Tema 2

Modelo de Regresión Lineal Simple

Contenido

2.1. Introducción. Un ejemplo

Supongamos que nos interesa conocer la relación que hay entre el precio de una vivienda y determinadas características de la misma. Empezaremos considerando el caso más sencillo, una única característica, la superficie. Se trata de cuantificar la influencia que tiene el tamaño de una vivienda en la determinación de su precio de venta mediante un modelo de regresión lineal simple.

En este capítulo vamos a especificar, estimar y analizar el ***modelo de regresión lineal simple***. La teoría necesaria para este fin será ilustrada mediante el estudio simultáneo del conjunto de datos *data3-1* disponible en Gretl dentro del conjunto de datos correspondiente a Ramanathan. Este fichero contiene el precio de venta y la superficie de 14 viviendas vendidas en el área de San Diego. Vamos a comenzar realizando un **análisis gráfico**.

1. Accedemos a este conjunto de datos en *Archivo* → *Abrir datos* →*Archivo de muestra* y en la carpeta de datos de *Ramanathan* seleccionamos *data3-1 House prices and sqft*:

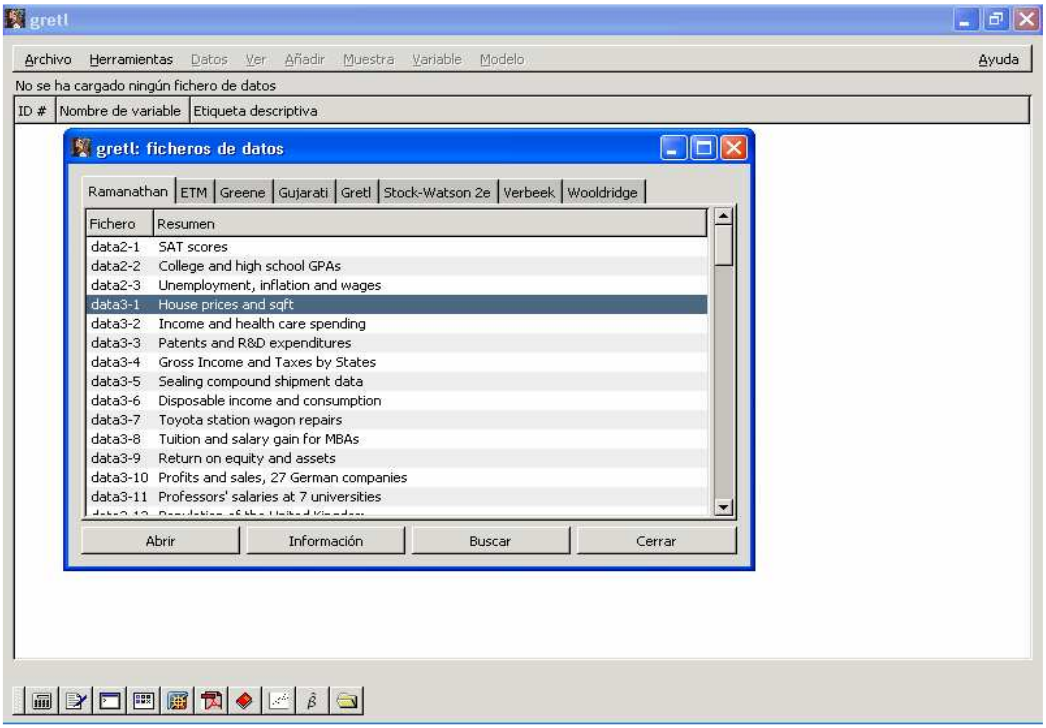


Gráfico 2.1: Selección de un fichero de muestra

Se abre un fichero que contiene tres variables, *const*, *price* y *sqft*. La Tabla 2.1 muestra los valores disponibles para cada variable.

2. En *Datos* →*Leer información* aparece la siguiente descripción del conjunto de datos:

```
DATA3-1: Precio de venta y superficie hábil de viviendas unifamiliares
    en la comunidad universitaria de  San Diego en 1990.
    price = Precio de venta en miles de dólares (Rango 199.9 - 505)
    sqft  = Pies cuadrados de área habitable (Rango 1065 - 3000)
```

i	P_i	F2	i	P	F2
1	199,9	1065	8	365,0	1870
2	228,0	1254	9	295,0	1935
3	235,0	1300	10	290,0	1948
4	285,0	1577	11	385,0	2254
5	239,0	1600	12	505,0	2600
6	293,0	1750	13	425,0	2800
7	285,0	1800	14	415,0	3000

Tabla 2.1: Conjunto de datos incluidos en *data3.1 House prices and sqft*

3. Seguidamente en *Variable* →*Editar atributos* cambiamos los nombres a las variables (*P* y *F2*), la descripción (*Precio de venta en miles de dólares* y *Pies cuadrados hábiles*) y el nombre a mostrar (*Precio, P* y *Superficie, F2*)

4. Guardamos los cambios en un fichero llamado *datos-cap3.gdt* con *Archivo* →*Guardar datos*.

5. Abrimos el diagrama de dispersión entre las dos variables (ver el Gráfico 2.2). En él observamos una relación lineal positiva entre P y $F2$.

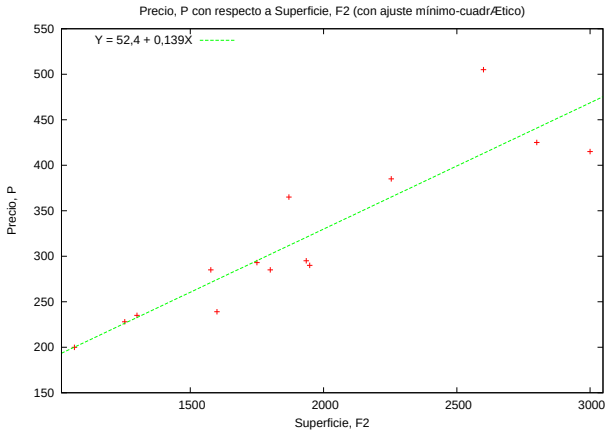


Gráfico 2.2: Diagrama de dispersión precio-superficie de viviendas

Un modelo sencillo que recoge una relación lineal causa-efecto entre superficie y precio es $P_i = \alpha + \beta F2_i$. Esto quiere decir que el precio de una vivienda depende *únicamente* de su superficie y, por lo tanto, dos viviendas de igual tamaño deben tener *exactamente* el mismo precio. Esta hipótesis es poco realista porque diferencias en otras características, como la orientación de la casa o su estado de conservación, también influyen en su precio. Debemos, por tanto, especificar un modelo econométrico que recoge esta característica: el modelo de regresión lineal simple.

2.2. Elementos del modelo de regresión simple

El modelo simple relaciona dos variables de forma lineal,

$$Y_i = \alpha + \beta X_i + u_i \qquad i = 1, \ldots, N \tag{2.1}$$

donde:

- Y es la ***variable a explicar***, ***variable dependiente o endógena***, es decir, la variable que estamos interesados en explicar.
- X es la ***variable explicativa***, ***variable independiente o exógena***.
- La ordenada α y la pendiente β del modelo son los ***coeficientes de la regresión***. Si definimos K como el *número de coeficientes desconocidos a estimar*, en el modelo de regresión simple tenemos $K = 2$ coeficientes a estimar.
- u es el término de error, variable aleatoria o ***perturbación***.
- El subíndice i denota ***observación***. En general, el subíndice i será empleado cuando la muestra contenga datos de sección cruzada y el subíndice t cuando tengamos observaciones correspondientes a series temporales, aunque esto no es de especial relevancia.
- N es el ***tamaño muestral***, número de observaciones disponibles de las variables de estudio (Y, X). Cuando tratemos con datos temporales T denotará el tamaño muestral[1].

El error u_i se introduce por varias razones, entre las cuales tenemos:

- Efectos impredecibles, originados por las características de la situación económica o del contexto de análisis, y efectos no cuantificables derivados de las preferencias y los gustos de los individuos o entidades económicas.
- Errores de medida producidos a la hora de obtener datos sobre las variables de interés.
- Errores de especificación ocasionados por la omisión de alguna variable explicativa o bien, por las posibles no linealidades en la relación entre X e Y.

Modelo para la relación precio-tamaño del piso. En este caso planteamos el siguiente modelo de regresión lineal:

$$P_i = \alpha + \beta\ F2_i + u_i \qquad i = 1, \ldots, N \tag{2.2}$$

donde

- P_i es la observación i de la variable dependiente (endógena o a explicar) *precio de venta* en miles de dólares.

[1] En este capítulo y los siguientes, por simplicidad, no reservaremos la letra mayúscula para variables aleatorias X y las minúsculas para realizaciones (x) sino que utilizaremos mayúsculas tanto para una variable aleatoria como como para su realización, es decir, para los datos.

- $F2_i$ es la observación i de la variable independiente (exógena o explicativa) *área habitable* en pies cuadrados.
- Los dos coeficientes a estimar son α y β, y sospechamos que al menos β tiene valor positivo ya que a mayor superficie habitable de la vivienda su precio lógicamente se esperará sea mayor.
- En este modelo el término de error o perturbación u_i recogería características específicas de los pisos: lugar en el que se sitúa, orientación de la casa, vistas, etc., es decir, características que diferencian el precio de los pisos que tienen la misma superficie habitable.

