Análisis de Regresión

Alfonso Novales
Departamento de Economía Cunatitativa
Universidad Complutense

20 de Septiembre de 2010

Contents

[1] Fichero de trabajo de EVIEWS: Bwght.wf1

[2] Fichero de trabajo: Bwages.wf1. La base de datos Bwages.txt está tomada de los archivos que acompañan a Kuleuven

[3] Fichero de trabajo: Ventas.wf1. Fichero de Excel: Ventas.xls.

[4] En el caso de una regresión múltiple, las expresiones de la varianza del residuo recursivo que aparecen en esta sección son más complejas. Sin embargo, la construcción de los estadísticos, su interpretación y la resolución de los contrastes de estabilidad son iguales a los que aquí se presentan.

1 Momentos poblacionales: momentos de una distribución de probabilidad.

Toda variable aleatoria está caracterizada por su distribución de probabilidad, que no es sino el conjunto de valores posibles de la variable aleatoria, acompañados de sus respectivas probabilidades. El modo en que se representa la distribución de probabilidad depende de que la variable aleatoria en cuestión sea de naturaleza discreta o continua.

Si denotamos por $P(x_i)$ la masa de probabilidad en cada punto x_i del soporte Ω de la distribución de probabilidad de una variable aleatoria X, (conjunto de valores posibles de la variable aleatoria X), y por $f(x_i)$ la función de densidad que la representa, cuando ésta existe (distribuciones de tipo continuo), la esperanza matemática de la variable X se define:

$$E(X) = \mu_x = \int_{-\infty}^{\infty} xf(x)dx;$$

si la medida de probabilidad es continua, o:

$$E(X) = \mu_x = \sum_{x_i} x_i dP(x_i)$$

si la medida de probabilidad es discreta. En este último caso, x_i denota cada uno de los valores posibles de la variable aleatoria X, en número finito o no.

La *mediana* m está definida por el punto del soporte valor numérico para el cual se cumple:

$$\int_{-\infty}^{m} f(x)dx = \frac{1}{2}$$

en el caso de una variable aleatoria o distribución de probabilidad continuas, y:

$$Med(X) = \inf\left\{m \mid \sum_{x_i}^{m} dP(x_i) = \frac{1}{2}\right\}$$

en el caso de una variable discreta. Esta formulación de la definición se debe a que en distribuciones discretas puede aparecer alguna ambigüedad en su cálculo.

La *moda* es el valor más probable de una distribución, es decir, el punto x_M del soporte Ω de la distribución, tal que:

$$P(X = x_M) \geq P(X = x)\ \forall x \in \Omega,$$

La moda puede no ser única. No existen condiciones bajo las cuales la mediana o la moda deban preferirse a la esperanza matemática como medida representativa de la distribución, pero hay que considerar tal posibilidad, dependiendo de las características de la distribución de probabilidad.

La esperanza matemática [suma de los valores numéricos ponderada por probabilidades] de las desviaciones entre los valores del soporte de la distribución y su esperanza matemática es igual a cero:

$$E(X - \mu_x) = E(X) - E(\mu_x) = \mu_x - \mu_x = 0$$

El valor numérico que minimiza la expresión: $E\left[(X-a)^2\right]$ es: $a=\mu_x$. El valor minimizado es la varianza de X.

El valor numérico que minimiza la expresión: E($\mid X-a \mid$) es: $a=m$.

La *varianza* de una variable aleatoria (cuando existe), es la esperanza matemática del cuadrado de las desviaciones entre los valores de la variable y su esperanza matemática:

$$\begin{aligned}\sigma_x^2 &= E\left(X-\mu_x\right)^2=\int_{-\infty}^{\infty}\left(x-\mu_x\right)^2 f(x)dx\\ \sigma_x^2 &= \sum_{x_i}\left(x_i-\mu_x\right)^2 dP(x_i)\end{aligned}$$

en distrib uciones continuas y discretas, respectivamente.

