Capítulo 13. INTRODUCCION AL ANALISIS DE REGRESIÓN

En el amplio espectro del estudio de las relaciones entre variables, tanto cuantitativas como cualitativas, que involucra el proceso de investigación econométrica, existe la posibilidad de establecer relaciones funcionales. En algunas de ellas, la variación de una variable dependiente es explicada por las variaciones de una o más variables independientes, mediante funciones lineales que involucran adicionalmente un término aleatorio. En particular, en este capítulo se inicia el estudio de las relaciones lineales entre dos variables cuantitativas. Para ello, se estudia la forma de especificación y estimación de esas relaciones lineales y la bondad de las mismas.

13.1. Especificación del modelo

En el análisis de la información existen, frecuentemente, un número de variables claves que se convierten en el centro de atención del estudio.

El análisis de regresión proporciona una herramienta que puede cuantificar tales relaciones y proporcionar un control estadístico. Es un análisis de dependencia puesto que involucra una **variable dependiente,** como el punto de atención del análisis, la cual es explicada por las **variables independientes** o variables explicativas.

El análisis de regresión está orientado hacia

- la descripción, obtención de las relaciones entre variables independientes y una variable dependiente para un conjunto de observaciones.
- la predicción, niveles a alcanzar por la variable dependiente bajo el supuesto de comportamiento de la variable explicativa.

La construcción de un modelo de regresión, generalmente, empieza con la especificación de la variable dependiente y de la variable, o variables, independientes.

$$Y = f(X)$$

Y : variable dependiente o endógena

X : variable independiente o exógena.

Como no se puede explicar en forma perfecta el comportamiento de Y sólo por X, se adiciona un término de error, ε:

$$Y = f(X) + \varepsilon$$

A ésta se la denomina *ecuación de regresión* de Y sobre X. ε es el residuo, o error, y es una variable aleatoria. Este término hace referencia a error de medición de las variables, error de especificación del modelo y omisión de otras variables explicativas que influyen sobre el comportamiento de Y. Al ser ε variable aleatoria, también lo es Y.

Se explicita una relación lineal para $f(X)$

$$f(X) = \sim + \quad X$$

Si se dispone de n observaciones de Y y de X, puede especificarse el siguiente modelo de regresión.

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad i = 1, \cdots, n$$

donde,

Y = variable dependiente

X = variable explicativa

ε = perturbación aleatoria o término de error

$\sim$, = parámetros del modelo.

El modelo está basado en las siguientes consideraciones:

La relación especificada es lineal. Esto conduce a tener que linealizar aquellas relaciones funcionales no lineales a efectos de poder aplicar el análisis.

El término de error es una variable aleatoria donde:

- El promedio es cero.

$$E(\varepsilon_i) = 0 \quad \forall i$$

- La varianza es constante. Esto significa que el error no es más grande para valores grandes de X que para valores pequeños de X. Esto se lo conoce como “homocedasticidad”.

$$V(\varepsilon_i) = E(\varepsilon_i \varepsilon_j) = \sigma^2 \quad \forall\ i = j$$

- Los errores son independientes entre sí. Esto significa que los errores de una observación no dependen de la observación anterior. Esto se lo conoce como “no autocorrelación”.

$$Cov(\varepsilon_i \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \quad \forall\ i \neq j$$

- El error tiene distribución normal de media cero y varianza constante.

$$\varepsilon_i \sim N(0, \sigma^2)$$

La variable explicativa X es fija, no estocástica e independiente de los errores

$$Cov(\varepsilon_i X_i) = E(\varepsilon_i X_i) = 0$$

Bajo el cumplimiento de este conjunto de supuestos, los estimadores minimocuadráticos son lineales, insesgados y óptimos (ELIO) y es posible construir intervalos de confianza y realizar test para $\hat{\alpha}$ y $\hat{\beta}$.

