Unidad 5

- Regresión Múltiple

12 Regresión múltiple

12.1 Introducción

La regresión múltiple es una extensión de la metodología vista en el capítulo 11 a más de una variable independiente y sus aplicaciones son interesantes y variadas. Hace algunos años, una revista comentaba la historia de un programador que trabaja en una compañía en los Estados Unidos que, sin perjuicio del buen desempeño de su trabajo, se interesó en la posible relación entre los precios "a futuro" de ciertos productos y algunas variables que le parecieron relacionadas. La revista, en su comentario sobre este programador, contaba que con la ayuda de la computadora de la compañía (y, por supuesto, algún conocimiento de regresión múltiple), el programador desarrolló unas ecuaciones de predicción en varias variables para los precios futuros de algunos de los productos. ¿Qué tan precisas fueron estas ecuaciones? El comentario hecho en la revista sobre este programador concluía dando la información de que había ganado más de un millón de dólares en pocos años y que ahora se dedicaba a dar consultoría sobre inversiones.

No todos los esfuerzos para construir ecuaciones de predicción en varias variables, para la respuesta bajo estudio llevan a soluciones tan buenas o espectaculares como la descrita en la revista referida. Sin embargo hay un número suficiente de aplicaciones exitosas que llevan a concluir que el análisis de regresión múltiple es una herramienta estadística muy poderosa que se utiliza en una gran variedad de áreas dentro de los negocios.

Como ilustración, piense en su área de especialización dentro de los negocios (o su futura área de especialización) y en alguna variable y que mida el éxito dentro de esa área de especialización. Por ejemplo si se dedicara a estudios del mercado, podría pensar en el volumen de ventas como una medida del éxito en esa área. En un negocio pequeño seguramente se usaría como medida del éxito la ganancia, y el jefe de seguridad de un gran almacén usaría como criterio de éxito, el valor de la mercancía robada.

Suponga que tiene una ecuación de predicción en varias variables (x_1, x_2, x_3, . . .) que le permite pronosticar con bastante precisión los valores de y a partir de los valores de las x_i. Piense en el beneficio que una tal herramienta podría proporcionar. Se podrían pronosticar valores para la variable que mide el éxito y a partir de varios juegos de valores de las variables x_i y así, al observar los valores que tomarían las variables x_i desarrollaría un mejor conocimiento sobre la manera de controlar finalmente los valores de y para que tome aquellos que sean más ventajosos.

El encontrar una ecuación de predicción en varias variables es el tema de este capítulo. El empleo de esta metodología—procedimientos de estimación y pruebas estadísticas—en un conjunto de datos es referido usualmente como **análisis de regresión múltiple.**

12.2 Modelos estadísticos lineales

Una ecuación de predicción en varias variables (o modelo de predicción) es una extensión de la ecuación del modelo lineal simple del capítulo 11. Algunos modelos típicos para variables de respuesta son:

$$(1) \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

$$(2) \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$$

$$(3) \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon$$

en donde x_1, x_2, y x_3 son variables predictoras y ϵ es un error aleatorio. En todos los modelos que se discutan en este capítulo, se supondrá que el error aleatorio ϵ se distribuye normalmente con media 0 y varianza σ^2. Más aún, se supondrá que los errores aleatorios asociados a cualquier par de valores para y, son independientes en el sentido probabilístico. (Estas mismas suposiciones sobre las propiedades de ϵ se hicieron al ver el modelo lineal simple en el capítulo 11.)

Como el valor esperado de ϵ es cero, se sigue que el valor esperado de y para valores específicos de las variables predictoras está dado por la parte determinística del modelo de predicción. De ahí que el valor esperado (media) de y para cada uno de los tres modelos anteriores sea*

$$(1) \quad E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$(2) \quad E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

$$(3) \quad E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

*Cuando el valor esperado de y es función de las variables de predicción x_1, x_2, y x_3 es usual denotarlo por $E(y/x_1, x_2, x_3)$. Esto es $E(y/x_1, x_2, x_3)$ representa el valor esperado de y dados los valores de x_1, x_2, y x_3. Como se verán diversos modelos con distintas variables predictoras, el símbolo para denotar el valor esperado de y se vuelve complicado, de ahí que, como simplificación, en este capítulo se utilizará solamente $E(y)$.

Los modelos (1), (2) y (3) se llaman **modelos estadísticos lineales** ya que en sus respectivas expresiones, el lado derecho de la igualdad que define a $E(y)$ resulta ser una función lineal en los parámetros β.

En contraste a los anteriores, el modelo

$$(4) \quad y = \beta_0 x^{\beta_1} + \epsilon$$

no es un modelo lineal porque el lado derecho de la ecuación de predicción no es una función lineal de los parámetros desconocidos β_0 y β_1. El análisis de regresión múltiple se basa en el supuesto de que y está representado por un modelo estadístico lineal. Sólo se verán modelos lineales en este capítulo.

12.3 Mínimos cuadrados para un modelo de predicción en varias variables

Una ecuación de predicción basada en las variables $x_1, x_2, \ldots, x_k$ se puede obtener por el método de mínimos cuadrados exactamente del mismo modo en que éste fue empleado para el modelo lineal simple. Por ejemplo, suponga que se desea ajustar el modelo

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

en donde y es el precio de las acciones de una empresa al final del mes, x_1 es el rendimiento de cada acción durante el pasado ejercicio fiscal, x_2 es el volumen de ventas de la empresa en el mes anterior, x_3 es la ganancia neta de la empresa también en el mes anterior y ϵ es un error aleatorio. (Nótese que se podrían adicionar otras variables o los cuadrados, cubos y productos cruzados de x_1, x_2, x_3.)

Para lo anterior se necesita una muestra aleatoria de los valores de y, x_1, x_2, x_3 registrados durante n meses seleccionados al azar de entre aquellos en los que la compañía ha estado en operación. El conjunto de registros o mediciones y, x_1, x_2, x_3 para cada uno de los n meses puede pensarse como las coordenadas de un punto en un espacio de cuatro dimensiones. Entonces, desde el punto de vista ideal, sería deseable tener una "regla" multidimensional que pudiéramos "mover" entre los n puntos hasta que las desviaciones entre los valores observados de y y los ajustados (esto es los de la regla multidimensional) fuesen en algún sentido mínimas. Aun cuando no se pueden graficar puntos en cuatro dimensiones, puede observar que un tal procedimiento es el que proporciona el método de mínimos cuadrados, que lo hace matemáticamente.

La suma de los cuadrados de las desviaciones entre los valores observados de y y los respectivos valores ajustados es

$$\text{SCE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i})]^2$$

en donde $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$ es el modelo ajustado y $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, y $\hat{\beta}_3$ son los estimadores de los parámetros del modelo. Se usa el cálculo para encontrar los estimadores $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, y $\hat{\beta}_3$ que hacen que SCE tome el valor mínimo. Los estimadores, al igual que en el caso lineal simple, se obtienen como solución a un sistema de ecuaciones lineales simultáneas conocidas como las **ecuaciones de mínimos cuadrados.** Estas ecuaciones también se conocen como **ecuaciones normales.**

En el caso mencionado de **tres** variables independientes x_1, x_2, x_3, las ecuaciones de mínimos cuadrados son **cuatro** ecuaciones lineales en las incógnitas $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, y $\hat{\beta}_3$. Las cuatro ecuaciones de mínimos cuadrados que no se han derivado aquí y que simplemente se establecen son

$$\begin{aligned}
\hat{\beta}_0 n &+ \hat{\beta}_1 \sum x_1 &+ \hat{\beta}_2 \sum x_2 &+ \hat{\beta}_3 \sum x_3 &= \sum y \\
\hat{\beta}_0 \sum x_1 &+ \hat{\beta}_1 \sum x_1^2 &+ \hat{\beta}_2 \sum x_1 x_2 &+ \beta_3 \sum x_1 x_3 &= \sum x_1 y \\
\hat{\beta}_0 \sum x_2 &+ \hat{\beta}_1 \sum x_1 x_2 &+ \hat{\beta}_2 \sum x_2^2 &+ \hat{\beta}_3 \sum x_2 x_3 &= \sum x_2 y \\
\hat{\beta}_0 \sum x_3 &+ \hat{\beta}_1 \sum x_1 x_3 &+ \hat{\beta}_2 \sum x_2 x_3 &+ \hat{\beta}_3 \sum x_3^2 &= \sum x_3 y
\end{aligned}$$

en donde cada sumatoria indica la cantidad que debe sumarse sobre todos los datos $i = 1, 2, \ldots, n$. Esto es,

$$\sum x_1 = x_{11} + x_{12} + \cdots + x_{1n}, \text{ etc.}$$

Observe el patrón que forman los distintos términos en las ecuaciones de mínimos cuadrados y habrá descubierto un resultado: que ese patrón es el mismo para cualquier número de variables independientes. Para el modelo de regresión

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

con **dos** variables independientes, las **tres** ecuaciones de mínimos cuadrados en las tres incógnitas $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$ son

$$\begin{aligned}
\hat{\beta}_0 n &+ \hat{\beta}_1 \sum x_1 &+ \hat{\beta}_2 \sum x_2 &= \sum y \\
\hat{\beta}_0 \sum x_1 &+ \hat{\beta}_1 \sum x_1^2 &+ \hat{\beta}_2 \sum x_1 x_2 &= \sum x_1 y \\
\hat{\beta}_0 \sum x_2 &+ \hat{\beta}_1 \sum x_1 x_2 &+ \hat{\beta}_2 \sum x_2^2 &= \sum x_2 y
\end{aligned}$$

Observe el lugar que estos términos ocupan en las ecuaciones del modelo de tres variables. Para el modelo lineal simple

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

con **una** variable independiente, las **dos** ecuaciones en las dos incógnitas $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtienen de las anteriores quitando los términos de x_2, y son

$$\begin{aligned}
\hat{\beta}_0 n &+ \hat{\beta}_1 \sum x &= \sum y \\
\hat{\beta}_0 \sum x &+ \hat{\beta}_1 \sum x^2 &= \sum xy
\end{aligned}$$

Al resolver estas dos ecuaciones para las incógnitas se obtienen *exactamente* los mismos valores para $\hat{\beta}_0$ y $\hat{\beta}_1$ que los que se obtienen al usar las fórmulas para $\hat{\beta}_0$ y $\hat{\beta}_1$ de la sección 11.3.

Extendiendo los argumentos previos, para un modelo de regresión de k variables independientes, hay $(k+1)$ ecuaciones de mínimos cuadrados en las $(k+1)$ incógnitas $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2, \ldots, \hat{\beta}_k$. La forma de las $(k+1)$ ecuaciones de mínimos cuadrados puede escribirse siguiendo el patrón dado para los casos de una, dos o tres variables independientes.

12.4 Solución a las ecuaciones de mínimos cuadrados

El resolver las ecuaciones de mínimos cuadrados para encontrar $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2, \ldots, \hat{\beta}_k$ garantiza que los estimadores resultantes, al ser sustituidos en la ecuación de predicción

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$$

minimizan la suma de los cuadrados de las desviaciones

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

Esto es, ningún otro juego de estimadores para las betas dá una SCE menor.

Un conjunto de m ecuaciones lineales en m incógnitas se resuelve usualmente por alguno de los siguientes procedimientos:

1. Se expresan las ecuaciones simultáneas en forma matricial y se resuelve el sistema por medio de álgebra de matrices (usando el inverso de una matriz y operaciones como el producto de matrices).
2. Se hace un proceso de eliminación, que permite encontrar el valor de cada incógnita individualmente.

Ambos procedimientos se vuelven tediosos si el número de ecuaciones (el de incógnitas) excede a tres. Existen rutinas preprogramadas para ser usadas en computadora y son éstas las que se deben emplear para resolver estos sistemas de ecuaciones.

Se ilustra la solución de las ecuaciones de mínimos cuadrados con tres incógnitas en el siguiente ejemplo.

Ejemplo 12.1 El dueño de una distribuidora de automóviles piensa que la relación entre el número y de autos nuevos vendidos por él en un mes dado y el número x de anuncios de su distribuidora en el diario local durante ese mes, está dada por el modelo

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

en donde $x_1 = x$ y $x_2 = x^2$. En la tabla que sigue aparecen datos correspondientes a los últimos seis meses.

	MES					
	1	2	3	4	5	6
y	10	10	15	20	30	40
x	0	1	2	2	3	4

Ajuste el modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ a los datos resolviendo las ecuaciones de mínimos cuadrados para así obtener los estimadores de los parámetros desconocidos β_0, β_1 y β_2.

Solución El modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, o equivalentemente el modelo $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, es un ejemplo de un polinomio de segundo grado. Requiere de la solución de tres ecuaciones de mínimos cuadrados con tres incógnitas, $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$. En la tabla 12.1 todas las cantidades requeridas para identificar las ecuaciones se obtienen a partir de los datos originales. Recuerde que para los cálculos de la tabla 12.1 $x_{1i} = x_i$ y $x_{2i} = x_i^2$, por lo que $x_{1i}^2 = x_i^2$, $x_{2i}^2 = x_i^4$, $x_{1i}x_{2i} = x_i^3$, $x_{1i}y_i = x_i y_i$, y $x_{2i}y_i = x_i^2 y_i$.

Tabla 12.1 *Cálculos con los datos del ejemplo 12.1*

MES i	y_i	x_{1i}	x_{2i}	x_{1i}^2	x_{2i}^2	$x_{1i}x_{2i}$	$x_{1i}y_i$	$x_{2i}y_i$
1	10	0	0	0	0	0	0	0
2	10	1	1	1	1	1	10	10
3	15	2	4	4	16	8	30	60
4	20	2	4	4	16	8	40	80
5	30	3	9	9	81	27	90	270
6	40	4	16	16	256	64	160	640
SUMAS	125	12	34	34	370	108	330	1060

Sustituyendo las sumas correspondientes de la tabla 12.1 en las ecuaciones de mínimos cuadrados, se obtiene

$$\text{(I)} \qquad 6\hat{\beta}_0 + 12\hat{\beta}_1 + 34\hat{\beta}_2 = 125$$

$$\text{(II)} \qquad 12\hat{\beta}_0 + 34\hat{\beta}_1 + 108\hat{\beta}_2 = 330$$

$$\text{(III)} \qquad 34\hat{\beta}_0 + 108\hat{\beta}_1 + 370\hat{\beta}_2 = 1060$$

Por conveniencia se han referido estas tres ecuaciones por (I), (II) y (III). Empleando el proceso de eliminación, se encontrará $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$.

Suponga que se resta 2(I) de la ecuación (II). Se tendría al haber hecho esa resta,

$$10\hat{\beta}_1 + 40\hat{\beta}_2 = 80$$

ó

$$\hat{\beta}_1 = 8 - 4\hat{\beta}_2$$

Si se resta $(\frac{34}{6})$(I) de (III), se obtiene

$$40\hat{\beta}_1 + 177.33\hat{\beta}_2 = 351.67$$

ó

$$\hat{\beta}_1 = 8.79 - 4.43\hat{\beta}_2$$

Con lo anterior se ha eliminado $\hat{\beta}_0$ y se tienen ahora dos ecuaciones con dos incógnitas. Igualando estas ecuaciones, se encuentra $\hat{\beta}_2$:

$$8 - 4\hat{\beta}_2 = 8.79 - 4.43\hat{\beta}_2$$

que da

$$.43\hat{\beta}_2 = .79$$

o sea $\hat{\beta}_2 = 1.84$.

De una de las ecuaciones previas, se obtiene $\hat{\beta}_1$:

$$\hat{\beta}_1 = 8 - 4\hat{\beta}_2 = 8 - 4(1.84) = .64$$

Regresando a, digamos, la ecuación (I), se puede reescribir ésta como

$$\hat{\beta}_0 = \tfrac{1}{6}(125 - 12\hat{\beta}_1 - 34\hat{\beta}_2)$$

De lo anterior

$$\hat{\beta}_0 = \tfrac{1}{6}[125 - 12(.64) - 34(1.84)] = \tfrac{1}{6}(54.76) = 9.13$$

La ecuación para predecir las ventas mensuales y a partir del número x de anuncios en el diario durante el mes en cuestión y su cuadrado $x_2 = x^2$ es

$$\hat{y} = 9.13 + .64x_1 + 1.84x_2$$

Los errores en la predicción (usualmente llamados **residuales**) se pueden obtener sustituyendo las variables de predicción x_1 y x_2 en la ecuación de predicción y evaluando $\hat{y}$, las ventas estimadas. Para el mes 1, se hubiera estimado

$$\hat{y}_1 = 9.13 + .64(0) + 1.84(0) = 9.13$$

De ahí que el error de predicción en el mes 1 sea

$$e_1 = y_1 - \hat{y}_1 = 10 - 9.13 = .87$$

Las ventas estimadas $\hat{y}_i$ y los errores de predicción para los 6 meses que sirvieron de base para el ajuste, se encuentran en la tabla que sigue.

y_i	10	10	15	20	30	40
$\hat{y}_i$	9.13	11.61	17.77	17.77	27.61	41.13
e_i	.87	−1.61	−2.77	2.23	2.39	−1.13

La suma de cuadrados de los errores es

$$\text{SCE} = \sum_{i=1}^{6} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{6} e_i^2 = 22.9838$$

Es una garantía que ningunos otros valores de los parámetros desconocidos β_0, β_1 y β_2 hubieran producido una SCE menor a 22.9838.

La gráfica de la ecuación

$$\hat{y} = 9.13 + .64x_1 + 1.84x_2 \qquad (x_1 = x,\ x_2 = x^2)$$

se muestra en la figura 12.1. Note que el modelo lineal simple

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

habría proporcionado un ajuste muy inferior al que resultó del uso del polinomio.

Para un modelo de predicción con tres o más variables independientes es prácticamente una obligación emplear una computadora electrónica para estimar los parámetros de la regresión $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$. Casi todos los centros de cálculo tienen disponible un "paquete" o programa de biblioteca de análisis de regresión con el cual el usuario sólo necesita ejecutar unas cuantas órdenes propias del paquete y alimentar los datos del problema. Se ilustra una solución obtenida por computadora en el siguiente ejemplo.