La varianza puede escribirse también:

$$\begin{aligned}\sigma_x^2 &= E\left[(X-\mu)^2\right]=E\left(X^2-2\mu X+\mu^2\right)=E\left(X^2\right)-\mu^2\\ \sigma_x^2 &= \sum_{x_i}\left(x_i-\mu_x\right)^2 dP(x_i)=\sum_{x_i}x_i^2 dP(x_i)-2\sum_{x_i}x_i\mu_x dP(x_i)+\sum_{x_i}\mu_x^2 dP(x_i)=\\ &= \sum_{x_i}x_i^2 dP(x_i)-2\mu_x\sum_{x_i}x_i dP(x_i)+\mu_x^2\sum_{x_i}dP(x_i)=E(x_i^2)-2\mu_x^2+\mu_x^2=E(x_i^2)-\mu_x^2\end{aligned}$$

Como en muchas ocasiones se quiere poner dicho indicador en relación con el valor medio de la variable, se prefiere un indicador que tenga unidades comparables a las de la rentabilidad por lo que, cuando hablamos de volatilidad solemos referirnos a la *desviación típica*: raíz cuadrada de la varianza, tomada con signo positivo:

$$DT(X)=\sigma_x=\sqrt{\sigma_x^2}$$

Otros momentos poblacionales son:

$$Coeficiente\ de\ variación=100\frac{\sigma_x}{\mu_x}$$

que considera la desviación típica (volatilidad) como porcentaje del nivel alrededor del cual fluctúa la variable, lo cual es útil al comparar la volatilidad de variables que tienen una esperanza matemática diferente; por ej., al comparar la volatilidad de dos índices bursátiles distintos.

$$Coeficiente\ de\ asimetría=\frac{E\left[\left(x-\mu_x\right)^3\right]}{\sigma_x^3}$$

que es positivo cuando la distribución es asimétrica hacia la derecha, en cuyo caso la moda es inferior a la mediana, y ésta es, a su vez, inferior a la media aritmética. El coeficiente de asimetría es negativo cuando la distribución es asimétrica hacia la izquierda, en cuyo caso la moda es mayor que la mediana, y ésta es, a su vez, superior a la media aritmética. Toda distribución simétrica tiene coeficiente de asimetría igual a cero.

$$Coeficiente\ de\ curtosis=\frac{E\left[\left(x-\mu_x\right)^4\right]}{\sigma_x^4}$$

también llamado *coeficiente de apuntamiento*, es un indicador del peso que en la distribución tienen los valores más alejados del centro. Toda distribución Normal tiene coeficiente de curtosis igual a 3. Un coeficiente de curtosis superior a 3 indica que la distribución es más apuntada que la de una Normal teniendo, en consecuencia, menos dispersión que dicha distribución. Se dice entonces que es *leptocúrtica*, o *apuntada*. Lo contrario ocurre cuando el coeficiente de curtosis es superior a 3, en cuyo caso la distribución es *platicúrtica* o *aplastada*. A veces se utiliza el Coeficiente de *exceso de curtosis*, que se obtiene restando 3 del coeficiente de curtosis.

La covarianza entre dos variables mide el signo de la asociación entre las fluctuaciones que experimentan ambas. Esencialmente, nos dice si, cuando una de ellas está por encima de su valor de referencia, p.ej., su media, la otra variable tiende a estar por encima o por debajo de su respectiva media:

$$Cov(X,Y) = E\left[(X - EX)(Y - EY)\right] = E(XY) - E(X)E(Y)$$

Siempre se cumple que:

$$Cov(X,Y) = E\left[X(Y - EY)\right] = E\left[(X - EX)Y\right]$$

Cuando alguna de las dos variables tiene esperanza cero, entonces:

$$Cov(X,Y) = E\left(XY\right)$$

El *coeficiente de correlación lineal* entre dos variables es el cociente entre su covarianza, y el producto de sus desviaciones típicas:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

Mientras que la covarianza puede tomar cualquier valor, positivo o negativo, el coeficiente de correlación solo toma valores numéricos entre -1 y +1. Esto ocurre porque, por la desigualdad de Schwarz, la covarianza está acotada en valor absoluto por el producto de las desviaciones típicas de las dos variables.

