6 Regresión lineal

Jorge Chica Olmo
Dolores M. Frías Jamilena

1. INTRODUCCIÓN

El objetivo fundamental de este capítulo es mostrar los aspectos básicos del modelo de regresión lineal y su aplicación en la investigación de marketing. Para ello se empieza con una introducción en la que se recoge el concepto y los objetivos de esta técnica. Posteriormente se desarrolla el modelo lineal de regresión y, por último, se presenta un ejemplo en el que se aplican los conceptos teóricos.

La regresión es una herramienta fundamental en el análisis de datos, tanto por su utilidad en sí misma como por servir de referente para otras técnicas. En términos generales, la metodología econométrica tradicional se realiza a través de las siguientes fases:

1. Planteamiento de la teoría económica que se desea analizar y de sus hipótesis.
2. Especificación del modelo econométrico apoyándose en la teoría.
3. Búsqueda y depuración de los datos.
4. Estimación de los parámetros del modelo.
5. Contraste de las hipótesis del modelo.
6. Explotación del modelo: predicción y utilización del modelo para fines de control o de política.

El término de regresión fue introducido por Francis Galton (1886) y corroborada su ley por Karl Pearson (1903). En términos generales se puede decir que el análisis de regresión trata del estudio de la dependencia de una variable a explicar con respecto a una o más variables explicativas.

Los objetivos que se pretenden conseguir con este análisis son varios:

1. Determinar la estructura o forma de la relación, es decir, la ecuación matemática que relaciona las variables independientes con la dependiente.

© Ediciones Pirámide

2. Verificar hipótesis deducidas de la teoría analizada.
3. Predecir los valores de la variable dependiente y realizar simulaciones.

La variable dependiente puede expresarse con diversos términos: variable explicada, predicha, regresada y respuesta y la terminología empleada para la variable independiente es como variable explicativa, predictor, regresor, variable de control estímulo.

Matemáticamente la relación entre la variable explicada y las variables explicativas se puede expresar como:

$$Y = f(X)$$

La letra Y representa la variable dependiente y las X (X_1, X_2, ..., X_k) representan las variables explicativas. Si el número de variables independientes es una nos encontramos ante un modelo de regresión simple; si son más de una se trata de un modelo de regresión múltiple.

Tipos de datos

Los datos que se utilizan en la aplicación de esta técnica pueden ser: series de tiempo, datos de corte transversal e información combinada.

Las series de tiempo son un conjunto de observaciones sobre los valores que toma una variable en diferentes momentos de tiempo. Tal información debe ser recogida en intervalos regulares, que pueden ser en forma diaria, mensual, trimestral, anual, etc. La información puede ser de carácter cuantitativo o cualitativo.

Los datos de corte transversal se refieren a observaciones de un conjunto de unidades o entes (unidades familiares, empresas, regiones, etc.). Este tipo de datos se conocen también como datos espaciales. El problema que presentan estas series es el de la heterogeneidad. Cuando incluimos unidades heterogéneas en un análisis estadístico, el efecto de tamaño o escala debe ser tenido en cuenta.

En la información combinada los datos agrupados tienen elementos de series de tiempo y de corte transversal reunidos. Hay un tipo especial de datos agrupados, la información de panel o longitudinal, también llamada información micropanel, en la cual la misma unidad de corte transversal es encuestada a través del tiempo.

2. EL MODELO DE REGRESIÓN LINEAL

2.1. Introducción al modelo de regresión simple

En el modelo de regresión lineal simple o modelo lineal simple (MLS), en el que figura una única variable explicativa, el comportamiento de la variable Y se puede explicar a través de una variable X, que representamos mediante:

$$Y = f(X)$$

© Ediciones Pirámide

Considerando que la relación f, que liga Y con X, es lineal se puede escribir de la siguiente forma:

$$Y_i = \beta_1 + \beta_2 X_i$$

donde

Y_i: Variable dependiente.

X_i: Variable independiente.

β_1: Ordenada en el origen o término independiente.

β_2: Pendiente de la recta.

Este tipo de relaciones raramente son exactas, más bien son aproximaciones en las que se han omitido muchas variables de importancia secundaria, lo que nos obliga a incluir un término de perturbación aleatoria, quedando la relación como sigue:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

donde

u_i: Término de perturbación aleatoria.

2.2. Estimación de los coeficientes del modelo de regresión lineal simple

El principal problema consiste en estimar, a partir de las observaciones disponibles, los valores de los parámetros β_1 y β_2. En primer lugar, se realiza una aproximación intuitiva utilizando la representación gráfica de las observaciones (X_i, Y_i, con $i = 1, 2, ..., n$). De tal forma que si la relación lineal de dependencia entre X e Y fuera exacta, las observaciones se situarían a lo largo de una recta (figura 6.1). Tomando u_i el valor 0 para todo i, y las estimaciones más adecuadas de β_1 y β_2, de hecho los verdaderos valores, serían, respectivamente, la ordenada en el origen y la pendiente de dicha recta.

En el caso de que la dependencia entre X e Y sea estocástica, en general las observaciones no se alinearán a lo largo de una recta, sino que formarán una nube de puntos, tal y como se muestra en la figura 6.2. Si designamos mediante $\hat{\beta}_1$ y $\hat{\beta}_2$ las estimaciones de β_1 y β_2, respectivamente, la recta vendrá dada por:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

Nuestro problema es hallar unos estimadores $\hat{\beta}_1$ y $\hat{\beta}_2$ tales que la recta se ajuste lo mejor posible a los puntos (X_i, Y_i). A la diferencia entre el valor observado de la variable dependiente y su valor ajustado o estimado se le denomina error o residuo:

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

© Ediciones Pirámide

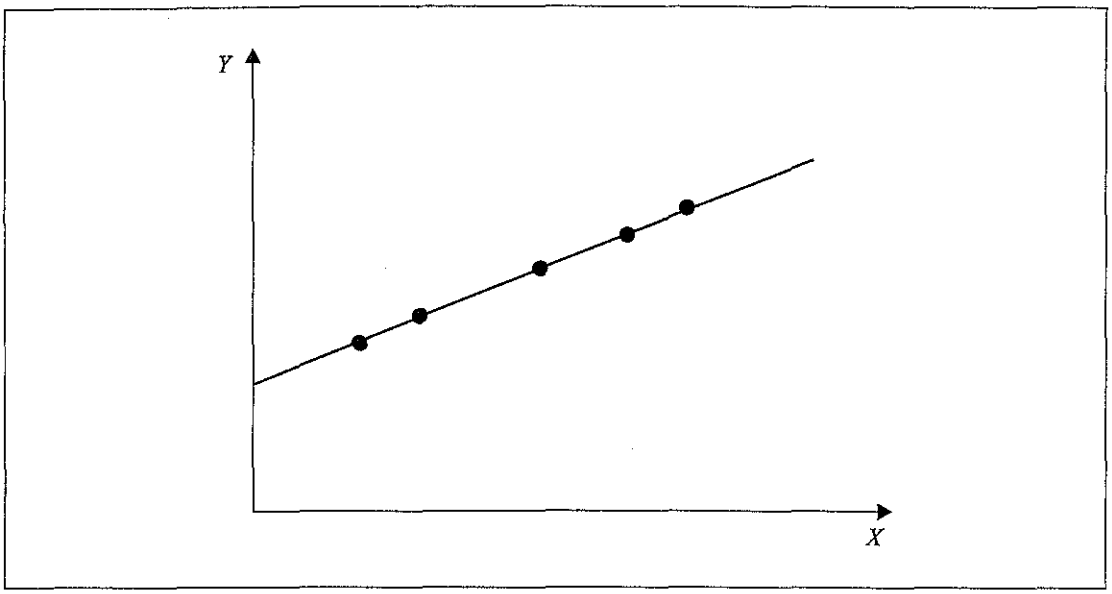

Figura 6.1.

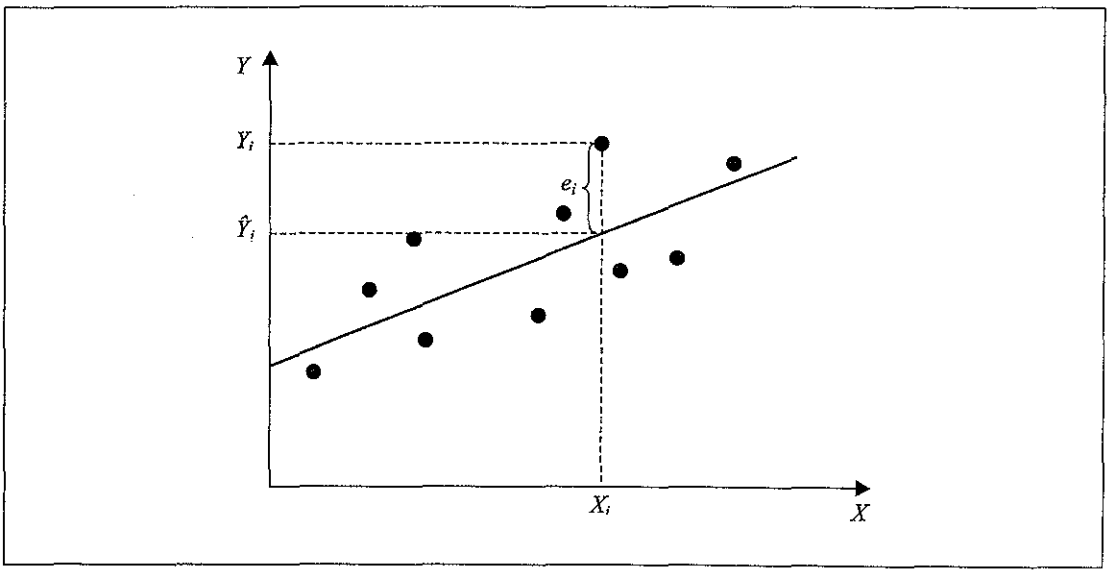

Figura 6.2.

Existen diversos criterios para el ajuste de la recta, siendo el más utilizado el *criterio de los mínimos cuadrados,* según el cual la mejor recta es aquella que haga mínimo la suma de los cuadrados de los residuos:

$$\text{Min } \Sigma e_i^2$$

© Ediciones Pirámide

Este criterio, al tomar los cuadrados de los residuos, evita la compensación de éstos; sin embargo, con este criterio estamos penalizando proporcionalmente más los residuos grandes frente a los pequeños (si un residuo es el doble que otro, su cuadrado será cuatro veces mayor).

Puesto que el MLS es un caso particular del modelo de regresión múltiple, el resto de las fases se desarrollan en el siguiente epígrafe.

3. EL MODELO DE REGRESIÓN MÚLTIPLE

El modelo de regresión general o modelo lineal general (MLG) se suele expresar, para la observación i-ésima, de la siguiente forma:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + u_i$$

para $i = 1, ..., n$; donde Y es la variable explicada, X_2, X_3, ..., X_k son las variables explicativas, u_i es el término de perturbación aleatoria, β_1, β_2, ..., β_k son los parámetros o coeficientes del modelo y n es el número de observaciones en la muestra que debe ser superior a k. El objetivo fundamental de dicho modelo es explicar lo mejor posible el comportamiento de la variable explicada Y a partir de las variables explicativas X.

