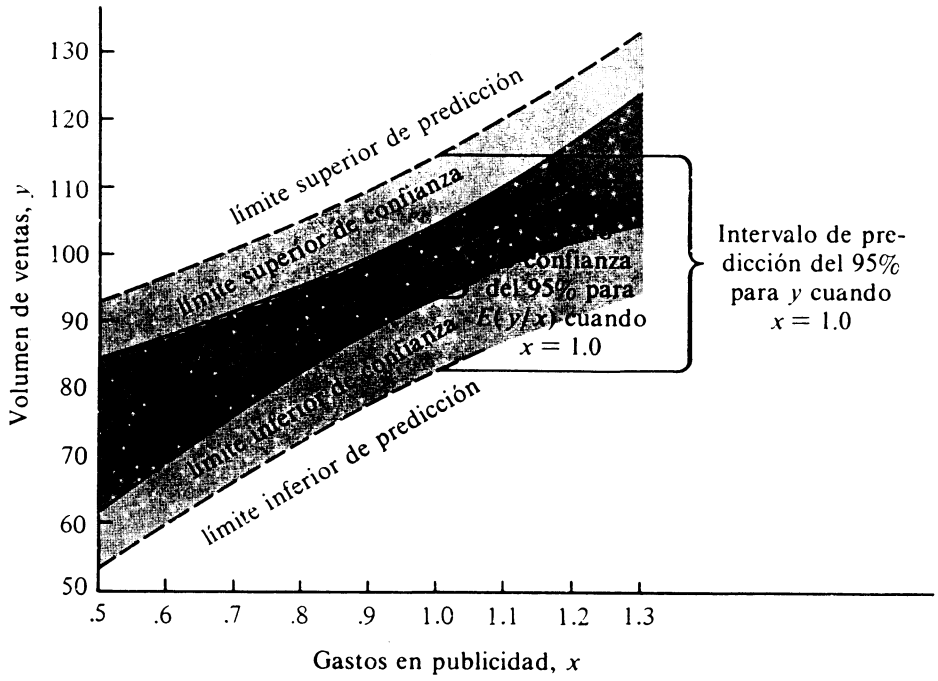

Figura 11.8 *Intervalos de confianza para $E(y/x)$ y de predicción para y basadas en la tabla 11.1*

error en la estimación de $E(y/x)$. Esto es evidente al observar las anchuras de los intervalos tanto de predicción como de confianza.

Una gráfica de la anchura del intervalo de confianza para $E(y/x)$ y de la anchura del intervalo de predicción para y, de los datos de la tabla 11.1, se muestra en la figura 11.8. La gráfica se construyó para ser usada con cualquier valor de x_p. La del intervalo de confianza aparece con líneas contínuas y la de predicción con líneas punteadas. Observe cómo la anchura de los intervalos crece cuando x_p se aleja de $\bar{x} = .94$. En particular, observe los intervalos de confianza y de predicción para $x_p = 1.0$ que fueron calculados en los ejemplos 11.5 y 11.6.

Ejercicios

11.17. Refiérase al ejercicio 11.5. Suponga que la producción en un cierto período es de 55,000 unidades. Encuentre un intervalo de predicción del 95% para y, el costo en el que se incurrirá en ese período. Compare este intervalo de predicción con el intervalo de confianza para $E(y/x = 5.5)$ calculado en el ejercicio 11.15.

11.18. Refiérase al ejercicio 11.13. Suponga que a un comprador, se le muestran tres presentaciones del producto. Encuentre un intervalo de predicción del 90% para y, el tiempo que requeriría el comprador hasta decidirse por una de las presentaciones.

11.19. Se condujo un experimento en un supermercado para estudiar la relación entre la cantidad de espacio destinado a un determinado café (marca A) y el volumen de ventas semanales de ese café. La cantidad de espacio destinado en la estantería se varió en exhibidores ("displays") de 3, 6 y 9 anaqueles aleatoriamente durante 12 semanas, mientras que para las otras marcas de café, se mantuvieron constantes en exhibidores de 3 anaqueles. Los datos del experimento se encuentran en la tabla.