Ejemplo 12.2 Considere un estudio diseñado para examinar el papel que juega la televisión en la vida de un grupo preseleccionado de personas de edades superiores a los 65 años. El propósito de dicho estudio es proporcionar información que permita hacer una programación adecuada a las necesidades de este grupo. Una muestra de $n = 25$ personas mayores, de edades superiores a los 65 años fue seleccionada y a cada persona le fue solicitada la siguiente información: y = el número promedio de horas diarias que pasa frente al televisor, x_1 = su estado civil ($x_1 = 1$ si vive con su cónyuge y $x_1 = 0$ si no), x_2 = su edad y x_3 = escolaridad del entrevistado en número de años de asistencia a la escuela.* Los datos aparecen en la tabla 12.2.

El objetivo de este estudio es el relacionar y, el número promedio de horas diarias que pasa un entrevistado frente al televisor, con las variables descriptivas x_1, x_2, y x_3. Por simplicidad suponga que se escoge el modelo de predicción

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

*La variable x_1 es un ejemplo de una **variable indicadora,** que se usa con frecuencia para incluir el efecto de un factor *cualitativo* en un modelo de regresión. Las variables indicadoras sirven para particionar un modelo de regresión en varias (dos en este caso) componentes

$$\hat{y} = (\hat{\beta}_0 + \hat{\beta}_1) + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

El primero de los modelos corresponde al de $x_1 = 1$, esto es, en el que el factor está presente y el otro modelo en el que está ausente.

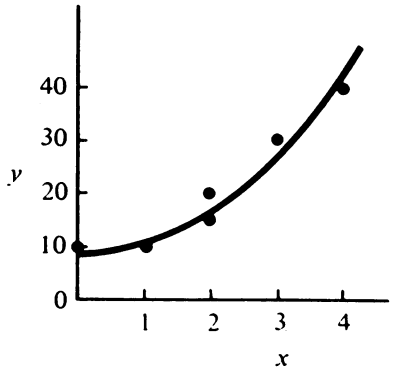


Figura 12.1 *Gráfica de los datos y el modelo de predicción del ejemplo 12.1*

Tabla 12.2 *Horas diarias pasadas frente al televisor, estado civil, edad y años de escolaridad de 25 personas mayores escogidas al azar. Ejemplo 12.2*

INDIVIDUO	HORAS y	ESTADO CIVIL x_1	EDAD x_2	ESCOLARIDAD x_3
1	.5	1	73	14
2	.5	1	66	16
3	.7	0	65	15
4	.8	0	65	16
5	.8	1	68	9
6	.9	1	69	10
7	1.1	1	82	12
8	1.6	1	83	12
9	1.6	1	81	12
10	2.0	0	72	10
11	2.5	1	69	8
12	2.8	0	71	16
13	2.8	0	71	12
14	3.0	0	80	9
15	3.0	0	73	6
16	3.0	0	75	6
17	3.2	0	76	10
18	3.2	0	78	6
19	3.3	1	79	6
20	3.3	0	79	4
21	3.4	1	78	6
22	3.5	0	76	9
23	3.6	0	65	12
24	3.7	0	72	12
25	3.7	0	80	6

Encuentre la ecuación de predicción de mínimos cuadrados para los datos de la tabla 12.2.

Solución El ajustar la ecuación de predicción a los datos de la tabla 12.2 por medio de computadora elimina la necesidad de plantear y resolver las

ecuaciones de mínimos cuadrados. Para que se convenza del ahorro de trabajo al usar una computadora, se muestran las ecuaciones de mínimos cuadrados:

$$\begin{aligned} 25\hat{\beta}_0 + 10\hat{\beta}_1 + 1{,}846\hat{\beta}_2 + 254\hat{\beta}_3 &= 58.5 \\ 10\hat{\beta}_0 + 10\hat{\beta}_1 + 748\hat{\beta}_2 + 105\hat{\beta}_3 &= 16.2 \\ 1{,}846\hat{\beta}_0 + 748\hat{\beta}_1 + 137{,}086\hat{\beta}_2 + 18{,}509\hat{\beta}_3 &= 4{,}376.0 \\ 254\hat{\beta}_0 + 105\hat{\beta}_1 + 18{,}509\hat{\beta}_2 + 2{,}892\hat{\beta}_3 &= 533.4 \end{aligned}$$

(Puede verificar que los coeficientes en las ecuaciones son correctos, calculando las sumas, sumas de cuadrados y sumas de productos cruzados dados en las ecuaciones de mínimos cuadrados para tres variables independientes de la sección 12.3.) Como podrá ver, el hacer los cálculos para encontrar las ecuaciones y el resolverlas después, es una tarea no sólo tediosa sino tardada. Sin embargo, esta tarea la hace rápidamente una computadora.

La tabla 12.3 reproduce un listado de computadora obtenido de un programa de análisis de regresión, de los comúnmente utilizados, aplicado a los datos de la tabla 12.2. En este ejemplo interesan solamente los estimadores de β_0, β_1, β_2 y β_3 que se han marcado en la tabla 12.3. La otra porción no marcada de la tabla se explicará en las secciones 12.5 y 12.7

Tabla 12.3 *Listado de computadora para los datos de la tabla 12.2 del estudio sobre las horas frente al televisor de personas mayores.*

```
R MULTIPLE              .7918
R CUADRADA              .6269
DESV. EST. ESTIM.       .7524

ANALISIS DE VARIANZA

                    SUMA DE    CUADRADO
             GL  CUADRADOS      MEDIO     COCIENTE F
REGRESION     3    19.972        6.657       11.760
RESIDUAL     21    11.888         .566
```

La porción marcada de la tabla 12.3, titulada ANALISIS INDIVIDUAL DE LAS VARIABLES, tiene cuatro columnas. La segunda, bajo el encabezado COEFICIENTE, contiene a los estimadores de β_0, β_1, β_2 y β_3 en orden de arriba hacia abajo. La columna 1 con encabezado VARIABLE, proporciona al programador la identificación de la variable y su correspondiente parámetro asociado. Así,

$$\hat{\beta}_0 = 1.41411 \qquad \hat{\beta}_1 = -1.17396 \qquad \hat{\beta}_2 = .03971 \qquad \hat{\beta}_3 = -.15106$$

y se sigue que la ecuación de predicción es

$$\hat{y} = 1.41411 - 1.17396x_1 + .03971x_2 - .15106x_3$$

Para este modelo particular, β_1, β_2, y β_3 representan el cambio en el valor esperado de y, $E(y)$, por un cambio unitario en x_1, x_2 y x_3 respectivamente. Por ejemplo, $\hat{\beta}_2 = 0.03971$ es el cambio medio estimado en el tiempo que se pasa diario frente al televisor si la edad x_2 del entrevistado aumenta un año. El coeficiente β_1 de la variable indicadora x_1 representa la diferencia en tiempos medios pasados frente al televisor entre entrevistados que viven con su cónyuge y aquellos que viven solos. El estimador de β_1 es -1.17396 horas. Esto es, se estima que los entrevistados que viven solos ven en promedio 1.17396 horas más al día que los que viven con su cónyuge.

12.5 Intervalos de confianza y pruebas de hipótesis para los parámetros

La parte marcada de la tabla 12.3 también proporciona la información necesaria para construir intervalos de confianza para los parámetros β del modelo lineal del ejemplo 12.2 y para pruebas de hipótesis concernientes a estos parámetros. Las desviaciones estándar estimadas $s_{\hat{\beta}_1}$, $s_{\hat{\beta}_2}$ y $s_{\hat{\beta}_3}$, de los estimadores $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\beta}_3$ de los coeficientes de regresión se dan en la columna con encabezado DESV. ESTANDAR. Estas cantidades permiten construir intervalos de confianza para los coeficientes de regresión, los parámetros β_1, β_2, β_3.

El procedimiento para construir intervalos de confianza para los parámetros β es idéntico al empleado en la sección 11.5 para el modelo lineal simple con excepción de que las fórmulas para $s_{\hat{\beta}_1}$, $s_{\hat{\beta}_2}$ y $s_{\hat{\beta}_3}$, son mucho más complejas que la correspondiente de la sección 11.5. La fórmula para el intervalo de confianza del $(1 - \alpha)100\%$ para el parámetro, digamos β_i, se da en el cuadro.

Intervalo de confianza del $(1 - \alpha)100\%$ para β_i

$$\hat{\beta}_i \pm t_{\alpha/2} s_{\hat{\beta}_i}$$

Los grados de libertad (g.l.) de la $t_{\alpha/2}$ son n, el número de datos, menos un grado de libertad por cada parámetro β del modelo.

Se ilustra el procedimiento de construcción de intervalos de confianza con un ejemplo.

Ejemplo 12.3 Refiérase al ejercicio 12.2. Encuentre un intervalo de confianza del 95% para β_1, la diferencia media en horas diarias pasadas frente al televisor entre entrevistados que viven con su cónyuge y entrevistados que viven solos.

Solución En la columna de DESV. ESTANDAR, se encuentra que $s_{\hat{\beta}_1} = 0.31445$. El valor de tablas para la $t_{.025}$ basada en (n — número de

parámetros β en el modelo) $= 25 - 4 = 21$ grados de libertad se obtiene de la tabla 4 del apéndice; se encuentra $t_{.025} = 2.080$. De lo anterior que el intervalo de confianza del 95% para β_1 es

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1} = -1.17396 \pm (2.080)(.31445)$$

o sea, entre -1.82802 y -0.5199. Ya que $x_1 = 1$ para los entrevistados que viven con su cónyuge y $x_1 = 0$ para los que viven solos, se estima entonces que los que viven solos ven televisión en promedio entre .52 y 1.83 horas diarias más que aquellos que viven con su cónyuge.

Una prueba de una hipótesis de que un parámetro particular, por ejemplo β_i, es igual a cero, se puede hacer por medio de una estadística t:

$$t = \frac{\hat{\beta}_i - 0}{s_{\hat{\beta}_i}} = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

El procedimiento es idéntico al empleado para probar una hipótesis acerca de la pendiente β_1 en un modelo lineal simple (sección 11.5) con la única excepción de que la fórmula para calcular $s_{\hat{\beta}_i}$ en el análisis de regresión múltiple es mucho más complicada.

La prueba anterior también se puede hacer por medio de una estadística F (introducida en la sección 9.7). Se puede usar una estadística F porque el cuadrado de una estadística t (con ν grados de libertad) es igual a una F con 1 grado de libertad en el numerador y ν grados de libertad en el denominador. Esto es

$$t^2_{\nu \text{g.l.}} = F_{1,\nu}$$

Algunos programas de computadora para regresión múltiple utilizan una estadística t para probar la hipótesis nula

$$H_0: \beta_i = 0$$

en contra de la hipótesis alternativa

$$H_a: \beta_i \neq 0$$

Sin embargo, el programa cuyo listado aparece en la tabla 12.3 utiliza la estadística F. Los valores calculados de la estadística F para probar las hipótesis nulas de que β_1, β_2 y β_3 son cero respectivamente, se muestran en la tabla 12.3 bajo el encabezado VALOR F.

Si la hipótesis nula es cierta, la estadística F resulta ser el cociente de dos estimadores insesgados (llamados *cuadrados medios*) de σ^2, la varianza del error aleatorio que aparece en el modelo lineal. De ahí

$$F = \frac{\text{cuadrado medio del numerador}}{\text{cuadrado medio del denominador}}$$

El cuadrado medio del denominador siempre es s^2, el estimador de σ^2. El cuadrado medio del numerador depende del parámetro, digamos β_i, bajo prueba. Si β_i es cero, el valor esperado del cuadrado medio del numerador es

σ^2. Si β_i es distinto de cero, el valor esperado del cuadrado medio del numerador es más grande que σ^2 y F será mayor que lo que sería si H_0 fuese cierta. De lo anterior que se rechace la hipótesis nula $H_0: \beta_i = 0$ para valores grandes de F. La región de rechazo para la prueba de F se encuentra en la cola superior de la distribución F como se muestra en la figura 12.2.

El valor crítico de la estadística F, F_α, depende de dos cantidades, ν_1 y ν_2 en donde

ν_1 = número de grados de libertad asociados al cuadrado medio del numerador

ν_2 = número de grados de libertad asociados a s^2

Para probar la hipótesis nula $\beta_1 = 0$, los grados de libertad ν_1 y ν_2 son

$\nu_1 = 1$

ν_2 = número de datos menos un grado de libertad por cada parámetro β que aparece en el modelo

Los valores críticos de F *para varias combinaciones de* ν_1 y ν_2 se encuentran tabulados en las tablas 6 y 7 del apéndice. La tabla 6 dá valores de $F_{.05}$, la tabla 7 dá valores de $F_{.01}$. (Para mayor información sobre estas tablas vea la sección 9.7.)

Ejemplo 12.4 Refiérase al ejemplo 12.2. Pruebe la hipótesis nula de que β_1 (la diferencia media entre el número de horas diarias que pasan frente al televisor los entrevistados que viven con su cónyuge y el número de horas diarias que pasan frente al televisor los entrevistados que viven solos) es igual a cero. Pruebe con un nivel de significancia $\alpha = .05$.

Solución El valor crítico para la prueba F basada en

$\nu_1 = 1$

$\nu_2 = n$ − el número de parámetros β en este modelo

$= 25 - 4 = 21$ grados de libertad

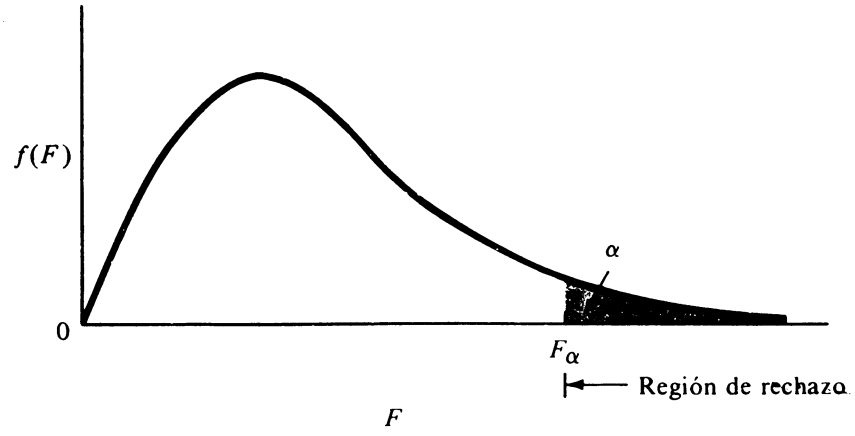


Figura 12.2 *Localización de la región de rechazo para la prueba F*

es $F_{.05} = 4.32$ que se obtiene de la tabla 6 con $\nu_1 = 1$ y $\nu_2 = 21$.

El valor calculado para la prueba F de la tabla 12.3 que corresponde a β_1 es 13.9380. Como este valor de F excede al valor crítico $F_{.05} = 4.32$, se rechaza la hipótesis nula. De lo anterior se dice que los datos proporcionan evidencia de una diferencia entre la cantidad de tiempo que ven televisión los entrevistados que viven con su cónyuge y la cantidad de tiempo que ven televisión aquellos que viven solos.

Ejemplo 12.5 Para mostrar la equivalencia entre las pruebas F y t para la hipótesis nula $H_0: \beta_1 = 0$, haga la prueba del ejemplo 12.4 pero usando la estadística t.

Solución Como se hizo notar anteriormente, se puede probar la hipótesis nula

$$H_0: \beta_i = 0$$

contra la hipótesis alternativa

$$H_a: \beta_i \neq 0$$

usando

$$t = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$$

Como la hipótesis alternativa implica una prueba de dos colas, se rechazaría la hipótesis nula si $t > t_{.025}$ o si $t < -t_{.025}$ en donde $t_{.025}$ (al igual que el denominador de la F) se basa en (n − el número de parámetros β del modelo) grados de libertad. El número de grados de libertad para la estadística t es 21 y de la tabla 4 del apéndice, $t_{.025} = 2.080$.

Para encontrar el valor de la estadística t, se requiere primero obtener $\hat{\beta}_1$ y $s_{\hat{\beta}_1}$ del listado de computadora de la tabla 12.3. Estos valores son

$$\hat{\beta}_1 = -1.17396 \qquad s_{\hat{\beta}_1} = .31445$$

De lo anterior que el valor observado de la estadística t sea

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \frac{-1.17396}{.31445} = -3.733$$

Como este valor es menor que el valor crítico $-t_{.025} = -2.080$, se rechaza la hipótesis nula de que $\beta_1 = 0$ (la misma conclusión a la que se llegó en el ejemplo 12.4).

Nótese ahora la equivalencia entre las pruebas F y t. Los valores críticos para estas pruebas son

$$F_{.05} = 4.32 \qquad \text{y} \qquad t_{.025} = 2.080$$

Al elevar al cuadrado $t_{.025}$, se obtiene $t^2_{.025} = (2.080)^2 = 4.326 \approx F_{.05}$. (La pequeña discrepancia se debe a que los valores de tablas para la F y la t aparecen redondeados hasta su tercera cifra decimal.) Como se lo habrá imaginado, la misma equivalencia existe para los valores calculados de F y t. Los valores para la F (extraídos de la tabla 12.3) y t (calculado) son

$$F = 13.9380 \quad \text{y} \quad t = -3.733$$

Elevando al cuadrado, $t^2 = (-3.733)^2 = 13.935$. Otra vez, observe que $F \approx t^2$.

Nótese que la prueba de F se puede utilizar para probar $H_0: \beta_i = 0$ sólo cuando la alternativa es $H_a: \beta_i \neq 0$. En contraste a esto, la prueba de t puede usarse para probar H_0 contra la alternativa (de un solo lado) $H_a: \beta_i > 0$ (ó $H_a: \beta_i < 0$). Por ejemplo, si se hubiera estado dispuesto a sostener la hipótesis alternativa de que los entrevistados que viven solos ven más televisión que aquellos que viven con su cónyuge, se hubiera utilizado la alternativa de un solo lado $H_a: \beta_1 < 0$. En ese caso, para probar $H_0: \beta_1 = 0$ contra la alternativa $H_a: \beta_1 < 0$, se hubiera rechazado H_0 sólo si $t < -t_\alpha$ (note que es t_α y no $t_{\alpha/2}$). La estadística F no hubiera sido apropiada para esta prueba con alternativa de un solo lado. Por esta razón, los paquetes de computadora que imprimen los valores de la estadística t son más versátiles que aquellos que proporcionan los de la estadística F exclusivamente.

Ejemplo 12.6 Refiérase al listado de computadora de la tabla 12.3 sobre el problema de las horas frente al televisor. Pruebe la hipótesis nula $H_0: \beta_1 = 0$ contra la hipótesis alternativa $H_a: \beta_1 < 0$. Esto es, se quiere ver si los datos proporcionan suficiente evidencia que indique que en promedio los entrevistados "solitarios" ven más televisión que los acompañados. Pruebe con un nivel de significancia $\alpha = .05$.