Existen diferentes razones para incluir en la expresión anterior el término de perturbación, pero la razón fundamental es que en muy escasas ocasiones[1] se puede establecer la relación exacta o determinista por la cual la variable dependiente viene puntualmente explicada por las variables explicativas. Si se supone que el modelo está bien especificado y no hay errores de medida, entonces la perturbación recogerá aquellas variables explicativas que de manera individual se consideran irrelevantes, pero que en conjunto afectan al comportamiento de la variable dependiente.

En términos matriciales el MLG se expresa:

$$Y = X\beta + u$$

donde Y es el vector que contiene las n observaciones de la variable explicada, X es una matriz $n \times k$ que contiene en la primera columna n unos[2] y en las $k-1$ columnas restantes están las observaciones de las variables explicativas, β es un vector con k parámetros constantes y u es el vector con n perturbaciones aleatorias.

Hipótesis básicas:

1. Se supone que la forma funcional que liga la variable explicada con las variables explicativas es de tipo lineal al menos en los parámetros.

[1] Mientras que en las ciencias físicas es más frecuente la relación exacta entre las variables del modelo, en las ciencias sociales, y en particular en las económicas, esto es más limitado.

[2] Si el modelo especificado no contiene término constante dicha columna de unos no aparecerá.

© Ediciones Pirámide

2. Las variables explicativas son fijas en el muestreo o al menos serán independientes de las perturbaciones. Además, los datos muestrales de las variables explicativas deben ser linealmente independientes, es decir, que no hay multicolinealidad exacta.
3. Las perturbaciones aleatorias se supone que son normales con:

   a) $E(u_i) = 0;\ \forall\ i.$

   b) $\text{Var}\ (u_i) = \sigma^2;\ \forall\ i.$

   c) $\text{Cov}\ (u_i, u_j) = 0;\ \forall\ i \neq j.$

Las hipótesis *b* y *c* implican, respectivamente, que la varianza de las perturbaciones es constante (homoscedasticidad) y la ausencia de autocorrelación entre las perturbaciones.

Cada una de estas hipótesis puede ser más o menos restrictiva, es decir, se cumplirá más o menos en la práctica dependiendo del fenómeno económico analizado. Así, por ejemplo, la hipótesis de linealidad no es demasiado restrictiva, ya que en la práctica este tipo de relación entre las variables se suele dar con frecuencia, aun cuando si se plantean modelos no lineales éstos lo suelen ser en las variables y, por tanto, fácilmente linealizables.

Para que los resultados obtenidos a partir del modelo estimado sean adecuadamente interpretados dependerá de que se cumplan las principales hipótesis básicas del modelo. En principio, supondremos que se cumplen dichas hipótesis y expondremos e interpretaremos las expresiones que se obtienen bajo el cumplimiento de dichas hipótesis y, posteriormente, se tratarán las más frecuentemente analizadas en la Econometría clásica: multicolinealidad, heteroscedasticidad y autocorrelación.

3.1. Estimación

Una vez especificado el modelo, la fase siguiente consiste en estimar los parámetros de dicho modelo, que lógicamente serán desconocidos. De los diferentes procedimientos de estimación los más avalados son el de mínimos cuadrados y el de máxima verosimilitud. Cuando se cumplen las hipótesis anteriores el método de mínimos cuadrados se conoce como mínimos cuadrados ordinarios (MCO) y consiste en minimizar la suma de los cuadrados de los residuos, los cuales vienen dados por la diferencia entre el verdadero valor de la variable explicada y su estimación.

El estimador mínimo cuadrático ordinario (EMCO) tiene la forma:

$$\hat{\beta} = (X'X)^{-1}X'Y$$

donde $\hat{\beta}$ es un vector columna que contiene las estimaciones MCO de los k parámetros del modelo. Este estimador posee una serie de propiedades estadísticas deseables. Así, dicho estimador es:

© Ediciones Pirámide

1. Insesgado o de sesgo nulo, es decir, la diferencia entre el verdadero valor del parámetro y el valor esperado de dicho estimador es cero: $\beta - E(\hat{\beta}) = 0$.
2. Varianza mínima. El EMCO cumple el teorema de Gauss-Markov ya que dicho estimador tiene varianza mínima dentro de la familia de los estimadores lineales e insesgados, por lo que dicho estimador se dice que es un estimador lineal, insesgado y óptimo (ELIO).
3. Consistente, ya que a medida que el tamaño de la muestra incrementa infinitamente dicho estimador converge hacia el verdadero valor del parámetro.

El modelo estimado se expresa:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki} + e_i$$

donde e_i son los residuos mínimo cuadráticos y los $\hat{\beta}_j$ son las estimaciones MCO de los k parámetros del modelo.

Puesto que tenemos k coeficientes estimados en el modelo, las varianzas y covarianzas de éstos se expresan mediante una matriz cuadrada que contiene en su diagonal principal las varianzas y a ambos lados de dicha diagonal están las covarianzas:

$$\text{Var}\,(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

en la práctica, puesto que σ^2, desviación típica del error, es desconocida, se estima mediante el estimador insesgado:

$$\hat{\sigma}^2 = \frac{\Sigma\, e_i^2}{n-k}$$

luego la varianza estimada de un $\hat{\beta}_j$ viene dada por:

$$\hat{\sigma}^2_{\hat{\beta}_j} = \hat{\sigma}^2 a_{jj}$$

donde a_{jj} es el elemento j-ésimo de la diagonal principal de la matriz $(X'X)^{-1}$.

3.2. Interpretación de los coeficientes estimados

Los coeficientes o parámetros del modelo lineal representan la derivada parcial de Y respecto de cada una de las variables explicativas. Por tanto, las estimaciones de dichos coeficientes se pueden interpretar como la variación esperada o promedio que se produce en Y (en las unidades en las que venga dada dicha variable) cuando incrementa en una unidad la variable explicativa correspondiente, suponiendo que el resto de variables explicativas permanecen constantes.

© Ediciones Pirámide

3.3. Intervalos de confianza y prueba de hipótesis

El EMCO nos proporciona una estimación puntual del valor desconocido de los parámetros. Esta estimación podrá variar con la muestra de datos usados, aunque si se tomaran diferentes muestras se esperaría que la media de dichas estimaciones fuera igual al verdadero valor de dichos parámetros. Por ello se suele obtener, además de la estimación puntual de los parámetros, la estimación por intervalos. Este tipo de estimación nos proporcionará un intervalo dentro del cual se encontrará el verdadero valor del parámetro dado un nivel de confianza o de probabilidad $1 - \alpha$, donde α es el nivel de significación. Dicho intervalo se obtiene a partir de las estimaciones de los parámetros, las cuales, como se ha indicado, variarán con la muestra usada. Por tanto, si se construyeran infinitos intervalos, en un $1 - \alpha$ de éstos estará el verdadero valor del parámetro. En la práctica el intervalo de estimación para un parámetro β_j viene dado por:

$$\hat{\beta}_j \pm \hat{\sigma}_{\hat{\beta}_j} t_{n-k,\alpha/2}$$

donde $t_{n-k,\alpha/2}$ es el valor de las tablas de la t-Student para $n - k$ grados de libertad y un nivel de significación $\alpha/2$.

Otro aspecto práctico importante que nos proporciona el modelo de regresión es la posibilidad de plantear y resolver hipótesis estadísticas relativas a los parámetros. En general, para verificar hipótesis relativas a un solo parámetro se usa el estadístico:

$$t = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

el cual sigue una distribución t-Student con $n - k$ grados de libertad.

Así, cuando se estima un modelo lo primero será plantearse si alguna de las variables X_j incluidas en el modelo no es significativa. Esto se traduce en plantear la siguiente hipótesis:

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

en cuyo caso el estadístico queda de la siguiente forma:

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

© Ediciones Pirámide

Comparando el valor de dicho estadístico, que se obtiene a partir de los datos muestrales, con el valor de las tablas de la *t*-Student, si aquél en términos absolutos es mayor que el de las tablas se rechazará la H_0, es decir, dicha variable es significativa, y en caso contrario la variable se considera no significativa.

Otra hipótesis que se suele plantear es si el modelo es significativo en su conjunto, esto es, si de manera conjunta el modelo explica o no las variaciones de la variable dependiente. En este caso la hipótesis nula es: H_0: $\beta_2 = \beta_3 = \cdots = \beta_k = 0$, frente a la hipótesis alternativa H_1: de que la hipótesis nula no se cumple. El estadístico para verificar dicha hipótesis es:

$$F = \frac{SCE/(k-1)}{SCR/(n-k)}$$

donde *SCE* es la suma de cuadrados de la explicada respecto de la media, es decir, $SCE = \Sigma(\hat{Y}_i - \bar{Y})^2$ y *SCR* es la suma de cuadrados de residuos, o sea, $SCR = \Sigma e_i^2$. Este estadístico sigue una distribución *F* con $k-1$ y $n-k$ grados de libertad en el numerador y denominador respectivamente. Este tipo de prueba, en el modelo de regresión, se conoce como análisis de la varianza o tabla ANOVA. Cuando el modelo de regresión tiene término constante se cumple la siguiente expresión: $SCT = SCE + SCR$, donde $SCT = \Sigma(Y_i - \bar{Y})^2$.

Como se verá más adelante, en las salidas de ordenador de los paquetes estadísticos, además de las estimaciones y valores de los estadísticos anteriores se suele añadir el valor α o nivel de significación, valor que nos indica el nivel de significación mínimo para rechazar la hipótesis nula.

3.4. Bondad del ajuste

Una medida de la bondad del ajuste del modelo estimado es el coeficiente de determinación R^2, que permite evaluar en qué medida el modelo estimado se ajusta a los datos muestrales disponibles. El coeficiente de determinación se define como el cociente entre la *SCE* y *SCT*, es decir, nos mide la proporción de las variaciones de la variable dependiente que vienen explicadas por el modelo.

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Si el modelo tiene término independiente los valores de este coeficiente están entre 0 y 1, de tal forma que a medida que se aproxima a 1 el ajuste será mejor, siendo perfecto si es igual a 1, en cuyo caso la *SCR* sería 0. Usualmente, este coeficiente se multiplica por 100, indicando entonces el porcentaje de las variaciones de

© Ediciones Pirámide

Y explicadas por el modelo ajustado. La interpretación adecuada de este coeficiente depende de que el modelo esté bien especificado.

Un inconveniente de este coeficiente es que a medida que se incluyen en el modelo más variables su valor incrementará o al menos no disminuirá, tendiendo a seleccionar aquellos modelos con mayor número de variables explicativas. Para comparar modelos con la misma variable explicada y distinto número de variables explicativas no es aconsejable usar el R^2; en su lugar sería más aconsejable utilizar el coeficiente de determinación ajustado:

$$\overline{R}^2 = 1 - \frac{SCR/(n-k)}{SCT/(n-1)}$$

este coeficiente, para $k > 1$, es menor que el R^2 y además puede tomar valores negativos, quedando su rango de valores fuera del intervalo (0, 1).

3.5. Importancia relativa de las variables explicativas

Cuando se está interesado en medir la importancia relativa de las variables explicativas que aparecen en el modelo se pueden usar diferentes instrumentos. Dicha importancia o grado de influencia de las variables explicativas puede ser analizada bajo diferentes enfoques.