Un caso importante es el de la covariación entre los valores de una variable con sus propios valores pasados. Así, tenemos, para cada valor entero de k:

$$\gamma_k = Cov(X_t, X_{t-k}),\ k = 0, 1, 2, 3, ...$$

sucesión de valores numéricos que configura la función de autocovarianza de la variable X_t, así como su función de autocorrelación:

$$\rho_k = \frac{Cov(X_t, X_{t-k})}{Var(X_t)} = \frac{\gamma_k}{\gamma_0}$$

El primer valor de la función de autocovarianza, $\gamma_{0,}$ es igual a la varianza de la variable. El primer valor de su función de autocorrelación, ρ_0, es siempre igual a 1.

Dos variables aleatorias son independientes si su función de densidad conjunta es igual al producto de sus funciones de densidad marginales:

$$f(x,y) = f_1(x).f_2(y)$$

dentro del rango de variación de ambas variables.

En el caso de distribuciones discretas (aquéllas en las que la variable en estudio toma valores en un conjunto discreto de puntos, que puede ser infinito), dos distribuciones son independientes si:

$$P(X = x, Y = y) = P(X = x).P(Y = y)$$

En general, en el caso continuo, la función de densidad de una variable Y, condicionada en otra variable X viene dada por:

$$f(y/x) = \frac{f(x,y)}{f_2(x)}$$

pudiendo definirse de modo similar la función de densidad de la variable X, condicionada por la variable Y.

En el caso discreto, se tiene:

$$P(Y = y/X = x) = \frac{P_{XY}(X = x, Y = y)}{P_Y(Y = y)}$$

Ver Ejemplo 1.

Es fácil probar que si dos variables aleatorias son independientes, entonces su covarianza es cero.

La varianza de una suma o de una diferencia de dos variables aleatorias es:

$$\begin{aligned} Var(X+Y) &= Var(X) + Var(Y) + 2Cov(X,Y) \\ Var(X-Y) &= Var(X) + Var(Y) - 2Cov(X,Y) \end{aligned}$$

de modo que solo si ambas variables son *independientes* se tiene que la varianza de su suma es igual a la varianza de su diferencia:

$$Var(X+Y) = Var(X) + Var(Y)$$

En tal caso, el riesgo (medido por la desviación típica) de una cartera sería función de las ponderaciones con que entran en ella cada uno de los activos que la configuran y del riesgo de cada uno de dichos activos, pero no dependería de si la posición adoptada en cada activo es *corta* o *larga*, es decir, de si estamos comprados o vendidos en cada uno de ellos.

Estas expresiones pueden extenderse análogamente a cualquier combinación lineal de$n$ variables. Un ejemplo sería la suma de dichas n variables.

Desigualdad de Chebychev:

$$E\left[g(X)\right] = \int_{-\infty}^{\infty} g(x)f(x)dx \geq \varepsilon^2 \int_S f(x)dx$$

siendo S el conjunto de puntos del soporte de X donde la función g es superior o igual a ε^2. Por tanto,

$$E\left[g(X)\right] \geq \varepsilon^2 \int_S f(x)dx = \varepsilon^2 P\left[g(X) \geq \varepsilon^2\right]$$

y, finalmente:

$$P\left[g(X) \geq \varepsilon^2\right] \leq \frac{E\left[g(X)\right]}{\varepsilon^2}$$

1.1 Distribuciones marginales y condicionadas: Un ejemplo

Consideremos la distribución de probabilidad bivariante,

		X_1				
		-2	-1	0	1	2
X_2	-1	2/24	0	2/24	4/24	0
	0	0	1/24	2/24	0	2/24
	2	0	3/24	2/24	0	6/24

donde X_1 puede tomar valores -2,-1,0,1,2, mientras que X_2 puede tomar valores -1, 0,2. El cuadro recoge probabilidades; por ejemplo, $P[X_1 = -1, X_2 = 0] = 1/24$. Las 15 probabilidades del cuadro suman 1.

La distribución marginal de X_1 es,

$Valores\ de\ X_1$	-2	-1	0	1	2
Pr$obabilidades$	2/24	4/24	6/24	4/24	8/24

con $E(X_1) = 1/2, Var(X_1) = 1/28$,siendo la distribución de X_2,

$Valores\ de\ X_2$	-1	0	2
Pr$obabilidades$	8/24	5/24	11/24

con $E(X_2) = 7/12, Var(X_2) = 263/144$.