Solución Como ésta es una prueba con una sola cola, se asigna la totalidad del $\alpha = .05$ a la cola inferior de la distribución. Así, se rechaza $H_0: \beta_1 = 0$ contra $H_a: \beta_1 < 0$ si $t < -t_{.05}$. El valor crítico para t basado en 21 grados de libertad es

$$-t_{.05} = -1.721$$

Ahora se compara el valor calculado para t, $t = -3.733$ (que se calculó en el ejemplo 12.5) con el valor crítico de t. Como el valor calculado es menor que $-t_{.05} = -1.721$, se rechaza la hipótesis nula. Se concluye que "parece ser" que los "solitarios" entrevistados ven más televisión que los que viven acompañados.

Antes de concluir esta sección hay que notar un punto importante. Al ajustar un modelo a un conjunto de observaciones, el número de éstas (datos) debe ser mayor que el número de parámetros del modelo para poder tener un número suficiente de grados de libertad asociados a s^2 (y en consecuencia a las estadísticas t y F). ¿Cuántos grados de libertad? Mientras más se tengan, mejor. Puede observarse de los valores tabulados para t y F, que estos son grandes si el número de grados de libertad es pequeño. Ciertamente será bueno contar con por lo menos 5 grados de libertad para la estimación de σ^2 y será preferible el contar con muchos más (10, 20 o más).

Si los datos originales se utilizaron para construir un modelo (encontrar un modelo que tenga un buen ajuste a los datos), se deberá contar con algunos

datos en dos partes, una parte la usan para construir el modelo y la otra para probar el modelo.
y luego se formula una teoría (se recojen datos y se encuentra el modelo que se ajusta mejor a esos datos). Después se prueba (confronta) la teoría observando nuevamente la naturaleza (se prueba el modelo ajustado con un nuevo conjunto de datos). Para hacer lo anterior, muchos investigadores dividen sus datos en dos partes, una parte la usan para construir el modelo y la otra para probar el modelo.

12.6 El problema de estimadores correlacionados: multicolinealidad

Si se ajusta un modelo de regresión múltiple

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

se debe tener mucho cuidado al interpretar los resultados de las pruebas de t sobre los parámetros β. Una de las razones para esta precaución es que los estimadores pueden estar correlacionados. Por ejemplo si uno de los parámetros β,se sobre-estima, a lo mejor eso hace que se tienda a subestimar otro de los parámetros. A este fenómeno se le llama **multicolinealidad.** Esto es, parte de la información que aportan dos o más de las variables independientes para predecir y puede ser distinta, pero parte puede ser idéntica.

Para ilustrar, suponga que se desea costruir un modelo para predecir el precio de un automóvil como función de una serie de variables independientes $x_1, x_2, \ldots, x_k$, dos de las cuales son

x_1 = peso del automóvil

x_2 = potencia del motor.

En general, es de esperarse que los automóviles más pesados y aquellos con motores más grandes (mayor potencia) cuesten más. Esto es, las dos variables x_1 y x_2, proporcionan información para la predicción del precio pero parte de esta información (no toda) es la misma ya que el peso y la potencia están correlacionados. Los autos pesados requieren de motores más potentes.

Cuando dos o más de las variables independientes están correlacionadas, no se puede determinar la contribución individual en la reducción de la SCE, la suma de cuadrados de las desviaciones entre los valores observados y ajustados de y. De ahí que la contribución en información hecha por una variable particular para la predicción de y dependa de las otras variables incluídas en el

modelo. Si dos variables contribuyen con información coincidente, la prueba para el parámetro β_1 para x_1 podría indicar significancia estadística (rechazo de la hipótesis nula $\beta_1 = 0$), mientras que una prueba del parámetro β_2 para x_2 podría indicar no-significancia. De hecho, la segunda variable podría estar inclusive causalmente relacionada con la primera y aun así la prueba de t (o F) no lleva a un rechazo de la hipótesis $\beta_2 = 0$ (que sería lógico si $\beta_1 = 0$ se rechazara). Cuando existe multicolinealidad, lo que importa es el modelo completo y ya no los parámetros β individuales.

Se puede probar si el modelo (en su totalidad) contribuye con información para la predicción de y, probando la hipótesis

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Se mostrará como hacer esta prueba en la sección 12.8. También se mostrará como medir la bondad con que se ajusta el modelo a los datos en la sección 12.7.

Un problema adicional aparece en ocasiones cuando el modelo de regresión múltiple se basa en variables medidas en el tiempo, esto es que utiliza datos de **series de tiempo,** como lo pueden ser las aplicaciones para pronósticos de ventas, análisis de demanda y en general estudios econométricos. Cuando con datos de series de tiempo se omite una o más de las variables independientes importantes en el modelo de regresión, los residuales dependen con frecuencia entre sí (se dicen **autocorrelacionados** o **correlacionadas serialmente**). Por ejemplo, suponga que el número de licencias para construcción para cada uno de los 60 meses anteriores se modela a través de una regresión considerando como única variable independiente a la tasa bancaria de interés. El omitir el tamaño de la población como otra variable independiente puede producir una correlación serial si el tamaño de la población está correlacionado con la tasa bancaria de interés.

La correlación afecta la precisión pero no la exactitud en la estimación de los parámetros β en un modelo de regresión múltiple. Los estimadores siguen siendo insesgados pero sus varianzas aparecen subestimadas cuando hay correlación serial.

Como resultado, la SCE puede subestimar por mucho la variación real no explicada por el modelo, produciendo valores de t y F mayores de los que debieran ser. Esto podría llevar a concluir que ciertos parámetros β son significativos cuando de hecho no lo son. Así el efecto de la correlación serial es el contrario al de la multicolinealidad.

La estadística más usada para probar la presencia de correlación serial es la de Durbin-Watson, que se describe en el libro de Neter y Wasserman en las páginas 358–360 (vea las referencias). Una prueba más sencilla que no tiene algunas de las limitaciones de la de Durbin-Watson es la prueba de rachas aplicada a los residuales. Esta prueba se discutirá en la sección 18.6 del texto.

Para resumir, sea cauteloso al interpretar las pruebas de t (o F) sobre los parámetros individuales β que aparecen en el modelo.

Ejercicios

12.1. El dueño de una distribuidora de automóviles realizó un estudio para determinar las relaciones en un mes determinado entre

y = número de automóviles vendidos en el mes por su distribuidora

x_1 = número de comerciales de 1 minuto sobre su distribuidora, televisados localmente en ese mes.

x_2 = número de anuncios sobre su distribuidora de página entera aparecidos en el diario ese mes.

	MES					
	1	2	3	4	5	6
y	10	10	20	30	40	40
x_1	0	1	2	2	3	4
x_2	1	0	2	3	3	3

Durante un período de 6 meses, el dueño anotó los resultados que se muestran en la tabla. Utilice el proceso de eliminación para ajustar el modelo $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ a los datos, resolviendo las ecuaciones de mínimos cuadrados para las incógnitas $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\beta}_2$.

Para resolver los ejercicios 12.2 hasta el 12.6, utilice el listado de computadora de la tabla 12.3.

12.2. Pruebe la hipótesis nula de que el incremento (o decremento) medio, en horas diarias pasado frente al televisor, al aumentar un año la edad de los entrevistados, es cero. Esto es, pruebe $H_0: \beta_2 = 0$ contra la alternativa $\beta_2 \neq 0$. Pruebe con un nivel de significancia $\alpha = .05$. Use una prueba de F.

12.3. Repita el ejercicio 12.2 pero ahora use una prueba de t. Muestre que el cuadrado del valor calculado de la t es igual al valor calculado de la F del ejercicio 12.2.

12.4. Encuentre un intervalo de confianza del 95% para el aumento medio, en horas diarias pasado frente al televisor, al aumentar un año la edad de los entrevistados.

12.5. Encuentre un intervalo de confianza del 95% para el aumento medio, en horas diarias pasado frente al televisor, al aumentar en un año la escolaridad de los entrevistados (en otras palabras encuentre un intervalo de confianza del 95% para β_3).

12.6. Suponga que tiene una teoría que respalda el hecho de que al aumentar la edad de los entrevistados, el tiempo medio en horas diarias pasado frente al televisor decrece. Pruebe la hipótesis nula $H_0: \beta_3 = 0$ contra la alternativa de un solo lado $H_a: \beta_3 < 0$. Pruebe con un nivel de significancia $\alpha = .05$.

12.7. En Estados Unidos, un urbanizador se interesó en crear un modelo para ser usado en la estimación del precio de venta de terrenos en la costa de Oregon. Para hacerlo registró las siguientes características para cada uno de 20 terrenos vendidos recientemente.

y = valor de venta del terreno (\$, en miles).

x_1 = superficie del terreno (en pies cuadrados)

x_2 = elevación del terreno (sobre el nivel del mar).

x_3 = inclinación del terreno (pendiente).

El urbanizador empleó un paquete de regresión múltiple en una computadora y obtuvo el listado siguiente.

```
R MULTIPLE              .8854
R CUADRADA              .7838
DESV. EST. ESTIM.       .6075

ANALISIS DE VARIANZA

                  SUMA DE    CUADRADO
            GL  CUADRADOS     MEDIO     COCIENTE F
REGRESION    3    21.409      7.136       19.345
RESIDUAL    16     5.903       .369

ANALISIS INDIVIDUAL DE LAS VARIABLES

   VARIABLE      COEFICIENTE   DESV.ESTANDAR   VALOR F
(CONSTANTE        -2.4911)
SUPERFICIE          .099          .058          2.935
ELEVACION           .029          .006         23.327
INCLINACION         .086          .031          7.841
```

a. Dé la ecuación de predicción para el modelo lineal que relaciona el valor de venta con la superficie, la elevación y la inclinación del terreno.
b. ¿Cuáles de las variables predictoras contribuyen con información para la predicción de y? Determine esto usando la prueba estadística apropiada. Use $\alpha = .05$.

12.8. Supónga que antes de haber obtenido la información del ejercicio 12.7, se aceptó la teoría de que los terrenos con mayor inclinación se prefieren a los de menor inclinación. ¿Proporcionan los datos suficiente evidencia como para afirmar que los precios aumentan si la inclinación aumenta? (Prueba $H_0: \beta_3 = 0$ contra la alternativa de un solo lado $H_a: \beta_3 > 0$.) Use $\alpha = .05$.

12.9. Refiérase al ejercicio 12.7. Encuentre un intervalo de confianza del 90% para aquel parámetro de la regresión que relaciona la superficie con el precio de venta del terreno; todo esto en presencia de la elevación y la inclinación. Para valores fijos de la elevación e inclinación proporcione una interpretación del intervalo.

12.7 Determinación de la bondad del ajuste de un modelo

En la sección 11.9 se vió una propiedad muy importante del análisis de regresión, y es que el total de la suma de cuadrados de las desviaciones entre los valores de y respecto a su media, se particiona en dos cantidades,

$$\text{SCE} = \sum_{i=1}^{n} (y_i - \hat{y})^2 = \text{suma de cuadrados del error.}$$

$$\text{SCR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \text{suma de cuadrados debida a la regresión}$$

En otras palabras,

$$\text{SC total} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \text{SCR} + \text{SCE}$$

La SCE, la suma de cuadrados de las desviaciones entre los valores de y y los pronosticados (aquellos que se calculan a partir de la ecuación de predicción), al dividirla por los grados de libertad apropiados, dá s^2, el estimador de σ^2.

Además, se mostró que r^2 mide la proporción de SC total que explica la variable independiente x. De ahí que r^2, que toma valores en el intervalo $0 \leq r^2 \leq 1$, mida la bondad del ajuste de un modelo lineal simple.

En el análisis de regresión múltiple, la SC total

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

se particiona exactamente del mismo modo,

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

y SCR y SCE se definen exactamente de la misma manera que para un modelo lineal simple. La única diferencia aquí, es que y es función de más de una variable predictora.

Suponga que se ajusta el modelo de regresión múltiple

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

a un determinado conjunto de datos. La cantidad

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

es la proporción de la SC total explicada por las variables predictoras x_1, $x_2, \ldots, x_k$. El resto de la SC total es explicada por la posible omisión de variables que contribuirían con información al modelo, por una formulación incorrecta del modelo y por un error experimental. Al igual que r^2, el coeficiente de determinación simple, R^2, el coeficiente de determinación múltiple, toma valores en un intervalo

$$0 \leq R^2 \leq 1$$

Un valor pequeño de R^2 quiere decir que las variables $x_1, x_2, \ldots, x_k$ contribuyen con poca información para la predicción de y; un valor de R^2 cercano a 1 quiere decir que $x_1, \ldots, x_k$ proporcionan casi toda la información necesaria para la predicción de y. Así, del mismo modo que r^2 proporciona una medida del ajuste del modelo lineal simple, R^2 proporciona una medida del ajuste de un modelo mucho más complejo.

Para ilustrar lo anterior, volvamos al listado de computadora (tabla 12.3) del ejemplo 12.2, el del análisis de la información de cuánta televisión ven los entrevistados. Se repite aquí en la tabla 12.4 el listado de computadora marcando la información que para este caso resulta pertinente.

Tabla 12.4 *Listado de computadora para los datos de la tabla 12.2 del estudio sobre las horas frente al televisor de personas mayores.*

ANALISIS DE VARIANZA

	GL	SUMA DE CUADRADOS	CUADRADO MEDIO	COCIENTE F
REGRESION	3	19.972	6.657	11.760
RESIDUAL	21	11.888	.566	

ANALISIS INDIVIDUAL DE LAS VARIABLES

VARIABLE	COEFICIENTE	DESV.ESTANDAR	VALOR F
(CONSTANTE	1.41411)		
ESTADO CIVIL	−1.17396	.31445	13.9380
EDAD	.03971	.03191	1.5480
AÑOS ESCOLARIDAD	−.15106	.05023	9.0456

El primer renglón del listado de computadora de la tabla 12.4, llamado R MULTIPLE, da el valor del coeficiente de correlación múltiple R. Este coeficiente mide la correlación entre y y la parte del modelo que tiene a $x_1, x_2, \ldots,$

x_k. Así, R es la generalización del coeficiente de correlación simple r. Observe que $R = .7918$ para el ejemplo de la televisión.

El segundo renglón del listado, llamado R CUADRADA, da el valor del coeficiente de determinación múltiple R^2. Este valor $R^2 = .6269$ es de más fácil interpretación para la bondad del ajuste del modelo. Nos dice que sólo el 62.69% de la variación total de los valores de y en relación a su promedio, puede ser explicada por medio del modelo. El resto, un 37.31%, queda no explicado. El ajuste relativamente pobre de este modelo puede deberse al hecho de que x_1, x_2 y x_3 no aparezcan como debieran en el modelo (quizás debieran haberse incluido términos con x_2^2, x_3^3, x_1x_2, x_1x_3, x_2x_3, etc.), o quizás y, el promedio de horas diarias pasadas frente al televisor, sea una función de muchas otras variables además de x_1, x_2, y x_3. Por ejemplo, podría haberse incluido la variable x_4 que mide la afición del entrevistado a la lectura y una variable indicadora x_5 que valga 1 si el entrevistado trabaja y 0 si no trabaja. Se puede pensar en otras variables que podrían afectar el tiempo que se pasa frente al televisor. El hecho de que $R^2 = .6269$ haya resultado tan bajo puede deberse a ambas de las razones descritas. Es probable que x_1, x_2, y x_3 no aparezcan en el modelo en la mejor manera (se hacen algunos comentarios al respecto de formulación de modelos en la sección 12.10) y que el modelo no incluya a un número adecuado de variables predictoras relacionadas con y.

El tercer renglón del listado de computadora de la tabla 12.4, llamado DESV. EST. ESTIM., es el valor de s para el análisis de regresión, la raíz cuadrada de s^2. Recuérdese que s^2 es el estimador de σ^2 (la varianza de los valores de y para valores fijos de $x_1, x_2, \ldots, x_k$) y s^2 es igual a SCE dividido por los grados de libertad apropiados. En el modelo lineal simple (que tiene dos parámetros β), se divide SCE por $(n - 2)$. En el caso general, s^2 se obtiene al dividir SCE por n, el número de datos, menos un grado de libertad por cada parámetro β que aparece en el modelo. De la tabla 12.4, puede verse que $s = .7524$ para los datos de la televisión.

Estimador de la varianza en regresión múltiple

$$s^2 = \frac{\text{SCE}}{n - (\text{número de parámetros } \beta \text{ en el modelo})}$$

En algunos listados de computadora aparece SCE, en otros s^2 y en otros s. Desde luego que una vez que se tiene una de estas cantidades, se puede calcular cualquiera de las otras dos. ¿Para qué sirven? La respuesta es que al igual que en el modelo lineal simple, en el caso general s aparece también en todas las fórmulas para intervalos de confianza y para probar hipótesis. No todos los intervalos de confianza ni todas las estadísticas para probar hipótesis aparecen en los listados. De ahí que s se incluya por si es requerida para construir un intervalo de confianza o para probar una hipótesis especial.

El uso directo de s es muy grande. Puede servir como verificación para detectar errores en los cálculos de la ecuación de predicción. Quizás algunos datos fueron alimentados incorrectamente a la computadora o quizás el programa es muy sensible a errores de redondeo y da respuestas equivocadas. Para

detectar este tipo de errores, se calculan las desviaciones entre los valores de y y los pronosticados por la ecuación. La mayoría de estas desviaciones deben ser menores que $2s$ y casi todas deben ser menores que $3s$. Si los valores de y no están de acuerdo a esta "receta," conviene que se revisen los cálculos.