Un primer objetivo del análisis econométrico es conocer α y β, que son los parámetros de la relación entre P y $F2$. Del total de viviendas del área objeto de estudio, tenemos una muestra con datos de N= 14 pisos. Por tanto, el objetivo del estudio es *inferir*, a partir de la muestra, la relación precio-tamaño de una vivienda en la población. Para llevar a cabo esta inferencia es necesario determinar la naturaleza aleatoria de las variables que intervienen en el estudio.

2.3. Hipótesis básicas

El modelo (2.1) debe completarse con la especificación de las propiedades estocásticas de la variable de interés Y. A partir de las propiedades de Y, es posible conocer las propiedades de los distintos métodos de estimación, elegir el mejor estimador en el modelo, realizar contrastes, etc. Las condiciones bajo las cuales vamos a trabajar en un principio se denominan ***hipótesis básicas***. Bajo estas hipótesis estimaremos y analizaremos el modelo para, finalmente, predecir Y. En una segunda etapa, podemos considerar otras situaciones, relajando algunas de estas hipótesis, analizando si los procedimientos de estimación y contraste anteriores siguen siendo válidos. Las hipótesis básicas se refieren a los distintos elementos de la regresión.

• *Sobre la forma funcional*

1. El modelo es lineal en los coeficientes. Los modelos a estimar a lo largo del curso son lineales en los coeficientes, $Y_i = \alpha + \beta X_i + u_i$. Sin embargo, podemos permitir no linealidades en las variables explicativas como puede ser la especificación:

$$P_i = \alpha + \beta \left(F2_i\right)^2 + u_i$$

en la que la superficie habitable de los pisos no influye de forma lineal sobre el precio, sino de forma cuadrática.

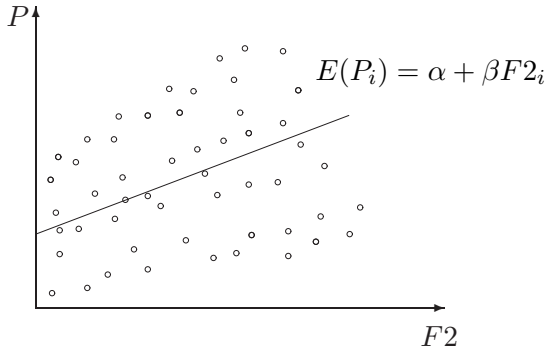
• *Sobre los coeficientes*

2. Los coeficientes α y β se mantienen constantes a lo largo de la muestra. Vamos a considerar que la influencia de las variables explicativas es estable a lo largo de la muestra. Supongamos que estamos interesados en analizar, en términos medios, el precio de los

pisos de Bilbao (P) en función de la superficie habitable en metros cuadrados ($F2$). En este caso interesaría estimar la *recta central* representada en el caso 1 del Gráfico 2.3.

No obstante, supongamos que algunos de estos pisos están localizados en el centro de Bilbao (representados en azul) y que otros están localizados en la periferia (en rojo). El caso 2 del Gráfico 2.3 muestra esta hipotética situación: en general, para una determinada superficie, los pisos del centro tienen mayor precio. Así, en el gráfico es posible distinguir dos nubes de puntos, cada una asociada a pisos de una determinada zona. Si este fuera el caso, estaríamos dispuestos a creer que existen (y debemos estimar) *dos rectas centrales* (la azul y la roja) permitiendo que tanto la ordenada como la pendiente cambien a lo largo de la muestra, dependiendo de la zona en la que se localice el piso.

Caso 1: Sin discriminar por localización



Caso 2: Discriminando por localización

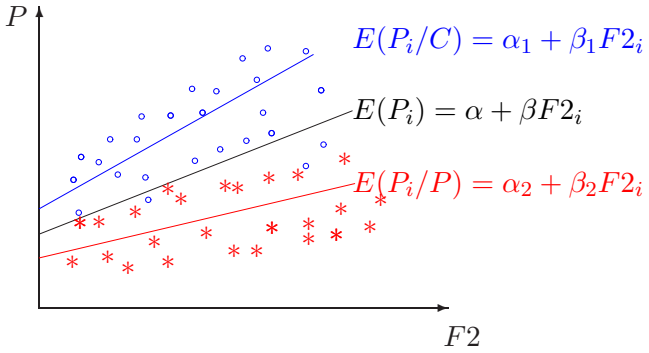


Gráfico 2.3: Precio pisos de Bilbao *vesus* superficie habitable

- *Sobre la variable endógena*

3. La variable endógena es cuantitativa. A lo largo de este curso básico vamos a suponer que la variable a explicar es cuantitativa. Lo contrario, una variable endógena cualitativa, requiere métodos de estimación alternativos al método que se analiza en este curso.

- *Sobre la variable explicativa*

4. La variable explicativa X tiene varianza muestral S_X^2 no nula y además $N \geq K = 2$. Estas hipótesis son necesarias para poder identificar los coeficientes (ordenada y pendiente). En primer lugar, si el número de coeficientes a estimar fuera mayor que el número de observaciones disponibles en la muestra, no tenemos suficiente información para poder llevar a cabo la estimación. Más adelante veremos que esta condición debe hacerse más estricta, $N > 2$, si además de estimar los dos parámetros α y β que determinan el valor medio de Y, nos interesa estimar su variabilidad.

   Por otra parte, si la variable explicativa tuviera varianza muestral nula ($S_X^2 = 0$), es decir, si la variable explicativa tomase un valor constante, por ejemplo, $X_i = 5\ \forall i$, la pendiente y la ordenada no podrían ser identificadas. Esto se debe a que la variable X es una combinación lineal del término constante, $X = 5 \times$ término constante $= 5 \times 1 =$

5. De hecho, tal y como se puede observar en el Gráfico 2.4, una situación de estas características no puede explicar las variaciones de la variable de interés Y.

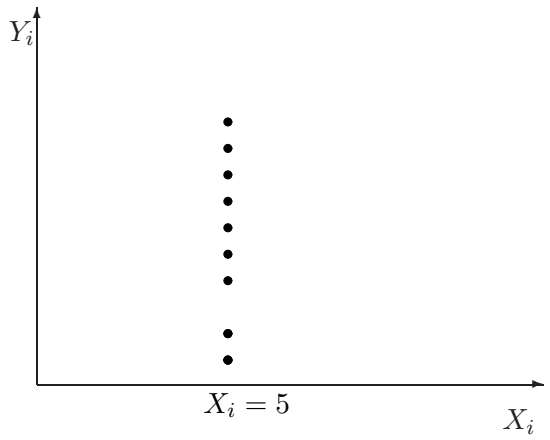


Gráfico 2.4: Modelo $Y_i = \alpha + \beta \times 5 + u_i$, con $S_X^2 = 0$

5. La variable exógena X es fija, no aleatoria. Las observaciones del regresor $X_1, \ldots X_N$ son valores fijos en muestras repetidas, es decir, suponemos que trabajamos en un contexto de experimento controlado. Esta condición implica que la variable explicativa X no podrá estar medida con error. En el caso práctico que estamos considerando, esto significa que los metros cuadrados habitables están medidos con exactitud. En muchos casos es un supuesto poco realista, pero lo utilizamos como punto de partida. El contexto en el que la variable explicativa X tiene carácter aleatorio se estudia en textos más avanzados, por ejemplo, Wooldridge (2003) o Alonso, Fernández & Gallastegui (2005).

6. El modelo está bien especificado. En general, esta hipótesis requiere que en el modelo no se incluyan variables irrelevantes ni que se omitan variables relevantes para explicar Y. En el contexto del modelo de regresión simple, esto significa que la variable explicativa X es la única variable relevante para explicar y predecir la variable de interés Y.

• *Sobre la perturbación*

El término de error recoge aquellos elementos que afectan a la variable de interés y que no observamos. Podemos hacer conjeturas sobre los valores que puede tomar, cuáles son más probables y cuáles menos. Así, consideramos que u_i es aleatorio y tiene las siguientes propiedades.

7. La perturbación tiene media cero. El error impredecible, la parte aleatoria del modelo, tiene media cero. Esto implica que la parte sistemática del modelo $(\alpha + \beta X_i)$ puede ser interpretada como el comportamiento medio a analizar, es decir, $E(Y_i) = \alpha + \beta X_i$.

8. La perturbación tiene varianza constante. Suponemos que la variabilidad del error se mantiene constante, $var(u_i) = \sigma^2$, $\forall i$ (ver caso 1 del Gráfico 2.5). De este modo, como puede verse en la distribución de la figura izquierda del Gráfico 2.6, dados unos valores específicos de la variable explicativa, el rango de posibles valores que puede tomar la variable endógena tiene la misma amplitud y la probabilidad de observar elementos alejados de la media no depende del valor que tome la variable explicativa X.