13.2. Estimación de los parámetros del modelo

Los parámetros α y β se estiman por el método de “mínimos cuadrados”. Esto es elegir $\hat{\alpha}$ y $\hat{\beta}$ como estimadores de α y β, respectivamente, de tal manera que la suma de las diferencias al cuadrado entre el valor observado y el valor estimado de la variable dependiente sea mínima. Es decir, que tenga la siguiente propiedad

$$Q = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \text{mínimo}$$

La gráfica muestra la dispersión de la información proveniente de la observación de dos variables X e Y. La línea recta simula el modelo estimado. La diferencia entre el valor observado y la línea estimada es el error. Lo que se “minimiza” es la suma de cuadrados entre los puntos y la línea, en forma vertical.

Los parámetros α y β son las características de la relación entre X e Y. Reemplazando en Q, $\hat{Y}$ por su igual:

$$Q = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} X_i \right)^2$$

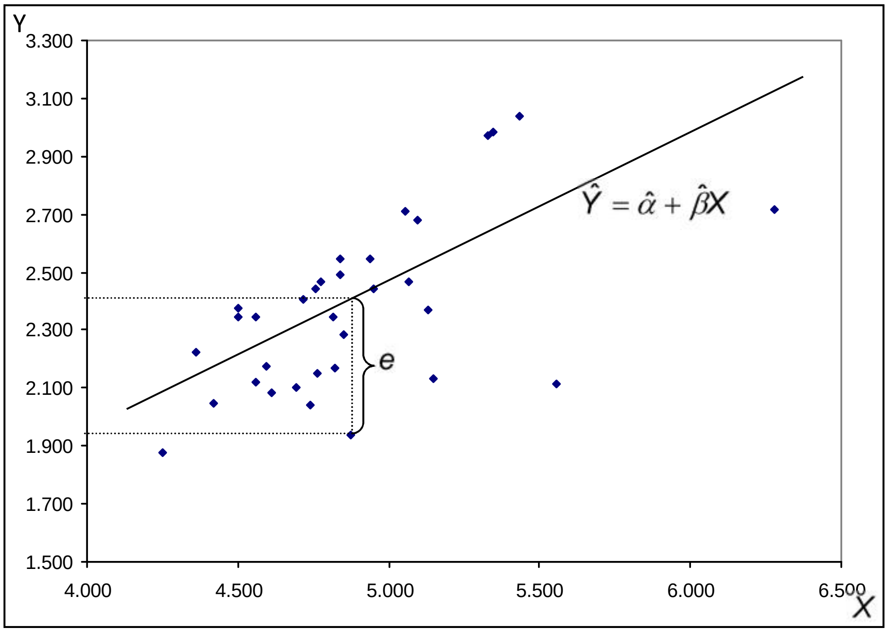

Figura 13.1 Gráfico de dispersión

Para que el valor de Q sea mínimo se debe derivar con respecto a $\hat{\alpha}$ y $\hat{\beta}$ e igualar a cero.

$$\frac{\partial Q}{\partial \hat{\alpha}} = 2\sum \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)(-1) = 0$$

$$\sum Y_i = n\hat{\alpha} + \hat{\beta}\sum X_i$$

$$\frac{\partial Q}{\partial \hat{\beta}} = 2\sum \left(Y_i - \hat{\alpha} - \hat{\beta} X_i\right)(-X_i) = 0$$

$$\sum X_i Y_i = \hat{\alpha}\sum X_i + \hat{\beta}\sum X_i^2$$

De esta manera tenemos un sistema de dos ecuaciones y dos incógnitas denominadas ecuaciones normales. De la primera se obtiene el valor de $\hat{\alpha}$

$$\sum Y_i = n\hat{\alpha} + \hat{\beta}\sum X_i$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$$

Reemplazando ésta en la segunda se obtiene $\hat{\beta}$

$$\sum X_i Y_i = \left(\bar{Y} - \hat{\beta}\bar{X}\right)\sum X_i + \hat{\beta}\sum X_i^2$$
$$\sum X_i Y_i = \left(\bar{Y} - \hat{\beta}\bar{X}\right)n\bar{X} + \hat{\beta}\sum X_i^2$$
$$\sum X_i Y_i - n\bar{X}\bar{Y} = \hat{\beta}\sum X_i^2 - \hat{\beta}n\bar{X}^2$$
$$\hat{\beta} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