12.8 Prueba de la utilidad de un modelo de regresión

El particionar la SC total en SCR y SCE es llamado un "análisis de varianza." Este nombre se usa porque en el caso en el que $x_1, x_2, \ldots, x_k$ no contribuyen con información alguna para la predicción de y (en otras palabras el modelo no sirve), entonces ambas cantidades, SCR y SCE, proporcionan un estimador independiente (en el sentido probabilístico) de σ^2, la varianza de y para valores fijos de $x_1, x_2, \ldots, x_k$. Estos estimadores son llamados **cuadrados medios.** Así,

$$\text{CMR} = \text{cuadrado medio de regresión} = \frac{\text{SCR}}{\nu_1}$$

$$\text{CME} = s^2 = \text{cuadrado medio del error} = \frac{\text{SCE}}{\nu_2}$$

en donde

ν_1 = el número de parámetros β en el modelo menos uno $= k$

$\nu_2 = n -$ (el número de parámetros β en el modelo) $= n - (k + 1)$

Para cualquier modelo de regresión múltiple con k variables predictoras,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

se usan CMR y CME para probar la hipótesis de que $x_1, x_2, \ldots, x_k$ no contribuyen con ninguna información para la predicción de y. Esto es equivalente a hipotetizar que

$$\beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Si los datos proporcionan evidencia suficiente para rechazar esta hipótesis, eso quiere decir que por lo menos una de las variables predictoras $x_1, x_2, \ldots, x_k$ contribuye con información para predicir y.

Para esta prueba se usa la estadística

$$F = \frac{\text{CMR}}{\text{CME}} = \frac{\text{CMR}}{s^2}$$

Esta estadística tiene una distribución F con ν_1 y ν_2 grados de libertad en donde, como ya se explicó,

ν_1 = grados de libertad del numerador

= número de parámetros en el modelo menos uno

$= k$

ν_2 = grados de libertad del denominador

= n − (número de parámetros β en el modelo)

= $n - (k + 1)$

Cuando la hipótesis nula es falsa (el modelo sí sirve para predecir y) SCR tenderá a ser mayor que lo que se esperaría que fuera si el modelo no sirve, y el valor de F sería mayor. De ahí que se rechace $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$ para valores de F que excedan $F_{.05}$, un valor en la cola superior de la distribución F (vea la figura 12.3). Los grados de libertad asociados con $F_{.05}$ son aquellos para CMR y s^2.

Se ilustra el procedimiento con un ejemplo.

Ejemplo 12.7 Pruebe la utilidad del modelo de regresión para predicir los hábitos de ver televisión (ejemplo 12.2). Use el listado de computadora de la tabla 12.5 al hacer la prueba.

Solución El listado de computadora para los datos de las personas mayores frente al televisor de la tabla 12.2 se muestran aquí en la tabla 12.5. Ahora se ha marcado la porción del listado relevante para este problema, la porción del análisis de varianza. En la tercera columna de la tabla, bajo el encabezado SUMA DE CUADRADOS, se dan SCR y SCE. Esto es,

$$\text{SCR} = 19.972 \qquad \text{SCE} = 11.888$$

En la segunda columna de la tabla, bajo el encabezado GL, se dan los grados de libertad asociados a cada una de las sumas de cuadrados. Así, ν_1, el número de grados de libertad asociados a SCR, es 3. Similarmente ν_2, el número de grados de libertad asociados a SCE, es 21.

En la cuarta columna, bajo el encabezado CUADRADO MEDIO, se dan CMR y CME. Así

$$\text{CMR} = \frac{\text{SCR}}{3} = \frac{19.972}{3} = 6.657$$

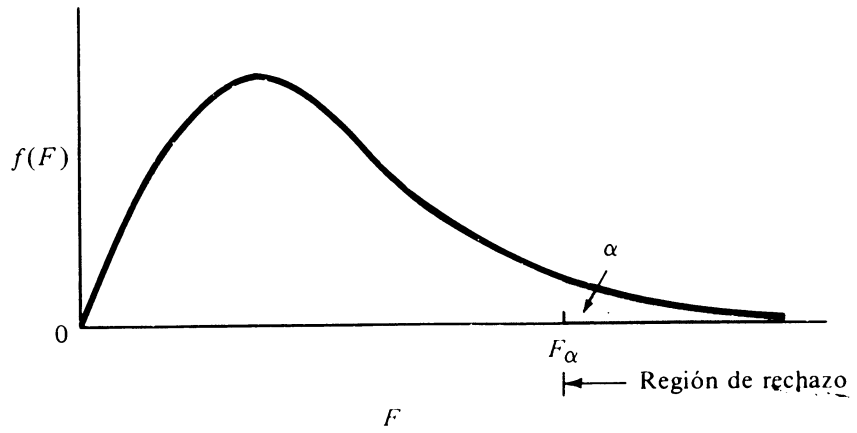


Figura 12.3 *Región de rechazo para la prueba F de* $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

$$\text{CME} = s^2 = \frac{\text{SCE}}{21} = \frac{11.888}{21} = .566$$

En la quinta columna de la tabla, bajo el encabezado COCIENTE F, se da el valor calculado para la estadística de prueba. Así

$$F = \frac{\text{CMR}}{\text{CME}} = \frac{\text{CMR}}{s^2} = \frac{6.657}{.566} = 11.760$$

Se compara este valor calculado de F con el valor crítico $F_{.05}$ basado en $\nu_1 = 3$ y $\nu_2 = 21$ grados de libertad y se rechaza H_0 si $F > F_{.05}$. Al buscar en la tabla 6 del apéndice, se encuentra $F_{.05} = 3.07$. Como el valor calculado de F a partir de los datos, $F = 11.760$, excede al valor de tablas, $F_{.05} = 3.07$, se rechaza la hipótesis nula

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

Se concluye que por lo menos una de las variables predictoras contribuye con información para la predicción de y.

Tabla 12.5 *Listado de computadora para los datos de la tabla 12.2 del estudio sobre las horas frente al televisor de personas mayores*

```
R MULTIPLE              .7918
R CUADRADA              .6269
DESV. EST. ESTIM.       .7524

ANALISIS INDIVIDUAL DE LAS VARIABLES

    VARIABLE       COEFICIENTE  DESV.ESTANDAR  VALOR F
(CONSTANTE          1.41411)
ESTADO CIVIL       -1.17396        .31445      13.9380
EDAD                .03971        .03191       1.5480
AÑOS ESCOLARIDAD   -.15106        .05023       9.0456
```

Ejercicios

12.10. Un representante de ventas de una compañía que vende soya como suplemento de la carne se interesa en construir un modelo para predecir las ventas de soya en distintas zonas comerciales. Se obtuvieron datos sobre ventas pasadas ($ en miles) para cada una de las 25 zonas comerciales de la compañía y se relacionaron con los valores, en cada zona de las siguientes variables:

x_1 = coeficiente de elasticidad cruzada entre soya y carne de res

x_2 = ingreso per capita ($ en miles)

x_3 = índice promedio de consumo con base en gasto familiar

x_4 = precio unitario de un paquete de soya

x_5 = proporción de gasto dedicado a publicidad por la compañía en esa zona

x_6 = 1 si la zona es productora de carne de res y 0 si no lo es

El representante utilizó un paquete de computadora de análisis de regresión que produjo el siguiente listado.

R MULTIPLE .9825
R CUADRADA .9654
DESV. EST. ESTIM. 1.7098

ANALISIS DE VARIANZA

	GL	SUMA DE CUADRADOS	CUADRADO MEDIO	COCIENTE F
REGRESION	6	1468.034	244.672	
RESIDUAL	18	52.614	2.923	

ANALISIS INDIVIDUAL DE LAS VARIABLES

VARIABLE	COEFICIENTE	DESV.ESTANDAR	VALOR F
(CONSTANTE	−51.034)		
ELAST-CRUZ	57.600	13.813	17.356
ING. PER CAP.	2.956	2.548	1.347
IND. CONSUMO	−.934	1.129	.682
PRECIO	−46.542	31.661	2.160
GASTO PUB.	46.355	9.499	23.843
ZONARES	−1.805	.775	5.431

a. Calcule el COCIENTE F que falta en la tabla y pruebe la hipótesis sobre si el modelo sirve para algo o no. Use $\alpha = .05$.
b. ¿Qué proporción de la variación en las ventas de soya se explica por las seis variables predictoras del modelo?
c. Dé la ecuación de regresión para la predicción de las ventas da soya en cualquier zona comercial.
d. En presencia de las otras variables, ¿cuál es la más significativa como predictora de las ventas? ¿Cuál es la menos significativa?
e. Suponga que es muy difícil obtener una medida precisa para el ingreso per cápita en cada zona comercial. ¿Pierde mucha información el modelo de predicción si se elimina como predictor al ingreso per cápita? Si es eliminado, ¿cómo debe plantearse el nuevo modelo?
f. Encuentre un intervalo de confianza del 90% para la diferencia en ventas entre zonas que sean productoras de carne de res y zonas que no lo sean (en otras palabras para β_6).

12.11. La tabla 12.6 lista los precios de venta y (en sustitución del valor) y lista 7 variables predictoras supuestamente relacionadas para cada una de 50 residencias uni-familiares.

Estos datos fueron efectivamente obtenidos por medio de muestreo de una zona residencial de Eugene, Oregon, E.U., durante 1974 por la Oficina Asesora del Condado de Lane. El objeto de su obtención fue para poder desarrollar un modelo para estimar el valor de las residencias. El modelo lineal $y = \beta_0 + \beta_1 x_1 + \cdots \beta_7 x_7 + \epsilon$ fue ajustado a los datos por medio de un programa estándar de análisis de regresión. Los resultados del análisis son

```
R MULTIPLE              .9490
R CUADRADA              .9006
DESV. EST. ESTIM.      3.0051

ANALISIS DE VARIANZA

                SUMA DE    CUADRADO
           GL  CUADRADOS     MEDIO     COCIENTE F
REGRESION   7  3438.183    491.169      54.389
RESIDUAL   42   379.291      9.031

ANALISIS INDIVIDUAL DE LAS VARIABLES

    VARIABLE        COEFICIENTE   DESV.ESTANDAR   VALOR F
(CONSTANTE          -13.4617)
SUPERFICIE            1.0207          .1110       84.5510
NUM. DORMS.           1.6270         1.4813        1.2065
NUM. BAÑOS           -3.8510         1.3634        7.9778
TOT. CUARTOS          2.8936          .9192        9.9095
EDAD                  -.0242          .1031         .0550
GARAGE                1.2497         1.1239        1.2363
VISTA                -1.2897         1.4965         .7428
```

Tabla 12.6 *Mediciones hechas en 50 residencias unifamiliares, ejercicio 12.11*

RESIDENCIA i	PRECIO DE VENTA y(×$1000)	SUPERFICIE PIES CUAD. x_1(× 100)	DORMITORIOS x_2	BAÑOS x_3	TOTAL CUARTOS x_4	EDAD x_5	GARAGE 1 = TIENE 0 = NO x_6	VISTA x_7
1	10.2	8.0	2	1	5	5	0	0
2	10.5	9.5	2	1	5	8	0	0
3	11.1	9.1	3	1	6	2	0	0
4	15.3	9.5	3	1	6	6	0	0
5	15.8	12.0	3	2	7	5	0	0
6	16.3	10.0	3	1	6	11	0	0
7	17.2	11.8	3	2	7	8	0	0
8	17.7	10.0	2	1	7	15	1	0
9	18.0	13.8	3	2	7	10	0	0
10	18.1	12.5	3	2	7	11	0	0
11	18.4	15.0	3	2	7	12	0	0
12	18.4	12.0	3	2	7	8	0	0
13	18.9	16.0	3	2	7	9	1	1
14	19.3	16.5	3	2	7	15	0	0
15	19.5	16.0	3	2	7	11	1	0
16	19.9	16.8	2	2	7	12	0	0
17	20.3	15.0	3	1	7	8	1	0
18	20.3	17.8	3	2	8	13	1	0
19	20.8	17.9	3	2	7	18	1	0
20	21.0	19.0	2	2	7	22	0	0
21	21.5	17.6	3	1	6	17	0	0
22	22.0	18.5	3	2	8	11	1	0
23	22.1	18.0	3	2	7	5	0	0
24	22.5	17.0	2	3	8	2	1	0
25	22.8	18.7	3	1	6	6	0	0
26	22.8	20.0	3	2	7	16	0	0
27	22.9	20.0	3	2	7	12	0	0
28	23.2	21.0	3	2	7	10	1	0
29	23.5	20.5	2	2	7	11	1	0
30	24.9	19.9	3	1	7	13	1	1
31	25.0	21.5	2	2	7	8	0	0
32	25.1	20.5	3	1	7	9	1	0
33	26.6	22.0	3	2	7	10	0	0
34	26.9	22.0	3	2	7	6	1	1
35	26.9	21.8	2	1	6	15	1	0

(continuada)

Tabla 12.6 (continuada)

RESIDENCIA i	PRECIO DE VENTA y(×$1000)	SUPERFICIE PIES CUAD. x_1(× 100)	DORMITORIOS x_2	BAÑOS x_3	TOTAL CUARTOS x_4	EDAD x_5	GARAGE 1 = TIENE 0 = NO x_6	VISTA x_7
36	27.8	22.5	3	2	7	11	1	0
37	28.0	24.0	3	2	7	17	0	0
38	28.7	23.5	3	2	8	12	0	0
39	29.0	25.0	3	2	7	11	1	0
40	30.1	25.6	3	2	7	15	1	0
41	32.0	25.0	4	2	8	12	1	0
42	33.8	25.0	2	2	8	8	0	1
43	35.3	26.8	3	2	7	6	1	0
44	37.1	22.1	3	2	8	18	1	0
45	37.5	27.5	3	2	8	12	1	0
46	38.0	25.0	4	2	8	10	1	0
47	38.4	24.0	3	2	8	13	1	1
48	39.0	31.0	4	3	9	25	1	0
49	43.0	21.0	4	2	9	18	1	0
50	55.0	40.0	5	3	12	22	1	0

Revise el listado de computadora para este análisis y discuta cada uno de los puntos que se fueron viendo en el ejemplo de las horas pasadas frente al televisor.

12.12. Refiérase al ejercicio 12.11. Utilice el modelo encontrado para obtener estimaciones del valor de cada una de las siguientes cinco residencias de Eugene. Los datos que las describen se dan en la tabla.

RESIDENCIA i	SUPERFICIE PIES. CUAD. x_1	DORMITORIOS x_2	BAÑOS x_3	TOTAL CUARTOS x_4	EDAD x_5	GARAGE x_6	VISTA x_7
1	22.4	4	2	7	18	1	1
2	15.3	3	2	7	6	0	0
3	17.2	4	1	7	4	1	0
4	31.7	5	3	9	24	0	0
5	20.0	4	2	8	11	1	1

12.13. Todos estamos concientes del efecto que tiene la inflación en el valor de los bienes raíces; en general tienden a aumentar su valor a la misma tasa que la de la inflación. Lo anterior hace que se tengan que actualizar los avalúos de las propiedades periódicamente. El encargado de actualizar los avalúos puede optar por cualquiera de los tres caminos:

i. Cada actualización puede hacerse aplicando la tasa de inflación al avalúo previo.
ii. Pueden obtenerse nuevos datos sobre la situación comercial y juntarse con los disponibles en el pasado para desarrollar un modelo de regresión para estimar el valor.
iii. Puede hacerse un modelo de regresión para estimar el valor basado sólo en datos nuevos sobre la situación comercial olvidando todos los datos anteriores.

¿Qué camino sugeriría usted? Explique.

12.14. ¿Existe una relación consistente entre la práctica de administrar por presupuestos y los rendimientos obtenidos? Si es así, la evidencia que se muestra a continuación respalda la práctica de programas de inversión apegados a una administración por programas presupuestales. Kim y Kwak* realizaron una regresión de y, el valor estimado de los rendimientos por acción, sobre las siguientes variables:

*S.H. Kim and N.K. Kwak, "Capital Budgeting Practices and their Impact on Earnings Performance," *Proceedings of the American Institute for Decision Sciences*, noviembre de 1976.

x_1 = grado de sofisticación del sistema de presupuesto (0–100)

x_2 = tamaño de la empresa (ventas anuales)

x_3 = intensidad de capital (depreciación/ventas anuales.

x_4 = riesgo (desviación estándar de los rendimientos anuales por acción)

x_5 = capitalización (deuda/ventas totales)

x_6 = cociente costo-beneficio

Se obtuvieron datos de cada una de $n = 114$ empresas dedicadas a la elaboración de maquinaria, con ingresos superiores a los \$50 millones en 1974. Los datos usados corresponden al período comprendido entre 1969 y 1974. Los resultados del análisis siguen:

$R^2 = .776$

$F = 61.991$

VARIABLE	CONSTANTE	X_1	X_2	X_3	X_4	X_5	X_6
COEFICIENTE	−1.613	.040	.001	.090	−.072	.018	.010
VALOR T		10.282	.464	2.232	−.559	4.143	.806

a. ¿Se concluye del estudio de Kim-Kwak que hay una relación significativa entre la práctica de presupuestar y los rendimientos para el tipo de empresas estudiadas?

b. ¿Cuáles de las variables consideradas en el análisis contribuyen con información para predecir el valor del rendimiento?

c. Explique e interprete la cantidad $R^2 = .776$.

12.9 Uso de la ecuación de predicción para estimación y predicción

Suponga que es usted un analista de una firma de corredores de bolsa que quiere investigar la relación entre el precio y de las acciones de una determinada compañía que proporciona energía eléctrica y un conjunto de variables independientes $x_1, x_2, \ldots, x_k$, en donde x_1, x_2, x_3 y x_4 son, por ejemplo

x_1 = tasa de interés preferencial

x_2 = (tasa de interés preferencial)$^2 = x_1^2$

x_3 = redituabilidad de las acciones

x_4 = tasa de dividendo de las acciones

Las variables predictoras restantes, $x_5, x_6, \ldots, x_k$ son otras variables que se piensan relacionadas o bien los cuadrados, cubos y productos cruzados de las cuatro primeras. Por ejemplo, para ajustar por el efecto del tiempo, se podría incluir el PIB (Producto Interno Bruto) como una de las variables predictoras. También, si se piensa que la tasa preferencial x_1 tiende a tener un efecto mayor (o menor) en el precio y dependiendo de la magnitud de x_1, como se muestra en la figura 12.4, se puede considerar un término en el modelo con $x_2 = x_1^2$. Así el modelo para el precio y de la acción de la empresa generadora de energía eléctrica podría ser:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_3 + \beta_4 x_4 + \cdots + \beta_k x_k + \epsilon$$

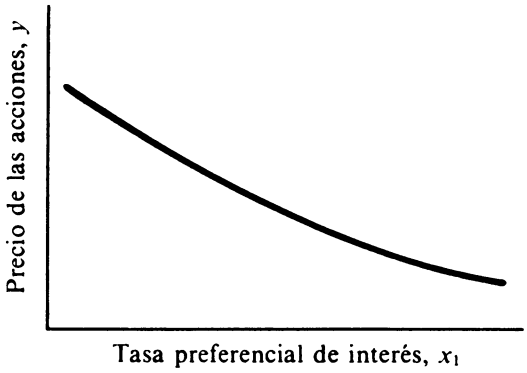


Figura 12.4 *Una gráfica del efecto hipotético de la tasa preferencial de interés x_1 en el precio de las acciones de una empresa generadora de energía eléctrica*

Suponga que se selecciona una muestra aleatoria de los precios de las acciones de la empresa en determinados momentos dentro de un período de dos años y que, con esas observaciones y los correspondientes valores para las variables independientes se ajusta un modelo, obteniéndose

$$\hat{y} = 18.7 - 0.6x_1 - 0.2x_1^2 + 1.1x_3 + 8.3x_4 + \cdots$$

¿Qué puede decirse al observar los estimadores $\hat{\beta}_1 = -0.6$, $\hat{\beta}_2 = -0.2$, $\hat{\beta}_3 = 1.1$, $\hat{\beta}_4 = 8.3$? La respuesta es que se puede decir muy poco, como ya se ha hecho notar en las secciones 12.7 y 12.8. Se puede probar la hipótesis nula de que un parámetro β particular es igual a 0 (en otras palabras, que la variable x correspondiente no contribuye con información alguna para la predicción de y en presencia de las otras variables x del modelo), pero esto en realidad ¿qué quiere decir?