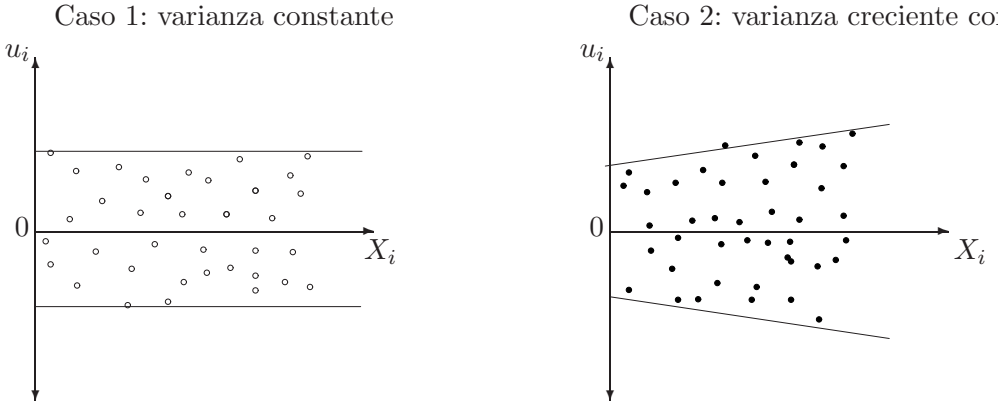


Gráfico 2.5: Ejemplos de realizaciones de u

En el caso contrario, estaríamos hablando de perturbaciones heterocedásticas, cuya dispersión puede variar a lo largo de la muestra (ver caso 2 del Gráfico 2.5). En el caso de los pisos, significaría, por ejemplo, que el rango de los precios de los pisos con menor superficie es más pequeño que el de los pisos con mayor superficie habitable (ver la figura derecha en el Gráfico 2.6). En otras palabras, los pisos pequeños y con la misma superficie tienen los precios bastante parecidos. Sin embargo, a medida que aumenta la superficie, la holgura crece y podemos encontrar pisos grandes de igual tamaño a diversos precios; es decir, $var(u_i)$ es una función creciente en X.

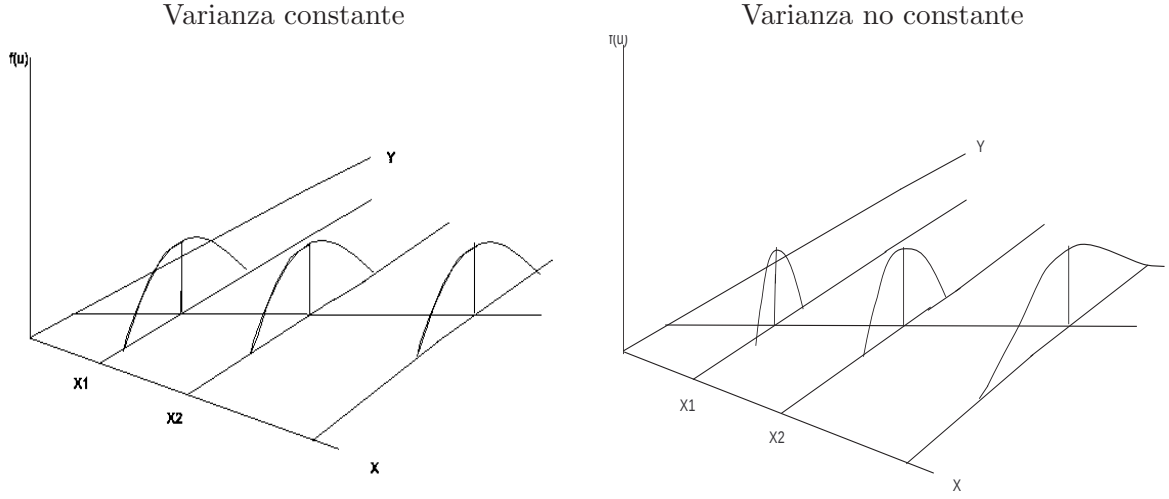


Gráfico 2.6: Ejemplos de distribución de Y

9. La perturbación no está autocorrelacionada. Por el momento vamos a suponer que la correlación entre dos observaciones distintas cualesquiera de la perturbación es cero, $corr(u_i, u_j) = r_{u_i,u_j} = 0;\ \forall i \neq j$. Esto implica que las covarianzas entre dos perturbaciones también es cero: $cov(u_i, u_j) = 0, \forall i \neq j$.

10. La perturbación sigue una distribución normal. Este último supuesto, como veremos más adelante, no se necesita para la estimación ni para la obtención de propiedades del estimador[2]. Sin embargo es necesario para poder realizar contraste de hipótesis o calcular intervalos de confianza.

2.3.1. Resumen: modelo de regresión lineal simple con hipótesis básicas

Abreviadamente, el modelo con las hipótesis básicas mencionadas se escribe:

$$Y_i = \alpha + \beta X_i + u_i, \quad X_i \text{ fija y } u_i \sim NID(0, \sigma^2) \quad \forall i$$

Es decir, $Y_i \sim NID(\alpha + \beta X_i, \sigma^2)$, siendo α, β y σ^2 parámetros desconocidos. En particular, nos interesamos por los parámetros de la media y su interpretación en este modelo es:

- $\alpha = E(Y_i|X_i = 0)$: valor medio o esperado de la variable endógena cuando el valor que toma la variable exógena es cero.

- $\beta = \dfrac{\Delta E(Y_i)}{\Delta X_i} = \dfrac{\partial E(Y_i)}{\partial X_i}$: un aumento unitario en la variable explicativa conlleva un aumento medio de β unidades en la variable endógena. La pendiente mide el efecto de un aumento marginal en la variable explicativa sobre $E(Y_i)$.

→ Así, volviendo a nuestro ejemplo tenemos que:

$\alpha = E(P_i|F2_i = 0)$ es el precio medio de venta en miles de dólares cuando el piso dispone de una superficie de cero pies habitables, que también puede ser considerado como precio mínimo de partida. En este caso, esperaríamos un coeficiente nulo dado que no tiene sentido hablar de un piso sin superficie hábil o bien un precio de partida positivo. No obstante, aunque en este contexto la ordenada no tiene en principio mucho sentido, no debemos de eliminarla a la ligera en aras de obtener resultados fáciles de interpretar.

$\beta = \dfrac{\Delta E(P_i)}{\Delta F2_i}$ indica que, cuando un piso aumenta su superficie hábil en un pie cuadrado, su precio medio aumenta en β miles \$.

2.4. Estimación por Mínimos Cuadrados Ordinarios

Una vez descrito el ámbito en el que nos vamos a mover, vamos a obtener un estimador adecuado de los coeficientes del modelo de regresión simple: el estimador de mínimos cuadrados ordinarios. En primer lugar, obtendremos el estimador y, a continuación, justificaremos su uso en base a sus propiedades. El modelo simple (2.1) nos indica que cada observación Y_i es una realización de una variable que tiene dos componentes: uno que depende del valor del regresor X_i, cuyo valor observamos, y un componente residual que no observamos. Esto significa que tenemos N igualdades con una misma estructura:

[2] Esto es así porque el método de estimación que se va a derivar es el de Mínimos Cuadrados Ordinarios. Sin embargo, si se estimase por máxima verosimilitud el supuesto de normalidad sobre la distribución de Y sí es necesario para la obtención del estimador.

$$\begin{aligned} Y_1 &= \alpha + \beta X_1 + u_1 \\ &\vdots \\ Y_i &= \alpha + \beta X_i + u_i \\ &\vdots \\ Y_N &= \alpha + \beta X_N + u_N \end{aligned}$$

El Gráfico 2.7 representa gráficamente una posible muestra. Los puntos (Y_i, X_i) se sitúan o distribuyen alrededor de la recta $\alpha + \beta X_i$. La desviación de cada punto respecto a esta *recta central* viene dada por el valor que tome el término de error no observable u_i. Por ejemplo, en el Gráfico 2.7, la perturbación es positiva para la primera observación, de modo que Y_1 se encuentra por encima de la recta central. Por otro lado, el punto (Y_2, X_2) se encuentra por debajo de la recta central, es decir, u_2 toma un valor negativo.

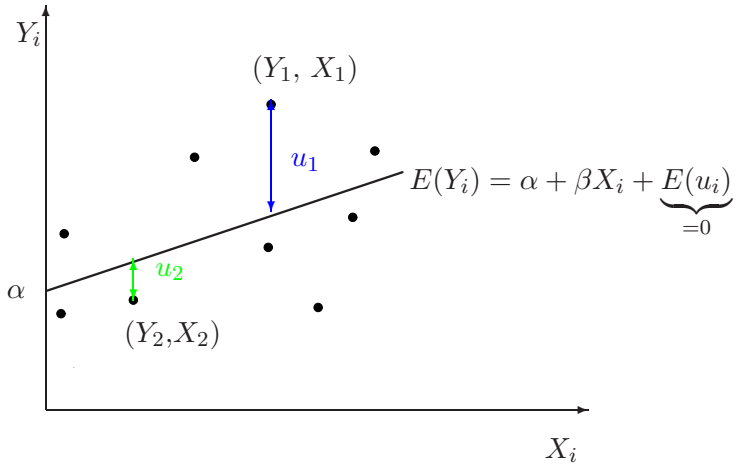


Gráfico 2.7: Modelo de regresión simple

Así, la recta central sería aquella recta que se obtiene cuando el valor de la perturbación es cero. Teniendo en cuenta que suponemos que la perturbación tiene media cero, es decir, que no tiene efectos sistemáticos sobre Y, la *recta central* recoge el comportamiento medio de la variable de interés. La ***estimación*** de un modelo de regresión pretende obtener una aproximación a esta recta central no observable. En términos econométricos, queremos calcular el comportamiento medio de la variable de interés, $\alpha + \beta X_i$, a partir de observaciones provenientes de una muestra $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_N, X_N)$. Gráficamente, la estimación consiste en calcular la pendiente y la ordenada que mejor se ajusta a la nube de puntos.