Coeficientes beta

Los coeficientes estimados del modelo no suelen ser buenos indicadores de la importancia relativa de cada variable explicativa, a no ser que todas las variables del modelo vengan dadas en las mismas unidades de medida. En estadística cuando se desea comparar variables que vienen en distintas unidades de medida previamente se suele tipificar dichas variables restándole la media aritmética y dividiendo por la desviación típica; pues bien, los coeficientes beta son los coeficientes del modelo de regresión, pero previamente tipificadas las variables:

$$\frac{Y_i - \overline{Y}}{S_y} = \beta_2^* \frac{X_{2i} - \overline{X}_2}{S_{x2}} + \beta_3^* \frac{X_{3i} - \overline{X}_3}{S_{x3}} + \cdots + \beta_k^* \frac{X_{ki} - \overline{X}_k}{S_{xk}} + v_i$$

Los coeficientes beta están relacionados con los coeficientes del modelo original mediante la expresión:

$$\beta_j^* = \beta_j \frac{S_{xj}}{S_y}$$

Lógicamente al sustituir β_j por su estimación obtendremos la estimación del coeficiente beta correspondiente, y dichos coeficientes nos permitirán ordenar en re-

© Ediciones Pirámide

lación a la importancia relativa de cada variable explicativa del modelo. En este caso la importancia viene medida en términos de unidades de desviación típica que cada variable explica.

Coeficientes de correlación parcial

Otra forma de ordenar en cuanto a importancia es usar los coeficientes de correlación parcial de la variable dependiente respecto de cada variable independiente, que miden la proporción de las variaciones de Y que vienen explicadas por dicha variable independiente y que no explica el resto de variables independientes del modelo. Así, supuesto un modelo con dos variables independientes X_2, X_3, el coeficiente de correlación parcial $r_{2,3}$ nos mediría el grado de asociación lineal entre la variable dependiente y la variable X_2, dejando fuera la influencia común que pudiera tener X_3 sobre la variable dependiente y sobre X_2. En este caso la importancia vendría medida por el grado de asociación que cada variable independiente tiene sobre la variable dependiente que no posee el resto de variables independientes.

Elasticidades

En economía suele ser frecuente medir la sensibilidad de una variable respecto de otra, para lo cual se usa la elasticidad. Puesto que los coeficientes del MLG representan las derivadas parciales de cada variable independiente respecto de la dependiente, se podrá obtener fácilmente la elasticidad de cada variable independiente respecto de la dependiente teniendo en cuenta la propia definición de la elasticidad:

$$E_{xj}^{y} = \frac{\partial Y}{\partial X_j}\frac{X_{ji}}{Y_i} = \beta\frac{X_{ji}}{Y_i}$$

Sustituyendo β por su estimación se podrá obtener la elasticidad en el punto i-ésimo; usualmente se obtendrá la elasticidad media reemplazando X_{ji} e Y_i por sus medias correspondientes. Estas elasticidades se pueden usar también para llevar a cabo una ordenación en las variables explicativas, ya que aquéllas no se ven afectadas por las unidades de medida. En este caso la importancia viene medida en términos de sensibilidad de la variable dependiente a las variaciones de las variables independientes. Las elasticidades también se pueden obtener estimando un modelo doblemente logarítmico linealizado:

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln X_{2i} + \cdots + \beta_k \ln X_{ki} + u_i$$

en cuyo caso las estimaciones de los β_j nos darán directamente las elasticidades de Y respecto de cada una de las variables X_j.

Posiblemente la ordenación dada por cualquiera de los tres apartados anteriores será parecida, indicando de esta forma la importancia relativa de cada variable independiente dentro del modelo.

© Ediciones Pirámide

3.6. Predicción

La última fase suele ser la explotación del modelo, y dentro de dicha fase lo frecuente es usar el modelo para hacer predicciones. Para realizar las predicciones se requiere conocer los valores que toman las variables independientes para el momento, si los datos son temporales, en el cual se desea predecir el valor que tomará la variable dependiente. La predicción podrá ser puntual o por intervalos. Se demuestra que el predictor lineal insesgado y óptimo (PLIO) es el que se obtiene sustituyendo en la expresión del MLG los parámetros por el EMCO:

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_{20} + \hat{\beta}_3 X_{30} + \cdots + \hat{\beta}_k X_{k0}$$

donde $\hat{Y}_0$ es la predicción puntual para el momento 0, dada por los valores de las variables independientes X_{20}, X_{30}, ..., X_{k0}.

La expresión para realizar la predicción por intervalos es:

$$\hat{Y}_0 \pm t_{n-k,\alpha/2}\hat{\sigma}\sqrt{1 + X_0'(X'X)^{-1}X_0}$$

donde $X_0' = 1, X_{20}, X_{30}, ..., X_{k0}$, es un vector que contiene los valores de las variables independientes para los cuales se desea realizar la predicción.

3.7. Variables ficticias

Es bastante frecuente en el análisis de marketing la necesidad de tratar con variables de tipo cualitativo, variables como el sexo, el estado civil, la región donde se encuentran ubicadas las empresas, etc. La inclusión de este tipo de variables en un modelo de regresión se realiza mediante la inclusión de variables ficticias. Este tipo de variables se denominan también binarias o dicotómicas, ya que se caracterizan porque pueden tomar dos valores: 1 o 0 dependiendo de que el individuo observado presente o no tal o cual característica. Las variables cualitativas pueden presentar dos o más categorías; así el sexo tiene dos categorías posibles pero el estado civil puede tener más de dos. Una regla fundamental a la hora de incluir las variables cualitativas en el modelo de regresión consiste en incluir $m - 1$ variable ficticia, donde m es el número de categorías que puede presentar la variable cualitativa. Si se incumple esta regla se incurre en la denominada trampa de las variables ficticias, lo que provocará que el modelo de regresión presente multicolinealidad perfecta y, por tanto, no será posible su estimación mediante MCO.

En el modelo de regresión se pueden incluir variables explicativas ficticias solamente o junto con variables de tipo cuantitativo. El primer caso se conoce como modelos de análisis de la varianza (ADV). La inclusión de las variables ficticias puede ser aditiva, multiplicativa o mixta, de forma que afecte al término constante, a la pendiente o a ambos. Para entender esto vamos a suponer que deseamos anali-

© Ediciones Pirámide

zar las ventas de un determinado producto (Y_i) de una empresa que posee diferentes sucursales que están repartidas en dos comunidades autónomas: Andalucía y Cataluña; y suponemos que las ventas, además de estar en función del precio (X_i), dependen de la comunidad en la que se encuentre la sucursal.

Inicialmente podemos plantear un modelo aditivo de la forma:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_{1i}$$

donde

$$D_i = \begin{cases} 0 \text{ si la sucursal está localizada en Andalucía.} \\ 1 \text{ si la sucursal está localizada en Cataluña.} \end{cases}$$

Suponiendo que se cumplen las hipótesis básicas del modelo de regresión, el valor esperado de las ventas de una sucursal que esté ubicada en una u otra comunidad autónoma y para un precio X_i sería para este caso:

— Sucursal localizada en Andalucía: $E(Y_i/D_i = 0, X_i) = \beta_1 + \beta_2 X_i$

— Sucursal localizada en Cataluña: $E(Y_i/D_i = 1, X_i) = (\beta_1 + \beta_3) + \beta_2 X_i$

La diferencia entre una y otra viene dada por β_3, que afecta al término independiente y que se recoge gráficamente en la figura 6.3.*a*. Por tanto, si se desea contrastar la existencia de un comportamiento diferencial en las ventas de las sucursales debido a la localización sólo se debería aplicar el test de la *t*-Student verificando la hipótesis nula H_0: $\beta_3 = 0$, de forma que si se acepta la H_0 no habría diferencia significativa en las ventas de las sucursales por razón de su localización.

Otra alternativa sería especificar el modelo de forma multiplicativa:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i X_i + u_{2i}$$

En este caso, el valor esperado sería:

— Sucursal localizada en Andalucía: $E(Y_i/D_i = 0, X_i) = \beta_1 + \beta_2 X_i$

— Sucursal localizada en Cataluña: $E(Y_i/D_i = 1, X_i) = \beta_1 + (\beta_2 + \beta_3) X_i$

Al igual que antes, la diferencia entre una y otra viene dada por β_3, pero en este caso afecta a la pendiente, lo que se representa gráficamente en la figura 6.3.*b*. Igual que antes, se podría plantear el contraste de la *t*-Student para analizar si existe o no un comportamiento diferencial en las ventas de unas sucursales respecto de otras.

Por último, el modelo se podría especificar de forma mixta:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta_4 D_i X_i + u_{3i}$$

De esta forma, el valor esperado de las ventas sería:

— Sucursal localizada en Andalucía: $E(Y_i/D_i = 0, X_i) = \beta_1 + \beta_2 X_i$

— Sucursal localizada en Cataluña: $E(Y_i/D_i = 1, X_i) = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) X_i$

© Ediciones Pirámide

Ahora la diferencia está tanto en la ordenada (β_3) como en la pendiente (β_4) (figura 6.3.*c*). Para contrastar el comportamiento diferencial en las ventas se haría mediante el estadístico F, verificando la hipótesis nula H_0: $\beta_3 = \beta_4 = 0$; si se acepta dicha hipótesis no habría diferencia significativa en las ventas entre sucursales debido a su localización. Si se rechaza dicha hipótesis se podría hacer el test individual mediante la t-Student para β_3 y β_4, lo que permitiría saber si la diferencia se debe a la ordenada o a la pendiente.

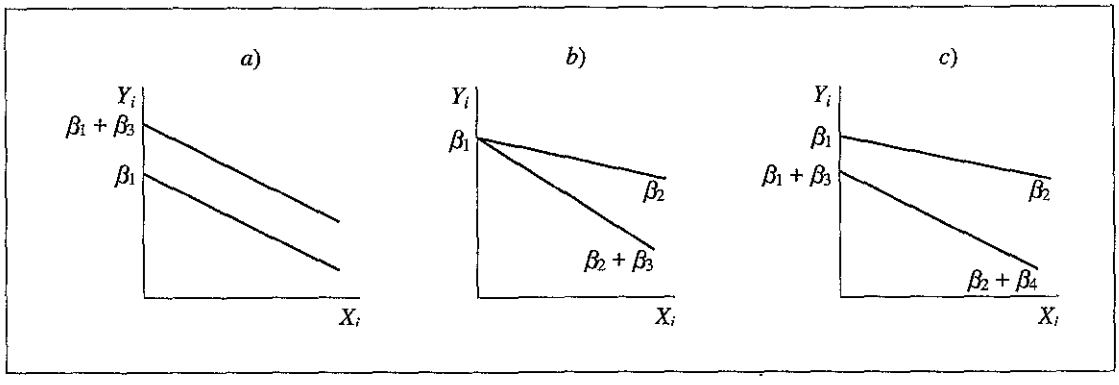

Figura 6.3.