La distribución d eprobabilidad de X_1 condicional en un valor numérico de X_2 es,

$Valores\ de\ X_1$	-2	-1	0	1	2
Si $X_2 = -1$	1/4	0	1/4	1/2	0
Si $X_2 = 0$	0	1/5	2/5	0	2/5
Si $X_2 = 2$	0	3/11	2/11	0	6/11

con $E(X_1/X_2 = -1) = 0, E(X_1/X_2 = 0) = 3/5, E(X_1/X_2 = 2) = 9/11$.

Luego $E(X_1/X_2)$ es una variable aleatoria que toma valores 0, 3/5, 9/11, con probabilidades respectivas: 8/24, 5/24, 11/24. Por tanto, su esperanza matemática es 1/2, que coincide con $E(X)$. Este es un resultado general, pues siempre se tiene,

$$E[E(X_1/X_2)] = E(X_1)$$

Las dos variables que hemos analizado no son independientes, pues ninguna de ellas satisface la condición de que su distribución marginal coincida con su distribución condicionada en cualquier valor de la otra. Dicho de otro modo, el valor que toma una variable X_2 es *informativo* acerca de los posibles valores de la otra variable X_1.

1.2 Media, Varianza, Desviación Típica, Covarianza y Coeficiente de correlación muestrales:

En general, contamos con observaciones históricas acerca de una o varias variables (precios, rentabilidades, etc.) y queremos calcular medidas de posición central, de dispersión y de correlación con el objeto de resumir las propiedades básicas de dichos datos.

El conjunto de datos observados define un histograma de frecuencias, o distribución muestral de frecuencias, que contiene *toda* la información disponible acerca de la variable considerada. Un histograma de frecuencias es similar a una distribución de frecuencias, pero es diferente de ella. Para entender la diferencia entre ambos, hemos de comprender el concepto de proceso estocástico, y el modo de utilizarlo en el análisis de datos de series temporales.

Un *proceso estocástico* $X_t, t = 1, 2, 3, ...$es una sucesión de variables aleatorias, indexadas por la variable tiempo. Las variables aleatorias pueden ser independientes entre sí o no, y pueden tener la misma distribución de probabilidad, o una distribución de probabilidad diferente.

Cada dato de una serie temporal debe interpretarse como una muestra de tamaño 1 de la distribución de probabilidad correspondiente a la variable aleatoria de ese instante. Por ej., el dato de cierre del IBEX35 (suponiendo que disponemos de datos de cierre diarios) de hoy es una *realización*, es decir, una muestra de tamaño 1 de la variable aleatoria "precio de la cesta IBEX35" (como índice) el día de hoy. La distribución de probabilidad de esta variable puede ser diferente de la variable aleatoria IBEX35 hace un año por tener, por ejemplo, una esperanza matemática menor, una volatilidad mayor, o no ser Normal, mientras que hace un año sí lo era.

Vamos a suponer inicialmente que las variables X_t tienen todas la misma distribución de probabilidad, y son independientes entre sí. Este es el caso más sencillo, y constituye un proceso de *ruido blanco*. Sólo en este caso está totalmente justificado la utilización de momentos muestrales como características de "la variable X". Esta observación debe servir como llamada de atención al lector, dada la excesiva frecuencia con que se calculan estadísticos muestrales, calculados con datos históricos, para representar características de una variable; por ej., la desviación típica de la rentabilidad bursátil de un determinado mercado.

Las medidas de posición central y dispersión análogas a la esperanza, varianza y desviación típica son:

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

mientras que la covarianza y coeficiente de correlación muestrales son:

$$Cov(X,Y) = \frac{1}{T}\sum_{t=1}^{T} (x_t - \bar{x})(y_t - \bar{y}) = \frac{1}{T}\sum_{t=1}^{T} x_t y_t - \bar{x}\bar{y}$$

La media, varianza, mediana, covarianza y coeficiente de correlación muestrales satisfacen propiedades similares a las ya mencionadas para sus análogos poblacionales. Entre ellas:

- La suma de las desviaciones de la variable respecto de su media, es igual a cero:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

- Como consecuencia de lo anterior, la media muestral de las diferencias $x_i - \bar{x}, i = 1, 2, ..., n$ es igual a cero.