$\hat{\alpha}$ y $\hat{\beta}$, se denominan **coeficientes de regresión** y se basan en la muestra aleatoria. Cada muestra aleatoria tendrá asociada un $\hat{\alpha}$ y un $\hat{\beta}$, *diferentes*. Una medida de esta variación de las estimaciones de los parámetros está dada por sus *errores estándar*. De este modo, así como la media muestral, $\bar{X}$, tiene un error estándar que es estimado por $\sigma_{\bar{X}}$; también $\hat{\alpha}$ y $\hat{\beta}$ tienen un error estándar asociado con cada uno de ellos, $s_{\hat{\alpha}}$ y $s_{\hat{\beta}}$.

$$s_{\hat{\alpha}} = \sqrt{\sigma_e^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum\left(X - \bar{X}\right)^2}\right]} \qquad s_{\hat{\beta}} = \sqrt{\frac{\sigma_e^2}{\sum\left(X - \bar{X}\right)^2}}$$

La varianza de los errores de la regresión, σ_e^2, se estima a partir de s_e^2

$$s_e^2 = \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum e^2}{n-2}$$

13.2.1 La interpretación de los parámetros estimados

Los parámetros tienen un significado muy preciso.

- El parámetro β indica que, si la variable X cambia en una unidad, la variable Y cambiará en β unidades.
- El parámetro α refleja el nivel que alcanza la variable dependiente cuando la variable explicativa asume el valor nulo.

Tal como se indicó anteriormente, la estimación de $\hat{\ }$, tiene una variación asociada con ella (medida por $S_{\hat{\ }}$), porque se basa en una muestra de individuos. Una forma de evaluar la magnitud de $\hat{\ }$, tomando en cuenta su variación, consiste en usar una **prueba estadística de hipótesis**. Para esto se calcula el estadístico t,

$$t = \frac{\hat{\ } -}{s_{\hat{\ }}}$$

Bajo la **hipótesis nula**, que será planteada como = 0, el estadístico t se reduce a,

$$t = \frac{\hat{\ }}{s_{\hat{\ }}} \sim t_{n-2}$$

Nuevamente, como se hizo en otras pruebas de hipótesis, este valor empírico de t se compara con el valor teórico. Si el valor empírico es superior, en valor absoluto, al valor teórico se rechaza la hipótesis nula y se decide que $\beta \neq 0$ y por lo tanto, la relación entre X e Y es significativa, a un valor de probabilidad α determinado.

13.2.2 Intervalos de confianza para los parámetros

Para construir intervalos de confianza para los parámetros de la regresión, se utiliza la distribución t, teniendo en cuenta que

$$\frac{\hat{\sim} - \sim}{s_{\hat{\sim}}} \sim t_{n-2} \qquad\qquad \frac{\hat{\ } -}{s_{\hat{\ }}} \sim t_{n-2}$$

Explícitamente, para el parámetro α es

$$P\left(t_{n-2;\alpha/2} \leq \frac{\hat{\alpha} - \alpha}{s_{\hat{\alpha}}} \leq t_{n-2;1-\alpha/2}\right) = 1 - \alpha$$

Resolviendo de igual manera a como se hiciera para estimar por intervalo el valor de μ, se tiene

$$P\left(\hat{\alpha} - s_{\hat{\alpha}} t_{n-2;\frac{\alpha}{2}} \leq \alpha \leq \hat{\alpha} + s_{\hat{\alpha}}\, t_{n-2;1-\alpha/2}\right) = 1 - \alpha$$

Para el parámetro β se procede de igual manera.

También es posible estimar el intervalo de confianza para la varianza de la estimación. Aquí hay que recordar que la varianza se distribuye chi cuadrado y que el intervalo se construye haciendo

$$P\left(\chi^2_{gl;\,\alpha/2} \leq \frac{gl\, s^2}{\sigma^2} \leq \chi^2_{gl;\,1-\alpha/2}\right) = 1 - \alpha$$

De modo que, el intervalo de confianza de la varianza es

$$P\left(\frac{gl\, s^2}{\chi^2_{gl;\,1-\alpha/2}} \leq \sigma^2 \leq \frac{gl\, s^2}{\chi^2_{gl;\,\alpha/2}}\right) = 1 - \alpha$$

Donde gl son los grados de libertad que, en regresión simple, es igual a $n-2$.