4. MULTICOLINEALIDAD

4.1. Concepto y consecuencias

La multicolinealidad es un problema de los datos y se produce cuando hay algún tipo de relación lineal entre las variables explicativas del modelo. Una de las hipótesis básicas del MLG es que entre las variables explicativas no puede darse una relación lineal exacta, ya que en otro caso la matriz ($X'X$) sería singular y, por tanto, no se puede invertir, lo que provocaría la imposibilidad de obtener los EMCO. En el caso anterior se diría que hay multicolinealidad exacta y sus consecuencias son la indeterminación del EMCO y sus varianzas serían infinitas. Pero en la práctica en raras ocasiones se presentará la multicolinealidad exacta y, por el contrario, sí que se presenta con frecuencia la multicolinealidad inexacta o imperfecta, en cuyo caso lo que se da es una relación lineal no exacta entre las variables explicativas. Dependiendo del grado de multicolinealidad las consecuencias sobre los resultados e interpretación de éstos serán más o menos graves. Desde el punto de vista teórico el EMCO, bajo presencia de multicolinealidad, seguirá cumpliendo las propiedades estadísticas deseables ELIO. Sin embargo, desde el punto de vista práctico se demuestra que a medida que el grado de multicolinealidad incrementa, la varianza de los EMCO también incrementa, teniendo como consecuencia que el estadístico t de un coeficiente o más tenderá a ser estadísticamente no significativo. Otras consecuencias

© Ediciones Pirámide

son que los EMCO y sus varianzas se vuelven muy sensibles a variaciones en los datos muestrales, y también las covarianzas de dichos estimadores se hacen grandes. Por ello, en presencia de multicolinealidad grave los tests estadísticos sobre la significación individual de los coeficientes nos pueden llevar a conclusiones erróneas, por lo que será conveniente detectar la posible presencia de multicolinealidad grave.

4.2. Detección

Una consecuencia práctica de la presencia de multicolinealidad grave es que algunos o todos los coeficientes del modelo sean no significativos de manera individual y, por el contrario, al verificar la significación global el modelo resulte significativo, o consecuentemente que el modelo tenga un coeficiente de determinación alto. Esta consecuencia, que resulta paradójica, se suele usar como un método para sospechar la posible presencia de multicolinealidad grave.

Otro método consiste en obtener los coeficientes de correlación simple entre las variables explicativas, de tal forma que nos permita apreciar la posible presencia de correlación lineal tomadas dos a dos las variables, y así valores de estos coeficientes superiores a 0,75 o 0,80 nos indicarían la presencia de colinealidad alta. Este procedimiento es una condición suficiente, pero no necesaria, ya que podrían ser bajos estos coeficientes y, sin embargo, presentar multicolinealidad grave, ya que en lugar de colinealidad por parejas podría existir entre grupos.

Realizar las regresiones auxiliares es otra forma para detectar la presencia de multicolinealidad grave. Estas regresiones consisten en regresar cada variable explicativa con el resto de variables explicativas del modelo original. El coeficiente de determinación de cada regresión auxiliar se denota por R_j^2 y se conoce como coeficiente de correlación múltiple; si el valor de dicho coeficiente es igual o superior a 0,75 se suele considerar la presencia de multicolinealidad grave. A partir de este coeficiente se han propuesto otros como el factor de agrandamiento de la varianza (FAV) o la tolerancia (TOL). Así:

$$FAV_j = \frac{1}{1 - R_j^2}$$

cuyo valor ideal, ausencia de multicolinealidad de la variable x_j con el resto, es 1 y valores superiores a 4 nos indicarán presencia de multicolinealidad grave.

Mientras que la tolerancia se define como: $TOL_j = (1 - R_j^2)$, de manera que valores próximos a 1 indican ausencia de multicolinealidad y próximos a 0 indicarían multicolinealidad muy grave.

Por último, otro procedimiento consiste en obtener el número de condición que se define como:

$$k = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

© Ediciones Pirámide

donde λ_{max} y λ_{min} son, respectivamente, las raíces características mayor y menor de la matriz $(X'X)$ normalizada. Se suele considerar la presencia de multicolinealidad grave cuando el número de condición está por encima de 20 o 25.

4.3. Soluciones

Como se ha indicado, la multicolinealidad es un problema de los datos, y entre las opciones para remediar las consecuencias negativas que provoca la presencia de multicolinealidad grave está la de eliminar aquellas variables más colineales con el resto, solución fácil, pero arriesgada, ya que existe la posibilidad de incurrir en un error de especificación. Dependiendo de la información disponible se aplican diferentes métodos: búsqueda de información a priori sobre la relación entre los parámetros; si se dispone, usar la combinación de datos de corte transversal y temporales e igualmente, si se puede, aumentar el tamaño de la muestra, aunque existen otros procedimientos mecánicos como la regresión alomada, la regresión con componentes principales o simplemente realizar algún tipo de transformación en las variables.

5. HETEROSCEDASTICIDAD

5.1. Concepto y consecuencias

Uno de los supuestos relativos al comportamiento de las perturbaciones es que éstas se consideran homoscedásticas, esto es, que tienen varianza constante. Esta hipótesis no siempre se cumplirá y así habrá fenómenos económicos en los cuales la varianza de dichas perturbaciones no será constante, en cuyo caso se dice que son heteroscedásticas:

$$E(u_i^2) = \sigma_i^2 \quad \text{para } i = 1, \ldots, n$$

Un ejemplo clásico en el que se espera este comportamiento es en el estudio del comportamiento del gasto de cualquier tipo de bien de lujo en función de los ingresos. Así, familias con rentas bajas tendrán un comportamiento similar y, por tanto, con varianza pequeña, mientras que entre las familias con rentas altas habrá mayor dispersión dependiendo de los gustos.

Si las perturbaciones son heteroscedásticas entonces el EMCO, aunque seguirá siendo lineal e insesgado, dejará de ser eficiente. Por tanto, en la práctica, si se aplica el EMCO en un modelo en el que existe heteroscedasticidad, las pruebas o tests que se realicen podrán llevar a conclusiones erróneas.

5.2. Detección

Puesto que la presencia de heteroscedasticidad puede invalidar los resultados de los tests realizados al aplicar MCO, se debe aplicar algún procedimiento para deter-

© Ediciones Pirámide

minar la posible presencia de heteroscedasticidad. Igual que para el problema de la multicolinealidad existen diferentes procedimientos para detectarla y ninguno es considerado de manera general como perfecto. Por tanto, se verán sólo algunos de los procedimientos que frecuentemente son aplicados.

Método gráfico

Puesto que el problema de la heteroscedasticidad depende de las perturbaciones y éstas son desconocidas, lo que se hace es observar los residuos. Así se representarán en unos ejes de coordenadas los residuos MCO al cuadrado en función de los valores estimados de la variable dependiente. Si dicho gráfico presenta algún patrón sistemático, como, por ejemplo, el de la figura 6.4.*b*, entonces se sospechará la presencia de heteroscedasticidad, mientras que si se observa un comportamiento aleatorio (figura 6.4.*a*), se podrá pensar en la presencia de homoscedasticidad.

Pero además de este gráfico también es conveniente realizar el gráfico de los residuos al cuadrado respecto de cada una de las variables explicativas. Este tipo de gráfico nos ayudará a identificar si alguna de las variables explicativas es la causante de la presencia de heteroscedasticidad.

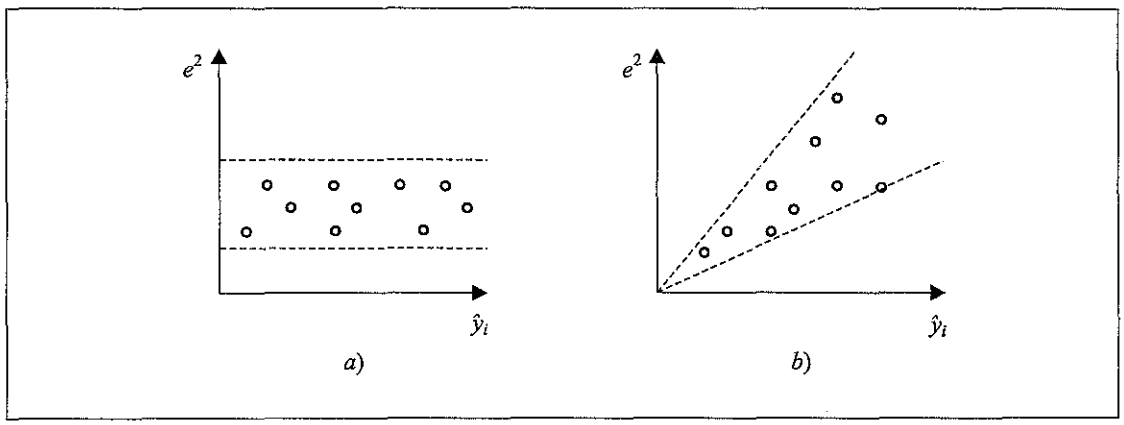

Figura 6.4.

Prueba de Park

Esta prueba se desarrolla en dos fases:

1. Especificar un modelo de regresión de los residuos MCO al cuadrado respecto de la variable explicativa que suponemos está provocando la heteroscedasticidad. Para especificar el modelo y seleccionar la variable que se supone puede provocar la heteroscedasticidad nos apoyaremos en los gráficos anteriores.
2. Posteriormente verificaremos la significación estadística del coeficiente del modelo así planteado, usando el test de la *t*-Student. Si no es significativa supondremos que dicha variable no provoca heteroscedasticidad.

© Ediciones Pirámide

El inconveniente fundamental de esta prueba radica en que las perturbaciones de este último modelo puede que no satisfagan los supuestos básicos y, en consecuencia, la prueba de significación no resulte válida.

Prueba de Goldfeld-Quandt

Es tal vez la más frecuentemente usada, y se realiza en las siguientes fases:

1. Ordenar las observaciones de manera creciente respecto de la variable que se supone provoca la heteroscedasticidad.
2. Omitir c datos centrales, donde c es un valor arbitrario que podría ser, por ejemplo, un 20 o 25% de los datos, dejando de esta forma dos grupos de datos de igual tamaño $(n-c)/2$.
3. Ajustar la regresión MCO por separado de los $(n-c)/2$ primeros y segundos datos, obteniendo en cada regresión anterior la SCR_1 y la SCR_2.
4. Se verifica la hipótesis nula H_0: homoscedasticidad usando el estadístico:

$$F = \frac{SCR_2}{SCR_1} \sim F_{(n-c-2k)/2,(n-c2k)/2}$$

Si el valor del estadístico es mayor que el valor de las tablas se rechaza la hipótesis de homoscedasticidad y, en caso contrario, no se puede rechazar. Esta prueba tiene los inconvenientes de que se elija adecuadamente el valor de c, y también de que la heteroscedasticidad dependa de una sola variable explicativa.

Prueba de White

Es una prueba robusta que no depende de los supuestos de normalidad de las perturbaciones ni de la ordenación en las variables. Esta prueba se realiza en tres pasos:

1. Se obtienen los residuos al cuadrado del modelo original.
2. Se lleva a cabo una regresión auxiliar entre dichos residuos al cuadrado, un término independiente y todas las variables explicativas, sus cuadrados y productos cruzados del modelo original, obteniéndose el R^2 de dicha regresión auxiliar.
3. Se verifica la hipótesis nula H_0: Homoscedasticidad usando el estadístico:

$$nR^2 \sim \chi^2_m$$

donde n es el número de observaciones y m el número de variables explicativas en la regresión auxiliar sin contar el término independiente. Si nR^2 es mayor que el valor de las tablas de la χ^2_m se rechaza la hipótesis de homoscedasticidad.