Antes de proceder a la estimación del modelo es preciso definir algunos nuevos conceptos. La recta central objeto de estimación se denomina ***Función de Regresión Poblacional (FRP)*** y depende de los coeficientes poblacionales desconocidos α y β. Se trata de la parte sistemática o predecible del modelo y corresponde al comportamiento medio o esperado de la variable a explicar:

$$E(Y_i) = E(\alpha + \beta X_i + u_i) \ = \ \alpha + \beta X_i + \underbrace{E(u_i)}_{=0} \ = \ \alpha + \beta X_i$$

La ***perturbación*** del modelo recoge todo aquello que no ha sido explicado por la parte sistemática del modelo y se obtiene como la diferencia entre la variable a explicar y la recta de regresión poblacional:

$$u_i = Y_i - \alpha - \beta X_i$$

El resultado final obtenido a partir de la información que ofrece una muestra dada se define como la ***Función de Regresión Muestral (FRM)***. Se obtiene una vez que los coeficientes de la regresión hayan sido estimados $(\hat{\alpha}, \hat{\beta})$ y también se conoce como ***modelo estimado***:

$$\hat{Y}_i = \widehat{E(Y_i)} = \hat{\alpha} + \hat{\beta} X_i$$

El ***residuo*** mide el error cometido al estimar la variable endógena y se define como la diferencia entre la variable a explicar y la recta de regresión muestral:

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta} X_i = \alpha + \beta X_i + u_i - \hat{\alpha} - \hat{\beta} X_i \\ &= (\alpha - \hat{\alpha}) + (\beta - \hat{\beta}) X_i + u_i \end{aligned} \tag{2.3}$$

Este error proviene de dos fuentes: la primera, por el hecho de no poder obtener los valores de la perturbación (u_i) y la segunda se debe a que la estimación de los coeficientes desconocidos (α, β) introduce un error adicional. Es importante, por tanto, diferenciar y no confundir el residuo con la perturbación.

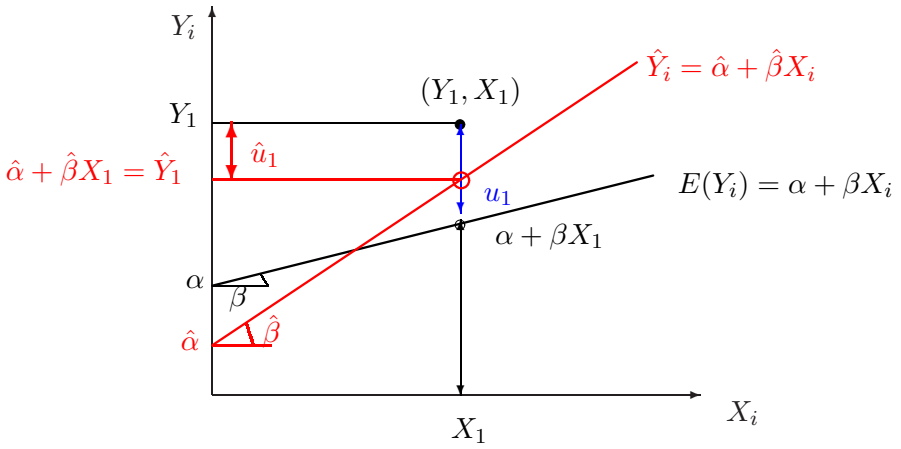


Gráfico 2.8: Función de regresión poblacional y función de regresión muestral

En el Gráfico 2.8 la función de regresión poblacional está trazada en color negro así como los coeficientes poblacionales, la ordenada (α) y la pendiente (β). Podemos ver que el valor Y_i se obtiene como la suma del valor que toma la parte sistemática $\alpha + \beta X_i$ (situada sobre la FRP) y del valor que toma la perturbación u_i, esto es, $Y_i = \alpha + \beta X_i + u_i$.

La función de regresión muestral y los coeficientes estimados ($\hat{\alpha}$ y $\hat{\beta}$) están representados en color rojo. La diferencia entre la FRP y la FRM se debe a los errores que se cometen en la estimación de los coeficientes de la regresión $(\hat{\alpha} \neq \alpha, \hat{\beta} \neq \beta)$. Basándonos en la FRM podemos obtener el valor del punto Y_i como la suma del valor estimado de la parte sistemática $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ (situado sobre la FRM) y del valor que toma el residuo $\hat{u}_i$, esto es, $Y_i = \hat{Y}_i + \hat{u}_i$.

2.4.1. El criterio de estimación mínimo-cuadrático

Dados el modelo y una muestra, debemos decidir cómo obtener la función de regresión muestral, es decir, cómo calcular las estimaciones $\hat{\alpha}$ y $\hat{\beta}$ a partir de los datos. Un método muy utilizado por su sencillez y buenas propiedades es el método de mínimos cuadrados ordinarios. El estimador de ***Mínimos Cuadrados Ordinarios***, o MCO, de los parámetros α y β se obtiene de minimizar la suma de los residuos al cuadrado:

$$\min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{N} \hat{u}_i^2 = \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \min_{\hat{\alpha},\hat{\beta}} \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 \tag{2.4}$$

Las expresiones del estimador de α y β se obtienen de las condiciones de primer orden, para lo cual igualamos las primeras derivadas a cero:

$$\frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\alpha}} = -2 \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0$$

$$\frac{\partial \sum_{i=1}^{N} \hat{u}_i^2}{\partial \hat{\beta}} = -2 \sum_{i=1}^{N} (Y_i - \hat{\alpha} - \hat{\beta} X_i) X_i = 0$$

Así, obtenemos un sistema de ecuaciones, llamadas ecuaciones normales, que vienen dadas por:

$$\sum_{i=1}^{N} \underbrace{(Y_i - \hat{\alpha} - \hat{\beta} X_i)}_{\widehat{u}_i} = 0 \tag{2.5}$$

$$\sum_{i=1}^{N} \underbrace{(Y_i - \hat{\alpha} - \hat{\beta} X_i) X_i}_{\widehat{u}_i X_i} = 0 \tag{2.6}$$

Las expresiones de los estimadores MCO para los coeficientes poblacionales α y β se obtienen de resolver las ecuaciones para $\hat{\alpha}$ y $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N} (X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2} \tag{2.7}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \tag{2.8}$$

2.4.2. Propiedades de los estimadores MCO

Necesitamos saber cuáles son las propiedades que justifican el uso de los estimadores MCO en el modelo de regresión simple bajo las hipótesis básicas. Los estimadores $\widehat{\alpha}$ y $\widehat{\beta}$ son **lineales** en la perturbación, es decir, pueden expresarse como una combinación lineal de las perturbaciones $u_1, \ldots, u_N$. En segundo lugar, los estimadores MCO son variables aleatorias cuya distribución está centrada alrededor del valor poblacional, esto es

$$E(\hat{\alpha}) = \alpha \qquad E(\hat{\beta}) = \beta$$

y, por tanto, son estimadores **insesgados**. Y en cuanto a la precisión, el Teorema de Gauss-Markov prueba que los estimadores MCO tienen **mínima varianza** dentro del conjunto de los estimadores lineales (en u) e insesgados. Las varianzas y covarianza para los estimadores son las siguientes:

$$var(\hat{\alpha}) = \sigma^2 \left(\frac{\sum_{i=1}^{N} X_i^2}{N \sum_{i=1}^{N} (X_i - \bar{X})^2} \right) = \sigma^2 \left(\frac{1}{N} + \frac{\bar{X}^2}{N\, S_X^2} \right) \tag{2.9}$$

$$var(\hat{\beta}) = \sigma^2 \left(\frac{1}{\sum_{i=1}^{N} (X_i - \bar{X})^2} \right) = \frac{\sigma^2}{N} \frac{1}{S_X^2} \tag{2.10}$$

$$cov(\hat{\alpha}, \hat{\beta}) = \sigma^2 \left(-\frac{\bar{X}}{\sum_{i=1}^{N} (X_i - \bar{X})^2} \right) = -\frac{\sigma^2}{N} \frac{\bar{X}}{S_X^2} \tag{2.11}$$

Ambas varianzas dependen de la dispersión de la perturbación $var(u_i) = \sigma^2$, del tamaño muestral y de la dispersión del regresor X. En ambos casos, cuanto mayor sea N o la variabilidad de X, S_x^2, menor es la varianza de los estimadores MCO. En cuanto a la covarianza será no nula a no ser que la media aritmética de la variable explicativa sea cero.