13.3 Predicción

El modelo de regresión, desde luego, puede ser usado como una herramienta predictiva. Dada la ecuación de regresión

$$Y = \hat{\sim} + \hat{\ } X$$

se pueden predecir valores futuros de Y (Y_F) dado un valor futuro de X (X_F) usando

$$\hat{Y}_F = \hat{\sim} + \hat{\ } X_F$$

El valor de Y_F viene dado por

$$Y_F = \sim + \ X_F + «_F$$

De aquí el **error de predicción** es

$$\hat{Y}_F - Y_F = (\hat{\sim} - \sim) + (\hat{\ } - \ )X_F - «_F$$

Como

$$\left.\begin{cases} E(\hat{\sim} - \sim) = 0 \\ E(\hat{\ } - \ ) = 0 \\ E(«_F) = 0 \end{cases}\right\} \Rightarrow E(\hat{Y} - Y_F) = 0$$

De donde: $E\left(\hat{Y}_F\right) = Y_F$

Esto significa que el predictor $\hat{Y}_F$ es <u>insesgado.</u>

La varianza del error de predicción es:

$$V\left(\hat{Y}_F - Y_F\right) = V\left(\hat{~} - \sim\right) + X_F^2 V\left(\hat{~} - \quad\right) - 2X_F \operatorname{cov}\left(\hat{~} - \sim, \hat{~} - \quad\right) + V\left(\ll_F\right)$$

$$= \pm_e^2 \left(\frac{1}{n} + \frac{\overline{X}^2}{\sum\left(X - \overline{X}\right)^2} \right) + X_F^2 \frac{\pm_e^2}{\sum\left(X - \overline{X}\right)^2}$$

$$- 2X_F \pm_e^2 \frac{\overline{X}}{\sum\left(X - \overline{X}\right)^2} + \pm_e^2$$

$$= \pm_e^2 \left[1 + \frac{1}{n} + \frac{\left(X_F - \overline{X}\right)^2}{\sum\left(X - \overline{X}\right)^2} \right]$$

Por lo tanto, $V\left(\hat{Y}_F - Y_F\right)$ se incrementa a medida que X_F se aleja de $\overline{X}$. Como $\pm^2$ es desconocida se estima por:

$$s_e^2 = \frac{SCR}{n-2} = \frac{\sum e^2}{n-2}$$

De aquí, el error estándar de la predicción es

$$s_{\hat{Y}_F} = \sqrt{\frac{\sum e^2}{n-2} \left[1 + \frac{1}{n} + \frac{\left(X_F - \overline{X}\right)^2}{\sum\left(X - \overline{X}\right)^2} \right]}$$

Esta última expresión se puede utilizar para construir intervalos de confianza para Y_F.

Para tener en cuenta:

- La predicción usando valores extremos de la variable independiente puede ser arriesgada. Recuerde que el supuesto lineal puede ser apropiado solo para un rango limitado de las variables independientes. Además, la muestra aleatoria no proporcionó información acerca de los valores extremos de peso.

- Si el contexto cambia, como el hecho de comunidades diferentes, entonces, los parámetros del modelo se verán afectados. Los datos provenientes de la muestra aleatoria fueron obtenidos bajo un conjunto de condiciones de la población. Si cambian, entonces el modelo puede verse afectado.

13.4 Coeficiente r^2

Para evaluar la capacidad predictiva del modelo se usa el coeficiente de determinación o r^2. Este coeficiente es la razón de la variación explicada sobre la variación total:

$$r^2 = \frac{\text{variación total} - \text{variación no explicada}}{\text{variación total}} = \frac{\text{variación explicada}}{\text{variación total}}$$

$$r^2 = \frac{\sum (Y_i - \overline{Y})^2 - \sum (\hat{Y}_i - Y_i)^2}{\sum (Y_i - \overline{Y})^2} = \frac{\sum (\hat{Y}_i - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2}$$

La variación total en Y se divide en la variación que explica el modelo de regresión, variación explicada, y la variación que no explica el modelo de regresión, variación no explicada.