Ventas semanales, y	526	421	581	630	412	560	434	443	590	570	346	672
Exhibidor de x anaqueles	6	3	6	9	3	9	6	3	9	6	3	9

a. Encuentre la recta de mínimos cuadrados para estos datos.
b. Grafique tanto los puntos como la recta como una verificación de sus cálculos.
c. Calcule s^2.
d. Encuentre un intervalo de confianza del 95% para la media de las ventas semanales del café (marca A) si se presenta en exhibidores de 6 anaqueles.
e. Suponga que durante la semana entrante se presentará el café en un exhibidor de 6 anaqueles. Encuentre un intervalo de predicción del 95% para las ventas semanales. Explique la diferencia encontrada entre este intervalo y el obtenido en *d*.

11.8 Coeficiente de correlación

Con frecuencia se requiere de un indicador o medida de la fuerza con la que dos variables y y x se encuentran linealmente relacionadas, de modo que el indicador no dependa de las escalas en las que cada una de las variables y y x se hayan medido. Un tal indicador o medida se conoce como una medida de la **correlación lineal** entre y y x.

La medida de correlación lineal comúnmente usada en la estadística es el llamado **coeficiente de correlación de Pearson** entre y y x. Esta cantidad, denotada por el símbolo r, se calcula como se indica en el cuadro.

Coeficiente de correlación de Pearson

$$r = \frac{SC_{xy}}{\sqrt{SC_x SC_y}}$$

Ejemplo 11.7 Calcule el coeficiente de correlación para los datos de gastos publicitarios y volúmenes de venta de la tabla 11.1.

Solución El coeficiente de correlación para los datos de la tabla 11.1 se obtiene utilizando la fórmula de r y las cantidades

$$SC_{xy} = 23.34 \qquad SC_x = .444 \qquad SC_y = 1600.9$$

que ya habían sido calculadas. De ahí,

$$r = \frac{SC_{xy}}{\sqrt{SC_x SC_y}} = \frac{23.34}{\sqrt{.444(1600.9)}} \approx .88$$

Un estudio sobre el coeficiente de correlación r proporciona resultados interesantes y entre ellos, la razón por la cual se escoge como medida de correlación lineal. Primero se observa que los denominadores, tanto de r como de $\hat{\beta}_1$ son siempre positivos por ser esencialmente sumas de cuadrados. También, se observa que el numerador de r y de $\hat{\beta}_1$ es el mismo. De lo anterior que el coeficiente de correlación r y $\hat{\beta}_1$ siempre tendrá el mismo signo y además r será cero sólo si $\hat{\beta}_1 = 0$. **Así que $r = 0$ implica la ausencia de correlación lineal entre y y x. Un valor de r positivo implica que la pendiente de la recta es positiva (la recta crece a la derecha); un valor de r negativo indica que la recta decrece a la derecha (pendiente negativa).**

La figura 11.9 muestra seis diagramas típicos y sus correspondientes coeficientes de correlación. Nótese que **$r = 0$ implica ausencia de correlación lineal y no ausencia de correlación.** Podría existir un patrón pronunciado de asociación curvilínea como el de la figura 11.9(b) con un coeficiente de correlación lineal igual a cero. En general, se puede decir que r mide la asociación lineal entre dos variables y y x. Cuando $r = 1$ ó $r = -1$, todos los puntos deben pertenecer a una línea recta; cuando $r = 0$, se encuentran dispersas sin mostrar evidencia alguna de relación lineal. Cualquier otro valor de r simplemente sugiere el grado de dependencia lineal.

La interpretación de los valores de r distintos de cero puede lograrse si se comparan los errores de predicción usando la ecuación de predicción

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

con los errores al predecir si se usa exclusivamente como predictor para y, su media muestral $\bar{y}$, que sería el predictor si se ignoraran por completo los valores de x.

Las figuras 11.10(a) y (b) muestran las rectas $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ y $\hat{y} = \bar{y}$ como dos ajustes para los mismos datos. Ciertamente que si x es de algún valor en