2.4.3. La estimación MCO en Gretl

→ Como ejemplo, calcularemos las estimaciones MCO del modelo para el precio de la vivienda, $P_i = \alpha + \beta F2_i + u_i$, con la muestra del fichero *datos-cap3.gdt*. Una forma sencilla de obtener la FRM mínimo-cuadrática es realizar el diagrama de dispersión en el cual la recta de regresión aparece en la parte superior izquierda. En el ejemplo que nos ocupa tenemos que $\hat{\alpha} = 52,4$ y $\hat{\beta} = 0,139$, como se puede ver en el Gráfico 2.2.

Vamos a ver cómo podemos obtener una tabla de resultados detallados. Una vez iniciada la sesión de Gretl y abierto el fichero *datos-cap3.gdt*, vamos a

Modelo →Mínimos cuadrados ordinarios...

Aparece la ventana donde se especifica la parte sistemática del modelo:

Gráfico 2.9: Ventana de especificación del modelo lineal

- Escogemos la variable dependiente, el precio de venta: en el cuadro izquierdo pinchamos sobre P y luego *Elegir* $->$.
- Elegimos la variable independiente, el tamaño: en el cuadro izquierdo pinchamos sobre $F2$ y luego *Añadir* $->$. La ventana de especificación aparece en el Gráfico 2.9.

Tras pinchar en *Aceptar* aparece la ventana de resultados del modelo (ver el Gráfico 2.10).

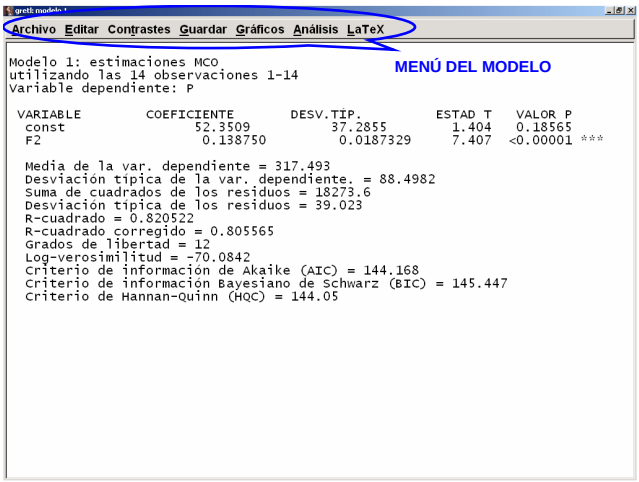


Gráfico 2.10: Ventana de resultados de estimación MCO

En esta ventana aparecen los resultados básicos para el análisis del modelo y que se explican detalladamente a lo largo del curso. La primera columna muestra las variables explicativas que se han incluido en el modelo, la constante (*const*) y la superficie que posee la vivienda ($F2$). En la segunda columna tenemos los coeficientes estimados por MCO correspondientes a cada una de las variables. Como ya vimos, la **estimación** de la ordenada es igual a $\widehat{\alpha}$= 52,35 miles de dólares y la estimación de la pendiente es $\widehat{\beta} = 0,138750$ miles \$ por pie cuadrado. Así la función de regresión muestral es:

$$\widehat{P}_i = 52,3509 + 0{,}138750\ F2_i \tag{2.12}$$

Es decir, cuando la superficie de la vivienda aumenta en un pie cuadrado, el precio medio de venta **estimado** aumenta en $\widehat{\beta} \times 1000 = 138,750$ dólares. Observar que esta interpretación corresponde a la estimación del coeficiente, no al parámetro poblacional β.

Esta ventana de resultados del modelo tiene un menú con siete opciones, *Archivo, Editar, Contrastes, Guardar, Gráficos, Análisis* y *Latex*, que sirven para mostrar otro tipo de resultados de estimación o guardarlos. Veamos algunas de estas utilidades.

Guardar resultados. Si en el menú de resultados del modelo vamos a *Archivo* $\rightarrow$*Guardar a sesión como icono*, el modelo queda guardado dentro de la carpeta `USER`. Así, podemos recuperarlo siempre que queramos; basta con pinchar sobre el botón *iconos de sesión*, cuarto por la izquierda de la barra de herramientas (ver el Gráfico 2.11), y en la ventana que aparece, pinchar dos veces sobre el icono llamado *Modelo 1*. Si posteriormente estimáramos otro modelo y lo guardáramos como icono, Gretl lo denominaría *Modelo 2*.

Algunos gráficos de interés. La opción *Gráficos* de la ventana de resultados del modelo incluye distintas representaciones gráficas tanto de la variable endógena de interés, como de

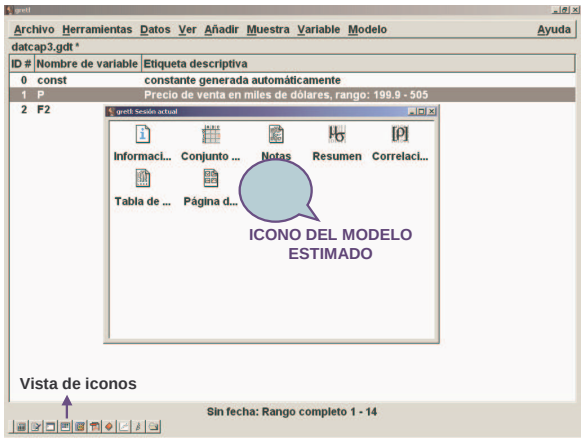


Gráfico 2.11: Ventana de iconos: recuperar resultados estimación

su ajuste y de los errores de su ajuste. Veamos algunos de los más utilizados en regresión con datos de sección cruzada.

- En *Gráficos* → *Gráfico de variable estimada y observada* →*contra F2* obtenemos el gráfico de dispersión de las observaciones reales P_i frente a la variable explicativa $F2_i$ junto con la función de regresión muestral (2.12). El resultado es la figura izquierda del Gráfico 2.12.

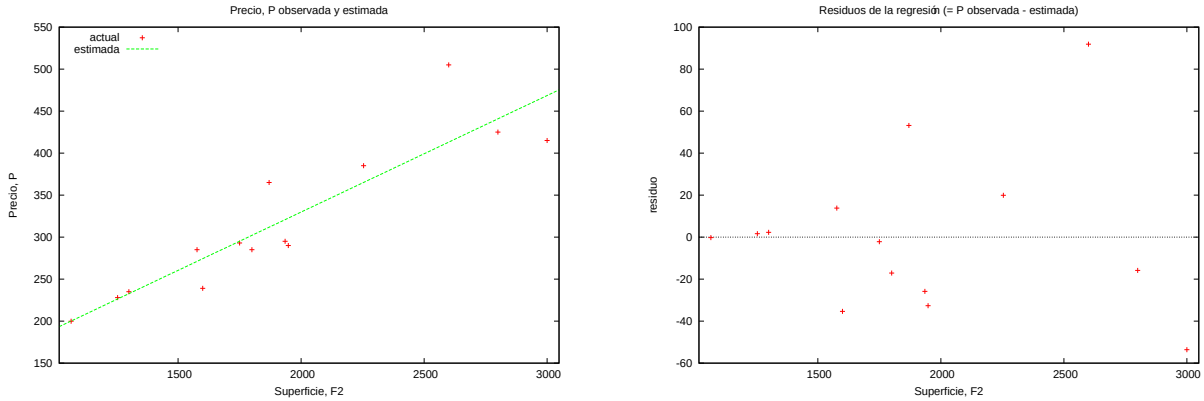


Gráfico 2.12: Gráficos de resultados de regresión MCO

- Si seleccionamos *Gráficos* →*Gráfico de residuos* →*contra F2*, se representan los errores de ajuste $\hat{u}_i$ sobre la variable explicativa $F2_i$, es decir, el diagrama de dispersión de los pares de puntos $(F2_1, \hat{u}_1), \ldots, (F2_{14}, \hat{u}_{14})$, como aparece en la figura derecha del Gráfico 2.12. Podemos apreciar que los residuos se distribuyen alrededor del valor cero $(\bar{\hat{u}} = 0)$ y que la variación con respecto a esta media crece a medida que aumenta el tamaño de los pisos. Este último resultado podría indicar que la hipótesis básica de varianza constante quizás no sea aceptable.

Variables asociadas a la regresión. Para ver los valores que toman los ajustes $\hat{Y}_i$ y los residuos $\hat{u}_i$, debemos seleccionar *Análisis* →*Mostrar variable observada, estimada, residuos.* El resultado que obtenemos es la tabla 2.2. Podemos guardar cualquiera de estos valores seleccionando la opción *Guardar* del menú del modelo, tal como muestra el Gráfico 2.13.

```
Rango de estimación del modelo: 1--14
Desviación típica de los residuos = 39,023
```

Observaciones	P	estimada	residuos	Observaciones	P	estimada	residuos
1	199,9	200,1	−0,2	8	365,0	311,8	53,2
2	228,0	226,3	1,7	9	295,0	320,8	−25,8
3	235,0	232,7	2,3	10	290,0	322,6	−32,6
4	285,0	271,2	13,8	11	385,0	365,1	19,9
5	239,0	274,4	−35,5	12	505,0	413,1	91,9
6	293,0	295,2	−2,2	13	425,0	440,9	−15,9
7	285,0	302,1	−17,1	14	415,0	468,6	−53,6

Tabla 2.2: Residuos de la regresión MCO.