© Ediciones Pirámide

5.3. Soluciones

Una vez detectada la presencia de heteroscedasticidad el paso siguiente consistirá en aplicar algún método de estimación que proporcione estimaciones eficientes. Se demuestra que el estimador mínimo cuadrático generalizado (EMCG) bajo presencia de heteroscedasticidad y/o autocorrelación es un estimador ELIO, pero para aplicar dicho estimador se requiere conocer la matriz de varianzas y covarianzas de las perturbaciones, la cual raramente se conoce. Desde el punto de vista práctico lo que se hace en presencia de heteroscedasticidad es aplicar mínimos cuadrados ponderados[3] (MCP), para lo cual es necesario establecer un supuesto sobre el comportamiento de las varianzas de las perturbaciones. En resumen, el procedimiento para estimar la presencia de heteroscedasticidad es:

1. Se estima el modelo original por MCO y a partir de los residuos de dicho modelo se establece un supuesto sobre el comportamiento de la varianza de las perturbaciones. Esto se puede realizar a partir de los gráficos de los residuos al cuadrado en función de las variables explicativas y también de los tests anteriores. Por ejemplo, a partir de dichos gráficos y test se sospecha que: $\sigma_i^2 = \sigma^2 x_{ji}^2$.
2. Se ponderan las observaciones de todas las variables dividiéndolas por la raíz cuadrada de la función que provocaba la heteroscedasticidad. En el ejemplo sería dividir cada observación (de la variable explicada y explicativas) por x_{ji}.
3. Se estima el modelo transformado, mediante dichas ponderaciones, por MCO.

6. AUTOCORRELACIÓN

6.1. Concepto y consecuencias

Clásicamente la autocorrelación se estudia para datos ordenados en el tiempo aunque, lógicamente, también ocurre en los datos de corte transversal o espaciales.

Se dice que existe autocorrelación entre las perturbaciones cuando la covarianza de éstas es distinta de cero:

$$\text{Cov}(u_i, u_j) \neq 0 \quad \text{para } i \neq j$$

Por tanto, existe autocorrelación en las perturbaciones cuando el valor que toma la perturbación en un momento depende del valor que toma ésta en otro momento. Lo usual es suponer que las perturbaciones siguen un proceso autorregresivo de pri-

[3] Que es un caso particular de MCG.

© Ediciones Pirámide

mer orden, esto es, que la perturbación en un momento t depende del valor de la perturbación en el momento anterior $t-1$:

$$u_t = \rho u_{t-1} + \varepsilon_t$$

donde ρ es un coeficiente que en términos absolutos es menor que 1, y ε es un término de perturbación bien comportado, esto es, con media 0, varianza constante y no autocorrelación.

La presencia de autocorrelación se debe a que:

— La variable dependiente dependa del tiempo o de la localización espacial donde se mide.

— Se hayan producido errores de especificación en el modelo, bien por la omisión de variables relevantes bien por la forma funcional que liga las variables.

— Existan comportamientos cíclicos en las variables.

— Se manipulen inadecuadamente los datos, etc.

Cuando se estima un modelo econométrico por MCO en presencia de autocorrelación en las perturbaciones el EMCO será, al igual que cuando había heteroscedasticidad, ineficiente.

6.2. Detección

Al igual que para la heteroscedasticidad, en el caso de la autocorrelación existen diferentes procedimientos o pruebas para detectarla, aunque aquí sí que existe un test generalmente usado y que la mayoría de los paquetes informáticos estadísticos proporcionan de manera automática: es el test de Durbin-Watson, aunque el método más intuitivo, y que primero debe aplicarse, consiste en la representación gráfica de los residuos frente al tiempo. Si dicho gráfico presenta algún patrón de comportamiento, como, por ejemplo, los que aparecen en *b)* y *c)* (figura 6.5), se sospechará la presencia de autocorrelación, y si, por el contrario, dicho gráfico presenta un comportamiento aleatorio como el de la figura *a)*, indicaría que no hay autocorrelación. Lógicamente este método gráfico debe ser contrastado con algún test, como, por ejemplo, el de Durbin-Watson.

Test de Durbin-Watson

El estadístico d de Durbin-Watson se define como:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

© Ediciones Pirámide

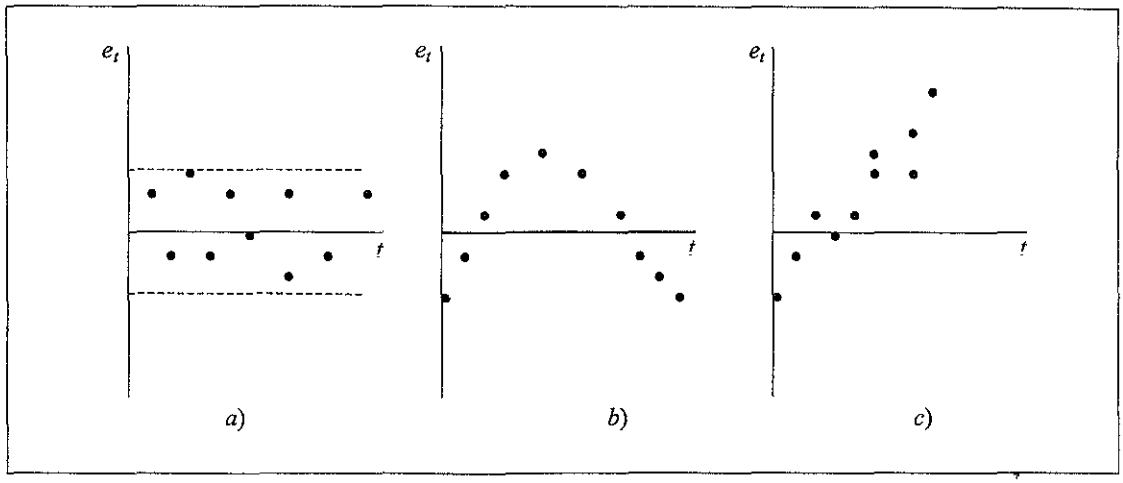

Figura 6.5. Patrones de autocorrelación.

Dicho test debe aplicarse bajo ciertas circunstancias, como son:

— Los datos deben ser temporales.

— Se supone que las perturbaciones siguen un proceso autorregresivo de primer orden.

— El modelo ha de tener término independiente.

— Las variables explicativas son no estocásticas.

— En el modelo no puede aparecer la variable dependiente retardada.

Para aplicar el test se requieren las tablas de la d de Durbin-Watson, en las cuales se encuentran los valores d_L y d_U para un nivel de significación, el n (número de datos) y k' (variables explicativas excluyendo el término independiente).

El test se aplica en las siguientes fases. Se estima el modelo original por MCO y se obtienen los residuos a partir de los cuales se obtiene el valor experimental del estadístico de la d de Durbin-Watson. Usualmente este valor lo proporcionan automáticamente los programas informáticos de estadística:

1. Se toman los valores de d_L y de d_U de las tablas.
2. Se representan los valores anteriores en la figura 6.6.
3. Se sigue la siguiente regla: si el valor experimental de la d cae en el intervalo representado por el signo más, hay autocorrelación positiva; si se sitúa en el intervalo representado por el signo menos hay autocorrelación negativa; si está comprendido en los intervalos con el signo de interrogación, no se puede afirmar ni negar la presencia de autocorrelación, y si cae en el intervalo representado por **No**, es que no hay autocorrelación.

© Ediciones Pirámide

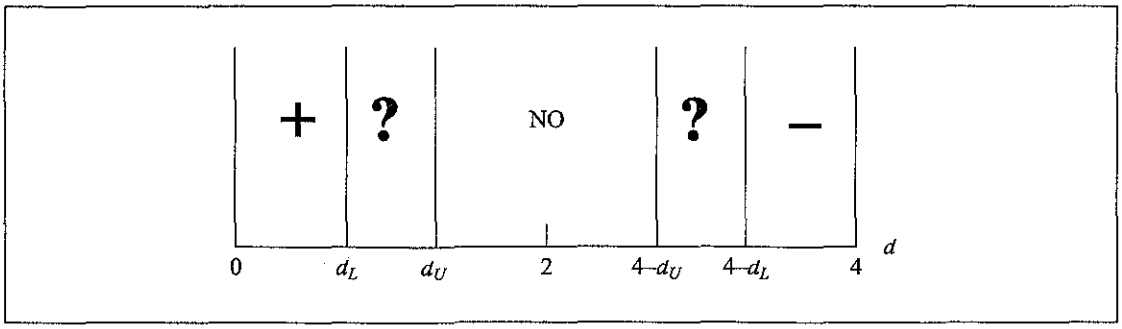

Figura 6.6.

Contraste de Breusch-Godfrey

Un contraste más general que el de Durbin-Watson para detectar la presencia de autocorrelación es el contraste de tipo ML (multiplicadores de Lagrange) desarrollado por Breusch-Godfrey, el cual permite detectar no sólo la presencia de autocorrelación de primer orden, sino de órdenes superiores. Este contraste se puede resumir en las siguientes etapas:

— Se obtienen los residuos MCO del modelo de regresión original: e_t.

— Se realiza la regresión entre los residuos obtenidos anteriormente y las variables explicativas del modelo original, incluyendo además como variables explicativas los residuos retardados: e_{t-1}, e_{t-2}, ..., e_{t-p}. El valor de p indica el orden de autocorrelación que se desea contrastar, obteniéndose de dicha regresión el coeficiente de determinación R^2.

— Se calcula el estadístico $ML = nR^2$ y se compara con el valor de las tablas de la chi-cuadrado con p grados de libertad; si el valor del estadístico es mayor que el de las tablas entonces se rechaza la hipótesis nula de no autocorrelación.

6.3. Soluciones

En primer lugar, se revisa el modelo por si la causa de la existencia de la autocorrelación se debiera a un error de especificación. Para ello nos apoyaremos en fundamentos teóricos en los que se base dicho modelo y también nos serán de utilidad los gráficos de los residuos. Si aun resolviendo el error de especificación persiste el problema, se debería aplicar MCG, que, como en el caso de la heteroscedasticidad, requiere conocer la matriz de varianzas y covarianzas de las perturbaciones, la cual raramente es conocida. Por tanto, habrá que realizar algún supuesto sobre el comportamiento que siguen las perturbaciones, y lo usual es suponer que siguen un proceso autorregresivo de primer orden. Bajo esta hipótesis se pueden aplicar diferentes procedimientos, siendo uno de ellos el procedimiento en dos etapas de Coch-

© Ediciones Pirámide

rene-Orcutt, el cual es fácil de implementar y ofrece buenos resultados. Dicho procedimiento se desarrolla en las dos etapas siguientes:

1. Se estima el modelo original por MCO y se obtienen los residuos, a partir de los cuales se resuelve el siguiente modelo de regresión:

$$e_t = \rho e_{t-1} + v_t$$

y se consigue una estimación de ρ, $\hat{\rho}$.

2. Usando el valor de ρ estimado, se estima por MCO la siguiente ecuación en primeras diferencias generalizadas:

$$(Y_t - \hat{\rho}Y_{t-1}) = \beta_1(1-\hat{\rho}) + \beta_2(X_{2t} - \hat{\rho}X_{2t-1}) + \beta_3(X_{3t} - \hat{\rho}X_{3t-1}) + \cdots$$
$$\cdots + \beta_k(X_{kt} - \hat{\rho}X_{kt-1}) + (u_t - \hat{\rho}u_{t-1})$$

o puesto resumidamente:

$$Y_t^* = \beta_1^* + \beta_2 X_{2t}^* + \beta_3 X_{3t}^* + \beta_k X_{kt}^* + u_t^*$$

donde:

$$Y_t^* = (Y_t - \hat{\rho}Y_{t-1})$$
$$\beta_1^* = (1-\hat{\rho})$$
$$X_{jt}^* = (X_{jt} - \hat{\rho}X_{jt-1})$$
$$u_t^* = (u_t - \hat{\rho}u_{t-1})$$

De esta forma, trabajando sobre las variables transformadas y aplicando MCO a dichas variables se puede obtener una estimación más eficiente de los parámetros del modelo original[4].