El término r^2 **es el cuadrado de la correlación entre** X **e** Y. Por consiguiente, se encuentra entre cero y uno.

Ejemplo. En un centro médico quieren estudiar, de la población de hombres adultos aparentemente sanos, la relación existente entre la concentración de glucosa en la sangre con el peso. Específicamente, están interesados en conocer si el peso de las personas determina el nivel de glucosa existente en la sangre.

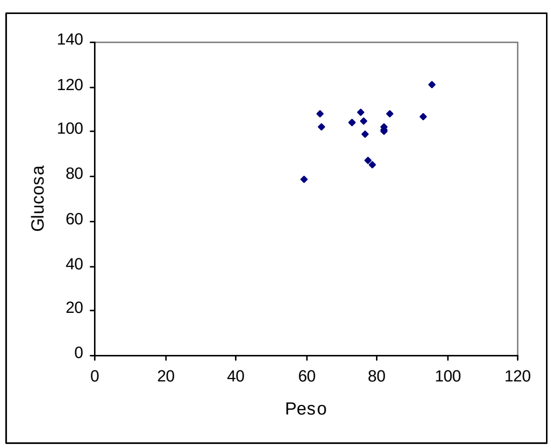

Figura 13.2 Gráfico de dispersión

Obser-vación	Peso (en kgrs.) X	Glucosa (mg/100ml) Y	XY	X^2	$\hat{Y}$	$Y-\hat{Y}$	$(Y-\hat{Y})^2$
1	64,0	108	6912,0	4096,00	94,520	13,480	181,710
2	75,3	109	8207,7	5670,09	100,283	8,717	75,986
3	73,0	104	7592,0	5329,00	99,110	4,890	23,912
4	82,1	102	8374,2	6740,41	103,751	-1,751	3,066
5	76,2	105	8001,0	5806,44	100,742	4,258	18,131
6	95,7	121	11579,7	9158,49	110,687	10,313	106,358
7	59,4	79	4692,6	3528,36	92,174	-13,174	173,554
8	93,4	107	9993,8	8723,56	109,514	-2,514	6,320
9	82,1	101	8292,1	6740,41	103,751	-2,751	7,568
10	78,9	85	6706,5	6225,21	102,119	-17,119	293,060
11	76,7	99	7593,3	5882,89	100,997	-1,997	3,988
12	82,1	100	8210,0	6740,41	103,751	-3,751	14,070
13	83,9	108	9061,2	7039,21	104,669	3,331	11,096
14	73,0	104	7592,0	5329,00	99,110	4,890	23,912
15	64,4	102	6568,8	4147,36	94,724	7,276	52,940
16	77,6	87	6751,2	6021,76	101,456	-14,456	208,976
Sumas	1237,8	1621	126128,1	97178,60			1204,648

Figura 13.3 Tabla de datos

Las observaciones fueron graficadas en el diagrama de dispersión de la Figura 13.2, donde, cada punto de la gráfica, representa un individuo para el cual se observó el peso y el nivel de glucosa. La dispersión de la nube de puntos presenta una asociación positiva.

El siguiente paso, consiste en obtener una línea que tenga el mejor "ajuste" para estos puntos. La línea se denomina línea de mínimos cuadrados y se obtiene a partir de la especificación del modelo de regresión:

$$Y = \alpha + \beta X + \varepsilon$$

donde:

Y , es la variable dependiente "nivel de glucosa"

X , es la variable explicativa "peso"

~ y , parámetros a estimar

« , término de error

Se deben calcular los valores de los estimadores de los parámetros ~ y a partir de los datos consignados en la tabla. De allí se obtienen los valores promedio de las variables Peso (X) y Concentración de glucosa (Y).

$$\overline{X} = \frac{\sum X}{n} = \frac{1237.8}{16} = 77.3625$$

$$\overline{Y} = \frac{\sum Y}{n} = \frac{1621}{16} = 101.3125$$

Se calculan los coeficientes de la regresión

$$\hat{\beta} = \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{\sum X_i^2 - n\overline{X}^2} = \frac{126128.1 - 16 * 77.3625 * 101.3125}{97178.6 - 16 * 77.3625^2} = 0.50975$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = 101.3125 - 0.50975 * 77.3625 = 61.8769$$

Con estos resultados la línea de regresión estimada es

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X = 61.88 + 0.51\, X$$

El valor 61.88 es una estimación del parámetro α, e indica el nivel de glucosa independiente del peso alcanzado por la persona. El valor 0.51, es una estimación del parámetro β que indica en cuánto cambia el nivel de glucosa cuando cambia el peso en una unidad.