Para almacenar $\hat{P}_i$ hay que elegir *Guardar* →*Valores estimados*. Sale una ventanilla en la que, por defecto, el valor ajustado o estimado de la variable endógena se llama *yhat1* y en la descripción aparece *valores estimados mediante el modelo 1*. Dado que nuestra variable dependiente es el precio de venta P, cambiamos de nombre a la variable y la renombramos como *phat1*. Si repetimos los pasos anteriores pero escogemos *Guardar* →*Residuos*, en la ventanilla correspondiente se nombra a los residuos como *uhat1* y la descripción es *residuos del modelo 1*. Una vez guardadas estas dos series, las encontramos en la ventana principal junto a la variable independiente P y la variable explicativa $F2$.

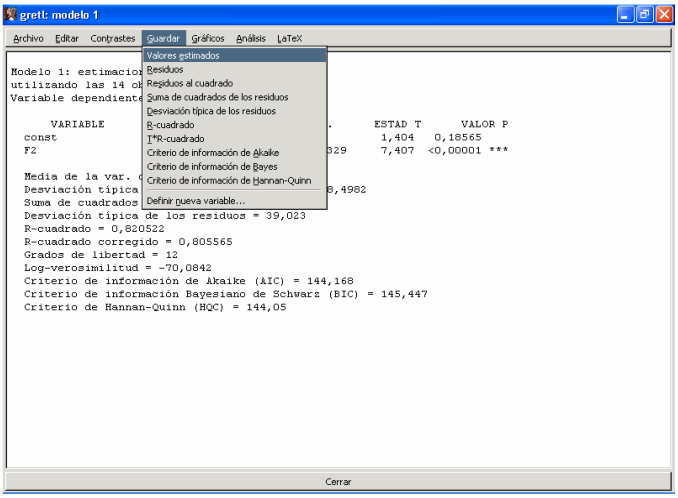


Gráfico 2.13: Residuos MCO

2.4.4. Propiedades de la recta mínimo-cuadrática

Vamos a realizar un pequeño análisis de las variables que intervienen en la regresión mínimo-cuadrática, con objeto de estudiar las similitudes y relaciones que pueden existir entre ellas. Finalmente, generalizaremos estos resultados, comprobando que estas propiedades se cumplen en cualquier regresión lineal mínimo-cuadrática.

Comenzaremos obteniendo los estadísticos descriptivos del regresor $F2$, la variable endógena P, su ajuste $\hat{P}$ y su residuo $\hat{u}$ en *Ver* →*Estadísticos principales* de la ventana inicial de Gretl:

Estadísticos principales, usando las observaciones 1 - 14

Variable	Media	Mediana	Mínimo	Máximo
P	317,493	291,500	199,900	505,000
F2	1910,93	1835,00	1065,00	3000,00
phat1	317,493	306,958	200,120	468,602
uhat1	0,000000	−1,1919	−53,601	91,8983

Variable	Desv. Típ.	C.V.	Asimetría	Exc. de curtosis
precio	88,4982	0,278741	0,653457	−0,529833
F2	577,757	0,302344	0,485258	−0,672125
phat1	80,1640	0,252491	0,485258	−0,672125
uhat1	37,4921	6,15597e+15	1,02687	0,817927

Tabla 2.3: Estadísticos descriptivos de variables de la FRM

Analizando esta tabla-resumen de los datos comprobamos que:

i) La media de los residuos (*uhat1*) es cero, $\bar{\hat{u}} = 0$.

ii) Las medias de la variable dependiente P_i y la estimada (*phat1*) coinciden, $\bar{P} = \bar{\widehat{P}}$.

iii) Los coeficientes de asimetría y curtosis de la variable dependiente ajustada $\widehat{P}_i$ coinciden con las de la variable independiente $F2_i$.

A continuación, vamos a analizar las relaciones lineales existentes entre estas variables. Mediante *Ver →Matriz de correlación* obtenemos la siguiente matriz de correlaciones:

Coeficientes de correlación, usando las observaciones 1 - 14
valor crítico al 5\% (a dos colas) = 0,5324 para n = 14

P	F2	uhat1	phat1	
1,0000	0,9058	0,4236	0,9058	P
	1,0000	−0,0000	1,0000	F2
		1,0000	−0,0000	uhat1
			1,0000	phat1

Tabla 2.4: Matriz de correlaciones

Podemos ver que:

iv) Los valores ajustados $\widehat{P}_i$ y el regresor $F2_i$ están perfectamente correlacionados, $r_{\widehat{P}F2} = 1$.

v) La correlación entre los valores observados P_i con los valores ajustados $\widehat{P}_i$ y la variable explicativa $F2_i$ es la misma, $r_{P\widehat{P}} = r_{PF2}$.

vi) Los residuos $\widehat{u}_i$ y la variable explicativa $F2_i$ están incorrelacionados, $r_{\widehat{u}F2} = 0$.

vii) Los residuos $\widehat{u}_i$ y la variable ajustada $\widehat{P}_i$ están incorrelacionados, $r_{\widehat{u}\widehat{P}} = 0$.

Justificación de estos resultados: La propiedad *i)* se deriva de la primera ecuación normal (2.5), que nos indica que la suma de los residuos ha de ser cero, por lo que $\bar{\hat{u}} = 0$. Notar que la primera ecuación normal existe sólo si el modelo tiene término independiente y no en otro caso. Por lo tanto, los resultados que se obtienen derivados de ella solo se cumplen en el caso

de que el término independiente exista. De $\bar{\hat{u}} = 0$ y como $\bar{Y} = \bar{\hat{Y}} + \bar{\hat{u}}$, se obtiene la propiedad *ii)*.

Las propiedades *iii)*, *iv)* y *v)* se deben a que los valores de $\widehat{P}$ se obtienen de un cambio de origen y escala de la variable $F2$, $\widehat{P} = \hat{\alpha} + \hat{\beta}F2$. Esta relación implica que sus distribuciones de frecuencias tienen las mismas las medidas de forma, están perfectamente correlacionadas entre sí y tienen la misma correlación lineal frente a terceras variables.

La propiedad *vi)* se deriva de las ecuaciones normales (2.5), que indica que $\bar{\hat{u}} = 0$, y (2.6), que implica que los residuos son ortogonales a la variable explicativa X, $\sum_i X_i \hat{u}_i = 0$. Como consecuencia, la covarianza muestral entre residuo y variable explicativa es cero:

$$S_{X\hat{u}} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(\hat{u}_i - \bar{\hat{u}}) = \frac{1}{N}\sum_{i=1}^{N} X_i\hat{u}_i - \bar{X}\bar{\hat{u}} = 0$$

y, por tanto, la correlación entre ambas variables es: $r_{\hat{u}X} = S_{\hat{u}X}/S_{\hat{u}}S_X = 0$. Esto nos viene a decir que en la parte del modelo que queda sin explicar, el residuo $\hat{u}$, ya no queda nada que la variable exógena X pueda explicar o aportar en términos lineales. Finalmente, basándonos en que $r_{\hat{u}X} = 0$ y que el ajuste $\hat{Y}$ es una transformación lineal de X, se demuestra la propiedad *vii)*, $r_{\widehat{u}\widehat{Y}} = 0$. De esta condición y dado que $Y_i = \hat{Y}_i + \hat{u}_i$, se deriva una última propiedad:

viii) La varianza muestral de Y puede descomponerse en dos términos: la varianza explicada por X y la varianza residual, es decir,

$$S_Y^2 = S_{\hat{Y}}^2 + S_{\hat{u}}^2$$

2.4.5. La precisión de la estimación y la bondad del ajuste

Una vez realizada las estimaciones de los coeficientes del modelo, la siguiente etapa del análisis consiste en el análisis y evaluación de los resultados. Por ejemplo,

1. Obtener una medida de la precisión en la estimación de α y β.
2. Evaluar la calidad del ajuste a los datos, es decir, si la función de regresión muestral, $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$, resume bien el comportamiento observado de la variable endógena.
3. Evaluar si el modelo propuesto es *correcto* o si hay algún error en la especificación del modelo, en las hipótesis planteadas.

Este apartado desarrolla los puntos 1 y 2. La respuesta al punto 3 es más compleja, de modo que el siguiente apartado introduce algunos aspectos de la evaluación del modelo.

La precisión de la estimación

En el apartado 7 del tema 1 vimos que la desviación típica de la distribución muestral de los estimadores era un buen indicador de la precisión. Sin embargo, habitualmente la desviación típica de los estimadores tiene algún elemento desconocido. Esto sucede en este caso, como puede comprobarse en la expresión de las varianzas (2.9) y (2.10), que dependen de la varianza

de la perturbación $var(u_i) = \sigma^2$. Podemos obtener una estimación de la desviación típica sustituyendo el parámetro poblacional σ por un estimador insesgado, $\widehat{\sigma}$. El resultado se conoce como **errores típicos de los coeficientes de la regresión**, es decir,

$$\begin{aligned} \textit{Error típico } (\hat{\alpha}) &= \widehat{des}(\hat{\alpha}) = \frac{\hat{\sigma}}{\sqrt{N}}\sqrt{1+\frac{\bar{X}^2}{N\,S_X^2}} \\ \textit{Error típico } (\hat{\beta}) &= \widehat{des}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{N}}\frac{1}{S_X} \end{aligned}$$

Un estimador insesgado de la varianza σ^2 es:

$$\hat{\sigma}^2 = \frac{1}{N-2}\sum_{i=1}^{N}\hat{u}_i^2 = \frac{1}{N-2}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

donde $\sum_i \hat{u}_i^2$ es la **suma de cuadrados residual**, (o SCR), y $N-2$ son los grados de libertad que tenemos tras estimar α y β. Su raíz cuadrada $\hat{\sigma}$ se conoce como **error típico** de los perturbaciones o **error típico** de la regresión. Por tanto, la precisión de las estimaciones de los coeficientes aumenta con el número de observaciones N y la dispersión del regresor S_X y disminuye cuando crece el error típico $\hat{\sigma}$.