Si no se rechaza la hipótesis nula de que ese parámetro es cero, eso no quiere decir que la correspondiente variable x no contribuye con información para predecir y. Es posible que no se tengan suficientes datos como para detectar la información proporcionada por x o es posible también que x proporcione información sobre y pero que esta misma información ya la hayan proporcionado otras variables x consideradas en el modelo.

En cualquier caso, parece ser que en la mayoría de las investigaciones en negocios, resulta ser de mucha más importancia el uso de toda la ecuación de predicción que el hacer estimaciones o pruebas de hipótesis sobre los parámetros individuales. La ecuación de predicción puede ser de valor en tres formas:

1. Puede utilizarse para estimar el valor medio de y para valores dados de dos variables predictoras.
2. Puede utilizarse para predecir algún valor futuro de y para valores dados de $x_1, x_2, \ldots, x_k$.
3. Si una ecuación de predicción proporciona un buen ajuste a los datos (R^2 es grande) y el número de variables predictoras no es muy grande entonces es posible que la ecuación en sí sea de ayuda para entender mejor el proceso bajo investigación.

Las estimaciones para el valor medio de y para valores dados de $x_1, \ldots, x_k$ o las predicciones para valores específicos de y para valores dados de $x_1, \ldots, x_k$ se obtienen sustituyendo los valores dados de $x_1, \ldots, x_k$ en la ecuación de predicción.

Por ejemplo, suponga que una compañía que vende por correo se interesa en relacionar el monto de sus ventas navideñas con dos variables predictoras (se mantienen sólo 2 para propósitos de ilustración), el número de folletos enviados por correo, x_1, y el periodo de tiempo (en meses) previo a la Navidad en que se enviaron los folletos, x_2. Más aún, suponga que de los registros de la compañía se conocen los valores que tomaron x_1, x_2, y y en 20 regiones distintas de ventas en las que opera la compañía. Después de lo anterior, se ajustó una ecuación de predicción a los datos que resultó ser la siguiente:

$$\hat{y} = 1.3 + 1.7x_1 - 0.1x_1^2 + 1.0x_1x_2 - 0.3x_1^2x_2$$

en donde

x_1, se expresa en cientos de miles, y se midió en el intervalo $0.5 \leq x_1 \leq 2.5$

x_2, se expresa en meses y se midió en el intervalo $1 \leq x_2 \leq 3$

y se expresa en \$, cientos de miles

El mejor estimador de las ventas medias $E(y)$ para una combinación dada de x_1 y x_2, por ejemplo $x_1 = 1$ (100,000 folletos enviados) y $x_2 = 2$ (enviados 2 meses antes de Navidad), se obtiene sustituyendo $x_1 = 1$ y $x_2 = 2$ en la ecuación de predicción.

$$\begin{aligned}\hat{y} &= 1.3 + 1.7x_1 - 0.1x_1^2 + 1.0x_1x_2 - 0.3x_1^2x_2 \\ &= 1.3 + 1.7(1) - 0.2(1)^2 + 1.0(1)(2) - 0.3(1)^2(2) \\ &= 4.3\end{aligned}$$

o sea una estimación de \$430,000 para la venta esperada.

Como se ha hecho notar en las secciones 11.6 y 11.7, $\hat{y}$ no sólo es el mejor estimador del valor medio de y para valores dados de x_1 y x_2, sino que también proporciona la mejor predicción para algún valor de y que se observará en el futuro para los mismos valores dados de x_1 y x_2. Esto es si se selecciona una región y se envían 100,000 folletos ($x_1 = 1$) con 2 meses de anticipación a la Navidad ($x_2 = 2$), las ventas pronosticadas para esa región son de \$430,000.

Se puede construir un intervalo de confianza para $E(y)$ y un intervalo de predicción para y con un procedimiento similar al empleado para el modelo lineal simple en las secciones 11.6 y 11.7. Sin embargo, las fórmulas para ellos son demasiado complejas como para presentarlas en este texto. Por suerte, en algunos paquetes de regresión para computadora, el cómputo de estos intervalos se da como una opción al usuario (véase la sección 12.12). Si no se tiene al alcance uno de estos paquetes, entonces se requiere familiarizarse con estas fórmulas, que debido a su complejidad siempre se expresan en notación matricial. Esta notación y las fórmulas, tanto del intervalo de confianza para $E(y)$ como para el intervalo de predicción para un valor futuro de y, aparecen junto con una explicación en el libro *An Introduction to Linear Models and the Design and Analysis of Experiments* de W. Mendenhall (vea las referencias).

Ejemplo 12.8 Aun cuando se supone que la demanda de un artículo disminuye al aumentar el precio de éste si hay artículos competitivos a un precio menor, parece que no siempre es éste el caso. De hecho se aproxima muchas veces la relación entre demanda y precio con un modelo de segundo orden; los aumentos pequeños de precio hacen que disminuye la demanda y los aumentos grandes de precio hacen que se perciba una aparente mejora en la calidad y por ende la demanda aumenta. Un esfuerzo tendiente a estudiar estas relaciones fue hecho por un distribuidor de licor, que normalmente vende el litro en $5.00. Realizó un experimento en 15 zonas distintas de ventas durante un período de 12 meses usando 5 niveles de precio para las botellas de un litro. Los resultados del experimento se muestran en la tabla.

NÚMERO DE CAJAS VENDIDAS AL MES POR CADA 10,000 HABITANTES, y	PRECIO POR LITRO, x
23, 20, 21	$5.00
19, 21, 18	5.50
18, 17, 20	6.00
21, 19, 20	6.50
25, 24, 22	7.00

a. Ajuste un modelo de segundo orden $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$, a los datos.
b. Pronostique y, el número de cajas que se venderán en un mes por cada 10,000 habitantes en una población en donde el precio por litro es $5.00. Pronostique y para x = $6.00, x = $7.00.
c. Construya un intervalo de predicción del 95% para y cuando x = $5.00, cuando x = $6.00; cuando x = $7.00.
d. Construya un intervalo de confianza del 95% para $E(y)$, el número medio de cajas vendidas en un mes por cada 10,000 habitantes cuando x = $5.00; cuando x = $6.00; cuando x = $7.00.

Solución Usando un programa de regresión estándar con la opción del cálculo de intervalos de confianza y de predicción, los siguientes resultados se obtuvieron en el listado de computadora.

```
R MULTIPLE              .8393
R CUADRADA              .7045
DESV. EST. ESTIM.     1.3292

ANALISIS DE VARIANZA

                    SUMA DE   CUADRADO
             GL  CUADRADOS     MEDIO    COCIENTE F
REGRESION     2    50.533     25.266      14.301
RESIDUAL     12    21.200      1.767
```

ANALISIS INDIVIDUAL DE LAS VARIABLES

VARIABLE	COEFICIENTE	DESV. ESTANDAR	VALOR F
(CONSTANTE	156.1306)		
PRECIO (X)	−46.9325	9.8564	22.6729
PRECIO CUAD (X2)	3.9999	.8204	23.7729

OBSERVACION	OBS. X VALOR	OBS. Y VALOR	PRONOSTIC. VALOR Y	LIM. INF. 95% PARA E(Y)	LIM. SUP. 95% PARA E(Y)
1	5.0000	23.0000	21.4681	18.9971	23.9391
2	5.0000	20.0000	21.4681	18.9971	23.9391
3	5.0000	21.0000	21.4681	18.9971	23.9391
4	5.5000	19.0000	19.0018	16.8414	21.1622
5	5.5000	21.0018	19.0018	16.8414	21.1622
6	5.5000	18.0000	19.0018	16.8414	21.1622
7	6.0000	18.0000	18.5356	16.3039	20.7673
8	6.0000	17.0000	18.5356	16.3039	20.7673
9	6.0000	20.0000	18.5356	16.3039	20.7673
10	6.5000	21.0000	20.0693	17.9089	22.2297
11	6.5000	19.0000	20.0693	17.9089	22.2297
12	6.5000	20.0000	20.0693	17.9089	22.2297
13	7.0000	25.0000	23.6031	21.1343	26.0719
14	7.0000	24.0000	23.6031	21.1343	26.0719
15	7.0000	22.0000	23.6031	21.1343	26.0719

OBSERVACION	OBS. X VALOR	OBS. Y VALOR	PRONOSTIC. VALOR Y	LIM. INF. 95% PARA Y	LIM. SUP. 95% PARA Y
1	5.0000	23.0000	21.4681	18.1736	24.7626
2	5.0000	20.0000	21.4681	18.1736	24.7626
3	5.0000	21.0000	21.4681	18.1736	24.7626
4	5.5000	19.0000	19.0018	15.9333	22.0703
5	5.5000	21.0000	19.0018	15.9333	22.0703
6	5.5000	18.0000	19.0018	15.9333	22.0703
7	6.0000	18.0000	18.5356	15.4166	21.6546
8	6.0000	17.0000	18.5356	15.4166	21.6546
9	6.0000	20.0000	18.5356	15.4166	21.6546
10	6.5000	21.0000	20.0693	17.0008	23.1378
11	6.5000	19.0000	20.0693	17.0008	23.1378
12	6.5000	20.0000	20.0693	17.0008	23.1378
13	7.0000	25.0000	23.6031	20.3102	26.8960
14	7.0000	24.0000	23.6031	20.3102	26.8960
15	7.0000	22.0000	23.6031	20.3102	26.8960

a. La ecuación de predicción para estimar y, el número de cajas que se venderán en un mes por cada 10,000 habitantes en una población en donde el precio por litro es x, es

$$\hat{y} = 156.1306 - 46.9325x + 3.9999x^2$$

b. Si $x = \$5.00$, se tiene

$$\hat{y} = 156.1306 - 46.9325(5.0) + 3.9999(25.0) = 21.4681$$

o sea más de 21 cajas. Similarmente para los otros precios por litro, se encuentra

$$\hat{y} = 18.5356 \quad \text{para} \quad x = \$6.00$$

$$\hat{y} = 23.6031 \quad \text{para} \quad x = \$7.00$$

El modelo para estimar un resultado particular (futuro) y o la respuesta media $E(y)$ es el mismo. Por ejemplo, el número medio de cajas que se estima que se venderán por cada 10,000 habitantes cuando el precio sea \$5.00 es $\hat{E}(y) = 21.4681$. Similarmente cuando $x = \$6.00$ y $x = \$7.00$, los estimadores de las ventas medias son $\hat{E}(y) = 18.5356$ y $\hat{E}(y) = 23.6031$.
c., d. Los intervalos de predicción para y del 95% y de confianza para $E(y)$ del 95% también, se leen directamente del listado de computadora. Se dan en la tabla.

PRECIO POR LITRO	INTERVALO DE PREDICCION DEL 95% PARA y	INTERVALO DE CONFIANZA DEL 95% PARA $E(y)$
\$5.00	18.1736 a 24.7626	18.9971 a 23.9391
6.00	15.4166 a 21.6546	16.3039 a 20.7673
7.00	20.3102 a 26.8960	21.1343 a 26.0719

Observe que en cada caso, el intervalo de predicción para y del 95% es más ancho que el intervalo de confianza del 95% para $E(y)$. Como se hizo notar en las secciones 11.6 y 11.7 esto es consecuencia de que la varianza del error al predecir un valor particular de y es superior a la varianza del error al estimar el valor medio $E(y)$. Además, como estas varianzas dependen de los valores que se fijen para las variables independientes al calcular $\hat{y}$, los tres intervalos de predicción y los tres intervalos de confianza tienen longitudes distintas.

Como se ha visto, la ecuación de predicción es útil para estimar $E(y)$ y para predecir algún valor futuro de y. Pero, también puede auxiliar en el entendimiento del proceso bajo estudio. Por ejemplo, considere la ecuación de predicción para el problema de la compañía que vende por correo. Una manera sencilla de estudiar la relación es graficando las ventas navideñas y, como una función del número de folletos enviados, x_1 con $x_2 = 1$, 2 ó 3 meses de anticipación. Por ejemplo si se sustituye $x_2 = 1$ en la ecuación de predicción, se obtiene

$$\begin{aligned}\hat{y} &= 1.3 + 1.7x_1 - .1x_1^2 + 1.0x_1x_2 - .3x_1^2x_2 \\ &= 1.3 + 1.7x_1 - .1x_1^2 + 1.0x_1(1) - .3x_1^2(1) \\ &= 1.3 + 2.7x_1 - .4x_1^2\end{aligned}$$

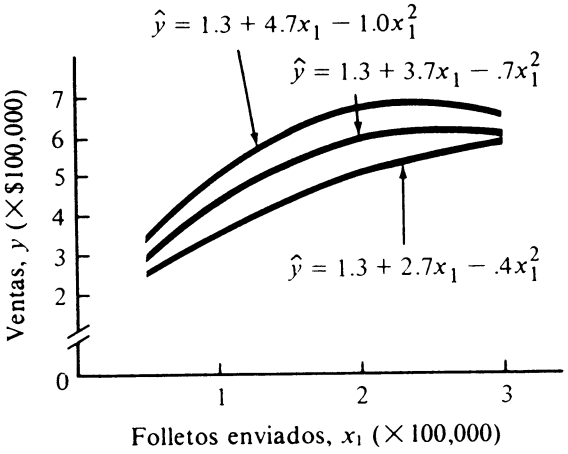


Figura 12.5 *Una gráfica de ventas navideñas y como función del número de folletos enviados*

Lo anterior nos da las ventas pronosticadas si los envíos de folletos se hacen con 1 mes de anticipación a la Navidad. Similarmente, para $x_2 = 2$ y $x_2 = 3$ meses de anticipación, se obtienen las siguientes ecuaciones de predicción.

$$\text{para} \quad x_2 = 2: \quad \hat{y} = 1.3 + 3.7x_1 - .7x_1^2$$

$$\text{para} \quad x_2 = 3: \quad \hat{y} = 1.3 + 4.7x_1 - 1.0x_1^2$$

Estas tres ecuaciones que predicen las ventas navideñas en función del número de folletos enviados x_1 aparecen graficadas en la figura 12.5.

Observe que las formas de estas curvas de ventas son distintas para los valores distintos de x_2. Esto quiere decir que la relación entre las ventas pronosticadas para y y el número de folletos enviados x_1, depende de cuándo fueron enviados estos folletos. Cuando esto ocurre, se dice que x_1 y x_2 **interaccionan** o, diciéndolo de otro modo, el efecto de x_1 en las ventas pronosticadas para y depende del valor de x_2 (y viceversa). Este ejemplo ilustra como las gráficas de $\hat{y}$ ayudan a entender la relación entre el valor que se pronostica para y y las variables predictoras.

Ejemplo 12.9 Se llevó a cabo un estudio para examinar la ganancia y medida porcentualmente, obtenida por una compañía constructora y la relación que guarda con el tamaño x_1 del contrato de construcción y con x_2, el número de años de experiencia del superintendente de obra. Un objetivo adicional del estudio era el investigar el posible efecto de interacción que el tamaño del contrato y la experiencia del superintendente tienen en las ganancias. Se obtuvieron datos para $n = 18$ proyectos de construcción que habia realizado la compañía en los últimos dos años. Estos datos se muestran en la tabla siguiente.

GANANCIA y	TAMAÑO DEL CONTRATO x_1	EXPERIENCIA DEL SUPERINTENDENTE (AÑOS) x_2	GANANCIA y	TAMAÑO DEL CONTRATO x_1	EXPERIENCIA DEL SUPERINTENDENTE (AÑOS) x_2
2.0	5.1	4	5.0	4.3	6
3.5	3.5	4	6.0	2.9	2
8.5	2.4	2	7.5	1.1	2
4.5	4.0	6	4.0	2.6	4
7.0	1.7	2	4.0	4.0	6
7.0	2.0	2	1.0	5.3	4
2.0	5.0	4	5.0	4.9	6
5.0	3.2	2	6.5	5.0	6
8.0	5.2	6	1.5	3.9	4

Ajuste el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \epsilon$$

a los datos sobre los contratos de construcción. Interprete cuidadosamente los resultados. Grafique la ganancia y como una función del tamaño del contrato x_1, para cada uno de los tres niveles de experiencia x_2, del superintendente. ¿Qué indican estas gráficas en relación al efecto de interacción entre el tamaño del contrato y la experiencia del superintendente?

Solución Por medio de un programa estándar de análisis de regresión, se obtuvieron los siguientes resultados.

```
R MULTIPLE              .9289
R CUADRADA              .8628
DESV. EST. ESTIM.       .9708

ANALISIS DE VARIANZA

                 SUMA DE    CUADRADO
            GL  CUADRADOS    MEDIO     COCIENTE F
REGRESION    4    77.026     19.256      20.432
RESIDUAL    13    12.252       .942

ANALISIS INDIVIDUAL DE LAS VARIABLES

    VARIABLE          COEFICIENTE   DESV.ESTANDAR   VALOR F
(CONSTANTE              19.3048)
TAMAÑO CONT. (X1)       -1.4874         1.1779        1.5944
EXPER. (X2)             -6.3707         1.0424       37.3506
TAMAÑO CUAD. (X1 CUAD)  -.7521           .2253       11.1460
TAMAÑO X EXP. (X1X2)     1.7169          .2538       45.7593
```

Se hace a continuación un análisis de las cantidades más importantes del listado.