7. APLICACIÓN

Son numerosas las aplicaciones conocidas de la regresión lineal, tanto en la economía, en general, como en la empresa y en el marketing. A modo ilustrativo y de forma muy breve, indicaremos algunas de las aplicaciones de esta técnica en el ámbito comercial.

1. Los profesores Pedret, Sagnier y Camp (1994) desarrollaron un modelo en el que el precio de un producto, ya definido en función del resto de elementos del marketing-mix: marca, envase, tamaño, promoción, distribución etc., se determinará según el valor que le es otorgado por el mercado. Los individuos indicaron su ranking de preferencias en los pares de combinaciones «producto-precio» susceptibles de configurar la oferta del mercado objeto de estudio. A la jerarquización efectuada

[4] Para obtener la estimación del término independiente habrá que deshacer la transformación: $\beta_1^*/(1-\hat{\rho})$.

© Ediciones Pirámide

por cada individuo, de todas y cada una de las posibles combinaciones le aplicaron un modelo de regresión múltiple en el que la variable a explicar serán las preferencias sobre el conjunto de combinaciones posibles y las variables explicativas los distintos productos y niveles de precio testados. La estimación del modelo les permitió obtener:

— La utilidad parcial que, en el proceso de compra, proporciona, a cada comprador, cada uno de los productos y cada uno de los niveles de precio testados.
— La utilidad total que, en el proceso de compra, proporciona, a cada comprador, cada combinación «producto-precio».

2. El profesor Rebollo (1992) analizó la variabilidad de precios que ocurre entre los autoservicios, los superservicios y los supermercados, clasificados por el tamaño de ventas. Los factores considerados y que influyen en la dispersión son: factores de demanda (segmentación del mercado, información costosa, costes de búsqueda de precios, etc.), factores de competencia (monopolios, competencia imperfecta, etc.) y factores empresariales (tamaño, poder de mercado, tecnología, etc.).

Los datos utilizados son los de una investigación (IRESCO, 1990) sobre los establecimientos con forma de venta en libreservicio, llevada a cabo por la Dirección General de Comercio Interior del MICT.

La hipótesis contrastada y que no se puede rechazar con los resultados obtenidos es que los comportamientos de los establecimientos en cuanto a precios difiere con el tamaño de la superficie de venta.

3. La profesora Yagüe (1992) realizó un estudio con el fin de explicar las diferencias de márgenes de beneficio de la industria española utilizando los datos de márgenes medios obtenidos para veintisiete sectores industriales españoles entre 1985 y 1989. El núcleo de la investigación estaba centrado en el estudio del efecto de la estructura de mercado sobre los márgenes, poniendo especial énfasis en el análisis de la relación entre el grado de concentración y el margen de beneficio sectorial, y en la existencia o no de estabilidad dinámica en dicha relación. La variable dependiente es el margen y las explicativas fueron las ventas de las cuatro mayores empresas/valor de la producción, variable *dummy* que toma valor 1 cuando el sector está controlado por el sector público, y 0 en caso contrario, activo/ventas (intensidad de capital), tasa de variación del consumo aparente real (efecto ciclo), exportaciones/valor de la producción (propensión exportadora), importaciones/consumo aparente (penetración de importaciones), variable *dummy* con valor 1 para producto de consumo, y 2 para producto industrial y el tamaño medio relativo medido sobre el empleo (economías de escala).

Se especificaron cinco modelos diferentes, de manera que cada modelo posterior incorpora nuevas variables con respecto al anterior. Con este procedimiento pretendían conocer, de un lado, los efectos individuales que ejercen cada uno de los indicadores

© Ediciones Pirámide

en la variabilidad de los márgenes, y de otro lado, los efectos que las nuevas variables ejercen sobre los coeficientes estimados de las variables anteriormente incorporadas.

7.1. Caso práctico[5] (Gastofarma)*

Para desarrollar un caso práctico de aplicación de la regresión lineal hemos analizado el comportamiento del gasto de especialidades farmacéuticas en España[6]. Para ello se ha tomado como referencia un período de 26 años (1970-1995), en el cual el gasto en especialidades farmacéuticas se ha incrementado un 100,24%. Este aumento genera una gran preocupación en la Administración Pública hasta el punto que le lleva a estudiar y poner en marcha medidas para su contención. Tales medidas están tanto orientadas a la demanda como a la oferta. La finalidad que se pretende conseguir con las primeras es intentar reducir la cantidad de productos farmacéuticos consumidos incidiendo sobre el médico-prescriptor, sobre las características de los productos y sobre el enfermo-consumidor, mientras que el objetivo principal de las medidas orientadas a la oferta es actuar sobre el precio de venta o sobre los márgenes, por ejemplo. Indagaciones previas nos llevan a considerar como variables causales del comportamiento del gasto farmacéutico el número de envases prescritos por persona protegida, el precio medio por envase, la renta disponible y la aportación del asegurado.

Conocidas las variables causales del comportamiento del gasto farmacéutico pretendemos probar las siguientes hipótesis:

— H_{01}: El número de envases prescritos por asegurado influye significativamente en el comportamiento del gasto farmacéutico.

— H_{02}: Las modificaciones en el precio medio por envase prescrito afectan significativamente al gasto farmacéutico.

— H_{03}: La renta disponible de los consumidores afecta significativamente al gasto farmacéutico.

Variables

— *Gasto en recetas (GASRECT):* Variable dependiente, medida en pesetas constantes de 1976 (millones). Fuente: Ministerio de Sanidad y Consumo.

— *Número de envases prescritos por asegurado (ENVASEPR):* Variable independiente, expresión individual del consumo en unidades físicas, recoge la tendencia hacia el consumo de productos farmacéuticos por parte de la población asegurada. Fuente: Ministerio de Sanidad y Consumo.

— *Precio medio por envase prescrito (PRECIENV):* Variable independiente, variable influida por la evolución de la política de revisiones de precios adhe-

[5] Para esta aplicación se ha utilizado el programa STATISTICA.

[6] Con anterioridad se había realizado algún estudio similar como el del profesor Cruz Roche (1984).

* Véase fichero en la dirección www.ugr.es/~tluque.

© Ediciones Pirámide

rida a las especialidades farmacéuticas registradas en España, así como por la política de comunicación que los laboratorios desempeñan. Medido en pesetas constantes de 1976. Fuente: Ministerio de Sanidad y Consumo.

— *Renta disponible de los consumidores (RENTDISP):* Variable independiente, medida en pesetas constantes de 1976. Fuente: Informe económico del BBV, 1990-1995.

Modelo teórico

$$GASRECT = \beta_1 + \beta_2 ENVASEPR + \beta_3 PRECIENV + \beta_4 RENTDISP + u_i$$

Modelo obtenido

Gasto recetas = −123.265,77 + 4.567,29*ENVASEPR* + 192,37*PRECIENV* + + 0,60*RENTDISP* + e_i

$R^2 = 0{,}97975086$

R^2 ajustado = 0,97698962

$F(3, 22) = 354{,}89$ $p < 0{,}00000$. Estimación de la desviación típica del error: 2.128,4

TABLA 6.1

Regresión

	Beta	Error estándar de beta	*B*	Error estándar de *B*	*t*(22)	*p*-nivel de significación
Término independiente			−123.265,77	7.490,42	−16,4564	0,000000
ENVASEPR	0,505426	0,035190	4.567,29	318	14,3624	0,000000
PRECIENV	0,499546	0,031237	192,37	12,02	15,9919	0,000000
RENTDISP	0,918701	0,035269	0,60	0,02	26,0477	0,000000

donde

Beta: Son los coeficientes beta, o coeficientes correspondientes a las variables estandarizadas.

B: Son los coeficientes estimados del modelo de regresión.

t(22): Valor experimental del estadístico de la *t*-Student para verificar la H_0: $B_i = 0$. Donde 22 son los grados de libertad $(n - k)$.

p-nivel de significación: Valor que nos indica el nivel de significación mínimo para rechazar la hipótesis nula.

© Ediciones Pirámide

En la tabla 6.1, observando la columna «*p*-nivel de significación» se advierte que para un nivel de confianza del 95% todas las variables son significativas. El modelo en su conjunto es significativo ($F = 354{,}82$) y el coeficiente de determinación es alto (0,98). Los coeficientes (B) nos indican la variación que se produce en el gasto en recetas ante la variación unitaria de alguna de las variables, suponiendo que el resto de variables permanecen constantes. Según el modelo cualquier incremento en el número de envases, precio medio o renta disponible implica un incremento en el gasto en recetas en la cantidad indicada por los coeficientes. Así, por ejemplo, si el número de envases incrementa en una unidad, los gastos en recetas incrementan en 4.567 millones de pesetas.

La relación de estas variables explicativas con la variable explicada es positiva, la Administración Pública, si quiere disminuir el gasto en recetas, tendría que actuar sobre el número de envases y el precio de los mismos.

Atendiendo a los coeficientes beta se podría dar la siguiente ordenación en cuanto a importancia relativa de las variables en el modelo: RENTDISP, PRECIENV, ENVASEPR.

En la figura 6.7 se representan los valores observados de la variable dependiente en función de sus valores estimados. Este gráfico nos indica que el ajuste es bueno (como ya lo recoge el valor del coeficiente de determinación), ya que la nube de puntos se encuentra próxima a la recta. Este gráfico también se usa para analizar la hipótesis de linealidad, la posible presencia de heteroscedasticidad y para detectar datos atípicos.

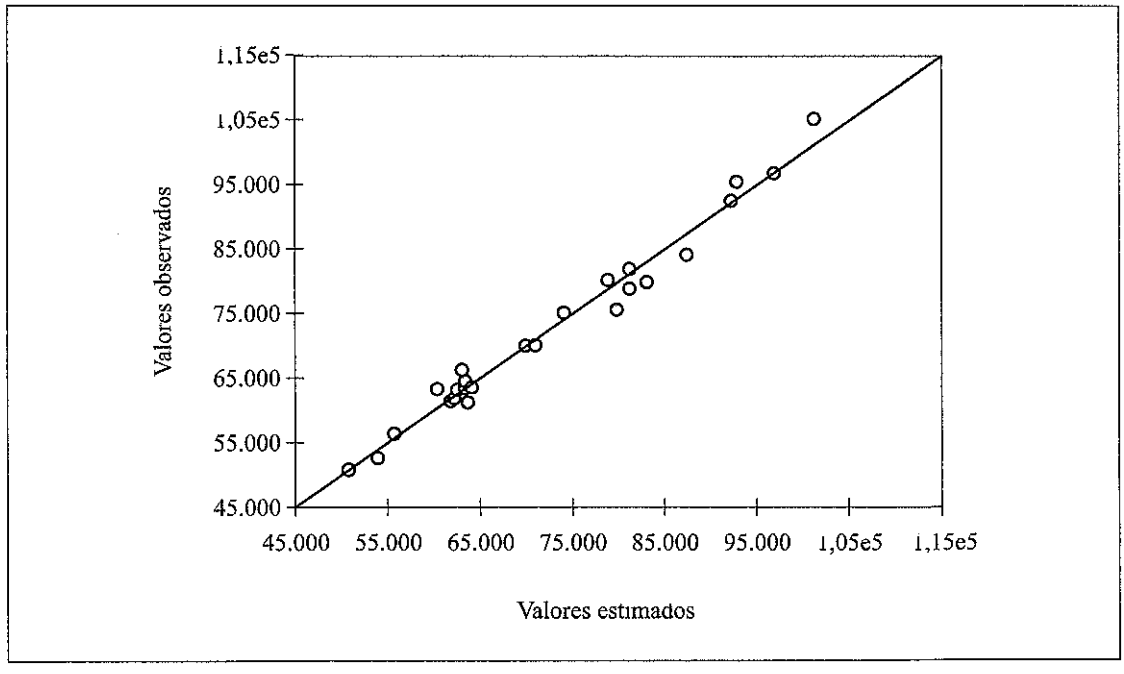

Figura 6.7. Análisis de los residuos. Valores observados de Y respecto de valores estimados. Variable: GASRECT.