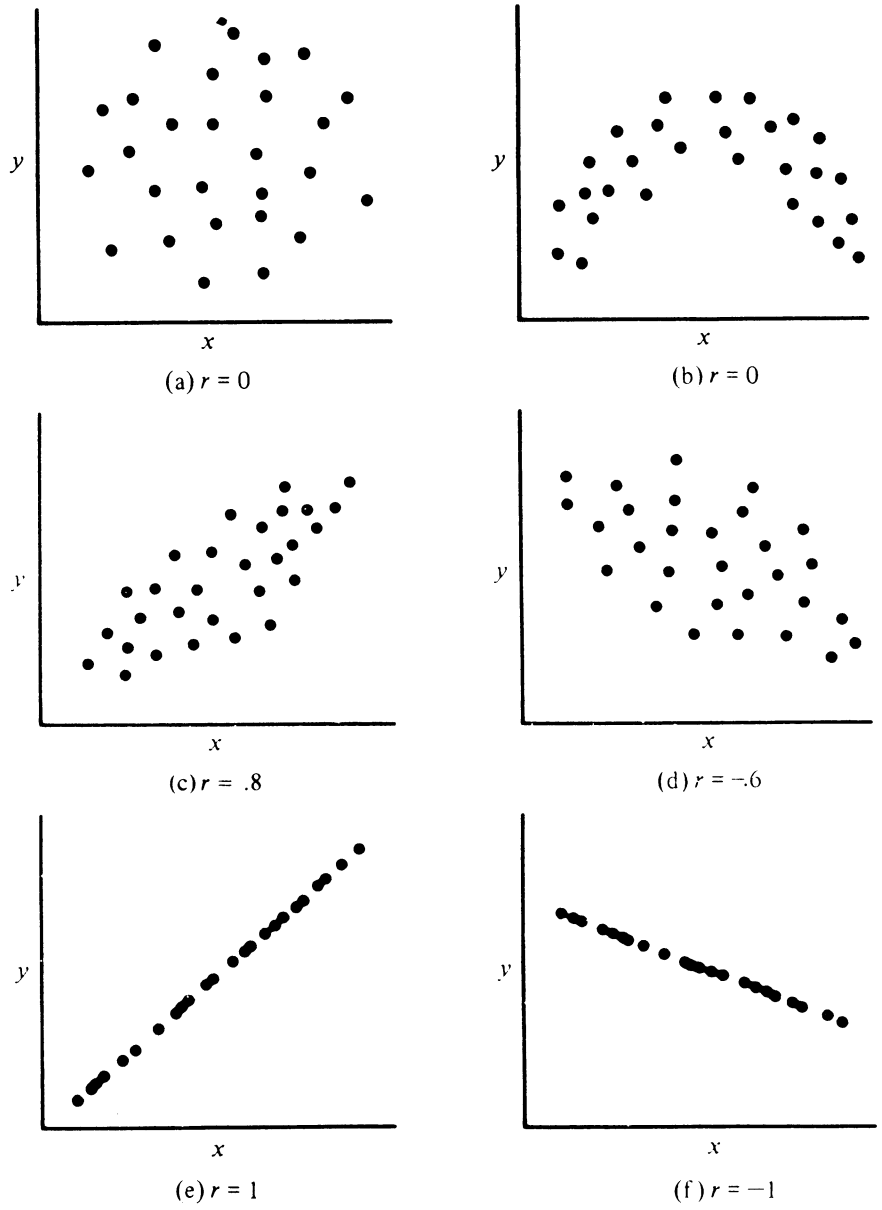

Figura 11.9 *Algunos diagramas típicos y sus correspondientes coeficientes de correlación*

la predicción de y, la suma de cuadrados SCE de desviación de y respecto a la recta $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ deberá ser menor que la suma de cuadrados de desviación de y respecto a su promedio $\bar{y}$, que es

$$SC_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

De hecho, puede verse que SCE no puede ser jamás mayor que

$$SC_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

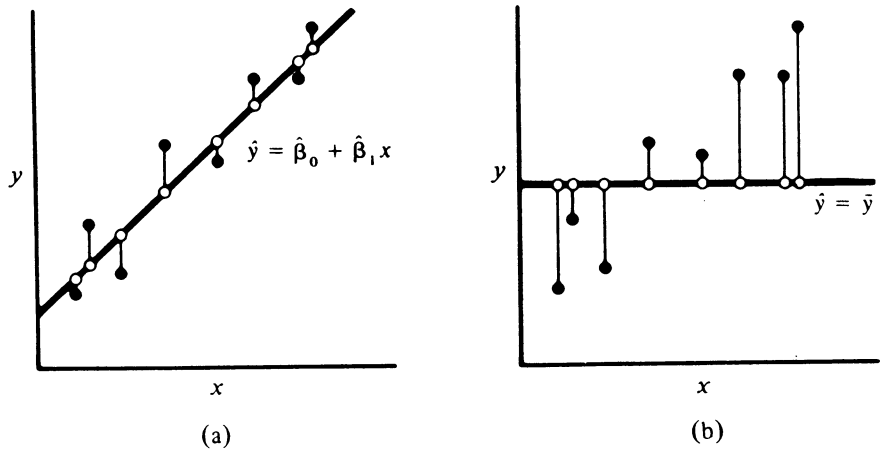

Figura 11.10 *Dos modelos para ajustar los mismos datos*

ya que

$$\text{SCE} = \text{SC}_y - \hat{\beta}_1 \text{SC}_{xy} = \text{SC}_y - \left(\frac{\text{SC}_{xy}}{\text{SC}_x}\right) \text{SC}_{xy} = \text{SC}_y - \frac{(\text{SC}_{xy})^2}{\text{SC}_x}$$