El término $\hat{Y}$ es una estimación del nivel de glucosa basada en el modelo de regresión, por ejemplo, cuando X es 64:

$$\hat{Y} = 61.88 + 0.51 * 64 = 94.520$$

De este modo, para un paciente de 64 Kg., la estimación del nivel de glucosa es de 94.52.

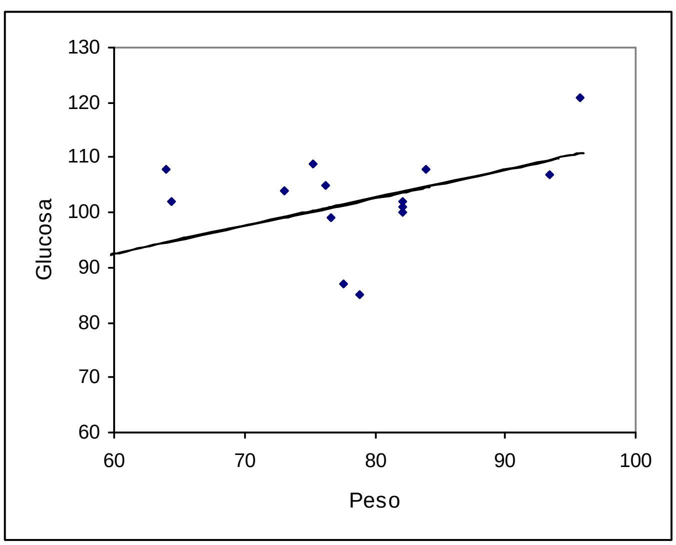

Figura 13.4. Gráfico de dispersión

Si se obtuviera *otra muestra* aleatoria de 16 varones adultos, contendría individuos diferentes. Como resultado, el gráfico de dispersión XY diferiría de la figura anterior y los coeficientes de regresión, $\hat{\sim}$ y $\hat{\ }$, serían diferentes.

La varianza de los errores y de los coeficientes se obtiene de la siguiente manera:

$$s_e^2 = \frac{\sum\left(Y_i - \hat{Y}_i\right)^2}{n-2} = \frac{\sum e^2}{n-2} = \frac{1204.648}{14} = 86.0463$$

$$s_{\hat{\sim}}^2 = \pm_e^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum\left(X - \overline{X}\right)^2}\right] = \pm_e^2\left[\frac{1}{n} + \frac{\overline{X}^2}{\sum X^2 - n\overline{X}^2}\right]$$

$$= 86.0463\left[\frac{1}{16} + \frac{77.3625^2}{97178.60 - 16 * 77.3625^2}\right]$$

$$= 368.2217$$

$$s_{\hat{\ }}^2 = \frac{\pm_e^2}{\sum\left(X - \overline{X}\right)^2} = \frac{\pm_e^2}{\sum X^2 - n\overline{X}^2}$$

$$= \frac{86.0463}{97178.60 - 16 * 77.3625^2} = 0.0606$$

Con estos valores los desvíos de la regresión y los coeficientes alcanzan los siguientes valores:

$$s_e = \sqrt{86.0463} = 9.2761$$

$$s_{\hat{\sim}} = \sqrt{368.2217} = 19.1891$$

$$s_{\hat{\ }} = \sqrt{0.0606} = 0.2462$$

Hasta aquí se ha realizado una estimación puntual del valor de los coeficientes de la regresión. Pero es posible, como ocurre con otros parámetros, realizar una estimación por intervalo asignándole una probabilidad de ocurrencia.