© Ediciones Pirámide

Análisis de la varianza

En la tabla 6.2 observamos el valor del estadístico *F* (354,8221), comentado anteriormente. También se muestra el valor de la suma de cuadrados de la explicada (*SCE*), la suma de cuadrados de residuos (*SCR*) y la suma de cuadrados totales (*SCT*).

TABLA 6.2
Análisis de la varianza

	Suma de los cuadrados	*Df*	Cuadrado de la media	*F*	*p*-nivel de significación
Regresión (*SCE*)	48.220.107E2	3	16.073.369E2	354,8221	0,000000
Residual (*SCR*)	99.659.566,	22	4.529.980,		
Total (*SCT*)	49.216.703E2				

Multicolinealidad

Para analizar la posible presencia de multicolinealidad o redundancia de las variables independientes el programa STATISTICA (opción *redundancy*) proporciona la tabla 6.3. En esta tabla la columna R^2 representa el coeficiente de correlación múltiple (R_i^2), que es el que se obtiene de realizar la regresión de cada variable independiente respecto del resto de variables independientes. Cuando el valor de algún R_i^2 es superior o igual a 0,75 se considera que hay un problema de multicolinealidad grave. En nuestro caso ningún R_i^2 toma dichos valores y, por tanto, concluimos que no hay un problema de multicolinealidad grave. También en dicha tabla aparece la tolerancia ($1 - R_i^2$), cuyo uso es similar al del R_i^2. La columna de correlación parcial nos permite dar de nuevo una ordenación en cuanto a la importancia relativa de las variables explicativas. En este caso la ordenación sería: RENTDISP, PRECIENV y ENVASEPR.

La columna de la tolerancia se obtiene a partir de las R_i^2: Tolerancia = $1 - R_i^2$; otra alternativa para descubrir la existencia de multicolinealidad.

TABLA 6.3

	Tolerancia	R_2	Correlación parcial
ENVASEPR	0,743237	0,256763	0,950593
PRECIENV	0,943269	0,056731	0,959578
RENTDISP	0,739904	0,260096	0,984171

© Ediciones Pirámide

Heteroscedasticidad

Para comprobar la presencia de heteroscedasticidad se suele comenzar usando el método gráfico. En este caso se ha representado en unos ejes de coordenadas los residuos al cuadrado en función de los valores estimados de la variable dependiente y de cada variable explicativa (figuras 6.8, 6.9, 6.10 y 6.11).

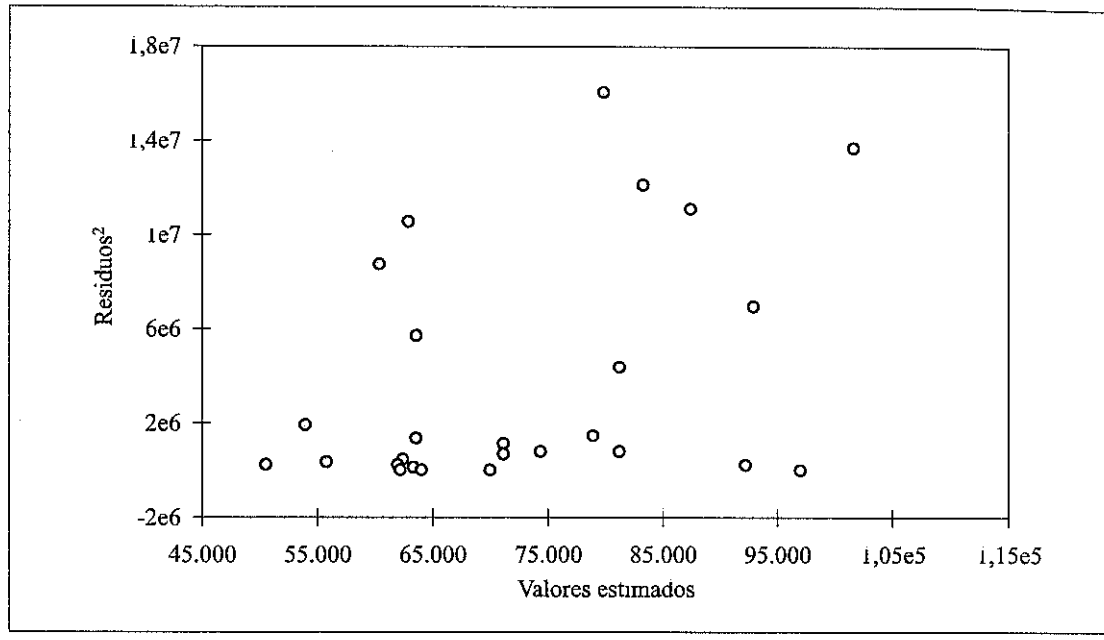

Figura 6.8. Variable dependiente: GASRECT.

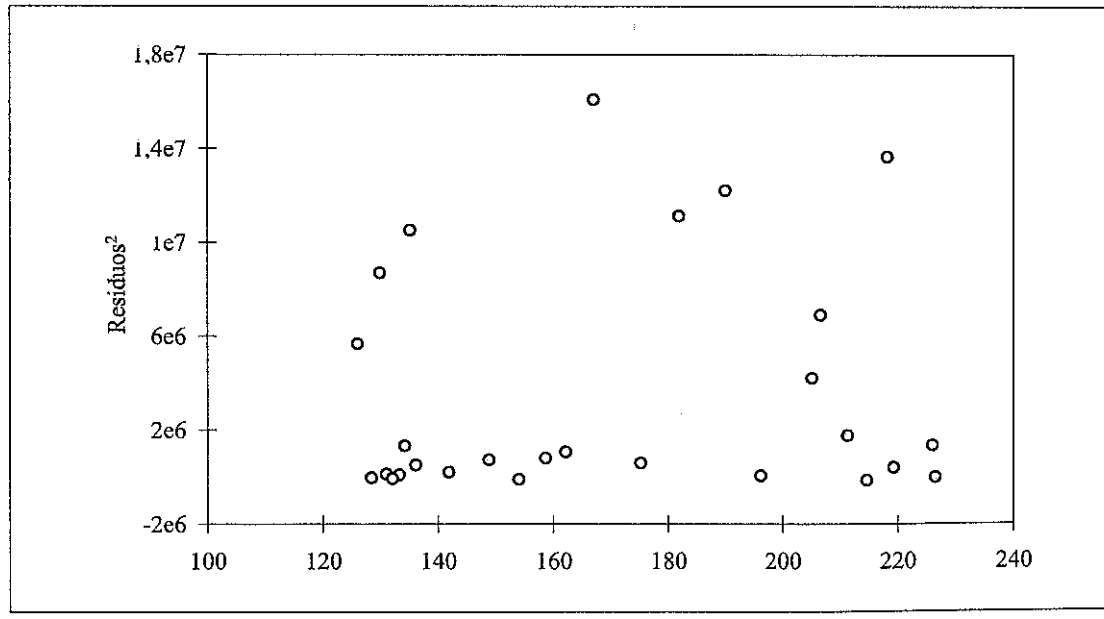

Figura 6.9. Variable independiente: PRECIENV.

© Ediciones Pirámide

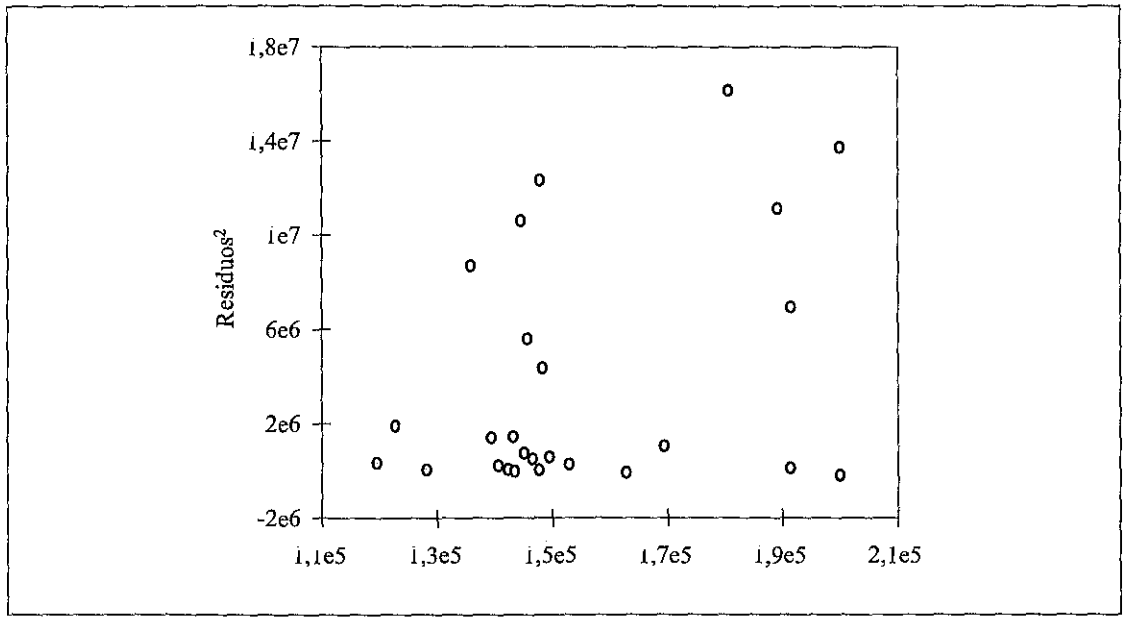

Figura 6.10. Variable independiente: RENTDISP.