En otras palabras, SCE es igual a SC_y menos una cantidad positiva y de ahí que no pueda ser mayor que SC_y.

Más aún, con una poca de manipulación algebraica, se puede mostrar que

$$r^2 = 1 - \frac{\text{SCE}}{\text{SC}_y} = \frac{\text{SC}_y - \text{SCE}}{\text{SC}_y}$$

En otras palabras, r^2 tiene que tomar valores en el intervalo

$$0 \leq r^2 \leq 1$$

y r valdrá $+1$ ó -1 sólo cuando todos los puntos del diagrama pertenezcan exactamente a la línea recta, que es lo implicado por la condición SCE igual a cero.

De hecho r^2 es igual al cociente entre la reducción en la suma de cuadrados de las desviaciones que se obtiene cuando se pasa del modelo $\hat{y} = \bar{y}$ al modelo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, y la suma de cuadrados total asociada a las desviaciones cuando se considera el modelo $\hat{y} = \bar{y}$. El cuadrado de r, r^2 es llamado el **coeficiente de determinación** y su interpretación como medida de la fuerza de la relación lineal es más sencilla de hacerse que para r.

El coeficiente de determinación

$$r^2 = \frac{\text{SC}_y - \text{SCE}}{\text{SC}_y} = \frac{\displaystyle\sum_{i=1}^{n} (y_i - \bar{y})^2 - \text{SCE}}{\displaystyle\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

Obsérvese que el coeficiente de correlación muestral r es un estimador del coeficiente de correlación poblacional, ρ (rho del griego), que se obtendría si se calculase el coeficiente de correlacion con todos los puntos de la población. Una discusión sobre el procedimiento para probar hipótesis acerca del valor de ρ o del cómo construir cotas para el error de estimación de ρ queda fuera del alcance de este texto. En general el interés sería probar la hipótesis nula de que $\rho = 0$ y una estadística comúnmente usada para ello es, otra vez, una t de Student con $(n - 2)$ grados de libertad:

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

Con álgebra común se puede mostrar que

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{\hat{\beta}_1}{s}\sqrt{SC_x}$$

De ahí que la prueba de t para la hipótesis $\rho = 0$ resulte equivalente a la prueba de t de la sección 11.5 para la hipótesis $\beta_1 = 0$. Aun cuando la prueba puede ser de interés en ocasiones, la mayoría de las investigaciones tendrán como objetivo el estimar la media de y o el predecir un valor particular de y para un valor dado de x.

El coeficiente de correlación r proporciona una medida razonable de qué tan bueno es el ajuste de los datos a la recta de mínimos cuadrados; pero su uso para inferir sobre ρ es dudoso en muchas ocasiones. Parece poco factible que en la física o la economía, el fenómeno observado y resulte ser función de una sola variable. De lo anterior que el coeficiente de correlación entre las ventas mensuales de una compañía y cualquier otra variable resulte ser pequeño y de utilidad cuestionable. Una mayor reducción en la suma de cuadrados de desviación SCE quizás podría lograrse construyendo un predictor para y basado en un conjunto de variables $x_1, x_2, \ldots$ y no solo una.

Otro aspecto en relación a la interpretación de r hay que tener presente, es la frecuencia con que algunos investigadores presumen con orgullo de valores de su coeficiente de correlación muestral en la vecindad de .5 (y en algunos casos hasta del orden de .1), como indicativos de la presencia de una relación entre y y x. Aún cuando estas estimaciones de ρ fuesen muy precisas, solo indicarían una relación débil. Un valor de r de .5 por ejemplo, implica que si x es utilizada para predecir y, la reducción porcentual lograda en la suma de cuadrados por usar x es simplemente $r^2 = .25$ o sea el 25%. Un valor del coeficiente de correlación de .1 indicaría una reducción porcentual en la suma de cuadrados del 1% por usar x como variable explicativa.