R^2 (*coeficiente de determinación*). La proporción de variación en las ganancias explicada por el modelo que incluye tamaño del contrato y experiencia del superintendente de obra es

$$R^2 = .8628$$

En otras palabras, aproximadamente el 14% de la variación en las ganancias se queda sin explicación, cantidad que quizás pueda reducirse al incluir en el modelo otras variables predictoras.

Cociente F. El cociente F permite probar la hipótesis

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

El valor calculado de la F a partir de los datos es $F = 20.432$, que excede el valor crítico $F_{.05}$ basado en $\nu_1 = 4$ y $\nu_2 = 13$ grados de libertad, que es $F_{.05} = 3.18$. Así que se rechaza la hipótesis nula y se concluye que el modelo escogido contribuye con información para predecir y.

Coeficientes de las variables. La ecuación de predicción que relaciona las ganancias con el tamaño del contrato x_1 y la experiencia del superintendente x_2 es

$$\hat{y} = 19.3048 - 1.4874x_1 - 6.3707x_2 - 0.7521x_1^2 + 1.7169x_1x_2$$

Valor F. Cada uno de los valores F calculados permite separadamente probar

$$H_0: \beta_j = 0 \qquad j = 1, 2, 3, 4$$

El valor tabulado $F_{.05}$ con $\nu_1 = 1$ y $\nu_2 = 13$ grados de libertad es $F_{.05} = 4.67$. De ahí que los parámetros β_2, β_3, y β_4 se tomen significativos (significativamente distintos de cero), pero β_1 no. Recuerde que esto no implica que x_1 no contribuya con ninguna información para predecir y. Simplemente quiere decir que la contribución de x_1 es pequeña cuando se analiza en presencia de las contribuciones de x_1^2, x_2 y x_1x_2.

El valor de F para el tamaño del contrato y la experiencia del superintendente (x_1x_2) es $F = 45.7593$, que por mucho excede el valor crítico $F_{.05} = 4.67$. Esto indica la presencia de una interacción fuerte entre estos dos factores de la ganancia y.

Lo anterior se hace evidente si se observa la figura 12.6, que proporciona una gráfica de las tres funciones distintas que relacionan la ganancia y con el tamaño del contrato x_1 al considerar los tres niveles de experiencia del superintendente por separado.

Las tres funciones separadas que aparecen en la figura 12.6 son:

$$\text{para } x_2 = 2: \quad \hat{y} = 6.5634 + 1.9464x_1 - .7521x_1^2$$

$$\text{para } x_2 = 4: \quad \hat{y} = -6.1780 + 5.3802x_1 - .7521x_1^2$$

$$\text{para } x_2 = 6: \quad \hat{y} = -18.9194 + 8.8140x_1 - .7521x_1^2$$

Nótese que la ganancia porcentual decrece rápidamente al aumentar el tamaño del contrato para superintendentes de $x_2 = 2$ años de experiencia. Lo

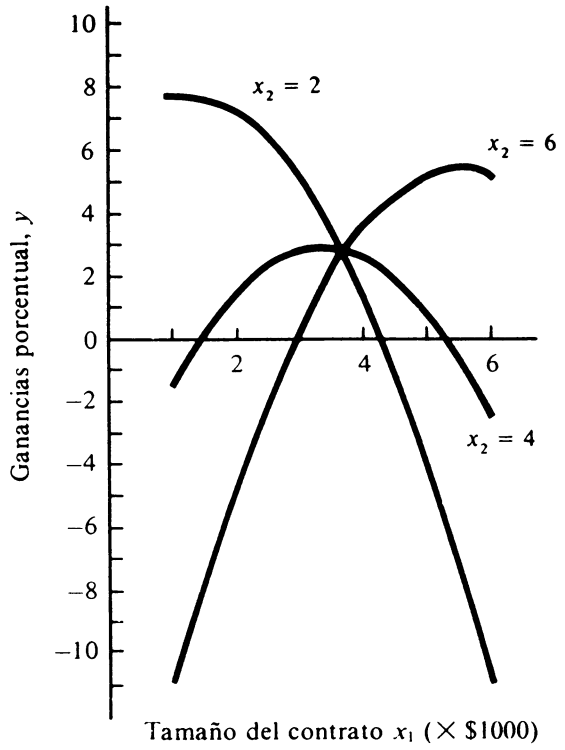


Figura 12.6 *Una gráfica de la ganancia y (en %) como función del tamaño del contrato* x_1

opuesto ocurre para los superintendentes de $x_2 = 6$ años de experiencia. Sobre la base de estas observaciones, parece razonable pensar que una política adecuada para la compañía será el asignar a contratos pequeños superintendentes con sólo unos cuantos años de experiencia y, por el contrario, asignar a los contratos grandes, superintendentes con el mayor número de años de experiencia posibles.

12.10 Algunos comentarios sobre la formulación de un modelo

Las variables independientes que se incluyen para contribuir con información para predecir y, pueden ser de dos tipos; cuantitativas o cualitativas. Como se verá más adelante, la manera de incorporar una variable en el modelo depende de su tipo.

Definición

Una **variable independiente cuantitativa** es una variable que puede tomar valores que corresponden a puntos de una recta. Las variables independientes que no sean cuantitativas se dirán **cualitativas.**

Una tasa de interés, una tasa de desempleo, el número de empleados y el número de máquinas son cuatro ejemplos de variables independientes cuantitativas. En contraste a éstas, suponga que tiene cuatro plantas similares de manufactura para un mismo producto y que se interesa en estudiar la ganancia de cada planta por unidad de tiempo. Cada planta tiene un supervisor; llámelos A, B, C y D. Parece lógico que el supervisor de la planta sea una variable independiente que puede afectar la ganancia de la planta. De lo anterior, que el supervisor sea una variable independiente cualitativa que se debe considerar dentro de la ecuación de predicción del modelo para la ganancia y de cada planta.

Definición

El nivel de intensidad de una variable independiente es llamado **nivel.**

Para variables independientes cuantitativas, los niveles corresponden a los valores que estas variables independientes pueden tomar y corresponden, entonces, a puntos de la recta. Por ejemplo si se piensa que una tasa de interés puede afectar la respuesta bajo estudio, y la respuesta y se registra para tres valores de la tasa de interés, 6%, 8% y 9.2%, entonces se habrá observado la variable independiente "tasa de interés" en tres niveles, 6%, 8% y 9.2%.

Los niveles de las variables independientes cualitativas no son cuantificables y por ello no corresponden a puntos de la recta. Sólo se pueden definir describiéndolos. Para el problema de la ganancia de las plantas de manufactura, la variable independiente "supervisor de planta" se observa en cuatro niveles, cada uno correspondiente a uno de los supervisores A, B, C o D.

Una buena manera de representar un modelo para una respuesta y que es función de una sola variable predictora cuantitativa es graficando $E(y)$, (o $\hat{y}$) como una función de x, como se hizo en el capítulo 11. La línea recta o curva que resulta es referida como **curva de respuesta.**

Del mismo modo en el que la ecuación para $E(y)$ (o $\hat{y}$) como función de x se representa trazando una curva en una hoja de papel, la correspondiente ecuación en dos (o más) variables cuantitativas para $E(y)$(o $\hat{y}$) se representa con una **superficie de respuesta** en un espacio de tres (o más) dimensiones.

Los dos modelos más usados que utilizan variables predictoras cuantitativas son los llamados **modelos lineales de primero y segundo orden** respectivamente. El modelo de primer orden, dado por la ecuación que sigue, se grafica como un plano de respuesta.

Un modelo lineal de primer orden

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

en donde $x_1, x_2, \ldots, x_k$ son variables predictoras cuantitativas y ϵ es un error aleatorio.

Una superficie ajustada de respuesta, de primer orden (correspondiente a un modelo de primer orden) que describe la relación entre el precio y de determinadas acciones y dos variables predictoras cuantitativas

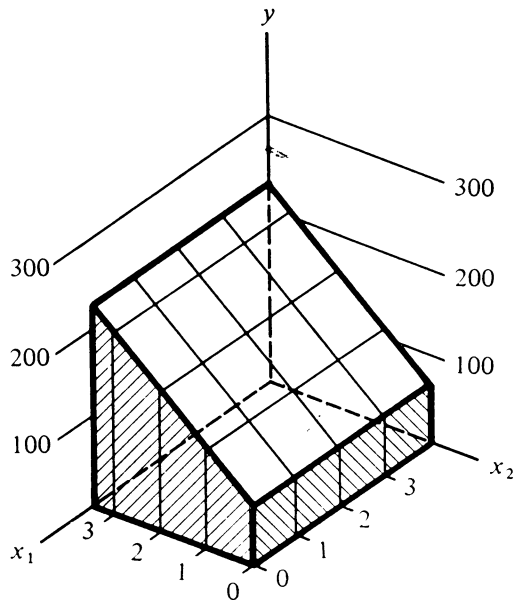


Figura 12.7 *La superficie de respuesta para un modelo lineal de primer orden*

x_1 = dividendos anuales de las acciones

x_2 = redituabilidad por acción

se muestra en la figura 12.7.

Los modelos lineales de segundo orden en k variables predictoras cuantitativas $x_1, x_2, \ldots, x_k$ incluyen todos los términos contenidos en el modelo de primer orden, además de todos los productos de dos variables $x_1x_2, x_1x_3, \ldots, x_2x_3, \ldots$ y todos los cuadrados $x_1^2, x_2^2, \ldots, x_k^2$. Por ejemplo un modelo de segundo orden en dos variables predictoras está dado por la siguiente ecuación.

Un modelo lineal de segundo orden en dos variables predictoras

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2 + \epsilon$$

en donde x_1, x_2 son variables predictoras cuantitativas y ϵ es un error aleatorio.

La superficie de respuesta de un modelo de segundo orden puede ser curva (la curvatura es inducida en primer lugar por $x_1^2, x_2^2, \ldots, x_k^2$) y también puede aparecer torcida. Lo torcido de la superficie es causado por los términos (llamados términos de interacción) que contienen a los productos cruzados $x_1x_2, x_1x_3, \ldots$. Una superficie de segundo orden que describe la relación entre el precio y de determinadas acciones y las dos variables predictoras x_1, los dividendos anuales, y x_2, la redituabilidad de la acción, se muestra en la figura 12.8.

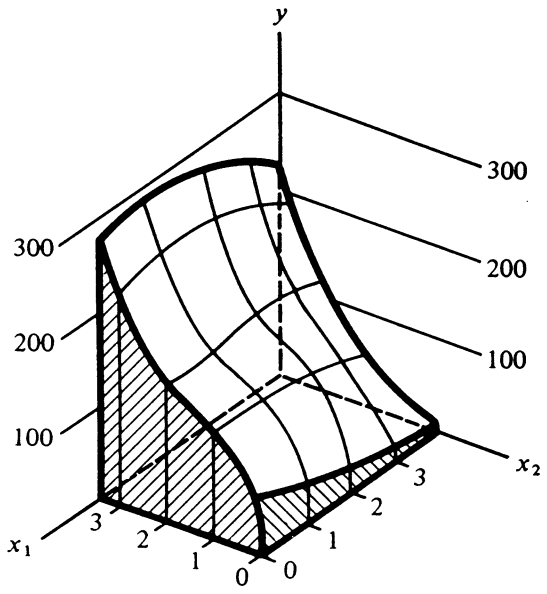


Figura 12.8 *La superficie de respuesta para un modelo lineal de segundo orden*

Las variables independientes cualitativas se incorporan al modelo por medio de variables indicadoras. Por cada variable independiente cualitativa se requieren de tantas variables indicadoras como niveles tenga la variable cualitativa, menos uno. Por ejemplo, si las ventas y de una compañía mayorista depende de la variable "localidad" y si tiene tres sucursales en tres localidades, A, B y C, los primeros términos del modelo son

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \left\{ \begin{array}{l} \text{términos asociados} \\ \text{a otras variables predictoras} \end{array} \right\} + \epsilon$$

en donde

$x_1 = 1$ si se trata de la localidad B, $x_1 = 0$ si no es así.

$x_2 = 1$ si se trata de la localidad C, $x_2 = 0$ si no es así.

LOCALIDAD	x_1	x_2
A	0	0
B	1	0
C	0	1

Las codificaciones para las tres localidades son las mostradas en la tabla. Si se mide una respuesta en la localidad A, se hacen $x_1 = 0$ y $x_2 = 0$. En el ejemplo 12.2 se usó una variable indicadora para denotar si se vivía con cónyuge o no, en otras palabras para la variable cuantitativa, estado civil, y en otros ejemplos de este capítulo ya han sido usadas variables indicadoras.

La formulación del modelo probabilístico es quizás la parte más importante en un análisis de regresión. ¿Por qué? Aun cuando se tenga toda la información con la que contribuyen las variables predictoras en el modelo, el ajuste puede ser malo si el modelo no se formula debidamente.

Por ejemplo, suponga que quiere ajustar un modelo lineal a los datos que se muestran de la figura 12.9(a) y se piensa que una variable predictora x contribuye con la gran mayoría de la información para la predicción de y. Si se ajusta un modelo de primer orden

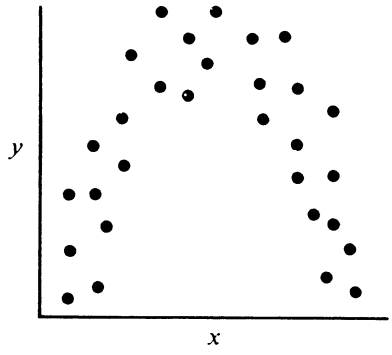
$$y = \beta_0 + \beta_1 x + \epsilon$$

a esos datos, puede verse en la figura 12.9(b) que se obtiene un ajuste muy pobre.

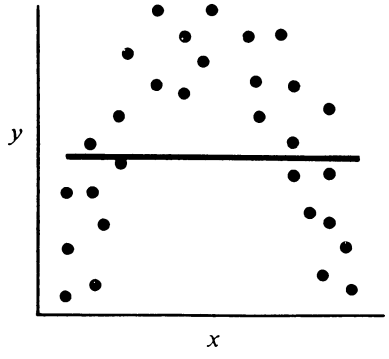
Si, por el contrario, se ajusta un modelo de segundo orden

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

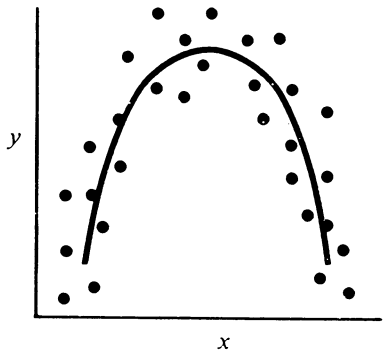
se obtiene un ajuste muy bueno (figura 12.9(c)).



(a) Los datos



(b) Un ajuste de un modelo de primer orden



(c) Un ajuste de un modelo de segundo orden

Figura 12.9 *Una comparación entre el ajuste de modelos de primero y segundo orden a los mismos datos*

Similarmente, si se ajusta un modelo lineal de primer orden a un conjunto de datos asociados a una superficie curva y torcida, se obtendrá un ajuste pobre. La única manera con la que se puede mejorar ese ajuste es usando un modelo de segundo orden (o mayor orden) que tenga la suficiente flexibilidad para ajustarse a la curvatura y la torsión de la verdadera superficie de respuesta.

Aprender a plantear el modelo apropiado en una situación particular requiere de experiencia. Se puede obtener un poco de experiencia dentro de la construcción de modelos, estudiando los ejemplos desarrollados de este capítulo. Para una información más completa en este tema puede consultar los textos que aparecen en las referencias.

Ejercicios

12.15. Para entender mejor los modelos para la respuesta media $E(y)$, grafique los siguientes polinomios de segundo orden:
a. $E(y) = 2x^2$
b. $E(y) = -2x^2$
c. $E(y) = 1 - 2x + x^2$
d. $E(y) = 1 + 2x + x^2$
e. $E(y) = 5 + 2x + x^2$

12.16. Grafique los polinomios de tercer orden:
a. $E(y) = 1 - 2x + x^2 - 3x^3$
b. $E(y) = 1 - 2x - x^2 + 3x^3$

12.17. Suponga que la respuesta media para valores dados de x_1 y x_2 está dada por

$$E(y) = 3 - x_1 + 2x_2$$

Grafique $E(y)$ como una función de x_1 para cada uno de los valores de x_2 iguales a 0, 1 y 2.

12.18. Suponga que la respuesta media para valores dados de x_1 y x_2 está dada por la ecuación

$$E(y) = 4 - 2x_1 + x_2 + 4x_1^2 + 2x_2^2$$

Grafique $E(y)$ como una función de x_1 (en el intervalo $0 \leq x_1 \leq 5$) para cada uno de los valores de x_2 iguales a 0, 1 y 2. Note que con excepción de translaciones verticales, las gráficas son partes de parábolas idénticas.

12.19. Continúe con el ejercicio 12.18 adicionando un término de producto cruzado al modelo.

Observe ahora que la forma de las gráficas depende del valor asignado a x_2. Por ejemplo, si

$$E(y) = 4 - 2x_1 + x_2 - 3x_1x_2 + 4x_1^2 + 2x_2^2$$

Grafique $E(y)$ para $x_2 = 0$, 1 y 2.

12.20. La relación entre tasas de interés e industria de construcción de viviendas es bien sabida. Las tasas de interés altas hacen que las mensualidades para pago de hipotecas sean altas. Por ejemplo, si se incrementa en 1% la tasa de interés se estará aumentando en \$20.83 el pago mensual de una hipoteca a 30 años por \$25,000. Por otro lado, las tasas de interés se ven afectadas por diversos factores económicos como lo son la disponibilidad de capitales y los rendimientos de los bonos estatales. Típicamente, cuando la oferta de capitales baja y los bonos aumentan en su rendimiento, las tasas de interés suben.

Suponga que le ha sido asignada la tarea de investigar la relación entre el número de construcciones que se inician y la tasa de interés x_1 para hipotecas convencionales, la oferta de capitales x_2 y el rendimiento x_3 de los bonos del estado. La investigación se haría con datos mensuales de los últimos tres años. Escriba un modelo de segundo orden que represente esta relación. ¿Espera usted que los términos de interacción sean de importancia en este análisis? Explique.