$$P\left(\hat{\sim} - t_{n-2} s_{\hat{\sim}} \le \sim \le \hat{\sim} + t_{n-2} s_{\hat{\sim}}\right) = 1 - 0.05$$
$$P(61.8769 - 2.1448 * 19.1891 \le \sim \le 61.8769 + 2.1448 * 19.1891) = 0.95$$
$$P(20.7201 \le \sim \le 103.0337) = 0.95$$

$$P\left(\hat{\ } - t_{n-2} s_{\hat{\ }} \le \quad \le \hat{\ } + t_{n-2} s_{\hat{\ }}\right) = 1 - 0.05$$
$$P(0.50975 - 2.1448 * 0.2462 \le \quad \le 0.50975 + 2.1448 * 0.2462) = 0.95$$
$$P(-0.01829 \le \quad \le 1.03779) = 0.95$$

Cuando se plantea el modelo de regresión, el primer supuesto es que el comportamiento de la variable X afecta el comportamiento de la variable Y. Para probar esto se contrasta la hipótesis de que el parámetro es nulo. Si esto fuera así; entonces, no habría efecto del peso sobre los niveles de glucosa y el modelo no debería ser usado para ningún propósito.

$$H_0 : \quad = 0$$

El estadístico de prueba es

$$t = \frac{\hat{\ } - \ _{H_0}}{s_{\hat{\ }}} = \frac{0.50975 - 0}{0.2462} = 2.07$$

Al nivel de significación de 0.05, $\sim = 0.05$, el valor de la distribución t con 14 grados de libertad (valor teórico de t) es de ± 2.1448. Esto significa que el valor de prueba cae en la zona de aceptación de la H_0, por lo que el peso alcanzado por los hombres adultos aparentemente sanos no afecta el nivel de glucosa observado en ellos.

Para predecir el valor del nivel de glucosa a partir del modelo de regresión, se supone un nivel de peso y se reemplaza en el mismo. Si se propone un peso de 80 Kg, un nivel de glucosa estimado en base al modelo, sería:

$$\hat{Y} = 61.88 + 0.51 * 80 = 102.68$$

El r^2 es igual a 0.23; por consiguiente, el 23% de la variación total de Y, está explicada por X

CASOS DE ESTUDIO, PREGUNTAS Y PROBLEMAS

Caso 13.1: Función consumo

Con información suministrada por el Ministerio de Economía de la Nación, especifica un modelo de regresión lineal para estimar la función consumo en Argentina.

Problemas

13.1. Replica el ejemplo desarrollado al final del capítulo utilizando algebra matricial.

13.2. El gerente de marketing de una cadena de autoservicio quiere determinar el efecto del espacio, en las estanterías, sobre las ventas de alimentos para animales domésticos. Se seleccionó una muestra aleatoria de 12 autoservicios, de igual tamaño, cuyos resultados se muestran en la tabla.

a) Grafique el diagrama de dispersión
b) Especifique el modelo y utilice el método de mínimos cuadrados para estimar los coeficientes de la regresión
c) Interprete el significado de la pendiente en este problema
d) Prediga las ventas semanales promedio, de alimentos para animales domésticos, para un autoservicio con 8 metros de estanterías para esos alimentos.
e) Calcule el error estándar de la estimación
f) Calcule el coeficiente de determinación e interprete su significado
g) Calcule el coeficiente de correlación
h) Encuentre una estimación de intervalo con el 90% de confianza en las ventas semanales promedio de un autoservicio que tiene 8 metros de estantería para alimentos de animales domésticos
i) Con un nivel de significación del 0.10, ¿hay relación lineal entre el espacio en la estantería y las ventas?