Por otro lado, si los coeficientes de correlación lineal entre y y cada una de las variables x_1 y x_2 fuesen .4 y .5 respectivamente, no se sigue de ahí que un predictor que use ambas variables produzca una reducción porcentual en SCE del $(.4)^2 + (.5)^2 = .41$, esto es, del 41%. Si x_1 y x_2 están altamente correlacionadas entre si, puede ser que contribuyan esencialmente lo mismo en el problema de predicción de y.

Finalmente, recuérdese que r es una medida de **correlación lineal** y que y y x pudieran estar perfectamente relacionadas por una función curvilínea con $\rho = 0$.

Ejercicios

11.20. Describa el significado del signo y la magnitud de r.

11.21. ¿Qué se implica si el valor de r es cercano a cero?

11.22. ¿Qué valor adquiere r si todos los puntos caen en una recta de pendiente positiva? ¿Y si la pendiente es negativa?

11.23. Calcule el coeficiente de correlación entre las redituabilidades anuales del fondo Penn Square Mutual y las redituabilidades promedio del mercado del ejercicio 11.10.

11.24. ¿Existe relación entre el consumo de energía de un país y su producto interno bruto (PIB)? Uno estaría dispuesto a suponer que un país con mayor ingreso per capita requeriría de mayor consumo de energía. Para examinar este problema se seleccionaron al azar 12 países y se han obtenido para ellos el consumo per capita (en libras) y el producto interno bruto per capita (en dólares). Los resultados se presentan en la tabla.

PAIS	CONSUMO DE ENERGIA (PER CAPITA)	PRODUCTO INTERNO BRUTO (PER CAPITA)	PAIS	CONSUMO DE ENERGIA (PER CAPITA)	PRODUCTO INTERNO BRUTO (PER CAPITA)
Estados Unidos	25,598	5,515	Canadá	23,715	4.704
Australia	12,568	3,370	Dinamarca	12,273	3,978
Noruega	10.227	3,779	Francia	9,156	3.810
Japón	7,167	2,757	Italia	6,164	2,170
Venezuela	5,452	1,291	Grecia	3,543	1,382
Brasil	1,173	513	India	410	98

Fuente: U.N. Statistical Yearbook, 1973, y *U.S. Statistical Abstract,* 1974.

a. Calcule el coeficiente de correlación r entre el consumo de energía per capita y el producto interno bruto per capita.

b. ¿Presentan los datos suficiente evidencia sobre la existencia de una relación lineal entre el consumo de energía per capita y el PIB per capita? Pruebe la hipótesis de que $\rho = 0$. Use $\alpha = .05$. (Recuerde que la hipótesis $\rho = 0$ es equivalente a la hipótesis $\beta_1 = 0$. Vea la sección 11.5.)

11.25. Refiérase al ejercicio 11.19. Calcule el coeficiente de determinación r^2 y discuta su significado.

11.26. Refiérase al ejercicio 11.5. ¿En qué porcentaje se reduce la suma de los cuadrados de las desviaciones de los costos si se utiliza la ecuación predictora $\hat{y}$ en lugar de $\bar{y}$?

11.27. Una variable independiente que muestre una asociación negativa fuerte con y es tan útil como una que muestre una asociación positiva. El hecho importante es que la relación sea fuerte, esto es, la magnitud absoluta del coeficiente de correlación entre y y x mas no el signo. Considere por ejemplo a las tasas de interés y el número de nuevas construcciones. Las tasas de interés (x) proporcionan un indicador clave para predecir el número de construcciones (y).

Si las tasas bajan, el número de construcciones aumenta y si suben, el número de construcciones disminuye. Suponga que los datos de la tabla representan a las tasas de interés en primeras hipotecas y el registro de nuevas construcciones iniciadas en los 8 años referidos.

Año	1969	1970	1971	1972	1973	1974	1975	1976
Tasa de interés (%)	6.5	6.0	6.5	7.5	8.5	9.5	10.0	9.0
Licencias de construcción	2165	2984	2780	1940	1750	1535	962	1310

a. Encuentre la recta de mínimos cuadrados para estimar el número de licencias de construcción a partir de las tasas de interés. Grafique los puntos y la recta de mínimos cuadrados como una verificación de sus cálculos.
b. Calcule el coeficiente de correlación para estos datos. ¿Es el coeficiente de correlación entre el número de licencias de construcción y la tasa de interés significativamente distinto de cero? Use $\alpha = .05$.
c. ¿En qué porcentaje se reduce la suma de cuadrados de las desviaciones del número de licencias al utilizar como predictor a la tasa de interés en lugar del simple promedio del número de licencias anuales?
d. Si los indicadores económicos indican que la tasa de interés para primeras hipotecas será del 8.5% el próximo año, pronostique el número de licencias de construcción que se otorgarán durante el año entrante por medio de un intervalo de predicción del 95%.