- Si una de las dos variables, X o Y tiene esperanza cero, tenemos:

$$Cov(X,Y) = \frac{1}{T}\sum_{t=1}^{T} x_t y_t = E(XY)$$

- La varianza de X puede escribirse:

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

Al igual que en el caso de una distribución de probabilidad, otras medidas utilizadas en la representación de una muestra son:

$$Coeficiente\ de\ variación = 100\frac{DT_x}{\bar{x}}$$

$$Coeficiente\ de\ asimetría = \frac{\frac{1}{T}\sum_{t=1}^{T}(x_t-\bar{x})^3}{DT_x^3}$$

$$Coeficiente\ de\ curtosis = \frac{\frac{1}{T}\sum_{t=1}^{T}(x_t-\bar{x})^4}{DT_x^4}$$

siendo T el tamaño muestral.

El *recorrido* o *rango* es la diferencia entre el mayor y el menor valor observados de una variable. Los cuartiles son los datos que dividen a la muestra, una vez ordenada crecientemente, en cuatro submuestras de igual tamaño (aproximadamente). El segundo cuartil es la *mediana*. El *rango intercuartílico* es la distancia entre los cuartiles primero y tercero. Estos estadísticos tienen la virtud de no verse afectados por la presencia de valores atípicos. De modo análogo se definen los *deciles* y *percentiles*.

En una variable temporal, las funciones de autocovarianza y autocorrelación muestrales se definen:

$$\gamma_k = Cov(X_t, X_{t-k}) = \frac{1}{T}\sum_{t=k+1}^{T}(x_t-\bar{x})(x_{t-k}-\bar{x})$$

$$\rho_k = Corr(X_t, X_{t-k}) = \frac{Cov(X_t, X_{t-k})}{\sqrt{S_x^2}\sqrt{S_x^2}} = \frac{\frac{1}{T}\sum_{t=k+1}^{T}x_t x_{t-k} - \bar{x}^2}{S_x^2}$$

siendo siempre: $\gamma_0 = Var(X_t)$ y $\rho_0 = 1$.

1.3 Distribuciones condicionales e incondicionales en procesos temporales: El caso del proceso autoregresivo

Especialmente interesante en el análisis de datos financieros es el modelo autoregresivo,

$$y_t = \phi_0 + \phi_1 y_{t-1} + u_t,\ -1 < \phi_1 < 1$$

donde suponemos que u_t es un proceso sin autocorrelación (correlación temporal consigo mismo). Es decir, $Corr(u_t, u_{t-k}) = 0\ \forall k$.

En estas condiciones, si u_t sigue una distribución Normal $u_t \sim N(0, \sigma_u^2)$, entonces y_t sigue una distribución

$$y_t \sim N(\frac{\phi_0}{1-\phi_1}, \frac{\sigma_u^2}{1-\phi_1^2})$$

Esta es la distribución marginal o incondicional, de y_t.

Por otra parte, condicional en la historia pasada de y_t, sin incluir el dato de fecha t, la distribu8ión de probabilidad condicional de y_t es,

$$y_t \sim N(\phi_0 + \phi_1 y_{t-1}, \sigma_u^2)$$

que tiene una menor varianza. De hecho, la varianza incondicional de y_t es tanto mayor cuanto más se acerque el parámetro ϕ_1 a 1, creciendo dicha varianza sin límite. Sin embargo, la varianza condicional es siempre σ_u^2, con independencia del valor numérico del parámetro ϕ_1.

La varianza condicional de y_t es igual a la varianza de u_t, σ_u^2, mientras que la varianza incondicional de y_t es siempre mayor que σ_u^2.