12.11 Construcción de modelos: prueba de una parte del modelo

Esta sección trata el problema de probar hipótesis consistentes en que uno o varios de los parámetros β son cero. Por ejemplo, una de las pruebas descritas

en el listado de computadoras de los datos sobre las horas frente al televisor de la sección 12.4, utiliza una estadística F para probar la hipótesis de que $\beta_1 = \beta_2 = \beta_3 = 0$. La estadística F también es utilizada para probar la hipótesis de que un β en particular es cero. En esta sección se explican con más detalle las razones que respaldan estas pruebas y se describen varias situaciones en las que estas pruebas son aplicables.

Suponga que tiene el modelo

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

o, equivalentemente

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

y que desea saber si ciertas variables contribuyen con información para la predicción de y. En otras palabras, se pregunta si los términos correspondientes a esas variables deben estar en el modelo.

Si un conjunto de variables x no contribuyen con información alguna para la predicción de y, entonces sus parámetros β debieran ser iguales a cero. En consecuencia, el probar si ciertas variables x deben incluirse en el modelo es equivalente a probar la hipótesis de que ciertos parámetros β son cero.

Suponga que se tienen dos modelos para $E(y)$, uno que es referido como "modelo completo" (llame a éste, el modelo 2) y otro que es referido "modelo reducido" (modelo 1). El modelo reducido incluye sólo parte de los términos del modelo completo. O dicha de otra forma, el modelo completo consta de los términos del modelo reducido y de algunos otros términos adicionales. El propósito de la prueba es el probar la hipótesis de que los parámetros β asociados a estos términos adicionales son cero. En otras palabras, se prueba si los términos adicionales contribuyen con información para la predicción de y.

Se representan los modelos reducido y completo por:

modelo 1 (reducido): $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_g x_g$

modelo 2 (completo): $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_g x_g + \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k$

Observe que el modelo completo consta, además de los términos del modelo reducido, de los términos adicionales, $\beta_{g+1}x_{g+1}$, $\beta_{g+2}x_{g+2}$, . . . , $\beta_k x_k$.

La prueba se describe intuitivamente. Se utiliza el método de mínimos cuadrados para ajustar el modelo reducido y calcular la suma de cuadrados del error SCE_1 (la suma de cuadrados de las desviaciones entre las observaciones y y los correspondientes valores que da el modelo ajustado). Después, se ajusta el modelo completo y se calcula la suma de cuadrados del error, SCE_2. Se comparan las dos sumas de cuadrados del error, SCE_1 y SCE_2. Si las variables $x_{g+1}, \ldots, x_k$ de verdad contribuyen con información para la predicción de y, entonces SCE_2 debiera ser mucho menor (significativamente menor) que SCE_1. Esto es, el incorporar estas variables al modelo produce una reducción en la suma de cuadrados de los errores de predicción. En consecuencia, mientras más grande resulte la diferencia $(SCE_1 - SCE_2)$, más grande será la evidencia de que los términos deben incluirse en el modelo. En otras palabras, habrá más evidencia que indique que al menos uno de los parámetros $\beta_{g+1}, \beta_{g+2}, \ldots, \beta_k$ difiere de cero.

Puede mostrarse en general, que cada vez que se adicionan términos a un modelo, se produce una reducción en la suma de cuadrados del error. La pregunta de interés es si esa reducción en la suma de cuadrados del error es debida al azar o efectivamente debida al hecho de que los nuevos términos contribuyen con información para la predicción de y. La respuesta viene dada en función de la magnitud de la reducción.

Para probar la hipótesis de que las variables $x_{g+1}, \ldots, x_k$ no proporcionan información para predecir y (esto es $\beta_{g+1} = \cdots = \beta_k = 0$), se utiliza la estadística

$$F = \frac{(\text{SCE}_1 - \text{SCE}_2)/(k-g)}{\text{SCE}_2/(n-k-1)} = \frac{(\text{Reducción en la SCE})/(k-g)}{s^2_{\text{del modelo completo}}}$$

Cuando se satisfacen las suposiciones usuales mencionadas en las primeras secciones del capítulo—que los valores de y se distribuyen normal e independientemente, con media $E(y)$ y varianza σ^2—entonces esta estadística F tiene una distribución F con $\nu_1 = (k-g)$ y $\nu_2 = (n-k-1)$ grados de libertad. Note que $\nu_1 = (k-g)$ es igual a la diferencia del número de parámetros entre el modelo completo y el modelo reducido. También obsérvese que $\nu_2 = (n-k-1)$ es igual al número de datos n, menos el número de parámetros en el modelo completo.

Como se hizo notar anteriormente mientras más grande sea la reducción en SCE (el numerador de la estadística F), más evidencia se tendrá para rechazar la hipótesis nula y aceptar la hipótesis alternativa de que por lo menos uno de los parámetros $\beta_{g+1}, \ldots, \beta_k$ es diferente de cero. De lo anterior que se rechace la hipótesis nula

$$H_0: \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$$

cuando la F es demasiado grande. Esto es, se usa una prueba de una cola y se rechaza H_0 cuando F es mayor que un valor crítico F_α, como se muestra en la figura 12.10.

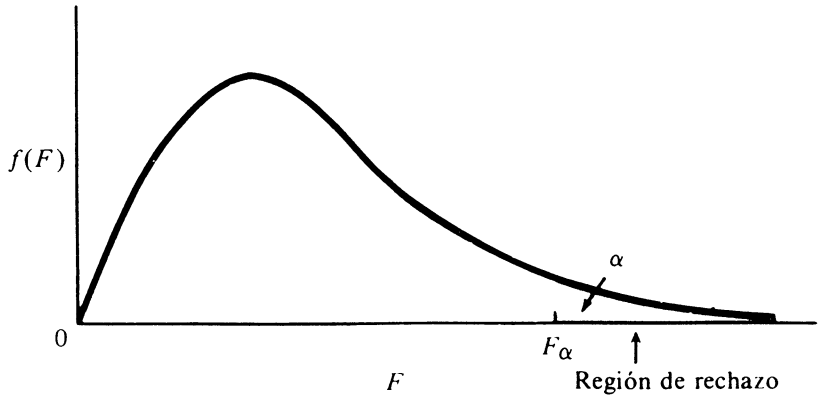


Figura 12.10 *Región de rechazo para la prueba F de* $H_0: \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$

Ejemplo12.10 Refiérase al ejemplo 12.9. El modelo de segundo orden

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \epsilon$$

fue ajustado a los datos que relacionan la ganancia porcentual y, con el tamaño de contrato de construcción x_1 y con los años de experiencia x_2 del superintendente de la obra para $n = 18$ proyectos de construcción, realizados por una compañía constructora. Pruebe la hipótesis

$$H_0: \beta_3 = \beta_4 = 0$$

esto es, pruebe que el modelo de segundo orden no proporciona en realidad ninguna mejora en relación al modelo de primer orden en cuanto a la predicción de la ganancia porcentual y.

Solución Se utilizó un programa estándar de análisis de regresión para ajustar el modelo de primer orden (el modelo reducido)

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

a los datos del ejemplo 12.9. Los resultados obtenidos son:

```
R MULTIPLE              .5731
R CUADRADA              .3284
DESV. EST. ESTIM.     1.9993

ANALISIS DE VARIANZA

                 SUMA DE   CUADRADO
            GL  CUADRADOS   MEDIO    COCIENTE F
REGRESION    2    29.321    14.660      3.668
RESIDUAL    15    59.957     3.997

ANALISIS INDIVIDUAL DE LAS VARIABLES

   VARIABLE      COEFICIENTE  DESV.ESTANDAR  VALOR F
(CONSTANTE          8.0136)
TAMAÑO CONT.       -1.3548        .5530        6.0012
EXPER.               .4626        .4346        1.1333
```

El listado de computadora para el modelo *reducido* muestra que

$$\text{SCE}_1 = 59.957$$

Del listado correspondiente al modelo *completo*, que aparece en la solución del ejemplo 12.9, se obtiene

$$\text{SCE}_2 = 12{,}252$$

Como $n = 18$, $k = 4$ y $g = 2$, la estadística para la prueba es

$$F = \frac{(\text{SCE}_1 - \text{SCE}_2)/(k - g)}{\text{SCE}_2/(n - k - 1)} = \frac{(59.957 - 12.252)/(4 - 2)}{12.252/13}$$

$$= \frac{23.853}{.942} = 25.322$$

Este valor calculado de la F excede al valor crítico de $F_{.05} = 3.81$ (con $\nu_1 = 2$ y $\nu_2 = 13$ grados de libertad), así que se rechaza la hipótesis nula y se concluye que el modelo de segundo orden sí proporciona una mejora sobre el de primer orden en cuanto a la predicción de la ganancia porcentual y como función del tamaño x_1 del contrato de construcción, y de la experiencia x_2 del superintendente de obra.

Ejercicios

12.21. El gerente de ventas de una compañía que surte hoteles y restaurantes, se interesa en la relación de tipo predictivo que pudiera tener tanto la publicidad como el tamaño de la fuerza de ventas, en las ventas mensuales. Para tal efecto, llevó un registro de las ventas mensuales y, la cantidad x_1 gastada mensualmente en publicidad directa y el número x_2 de representantes de ventas. Lo anterior lo hizo para cada una de doce regiones (territorios) de ventas seleccionadas al azar. Los datos de su registro se muestran en la tabla.

y (× \$10,000)	x_1 (× \$1,000)	x_2	y (× \$10,000)	x_1 (× \$1,000)	x_2
25.9	9.75	3	30.3	12.20	3
27.1	9.20	2	31.0	13.86	4
27.9	10.00	3	31.7	13.50	3
29.0	11.04	3	33.0	14.20	4
29.8	10.80	2	34.1	16.30	4
30.2	12.00	3	36.2	17.50	4

El modelo

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

fue ajustado a los datos usando el programa BMDP1R (vea la sección 12.12) que dió los siguientes resultados:

```
    NOMBRE REGRESION ..... REGRESION BMD #3
    VARIABLE DEPENDIENTE ..... 4
    TOLERANCIA ..... .0100

TODOS LOS DATOS CONSIDERADOS COMO UN SOLO GRUPO

R MULTIPLE          .9809
R CUADRADA          .9621
DESV. EST. ESTIM.   .6734

ANALISIS DE VARIANZA

            SUMA DE       CUADRADO
           CUADRADOS  GL    MEDIO    COCIENTE F  P(COLA)
REGRESION   92.110    3    30.703      67.717      .0001
RESIDUAL     3.627    8      .453
```

ANALISIS INDIVIDUAL DE LAS VARIABLES

VARIABLE		CO-EFICIENTE	DESV. ESTANDAR	COEF. ESTANDAR DE RE-GRESION	T	P (2 COLAS)
ORDENADA		14.9600				
X1	1	1.5231	.5910	1.3582	2.5925	.0345
X2	2	−.4323	1.7964	−.1052	.2407	.8870
X1X2	3	−.0553	.1554	−.3165	.3557′	.7146

a. Haga un análisis completo de los resultados del listado.

b. Use los métodos del capítulo 11 para ajustar el modelo simple de primer orden

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

a los datos de y y x_1. Calcule SCE para este modelo.

c. Con los resultados de (b) y los del análisis del listado de (a), determine si las ventas en realidad pueden describirse satisfactoriamente sólo por la cantidad gastada directamente en publicidad. En otras palabras, en el modelo

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

pruebe la hipótesis $H_0: \beta_2 = \beta_3 = 0$.

12.22. Suponga que se desea probar la hipótesis de que determinadas variables de un modelo de regresión son insignificantes (no significativas), en cuanto a su capacidad predictora para la variable dependiente y en presencia de otras variables. ¿Por qué puede uno llegar a conclusiones falsas si se basa, para lo anterior, en los valores F (o valores t) asociados a cada uno de los parámetros por separado?

12.12 Un resumen de procedimientos para regresión múltiple

En las secciones anteriores se vió que con excepción de la prueba descrita en la sección 12.11, las mismas pruebas y procedimientos de estimación y predicción disponibles para el modelo de regresión múltiple, lo están también para el de regresión lineal simple. Para un modelo lineal

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

los procedimientos de prueba, estimación y predicción son los siguientes:

1. *Prueba de la hipótesis de que uno de los parámetros, digamos β_i, es igual a cero.* Esto puede hacerse con una prueba t del tipo de la usada en el capítulo 11 en donde t se basa en $(n - k - 1)$ grados de libertad. (Note que n es el número de observaciones o datos y k es el número de variables independientes en el modelo.) También puede usarse una prueba F como se indica en el listado de computadora para los datos del ejemplo de las horas frente al televisor. La estadística F se basa en $\nu_1 = 1$ y $\nu_2 = (n - k - 1)$ grados de libertad (vea los comentarios de precaución relativos a la interpretaeión en la sección 12.6).

2. *Un intervalo de confianza para un parámetro de regresión individual, digamos* β_i. El intervalo de confianza es de la forma

$$\hat{\beta}_i \pm t_{\alpha/2} s_{\hat{\beta}_i}$$

en donde la t de tablas está basada en $(n - k - 1)$ grados de libertad y $s_{\hat{\beta}_i}$ es la desviación estándar estimada de $\hat{\beta}_i$. Esta última cantidad se muestra en el listado de cualquier programa de regresión.

3. *Un intervalo de confianza para el valor medio de y para valores dados de* $x_1, x_2, \ldots, x_k$. El estimador para $E(y)$, $\hat{y}$, se obtiene de la ecuación de predicción sustituyendo los valores dados de $x_1, x_2, \ldots, x_k$. El intervalo de confianza para $E(y)$ se da por una fórmula complicada que no se trata en este texto. Muy pocos de los programas de regresión proporcionan este intervalo de confianza opcionalmente.

4. *Un intervalo de predicción para un valor futuro de y.* El valor pronosticado para y, para valores específicos de $x_1, x_2, \ldots, x_k$ se obtiene de la ecuación de predicción sustituyendo los valores dados de $x_1, x_2, \ldots, x_k$. La ecuación de predicción la dan la mayoría de los programas de regresión. La fórmula para el intervalo de predicción para y es parecida a la del intervalo de confianza para $E(y)$ (párrafo 3) y es muy complicada. Desafortunadamente sólo unos cuantos programas de regresión para computadora tienen este intervalo como opción.

5. *Una prueba para la hipótesis de que uno o varios de los parámetros* β *son cero simultáneamente.* Esta prueba F, que puede usarse para probar la hipótesis de que un parámetro es cero (véase el párrafo 1), tiene su uso principal en la construcción de modelos (sección 12.11).

Tabla 12.7 *Opciones disponibles para algunos programas de regresión*

PROGRAMA Y PROVEEDOR	LISTADO SECUENCIAL O ESTANDAR	t o F (párrafo 1)	INTERVALO DE CONFIANZA PARA MEDIA DE y (párrafo 3)	INTERVALO DE PREDICCION PARA y_i (párrafo 4)
BMD-P1R (UCLA Biomedical Computing Facilities)	estándar	t	no	no
BMD-P2R (UCLA Biomedical Computing Facilities)	secuencial	t	no	no
BMD-02R (UCLA Biomedical Computing Facilities)	secuencial	F	no	no
SSP-MULTR (IBM Corp.)	estándar	t	no	no
SSPS-REGRESSION University of Chicago	opcional	F	no	no
SAS (SAS Institute Raleigh, N.C.)	opcional	t	sí	sí
MINITAB Pennsylvania State University	estándar	t	sí	sí

Existe un buen número de programas de regresión para computadora (paquetes) para calcular las cantidades mencionadas y todos son fáciles de usar. Debe usted averiguar primero cuáles de ellos le están disponibles y después familiarizarse con su uso. Los listados de los distintos paquetes no son idénticos, por lo que debe decidir cuál de ellos usar dependiendo de sus necesidades. Algunos imprimen el valor de t para probar la hipótesis de que uno de los parámetros β_i es cero, mientras que otros imprimen el valor F; algunos imprimen un intervalo de confianza para el valor medio de y para valores fijos de $x_1, x_2, \ldots, x_k$ y otros simplemente no lo hacen. La tabla 12.7 le presenta un resumen de las distintas opciones que cada uno de los paquetes más usuales, le ofrece.

12.13 Ejemplos resueltos

Para ganar una poca de experiencia adicional en la interpretación de resultados de un análisis de regresión, se han incluido en esta sección varios ejemplos resueltos. Los análisis (y listados) se hicieron usando distintos programas de regresión para ilustrar la variedad de presentaciones que puede uno encontrar en la práctica.

Ejemplo 12.11 El gerente de ventas de una compañía farmacéutica está preocupado por un aparente rendimiento menor de sus agentes más experimentados. Ha observado que mientras más años de experiencia tengan sus agentes las ventas hechas por ellos no sólo se estabilizan sino que en algunos casos decrecen. Para estudiar este problema, el gerente de ventas ha registrado las ventas territoriales (por territorio de ventas) habidas en los últimos tres meses y los años de experiencia de cada uno de los agentes responsables de cada uno de los diez territorios estudiados.

VENTAS, y (× \$1000)	EXPERIENCIA, x (años)	VENTAS, y (× \$1000)	EXPERIENCIA, x (años)
36.7	2.0	41.2	4.5
22.9	1.5	18.5	1.0
30.5	4.5	43.4	3.0
9.2	.8	25.5	2.3
38.4	3.5	28.4	5.5

a. Grafique la relación entre ventas territoriales y experiencia del agente.
b. Ajuste y a x usando un modelo polinomial de segundo orden. ¿Proporciona el modelo de segundo orden un buen ajuste a estos datos?