11.9 La aditividad de sumas de cuadrados

Una propiedad interesante del análisis de regresión es que particiona la suma de cuadrados totales

$$SC_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

en dos partes. Una de ellas es atribuible a la suma de cuadrados de las desviaciones de los valores de y respecto a la recta ajustada y que se denota SCE. La otra parte representa a la reducción (absoluta, no porcentual) en la suma de cuadrados de desviaciones que resulta de incorporar la información con la que contribuye la variable auxiliar x.

Para entender mejor la partición, note que el análisis tiende a particionar a la desviación misma de cada observación respecto a su media (muestral), $(y_i - \bar{y})$, en dos partes, así

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

La partición de $(y_i - \bar{y})$ puede verse en la figura 11.11.

Al tomar la suma de las observaciones al cuadrado, sobre todas las observaciones, puede mostrarse llevando a cabo las manipulaciones algebraicas que

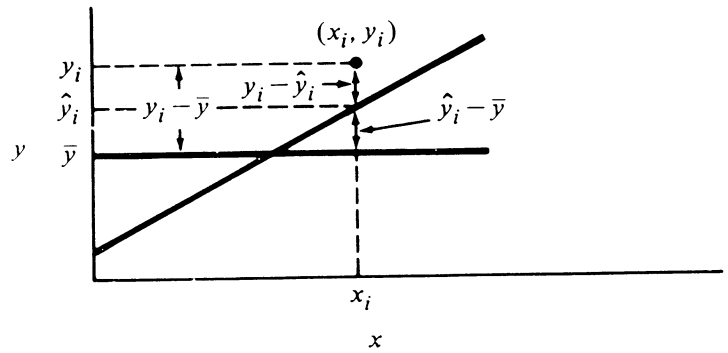

Figura 11.11 *Partición de* $(y_i - \bar{y})$ *en* $(y_i - \hat{y}_i)$ y $(\hat{y}_i - \bar{y})$

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

así, la suma de cuadrados de y, denotada SC_y (también llamada suma de cuadrados total, SC total) se particiona en dos componentes.

Partición de la suma de cuadrados de y

$$\text{SC total} = \text{SCR} + \text{SCE}$$

en donde

$\text{SCR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 =$ suma de cuadrados debida a la regresión, que es la cantidad de variación total explicada por la variable auxiliar x, y

$\text{SCE} = \sum_{i=1}^{n} (y_i - \hat{y}_1)^2 =$ suma de cuadrados del error, que es la cantidad de variación total que no pudo ser explicada por x.

La partición de la suma de cuadrados es importante por dos razones. Proporciona una relación aditiva de importancia entre las sumas de cuadrados

$$\text{SC total} = \text{SCR} + \text{SCE}$$

que con frecuencia reduce el esfuerzo computacional en el análisis de regresión (se puede obtener cualquiera de ellas si se conocen las otras dos). La otra razón es que ayuda en la interpretación de la contribución hecha por x para la predicción de y. Finalmente, nótese que

$$r^2 = \frac{SC_y - \text{SCE}}{SC_y} = \frac{\text{SC total} - \text{SCE}}{\text{SC total}} = \frac{\text{SCR}}{\text{SC total}}$$

11.10 Resumen

Aunque no se enfatizó, el problema de predecir el valor de una variable aleatoria y se consideró solo en un contexto muy elemental en los capítulos 8 y 9. Si no se tiene información de variables relacionadas con y, la única información para predecir valores de y es la que proporciona su distribución de probabilidad. En el capítulo 5 se vió que la probabilidad de que y ocurra (caiga) entre dos valores específicos y_1 y y_2 es igual al área bajo la curva de su distribución de probabilidad sobre el intervalo: $y_1 \leq y \leq y_2$. Si se tuviese que elegir al azar un valor de la población (que es el proceso de observar un valor de y), seguramente que para pronosticar, se escogerían algunos de los más probables, como por ejemplo alrededor de μ o alguna otra medida de tendencia central, y la estimación de μ se hizo en los capítulos 8 y 9.