Además,

$$E(y_t/y_{t-1}) = \phi_0 + \phi_1 y_{t-1};\ E(y_t) = \frac{\phi_0}{1-\phi_1}$$

2 El modelo de regresión lineal

El objeto básico de la Econometría consiste en especificar y estimar un modelo de relación entre las variables económicas relativas a una determinada cuestión conceptual. Por ejemplo, para conocer en profundidad el comportamiento del consumo privado agregado de un país, será preciso especificar y estimar un modelo de relación entre observaciones temporales de consumo privado y renta disponible. De modo similar, para analizar si la expansión monetaria en un país ha sido inflacionista, será preciso especificar y estimar un modelo de relación entre las tasas de inflación y las tasas de crecimiento históricas de algún agregado monetario. En su forma más general y, por tanto, más abstracta, tal modelo de relación puede representarse como:

$$Y = f(X_1, X_2, X_3, ..., X_k; \beta)$$

donde Y es la variable cuyo comportamiento se pretende explicar, y $X_1, X_2, ..., X_k$ son las distintas variables que se suponen potencialmente relevantes como factores explicativos de la primera. El vector denota una lista de parámetros que recogen la magnitud con que las variaciones en los valores de las variables X_i se transmiten a variaciones en la variable Y.

Vamos a limitarnos aquí al estudio de modelos de relación o modelos de regresión lineales, es decir, del tipo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k$$

en el que resulta evidente que los parámetros transmiten directamente efectos inducidos por los valores de las variables X_i sobre la variable Y, que se pretende explicar.

La estimación de tales relaciones se efectúa a partir de información muestral acerca de los valores tomados por $Y, X_1, X_2, ..., X_k$, y trata de cuantificar la magnitud de la dependencia entre ellas.

Con objeto de ganar precisión y aprender más acerca del proceso de relación entre las variables consideradas querremos evaluar críticamente la validez de las hipótesis propuestas por la Teoría Económica acerca de las relaciones estimadas que, en ocasiones, consistirán en si determinada variable explicativa entra o no en la relación que se analiza, o si aparece con un determinado coeficiente, por ejemplo, 1 ó -1. Ejemplos del primer tipo son las cuestiones:

1) ¿Influye el déficit sobre los tipos de interés?

2) ¿Afecta el precio de la competencia a la demanda de nuestro producto?

mientras que ejemplos del segundo tipo son:

3) ¿Es el crecimiento monetario neutral, es decir, incide con coeficiente unitario sobre la inflación?

4) ¿Tiene la demanda de nuestro producto elasticidad-precio unitaria? es decir, ¿el efecto de un aumento de un 10% en el precio es una caída del 10% en la demanda?

Estos son problemas de inferencia estadística, similares a los que resolvimos para contrastar hipótesis acerca de la esperanza o la varianza, desconocidas, de una determinada distribución de probabilidad. Por último, especialmente en cuestiones macroeconómicas, estaremos interesados en efectuar un ejercicio de seguimiento coyuntural y de previsión de las variables analizadas. Todo ello puede realizarse de modo riguroso mediante la utilización de procedimientos econométricos que vamos a estudiar en éste y en los dos próximos capítulos.

Así, mediante métodos econométricos, el analista económico puede tratar de responder a preguntas como:

1) ¿cuáles son los determinantes de la tasa de inflación?

2) sobre la base de la información histórica disponible, ¿cuál es la importancia cuantitativa de cada uno de dichos determinantes?

3) ¿podemos contrastar algunas de las implicaciones de la Teoría Económica acerca del efecto que variables como el crecimiento monetario tienen sobre la tasa de inflación?

4) ¿qué sugiere el modelo que hemos estimado para la tasa de inflación acerca del comportamiento de esta variable durante el próximo año?

Es crucial que el analista económico:

a) comience delimitando muy claramente la cuestión teórica que va a ser el centro de su ejercicio empírico,

b) a continuación, debe tratar de identificar cuál es la variable cuyo comportamiento pretende explicar, y cuáles son sus determinantes potenciales. Denominamos a este proceso especificación de un modelo de relación entre variables económicas. Como parte del proceso de especificación, el investigador toma posición acerca de qué variable influye sobre cuál, es decir, propone una relación causal. A diferencia del análisis que pudo efectuarse mediante un coeficiente de correlación, que no descansa en una determinada dirección en la relación entre dos variables, un análisis de regresión en Econometría supone que una variable X influye sobre otra variable Y, y no al revés;

c) luego, el analista debe escoger cuidadosamente la información estadística relevante para cuantificar tal relación, y

d) debe proceder a su cuantificación, es decir, debe estimar los parámetros desconocidos que aparecen en la relación antes especificada;

e) por último, utilizará el modelo de relación estimado, ya sea a efectos de contrastación de algún supuesto teórico, mediante un proceso de inferencia, o como elemento de análisis y seguimiento de la variable cuyo comportamiento escogió explicar.