De forma similar, se construye el siguiente estimador insesgado de la matriz de las varianzas y la covarianza de los estimadores MCO:

$$\widehat{V}\begin{pmatrix}\widehat{\alpha}\\ \widehat{\beta}\end{pmatrix} = \begin{pmatrix}\widehat{var}(\hat{\alpha}) & \widehat{cov}(\hat{\alpha},\hat{\beta})\\ \widehat{cov}(\hat{\alpha},\hat{\beta}) & \widehat{var}(\hat{\beta})\end{pmatrix} = \hat{\sigma}^2\begin{pmatrix}\left(\frac{1}{N}+\frac{\bar{X}^2}{\sum_i(X_i-\bar{X})^2}\right) & \frac{-\bar{X}}{\sum_i(X_i-\bar{X})^2}\\ & \frac{1}{\sum_i(X_i-\bar{X})^2}\end{pmatrix}$$

→ **Errores típicos de estimación y estimación de las varianzas en Gretl.** En los resultados de estimación del caso práctico aparecen los siguientes valores relacionados con la precisión:

```
Modelo 1: estimaciones MCO utilizando las 14 observaciones 1-14
Variable dependiente: P

  VARIABLE       COEFICIENTE     DESV.TÍP.       ESTAD T     VALOR P
  const           52,3509        37,2855         1,404      0,18565
  F2               0,138750       0,0187329      7,407     <0,00001 ***

Suma de cuadrados de los residuos = 18273,6
Desviación típica de los residuos = 39,023
```

La columna encabezada por *DESV. TÍP.* proporciona los errores típicos de estimación, es decir, $\widehat{des}(\hat{\alpha})$ y $\widehat{des}(\hat{\beta})$. Se observa que es más precisa la estimación del efecto marginal de la superficie del piso β que la de la ordenada α ya que su varianza estimada es menor. La *desviación típica de los residuos* es el error típico $\hat{\sigma}$ y *Suma de cuadrados de los residuos* es $SCR = \sum_i \hat{u}_i^2$.

En esta tabla no aparece la estimación de la varianza de la perturbación, pero se puede calcular:

- De su relación con la desviación típica de los residuos: $\hat{\sigma}^2 = 39,0230^2 = 1522,8$.
- Dividiendo la suma de cuadrados de los residuos entre los grados de libertad $N-2$, así

$$\hat{\sigma}^2 = \frac{18273,6}{14-2} = 1522,8$$

También es posible obtener la estimación de la matriz de varianzas y covarianzas de los coeficientes de regresión seleccionando en el menú del modelo *Análisis →Matriz de covarianzas de los coeficientes*. El resultado para el conjunto de 14 observaciones es:

```
Matriz de covarianzas de los coeficientes de regresión
        const          sqft
      1390,21     -0,670583     const
                   3,50920e-04  sqft
```

Tabla 2.5: Estimación de varianzas y covarianza de $\hat{\alpha}$ y $\hat{\beta}$.

es decir, $\widehat{var}(\hat{\alpha}) = 1390,21$, $\widehat{var}(\hat{\beta}) = 3,5092 \times 10^{-4}$ y $\widehat{cov}(\hat{\alpha},\hat{\beta}) = -0,670583$.

Los errores típicos de estimación y de la regresión dependen de las unidades de medida, es decir, las podemos reducir o agrandar cuanto queramos con sólo cambiar de escala las variables dependiente e independiente. Por otro lado, interesa tener una medida que nos indique, en la medida de lo posible, si estamos ante unos buenos resultados de ajuste a los datos de la función de regresión muestral.

Bondad del ajuste

La medida de la bondad del ajuste que vamos a utilizar es el coeficiente de determinación, R^2 ó R-cuadrado. Este coeficiente, descrito al final de la primera práctica, tiene la siguiente expresión en el modelo de regresión lineal simple:

$$R^2 = r_{XY}^2 = 1 - \frac{\sum_i \hat{u}_i^2}{\sum_i (Y_i - \bar{Y})^2} = \frac{\sum_i (\hat{Y}_i - \bar{\hat{Y}}_i)^2}{\sum_i (Y_i - \bar{Y})^2} \tag{2.13}$$

Este coeficiente mide la ganancia obtenida al pasar de un modelo sin variable explicativa X:

$$Y_i = \alpha + u_i$$

a otro en el que se incluye esta variable: $Y_i = \alpha + \beta X_i + u_i$

Por tanto el R-cuadrado mide la proporción de la variabilidad observada de la variable dependiente Y que se ha podido explicar por incluir de forma lineal en el modelo la variable explicativa X. Normalmente se interpreta en porcentajes, por ejemplo, se dice que la regresión explica el $100 \times R^2$ por ciento de la variación observada en Y. Es fácil comprobar que:

- El criterio mínimo-cuadrático equivale a maximizar R^2.
- $R^2 = r^2_{Y\hat{Y}}$, mide la correlación entre el valor observado y el valor predicho o ajustado con la regresión. Como $0 \leq r^2_{Y\hat{Y}} \leq 1$, si $R^2 \simeq 0$ diremos que el ajuste es pobre y, por el contrario, será un buen ajuste cuando este estadístico esté próximo a la unidad.
  Esta propiedad no se cumple en modelos sin término independiente, es decir, $Y_i = \beta X_i + u_i$.

→ Si analizamos el caso práctico, vemos que el coeficiente de determinación aparece en la tabla de resultados básicos de estimación, `R-cuadrado = 0,820522`. Podemos decir que este ajuste es bueno, ya que la variabilidad muestral de la superficie de la vivienda ($F2$) ha explicado el 82 % de la variabilidad muestral de los precios de venta de dichas viviendas (P).

2.5. Contrastes de hipótesis e intervalos de confianza

Al proponer un modelo para el precio de los pisos hemos asumido que el tamaño del piso es el factor más relevante en la fijación de su precio. Las conclusiones que obtengamos de la estimación y predicción dependerán del cumplimiento de esta hipótesis. Por tanto, conviene valorar si este supuesto es sensato. Para ello vamos a utilizar los contrastes de hipótesis y los intervalos de confianza sobre la distribución de los estimadores. El planteamiento es el siguiente:

- Si el precio de un piso no se ve afectado por su superficie, entonces su efecto marginal es cero, luego $\beta = 0$, y diremos que la variable explicativa no es significativa o relevante para explicar Y. Si esto es cierto, el modelo propuesto no tiene sentido y debemos reformularlo.
- Por el contrario, si el precio está relacionado con la superficie del piso, entonces $\beta \neq 0$ y decimos que el regresor X es significativo o relevante para explicar (y predecir) Y.

2.5.1. Contrastes de hipótesis sobre β

Contraste de significatividad individual de X. Para verificar si la variable independiente $F2$ es significativa para determinar el precio medio de la vivienda, podemos realizar un contraste. Planteamos las siguientes hipótesis a contrastar:

$$\begin{cases} H_0 \colon \beta = 0 & (X \text{ no es significativa o relevante para explicar } Y) \\ H_a \colon \beta \neq 0 & (X \text{ es significativa o relevante para explicar } Y) \end{cases}$$

Para obtener un estadístico de contraste partimos de la siguiente variable aleatoria:

$$\frac{\widehat{\beta} - \beta}{\widehat{des}(\widehat{\beta})} \sim t_{(N-K)} \tag{2.14}$$

El estadístico del contraste se obtiene sustituyendo en esta variable el valor recogido en H_0:

$$t = \frac{\widehat{\beta} - 0}{\widehat{des}(\widehat{\beta})} \overset{H_0}{\sim} t_{(N-K)}$$

Se trata de un estadístico tipo t similar al visto en el apartado 7.2 del tema 1. Es un contraste bilateral, como se observa en el siguiente gráfico de la distribución del estadístico bajo H_0:

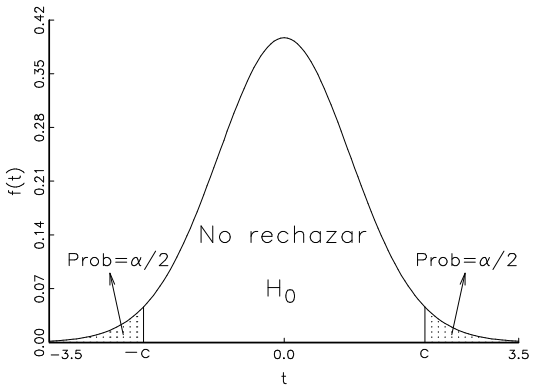


Gráfico 2.14: Criterio de decisión del contraste de significatividad individual

por lo que la regla de decisión es la siguiente: fijado un nivel de significación α,

- Rechazamos H_0 si el valor muestral del estadístico t^m pertenece a la región crítica, es decir, si es menor que $-c = -t_{(N-K)\alpha/2}$ o bien mayor que $c = t_{(N-K)\alpha/2}$ y concluimos que la *variable explicativa es relevante*.
- No rechazamos H_0 en otro caso, es decir, si el valor muestral t^m se sitúa en el intervalo $[-c, c]$ con $c = t_{(N-K)\alpha/2}$. Concluimos que la variable X *no es relevante* o significativa para explicar la variable dependiente Y.