Relación entre ventas y espacio en estantería		
Autoservicio	Espacio X (metros)	Ventas semanales Y (miles de pesos)
1	5	1.6
2	5	2.2
3	5	1.4
4	10	1.9
5	10	2.4
6	10	2.6
7	15	2.3
8	15	2.7
9	15	2.8
10	20	2.6
11	20	2.9
12	20	3.1

13.3. En una fábrica de automóviles, un estadístico quiere desarrollar un modelo para predecir el tiempo de entrega (número de días entre la fecha del pedido y la fecha de entrega) de automóviles nuevos cuyo pedido contempla numeroso equipo opcional. El estadístico cree que hay una relación lineal entre el número de opciones pedidas y el tiempo de entrega. Se selecciona una muestra aleatoria de 16 automóviles que arroja los resultados de la tabla

a) Grafique el diagrama de dispersión

b) Especifique el modelo y utilice el método de mínimos cuadrados para estimar los coeficientes de regresión
c) Interprete el significado de la pendiente en este problema
d) Si se ordena un automóvil que tenga 16 opciones, cuántos días cree usted que tardarían para la entrega?.
e) Calcule el error estándar de la estimación
f) Calcule el coeficiente de determinación e interprete su significado
g) Calcule el coeficiente de correlación
h) Encuentre una estimación de intervalo con el 95% de confianza para el tiempo medio de entrega de un automóvil equipado con 16 opciones
i) Con un nivel de significación del 0.05, ¿hay relación lineal entre el número de opciones y el tiempo de entrega?

Relación entre tiempo de entrega y opciones ordenadas		
Automóvil	Número de opciones X	Tiempo de entrega Y
1	3	25
2	4	32
3	4	26
4	7	38
5	7	34
6	8	41
7	9	39
8	11	46
9	12	44
10	12	51
11	14	53
12	16	58
13	17	61
14	20	64
15	23	66
16	25	70

13.4. Un ingeniero agrónomo quiere determinar el efecto de aplicar un fertilizante orgánico natural sobre el rendimiento de tomates. Se van a utilizar cinco cantidades diferentes de fertilizantes en 10 lotes equivalentes. Los niveles de fertilizante se asignan en forma aleatoria a los lotes arrojando los resultados de la tabla.

a) Grafique el diagrama de dispersión
b) Especifique el modelo y utilice el método de mínimos cuadrados para estimar los coeficientes de regresión
c) Interprete el significado de la pendiente en este problema
d) Prediga el rendimiento de tomates para un lote al cual se le han aplicado 15 kilos por

hectárea de fertilizante orgánico natural.

e) Calcule el error estándar de la estimación
f) Calcule el coeficiente de determinación e interprete su significado
g) Calcule el coeficiente de correlación
h) Encuentre una estimación de intervalo con el 90% de confianza para el rendimiento promedio de tomates fertilizados con 15 kilos por hectárea de fertilizante orgánico natural
i) Con un nivel de significación del 0.05, ¿hay relación lineal entre la cantidad de fertilizante utilizado y el rendimiento de tomates?

Relación entre rendimiento y nivel de fertilizante aplicado		
Lote	Cantidad de fertilizante X (kilos/hectárea)	Rendimiento Y (kilos)
1	0	6
2	0	8
3	10	11
4	10	14
5	20	18
6	20	23
7	30	25
8	30	28
9	40	30
10	40	34

Bibliografía

- **Berenson, Mark y Levine, Daniel.** *Estadística Básica En Administración*. México: Prentice Hall, 1996.
- **____.** *Estadística Para Administración Y Economía. Conceptos Y Aplicaciones.* México: Mc. Graw Hill, 1993.
- **Dagum, C y Bee de Dagum E.** *Introducción a La Econometría*. México: Editorial Siglo XXI, 1971.
- **Daniel, W.** *Bioestadística, Base Para El Análisis De Las Ciencias De La Salud.* México: Editorial Limussa, 1999.
- **Gujarati, Damodar.** *Econometría*. México: Mc.Graw Hill, 2004.
- **Hildebrand, D. y Lyman Ott R.** *Estadística Aplicada a La Administración Y a La Economía*. Wilmington, USA: Addison Wesley Iberoamericana, 1997.
- **Johnston, J. y Dinardo, J. .** *Métodos De Econometría.* Barcelona: Editorial Vicens Vives, 2001.
- **Kazmier, L y Diaz Mata, A.** *Estaística Aplicada a La Administración Y a La Economía*. México: McGraw Hill, 1993.
- **Maddala, G.S.** *Introducción a La Econometría*. Naulcapan de Juarez, México: Prentice Hall, 1996.