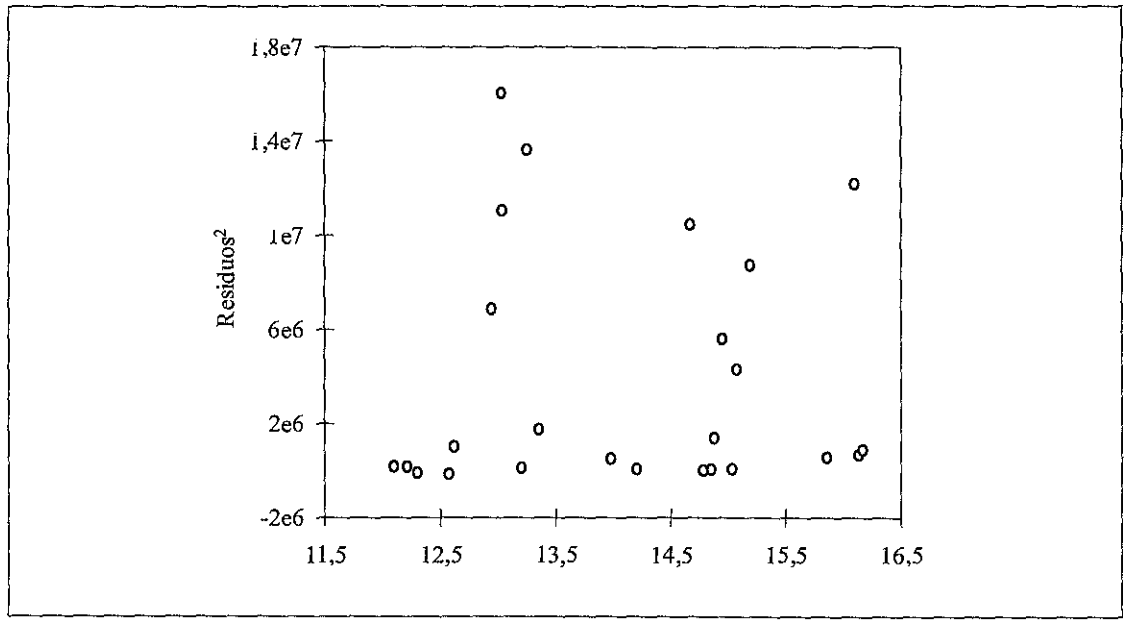

Figura 6.11. Variable independiente: ENVASEPR.

Según la figura 6.8 cabría sospechar de una posible presencia de heteroscedasticidad, puesto que aparentemente en dicho gráfico existe un patrón de comportamiento. Tal gráfico se obtiene en la opción de análisis de residuos. Para determinar si alguna de las variables es la que está provocando la heteroscedasticidad se han obte-

© Ediciones Pirámide

nido las figuras 6.9, 6.10 y 6.11, para lo cual los residuos se han llevado a una hoja de cálculo, se han calculado los cuadrados y posteriormente se han representado.

Además del método gráfico se han aplicado dos tests o pruebas para detectar la heteroscedasticidad. En primer lugar se ha aplicado el test de Park a cada variable explicativa:

1. ENVASEPR

$$R^2 = 0{,}00003919$$

$F(1, 24) = 0{,}00094$ $p < 0{,}97579$. Error estándar del estimador: 5.172E3.

TABLA 6.4

	Beta	Error de beta	*B*	Error de *B*	*t*(24)	*p*-nivel de significación
Término independiente ENVASEPR	–0,006260	0,204120	4.131.600, –20.430,	9.787.080, 666.173,	0,422148 –0,030668	0,676674 0,975788

2. PRECIENV

$$R^2 = 0{,}00961035$$

$F(1, 24) = 0{,}23289$ $p < 0{,}63376$. Error estándar del estimador: 5.147E3.

TABLA 6.5

	Beta	Error de beta	*B*	Error de *B*	*t*(24)	*p*-nivel de significación
Término independiente PRECIENV	0,098032	0,203141	1.511.411, 13.634,	4.915.639, 28.253,	0,307470 0,482583	0,761138 0,633764

3. RENTDISP

$$R^2 = 0{,}13477123$$

$F(1, 24) = 3{,}7383$ $p < 0{,}06506$. Error estándar del estimador: 4.811E3.

TABLA 6.6

	Beta	Error de beta	*B*	Error de *B*	*t*(24)	*p*-nivel de significación
Término independiente RENTDISP	0,367112	0,189872	–10.062.554, 87,	7.248.522, 45,	–1,38822 1,93348	,177822 0,065057

© Ediciones Pirámide

Para un nivel de confianza del 95% en ninguno de los tres modelos el coeficiente de la variable explicativa es significativo y, por tanto, no se admite la presencia de heteroscedasticidad, al menos de tipo lineal.

Además se ha utilizado la prueba de White; para ello hemos obtenido el siguiente modelo:

$$e_t^2 = \beta_0 + \beta_1 ENVASEPR + \beta_2 PRECIENV + \beta_3 RENTDISP + \beta_4 ENVASEPR^2 + \\ + \beta_5 PRECIENV^2 + \beta_6 RENTDISP^2 + \beta_7 ENVASEPR \times \text{PRECIENV} + \\ + \beta_8 ENVASEPR \times RENTDISP + \beta_9 PRECIENV \times RENTDISP$$

con $R^2 = 0{,}45$.

Luego el estadístico $= nR^2 = 26 \times 0{,}45 = 11{,}68$. Comparándolo con el valor para un $\alpha = 0{,}05$ de la $\chi_9^2 = 16{,}92$, podemos indicar que se acepta la H_0 = Homoscedasticidad.

Autocorrelación

Para analizar la autocorrelación se ha usado el procedimiento gráfico, representando los residuos en función del tiempo, usando la opción *residual analysis, plots of residuals, raw residuals* (programa STATISTICA). En la figura 6.12 no se observa claramente un patrón de comportamiento que nos indique la posible presencia de autocorrelación. Un procedimiento más formal para detectar la autocorrelación de primer orden es usar el estadístico Durbin-Watson (tabla 6.7). Comparando el valor de dicho estadístico con el valor teórico de las tablas de Durbin-Watson para un nivel de significación del 5%, $n = 26$ y $k' = 3$, se obtienen los valores de $d_L = 1{,}14$ y $d_U = 1{,}65$, puesto que el valor del estadístico está entre d_L y d_U, caería en la zona de incertidumbre, por lo que no se podría concluir la existencia o no de autocorrelación de primer orden.

TABLA 6.7

Durbin-Watson y correlación serial de los residuos

	Durbin-Watson	Correlación serial
Estimado	1,59682	0,142880

Otro procedimiento para detectar la autocorrelación es el contraste de Breusch-Goldfrey. El modelo obtenido incluyendo como variable explicativa un residuo retardado es el siguiente:

$$e_t = -830{,}48 + 1{,}03 PRECIENV + 22{,}04 ENVASEPR + 0{,}002 RENTDISP + 0{,}15\ e_{t-1}$$

© Ediciones Pirámide

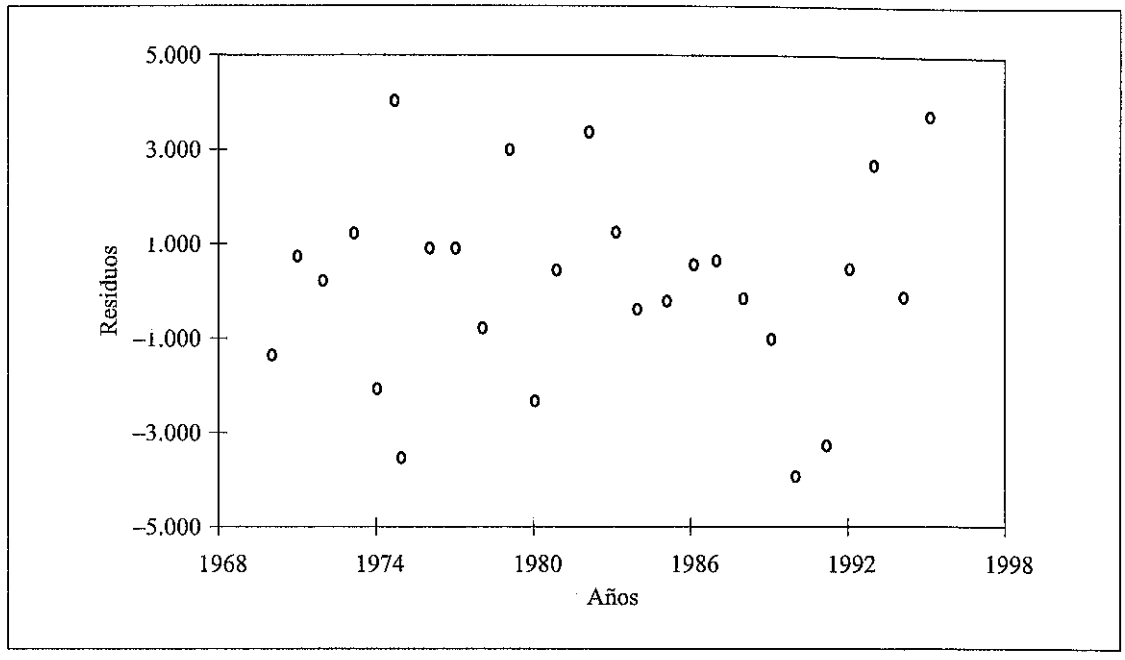

Figura 6.12.

donde:

$$R^2 = 0{,}018$$
$$ML = 0{,}48$$

Si comparamos el $ML = 0{,}48$ con el valor de las tablas de la $\chi^2_1 = 3{,}84$ no podemos rechazar la hipótesis de no autocorrelación.

INVENTARIO DE TÉRMINOS Y CONCEPTOS

- Regresión lineal simple.
- Regresión lineal múltiple.
- Mínimos cuadrados.
- Multicolinealidad.
- Homoscedasticidad.
- Autocorrelación.
- Mínimos cuadrados ordinarios (MCO).
- Maximoverosimilitud.
- Estimados mínimos cuadráticos ordinarios (EMCO).
- Estimador lineal insesgado y óptico (ELIO).

© Ediciones Pirámide

- Suma de cuadrados explicada (SCE).
- Suma de cuadrados de los residuos (SCR).
- Suma de cuadrados de los totales (SCT).
- Coeficiente de determinación.
- Coeficiente de determinación ajustado.
- Coeficientes beta.
- Coeficiente de correlación parcial.
- Elasticidades.
- Predictor lineal insesgado y óptimo (PLIO).
- Coeficiente de correlación múltiple.
- Factor de agrandamiento de la varianza.
- Tolerancia.
- Número de condición.
- Prueba de Park.
- Prueba de Goldfeld-Quandt.
- Prueba de White.
- Mínimos cuadrados ponderados.
- Test de Durbin-Watson.
- Procedimiento de Cochrene-Orcutt.

BIBLIOGRAFÍA

Abascal, E., y Grande, I. (1989): *Métodos multivariantes para la investigación comercial,* Ariel, Barcelona.

Cruz, I. (1984): *Uso racional y financiación pública de los medicamentos en Europa,* Organización Mundial de la Salud, Madrid, 22-26 de octubre.

Frías Jamilena, D. M. (1996): *Comercialización de productos farmacéuticos. Análisis del gasto en productos farmacéuticos,* tesis doctoral, Universidad de Granada.

Gujarati, D. (1997): *Econometría,* McGraw-Hill.

Intriligator, M. (1996): *Econometric Models, Techniques, and Applications,* Prentice-Hall.

Johnston, J. (1987): *Métodos de econometría,* Vicens Universidad.

Lambin, J. J. (1993): *La recherche marketing,* Ediscience.

Malhotra, N. (1997): *Investigación de mercados. Un enfoque práctico,* Prentice-Hall.

Pedret, R., Sagnier, D., y Camp, F. (1994). «Fijación del precio a partir de la utilidad percibida por el mercado», *VI Encuentro de Profesores Universitarios de Marketing.*

Rebollo, A. (1992): «La dispersión de precios en las formas comerciales de libre servicio en España», *IV Encuentro de Profesores Universitarios de Marketing.*

Yagüe, M. J. (1992): «Estructura de mercado y márgenes precio-coste en los sectores industriales españoles», *IV Encuentro de Profesores Universitarios de Marketing.*

© Ediciones Pirámide