→ Veamos si la superficie de la vivienda es un factor relevante para determinar su precio:

$$\begin{cases} H_0 \colon \beta = 0 \\ H_a \colon \beta \neq 0 \end{cases} \qquad t = \frac{\widehat{\beta}}{\widehat{des}(\widehat{\beta})} \overset{H_0}{\sim} t_{(14-2)}$$

El valor muestral del estadístico t^m se incluye en los resultados de estimación, es la cuarta columna, encabezada por *ESTAD T*. Es decir,

$$ESTAD\ T = t^m = 7,4068 = \frac{columna\ COEFICIENTE}{columna\ DESV.TIP.} = \frac{0,13875}{0,0187329}$$

El valor crítico del contraste para el nivel de significación del 5 % es $c = t_{(14-2)0,05/2} = 2,179$. Como resultado tenemos que $7,4068 > 2,179$, por lo que t^m pertenece a la región crítica y, en consecuencia, rechazamos H_0 a un nivel de significación del 5 %. Podemos concluir que la variable $F2$ es significativa o relevante para determinar el precio medio de la vivienda. En el tema siguiente, veremos cómo la columna *VALOR P* de la tabla de resultados de Gretl informa sobre la conclusión del contraste.

Otros contrastes sobre β. Como hay evidencia estadística de que β es distinto de cero y, por lo tanto, la variable explicativa X es significativa, nos puede interesar saber qué valor puede tomar. Vamos a generalizar el procedimiento de contraste anterior. Veamos dos ejem-

plos.

→ **Ejemplo 1**. Ante un aumento de la superficie de la vivienda de un pie cuadrado, ¿podría el precio medio de venta de la vivienda aumentar en 100 dólares? Planteamos el contraste:

$$\begin{cases} H_0 : \beta = 0,1 \\ H_a : \beta \neq 0,1 \end{cases}$$

Sustituyendo en la variable (2.14) el valor bajo H_0, obtenemos el estadístico de contraste:

$$t = \frac{\widehat{\beta} - 0,1}{\widehat{des}(\widehat{\beta})} \overset{H_0}{\sim} t_{(N-K)}$$

Hay que tener en cuenta que la columna *ESTAD T* de los resultados de estimación de Gretl, corresponde al valor muestral del estadístico para H_0: $\beta = 0$. Por tanto, tenemos que calcular el valor muestral del estadístico de contraste, que en este caso es:

$$t^m = \frac{0,138750 - 0,1}{0,0187329} = 2,068$$

El valor crítico para $\alpha = 5\,\%$ es $c = t_{(14-2)0,05/2} = 2,179$. Como el valor calculado cae fuera de la región crítica, $-2,179 < 2,068 < 2,179$, no rechazamos la H_0 a un nivel de significación del $5\,\%$. Por tanto, es posible un incremento de 100 dólares en el precio medio de la vivienda ante un aumento unitario en la superficie.

→ **Ejemplo 2**. Ante el mismo aumento unitario en la superficie, ¿podría el precio medio de venta de la vivienda aumentar en 150 dólares? Planteamos el contraste y, al igual que en el caso anterior, llegamos al estadístico de contraste:

$$\begin{cases} H_0 : \beta = 0,15 \\ H_a : \beta \neq 0,15 \end{cases} \qquad t = \frac{\widehat{\beta} - 0,15}{\widehat{des}(\widehat{\beta})} \overset{H_0}{\sim} t_{(N-K)}$$

El estadístico de contraste en este caso toma el valor

$$t^m = \frac{0,138750 - 0,15}{0,0187329} = -0,6005 \Rightarrow -c = -2,179 < -0,6005 < 2,179 = c$$

con $c = t_{(12)0,025}$. Así, no rechazamos H_0 a un nivel de significación del $5\,\%$ y también es posible que si $\Delta F2 = 1$, entonces el precio medio de la vivienda aumente en 150$.

Si observamos los contrastes anteriores, siempre y cuando el valor del estadístico calculado t^m esté fuera de la región crítica, es decir, en el intervalo $[-2,179; 2,179]$ no rechazaremos la hipótesis nula propuesta.

2.5.2. Intervalos de confianza

Un intervalo de confianza está definido por dos valores entre los cuales se encuentra el valor del parámetro con un determinado nivel de confianza que se denota $(1 - \alpha)$. Para obtener el intervalo de confianza del coeficiente β, definimos el intervalo de valores que tiene una probabilidad $(1 - \alpha)$ en la distribución (2.14) asociada al estimador. Así

$$Prob\left[-t_{(N-2)\alpha/2} \leq \frac{\widehat{\beta}-\beta}{\widehat{des}(\widehat{\beta})} \leq t_{(N-2)\alpha/2}\right] \;=\; 1-\alpha$$

Reordenamos:

$$Prob\left[\widehat{\beta}-t_{(N-2)\alpha/2}\,\widehat{des}(\widehat{\beta}) \leq \beta \leq \widehat{\beta}+t_{(N-2)\alpha/2}\,\widehat{des}(\widehat{\beta})\right] \;=\; 1-\alpha$$

y obtenemos el intervalo de confianza $(1-\alpha)$ para el parámetro β. Observamos que está centrado en la estimación puntual y que se desvía en una cantidad que está dada por $t_{(N-K)\alpha/2}$ veces su error típico de estimación, $\widehat{des}(\widehat{\beta})$. Si estimamos con muy poca precisión, este intervalo será amplio. Esto quiere decir que la variabilidad muestral del estimador acota a β en un intervalo más amplio. En lo que sigue del curso emplearemos la siguiente notación para expresar el intervalo de confianza:

$$IC(\beta)_{1-\alpha} = \left[\hat{\beta} \pm t_{(N-2)\alpha/2}\;\widehat{des}(\hat{\beta})\right]$$

El correspondiente intervalo de confianza para α se obtiene de forma similar:

$$IC(\alpha)_{1-\alpha} = \left[\hat{\alpha} \pm t_{(N-2)\alpha/2}\;\widehat{des}(\hat{\alpha})\right]$$

→ Continuando con la práctica, vamos a obtener los intervalos de confianza para los dos coeficientes de regresión. Para ello, vamos a *Análisis →Intervalos de confianza para los coeficientes*. El resultado es:

```
t(12, .025) = 2,179

      VARIABLE      COEFICIENTE      INTERVALO DE CONFIANZA 95%
           const         52,3509         (-28,8872, 133,589)
              F2          0,138750       (0,0979349, 0,179566)
```

Tabla 2.6: Estimación por intervalo

En esta tabla de resultados, la segunda columna ofrece las estimaciones por punto, esto es, $\widehat{\alpha} = 52,3509$ y $\widehat{\beta} = 0,138750$. La tercera indica los límites de los intervalos a una confianza del 95 %, esto es:

$$IC(\alpha)_{0,95} = [-28,887\ ;\ 133,587]$$

$$IC(\beta)_{0,95} = [0,0979349\ ;\ 0,179566]$$

Por tanto, podemos afirmar con un nivel de confianza del 95 % que, ante un aumento de la superficie de la vivienda de un pie cuadrado, el precio medio de venta de dicha vivienda aumentará entre 97,9349 y 179,566 dólares.

2.6. Resumen. Presentación de los resultados

Los resultados de la estimación de un modelo se suelen presentar de forma resumida, incluyendo tanto la recta de regresión como un conjunto de estadísticos útiles para evaluar los resultados. Una forma habitual de presentar la estimación es la siguiente:

$$\underset{(\widehat{des})}{\widehat{P}} = \underset{(37{,}285)}{52{,}3509} + \underset{(0{,}018733)}{0{,}138750}\, F2$$

$$N = 14 \qquad R^2 = 0{,}82 \qquad \hat{\sigma} = 39{,}023$$

Bajo cada coeficiente estimado aparece su error típico de estimación. Otra opción es incluir los estadísticos t^m de significatividad individual o los grados de libertad. Por ejemplo,

$$\underset{(estad.\,t)}{\widehat{P}} = \underset{(1{,}404)}{52{,}3509} + \underset{(7{,}407)}{0{,}138750}\, F2$$

$$\text{Grados libertad} = 12 \qquad R^2 = 0{,}82 \qquad \hat{\sigma} = 39{,}023$$

Bibliografía

Alonso, A., Fernández, F. J. e I. Gallastegui (2005), *Econometría*, Prentice-Hall.

Ramanathan, R. (2002), *Introductory Econometrics with Applications*, 5ª edn., South-Western.

Wooldridge, J. M. (2003), *Introductory Econometrics. A Modern Approach*, 2ª edn., South-Western.