Solución a. Una gráfica que describe la relación entre las ventas y y la experiencia x se muestra en la figura 12.11. Esta gráfica sugiere que un modelo de segundo orden

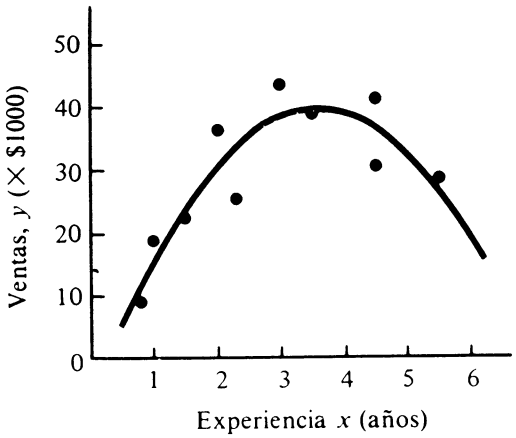


Figura 12.11 *Una gráfica de los datos de ventas y experiencia de trabajo del ejemplo 12.11 con un modelo de segundo orden ajustado*

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

proporcionará un buen ajuste a los datos.

b. El programa de regresión BMD-P1R fue usado para el ajuste del modelo de segundo orden. Los resultados son:

```
    NOMBRE REGRESION ..... REGRESION BMD #2
    VARIABLE DEPENDIENTE ..... 3Y
    TOLERANCIA ..... .0100

TODOS LOS DATOS CONSIDERADOS COMO UN SOLO GRUPO

R MULTIPLE              .8961
R CUADRADA              .8029
DESV. EST. ESTIM.     5.4526

ANALISIS DE VARIANZA

            SUMA DE          CUADRADO
           CUADRADOS  GL      MEDIO    COCIENTE F  P(COLA)
REGRESION   847.880    2     423.940      14.259     .00340
RESIDUAL    208.120    7      29.731

ANALISIS INDIVIDUAL DE LAS VARIABLES

                                  COEF.
                                ESTANDAR
                    CO-    DESV.  DE RE-              P
VARIABLE        EFICIENTE ESTANDAR GRESION      T   (2 COLAS)
ORDENADA        -6.173
X1      1       25.332    5.439    3.770    4.657     .002
X2      2       -3.499     .872   -3.249   -4.014     .005
```

Las cantidades más importantes del listado se interpretan en los párrafos siguientes:

R^2 (*coeficiente de determinación*). La proporción de la variación en las ventas explicada por el modelo de segundo orden con la variable explicativa "experiencia de trabajo" es

$$R^2 = .8029$$

Esto es, aproximadamente el 80% de la variación total la explica el modelo.

Cociente F. El cociente F es 14.259 y se asocia a una probabilidad, P(COLA) de .0034. Esta es la probabilidad de observar un valor igual o mayor que el de la $F = 14.259$; en consecuencia, la probabilidad, P(COLA) es la significancia del valor observado de la F. De lo anterior se rechaza la hipótesis nula

$$H_0: \beta_1 = \beta_2 = 0$$

con un $\alpha = .0034$ (o cualquier otro nivel mayor, por ejemplo $\alpha = .05$). Lo anterior proporciona evidencia clara de que por lo menos uno, o los dos, de los términos $\beta_1 x$ y $\beta_2 x^2$ contribuyen con información para predecir y.

Coeficientes de las variables. El modelo estimado de segundo orden que relaciona ventas y con la experiencia de trabajo x es

$$\hat{y} = -6.173 + 25.332x - 3.499x^2$$

Valor t. Los valores para la t listados dan los valores calculados para la estadística t y las probabilidades asociadas, P(2COLAS) (significancias), para las pruebas de significancia sobre los parámetros β individuales. Los primeros valores $t = 4.657$ y P(2COLAS) = .002 indican que se rechazaría la hipótesis

$$H_0: \beta_1 = 0$$

con un nivel de significancia aun tan chico como $\alpha = .002$. El segundo valor de t tiene una interpretación análoga para β_2.

Como se hizo notar previamente, la parte más importante de un listado de computadora es el análisis de varianza y la prueba F. Esta prueba nos dice que el modelo usado contribuye con información para la predicción de y. Al mismo tiempo un valor de R^2 tan bajo como .8029 indica que todavía se puede mejorar el modelo. Parece dudoso que el incluir más términos de mayor orden en x pueda mejorar mucho el ajuste; lo que sí parece factible es que el ajuste se mejore al adicionar al modelo otras variables relacionadas con las ventas.

Ejemplo 12.12 En un determinado condado de los Estados Unidos, se ha desarrollado un modelo que relaciona el valor comercial de viviendas unifamiliares con el tamaño (en superficie construida) y el número de dormitorios de que consta la residencia. El uso que se le da a la ecuación de predicción resultante es el de estimar el valor comercial de las viviendas del condado para estimar así

los impuestos correspondientes. Para ello, se registraron los valores comerciales y, la superficie construida x_1, en pies cuadrados, y el número de dormitorios x_2, de 20 residencias unifamiliares vendidas recientemente a un precio que se consideró razonable.

Los datos se dan en la tabla.

y (× \$1000)	x_1 (× 100)	x_2	y (× \$1000)	x_1 (× 100)	x_2
20.9	8.40	2	32.6	14.75	3
21.7	8.65	2	34.7	17.50	4
22.5	10.25	3	35.8	19.20	4
24.0	9.60	2	38.9	18.70	4
26.1	10.50	2	40.0	18.50	3
27.3	11.25	3	42.1	19.00	4
27.5	14.70	3	44.6	17.50	3
29.2	12.80	2	47.9	19.50	4
30.4	13.85	3	49.3	19.00	4
30.5	15.00	4	52.4	19.45	4

El programa de regresión SAS fue usado para ajustar a los datos un modelo de segundo orden,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

El listado de computadora obtenido de dicho análisis se muestra a continuación.

```
                                SAS
                        REGRESION MULTIPLE
              PROCEDIMIENTO GENERAL PARA MODELOS LINEALES

VARIABLE DEPENDIENTE:  PRECIO DE VENTA

                                SUMA DE      CUADRADO
     FUENTE                GL  CUADRADOS     MEDIO       VALOR F
MODELO                      4   1519.791     379.948      23.495
ERROR                      15    242.569      16.171
TOTAL CORREGIDO            19   1762.360

R CUADRADA      C.V.       DESV. EST.    MEDIA PRECIO VENT.
   .8623      11.8555        4.0214           33.9200

                                T PARA H0:
                                PARAMETRO                  DESV. EST.
     PARAMETRO          ESTIMADO   = 0      PR > |T|      DEL ESTIM.
ORDENADA                24.6970
SUPERFICIE (X1)         -1.2520    -.5096     .5873          2.4569
X1 CUADR.                .0933     .8771     .4250           .1064
X2 CUADR.               -1.1157   -1.1225     .2465          .9939
(X1)(X2)                 .3565     .8317     .4345           .4286

VARIABLE NUM. DORMITORIOS (X2) REDUNDANTE Y QUITADA
DEL LISTADO
```

Dé un análisis del listado y bosqueje una gráfica para el modelo de segundo orden ajustado por la computadora.

Solución Es interesante hacer notar que el usuario del programa SAS (y de muchos otros programas) puede controlar el hecho de que las variables que resulten claramente redundantes no queden incluidas en el modelo. En este listado, la variable x_2, número de dormitorios, fue eliminada del análisis pues su valor de t no fue mayor que un valor de control predeterminado, $t = .1$. Aún cuando la selección del valor predeterminado para t o F es arbitraria, usualmente se eligen $t = .1$ y $F = .01$ como los valores de control para identificar variables como redundantes si sus valores correspondientes de t o F no exceden a estos valores de control.

La variable x_2 no quedó como variable independiente en el modelo, sin embargo los términos de segundo orden en x_2, esto es x_2^2 y x_1x_2 si quedaron incluidos como variables predictoras no redundantes.

R CUADRADA (*coeficiente de determinación*). Como el coeficiente de determinación R^2 es 0.8623, se dice que el 86.23% de la variabilidad de los precios de venta de estas 20 residencias se explica por un modelo de segundo orden en x_1 = superficie construída y x_2 = número de dormitorios.

Valor F. El valor $F = 23.495$ indica que se debe rechazar la hipótesis de que todos los parámetros $\beta_1, \beta_2, \ldots, \beta_5$ son cero en el modelo de segundo orden

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$$

ya que el valor tabulado para la F con 4 y 15 grados de libertad es 4.89 para $\alpha = .01$. Esto es, por lo menos uno de los parámetros es significativamente distinto de cero.

Parámetros estimados. El modelo de segundo orden ajustado es

$$\hat{y} = 24.6970 - 1.2520x_1 + .0933x_1^2 - 1.1157x_2^2 + .3565x_1x_2$$

T PARA HO: PARAMETRO = 0. Cada valor de t es para probar la hipótesis

$$H_0: \beta_j = 0$$

para el parámetro de regresión correspondiente, β_j. Recuerde sin embargo, que ésta no es una prueba de significancia de la variable correspondiente como predictor para y. Dado que el valor tabulado para t con 15 grados de libertad y $\alpha = .01$ es 2.602, no se rechaza la correspondiente H_0 para ninguno de los cuatro parámetros, lo cual es una aparente contradicción con el resultado de la F sobre la significancia de los parámetros en conjunto. Lo que en realidad ocurre con el valor de t de cada parámetro es que ninguno de los términos x_1, x_1^2, x_2^2 ó x_1x_2 resultan significativos como predictores para y *en presencia de los restantes tres predictores* (posiblemente por la duplicidad en la información que contienen). Sin embargo, cuando se consideran individualmente (en ausencia de los otros) cualquiera de los términos o quizás todos pueden ser predictores valiosos para y. Una gráfica del modelo ajustado

$$\hat{y} = 24.6970 - 1.2520x_1 + .0933x_1^2 - 1.1157x_2^2 + .3565x_1x_2$$

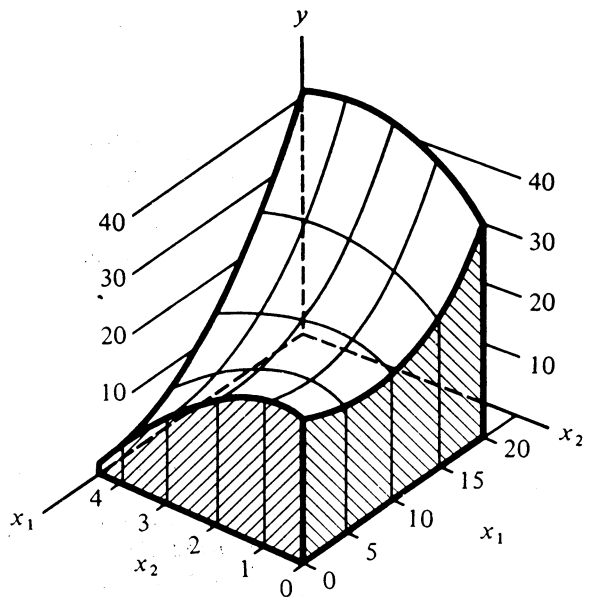


Figura 12.12 *Gráfica del modelo de segundo orden ajustado a los datos del ejemplo 12.12*

que relaciona el precio de venta con la superficie y el número de dormitorios se presenta en la figura 12.12.

12.14 Resumen

Un análisis de regresión múltiple es una extensión del análisis del modelo de regresión simple del capítulo 11 al caso en el que la variable de respuesta y está relacionada a varias variables predictoras $x_1, x_2, \ldots, x_k$. Todas las pruebas y procedimientos de estimación y predicción del capítulo 11 se aplican al modelo general de regresión lineal del capítulo 12. Aun los coeficientes múltiples tanto de correlación R, como de determinación R^2, tienen significados similares a los coeficientes de correlación r y de determinación r^2, vistos en el capítulo 11. La principal diferencia entre los modelos de regresión lineales simple y múltiple es la aplicabilidad de este último. Muy pocas variables de respuestas en las aplicaciones a negocios quedan adecuadamente modeladas por el modelo probabilístico simple

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

del capítulo 11. Por el contrario, el modelo de regresión múltiple, cuando se construye con cuidado, proporciona un modelo muy bueno para muchas de las aplicaciones en los negocios. Con frecuencia, la ecuación de predicción resultante proporciona muy buenos estimadores para la respuesta media asociada a valores fijos de las variables predictoras y también proporciona pronósticos adecuados para valores futuros de la respuesta.

Ejercicios complementarios

12.23. Comente la relación que guardan los siguientes conceptos con el modelo de predicción en varias variables:
a. las ecuaciones de mínimos cuadrados
b. variable independiente cuantitativa
c. variable independiente cualitativa
d. términos de interacción en un modelo de regresión
e. modelo de regresión de segundo orden

12.24. ¿Qué se quiere decir en el análisis de regresión por construcción de modelos? ¿Por qué es importante la construcción de modelos?

12.25. En Estados Unidos, el mercado de valores ha cambiado considerablemente en los últimos diez años. Como resultado de ello y para crear un mayor interés en los inversionistas, se ha hecho mucho más eficiente el servicio que les proporcionan los agentes de bolsa. Un aspecto interesante al respecto es la relación entre el grado de participación de un agente y la diferencia entre el número de acciones solicitadas y ofrecidas. Algunos investigadores* han estudiado lo anterior haciendo la regresión de

y = grado de participación del agente medido en número de operaciones de compra-venta hechas a través de él.

respecto a

x_1 = diferencias entre solicitudes y ofrecimientos ($\times$ 1000)

x_2 = valor promedio de la acción en 1961

x_3 = (precio máximo − precio mínimo)/x_2

x_4 = un cierto factor de actividad

El ajuste se hizo con $n = 65$ tipos de acciones. El análisis de regresión correspondiente se da en la tabla.

VARIABLE	COEFICIENTE	DESVIACION ESTANDAR	VALOR t
constante	12759.050		
x_1	−.905	.354	−2.557
x_2	33.227	6.893	4.821
x_3	298.117	1181.811	.252
x_4	706.343	47.621	14.833

$R^2 = .812$ *Error Estándar* = 2730.070 $F = 64.760$

*S. Tinic and R. West, "Competition and the Pricing of Dealer Service in the Over-the-Counter Stock Market," *Journal of Financial and Quantitative Analysis* (junio de 1972).

a. ¿Afecta la diferencia entre solicitudes y ofertas el grado de participación del agente en presencia de las otras variables? (Establezca la hipótesis apropiada y pruébela).
b. ¿Proporciona un buen ajuste a los datos el modelo de primer orden?

12.26. Un representante de un sindicato regional quiere hacer un modelo de predicción en varias variables para predecir el salario por hora de los trabajadores usando su edad, años de experiencia y el número de años que llevan de pertenecer al sindicato. Si se tuvieran datos sobre 75 de los trabajadores, tanto hombres como mujeres, de cinco empresas distintas ubicadas en dos estados diferentes, ¿qué recomendaría usted en cuanto a otras variables independientes, su tipo y sus niveles?

12.27. ¿Produce la competencia entre productos siempre menores precios al consumidor? Algunos especialistas en mercados sugieren que la relación entre el precio al consumidor y el número de productos competitivos es de segundo orden; esto es, que los precios son altos en presencia de muy pocos productos competitivos o de muchos productos competitivos, y que los precios son bajos en presencia de un número moderado de productos competitivos. Para estudiar lo anterior, se hizo un registro de los precios promedio de las cervezas (en presentación de 6) y el número de marcas distintas ofrecidas en una cierta cadena de supermercados en 12 ciudades. Los datos registrados son:

PRECIO, y	NUMERO DE MARCAS COMPETITIVAS, x	PRECIO, y	NUMERO DE MARCAS COMPETITIVAS, x
$1.25	4	$1.56	10
1.40	7	1.34	5
1.60	2	1.45	8
1.52	9	1.41	2
1.40	3	1.27	7
1.35	6	1.55	8

Ajuste un modelo de segundo orden $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ a estos datos resolviendo las ecuaciones de mínimos cuadrados, como se hizo en el ejemplo 12.1, o por medio de algún programa de regresión para computadora al cual tenga acceso. Interprete los resultados con cuidado. ¿Sugieren estos datos que el número x de marcas competitivas se puede usar para predecir el precio y del producto bajo estudio?

12.28. La investigación en el desarrollo de nuevos productos es vital para la mayoría de las empresas

de manufactura. ¿Pero qué tanto se debe gastar en el desarrollo y promoción de nuevos productos? La respuesta depende de la redituabilidad esperada de estas operaciones y la cantidad invertida en el desarrollo y promoción de cada producto. El director de estudios de mercado de una empresa alimenticia registró la redituabilidad porcentual y la cantidad invertida en el desarrollo y promoción de 10 productos distintos lanzados al mercado por su compañía en los últimos años. Los datos se muestran en la tabla. Ajuste un modelo de segundo orden $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ resolviendo las ecuaciones de mínimos

REDITUABILIDAD PORCENTUAL, y	INVERSION, x (×$100,000)	REDITUABILIDAD PORCENTUAL, y	INVERSION, x (×$100,000)
4.0	1.1	5.6	2.5
9.2	3.5	8.0	3.1
6.0	4.7	6.1	5.9
1.8	1.0	3.2	2.1
7.7	5.2	7.8	4.6

cuadrados, como en el ejercicio 12.1 o mediante el uso de un programa de computadora. Interprete cuidadosamente los resultados de su análisis.

12.29. El gerente de producción de una planta que produce un fertilizante químico ha registrado los costos marginales de producción para varios niveles de producción que se observan en doce meses escogidos al azar. Los datos se muestran en la tabla.

COSTO, y (100 kg)	PRODUCCION, x (× 1000 kg)	COSTO, y (100 kg)	PRODUCCION, x (× 1000 kg)
30.0	1.9	40.5	5.5
29.6	4.5	23.4	3.9
15.1	3.5	15.0	2.4
41.2	7.1	34.9	6.0
22.6	2.6	20.5	1.1
21.0	1.9	34.6	4.5

Un modelo de tercer orden

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon$$

se seleccionó para describir el costo marginal y como función de la producción x. El análisis con el BMP-P1R produjo el siguiente listado:

```
NOMBRE DE REGRESION ..... REGRESION BMD #4
VARIABLE DEPENDIENTE ..... 4 Y
TOLERANCIA ..... .0100

TODOS LOS DATOS CONSIDERADOS COMO UN SOLO GRUPO

R MULTIPLE              .7840
R CUADRADA              .6146
DESV. EST. ESTIM.     6.2573

ANALISIS DE VARIANZA

            SUMA DE        CUADRADO
           CUADRADOS  GL    MEDIO    COCIENTE F  P(COLA)
REGRESION   562.002   2   281.001      7.177     .01369
RESIDUAL    352.383   9    39.154

                                 COEF.
                                ESTANDAR
                CO-      DESV.  DE RE-
VARIABLE      EFICIENTE ESTANDAR GRESION      T     P (2 COLAS)
ORDENADA      21.495
X1      1       .019    2.466     .004     .008       .994
X2      2                       VARIABLE REDUNDANTE
X3      3       .065     .044     .780    1.482       .172
```

¿Proporciona el modelo de tercer orden un ajuste adecuado a los datos? (En su respuesta refiérase a los mismos argumentos empleados en las secciones precedentes.)

12.30. El conocer la potencialidad de los territorios de venta permite planear sistemas de control y de incentivos para los