En el capítulo 11, se planteó el problema de predecir y con base en la información auxiliar proporcionada por otras variables $x_1, x_2, \ldots$ que se su-

ponen relacionadas con y y que por ello son útiles en su predicción. La atención se centró en la predicción de y como una función lineal de una sola variable x, que es la extensión inmediata al problema de predicción considerado en los capítulos 8 y 9. El caso más interesante aún es el que considera a y como una función lineal de un conjunto de variables independientes, y ese caso es el tema del capítulo 12.

Ejercicios complementarios

11.28. ¿Qué suposiciones se requieren para poder usar la ecuación de predicción $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ en la predicción del valor de la variable dependiente y a partir de la variable predictora x?

11.29. ¿Para que tipo de configuraciones de los puntos muestrales se tendrá $s^2 = 0$?

11.30. ¿Para qué parámetro resulta ser s^2 un estimador insesgado? Explique en qué forma aparece este parámetro en el modelo probabilístico.

11.31. Para la ecuación lineal $y = 6 + 3x$.

a. Obtenga la ordenada al origen y la pendiente.

b. Grafique la recta correspondiente a la ecuación.

11.32. Repita lo pedido en el ejercicio 11.31 para la ecuación lineal $2x - 3y - 5 = 0$.

11.33. Suponga que le son dados cinco puntos con coordenadas.

y	0	0	1	1	3
x	−2	−1	0	1	2

a. Encuentre la recta de mínimos cuadrados para estos datos.

b. Como una verificación de sus cálculos, grafique los puntos y la recta obtenida.

c. Calcule s^2.

d. ¿Presentan los datos suficiente evidencia de una relación lineal entre y y x? (Pruebe la hipótesis $\beta_1 = 0$ usando $\alpha = .05$.)

11.34. Los siguientes datos del año pasado representan a los días laborables en los que estuvo ausente (y), y la antigüedad en años en la compañía (x) de cada uno de 7 empleados escogidos al azar.

y	2	0	5	6	4	9	2
x	7	8	2	3	5	3	7

a. Encuentre la recta de mínimos cuadrados para estos datos.

b. Como una verificación para sus cálculos de (a), grafique los siete puntos y la recta obtenida.

c. Calcule s^2.

d. ¿Presentan los datos suficiente evidencia de que y y x están linealmente relacionadas? (Pruebe la hipótesis de que $\beta_1 = 0$ usando $\alpha = .05$.)

11.35. Encuentre un intervalo de confianza del 90% para la pendiente de la recta del ejercicio 11.33.

11.36. Encuentre un intervalo de confianza del 95% para la pendiente de la recta del ejercicio 11.34.

11.37. Refiérase al ejercicio 11.34. Encuentre un intervalo de confianza del 95% para la media de los días laborales en los que hay ausentismo de los trabajadores con 5 años de antigüedad. (Esto es, encuentre un intervalo de confianza del 95% para el valor esperado de y cuando $x = 5$.)

11.38. Refiérase al ejercicio 11.34. Encuentre un intervalo de predicción del 95% para el número de días laborables que estará ausente un empleado determinado que se sabe que tiene 5 años de antigüedad. Compare este intervalo con el calculado en el ejercicio 11.37. Explique la diferencia entre los objetivos al hacer inferencias de éste y el ejercicio 11.37.

11.39. Calcule el coeficiente de correlación r para los datos del ejercicio 11.34. ¿Presentan los datos evidencia como para suponer que el coeficiente de correlación poblacional ρ es significativamente distinto de cero? Use $\alpha = .05$.

11.40. ¿En qué porcentaje se reduce la suma de cuadrados de las desviaciones para los días laborables con ausentismo al incorporar la información auxiliar proporcionada por la antigüedad de los empleados en lugar de usar simplemente $\bar{y}$ como predictor en el ejercicio 11.37?

11.41. Se ha observado que para predecir la demanda (consumo) de combustibles para calefacción, resulta ser más preciso el pronóstico a largo plazo de las temperaturas y el uso de la relación temperatura-consumo que el tratar de pronosticar directamente demanda analizando las ventas de combustibles. Un distribuidor de combustibles mantiene un registro de ventas mensuales de combustible y de temperaturas máximas en esos meses. A continuación aparecen los datos de nueve de estos meses seleccionados al azar.