2.1 El modelo de regresión lineal simple.

Vamos a limitarnos inicialmente al estudio del denominado modelo de regresión lineal simple, que considera una sola variable explicativa X:

$$Y = \beta_0 + \beta_1 X \tag{1}$$

En aplicaciones prácticas disponemos de una muestra de observaciones de ambas variables, y el modelo anterior sugiere que la relación entre las dos variables se satisface para cada una de las observaciones correspondientes. En algunas ocasiones especificaremos modelos de relación

como (1) con el objeto de estimar el comportamiento de determinados agentes económicos. Un ejemplo importante consiste en entender la evolución del consumo agregado del sector privado de una economía real. En algunos casos se tratará de una muestra de datos temporales, y tendremos una relación del tipo (1) para cada instante de tiempo. Para ello, consideraríamos el modelo:

$$C_t = \beta_0 + \beta_1 Y_t,\ t = 1, 2, ..., T$$

donde Y_t denota el PIB del país, o la renta disponible del sector privado (renta total, menos impuestos, más transferencias), según el alcance que se quiera dar al análisis. Los subíndices t hacen clara referencia al hecho de que éste será un modelo a estimar con datos de series temporales. El coeficiente β_1 indica la variación que experimenta el consumo privado del país al variar, a lo largo del ciclo económico, la variable renta que hayamos incorporado como variable explicativa en (1).

En otros casos se dispondrá de una muestra de sección cruzada o de datos transversales, y tendremos una relación como (1) para cada una de las unidades muestrales que, en datos transversales, están constituidas por familias, empresas, países, comunidades autónomas, etc.. Por ejemplo, si disponemos de datos de observaciones de consumo y renta disponible de un conjunto de familias, podríamos especificar:

$$C_i = \beta_0 + \beta_1 Y_i,\ i = 1, 2, ..., n \tag{2}$$

siendo éste un modelo en que la interpretación del coeficiente β_1 sería ahora diferente de la que hicimos con datos de series temporales; en tal caso, β_1 nos proporciona el incremento que se produce en el gasto en consumo de una familia cuando aumenta su renta. No tendría ninguna connotación temporal, pues no hemos utilizado datos de tal tipo. De hecho, si dispusiésemos de dos muestras de sección cruzada, de las mismas familias, pero obtenidas en distintos momentos de un ciclo económico, bien podría ocurrir que la estimación del coeficiente β_1 variase significativamente entre ambas muestras.

En otras ocasiones, se pretende estimar una relación que no es de comportamiento, sino que refleja, más bien, un determinado proceso económico, como pueda ser la producción de bienes. Así, un modelo como:

$$C_t = \beta_0 + \beta_1 K_t + \beta_2 L_t,\ t = 1, 2, ..., T$$

podría interpretarse como la linealización de una función de producción agregada del tipo Cobb-Douglas para una determinada economía real, en la que los coeficientes β_1 y β_2 serían las elasticidades de producción de ambos inputs. En este caso, necesitaríamos un modelo de regresión algo más complejo que el modelo de regresión simple, que incluya varias variables explicativas.

El problema que nos interesa en economía estriba en la estimación de los valores numéricos de los dos coeficientes del modelo de regresión, por ejemplo, β_0 y β_1 en (2), así como en la posibilidad de contrastar hipótesis acerca de sus verdaderos valores numéricos, que son desconocidos.

2.2 Componentes del modelo de regresión

Por razones de exposición, y sin pérdida alguna de generalidad, suponemos en lo sucesivo que disponemos de una muestra de sección cruzada, y mantenemos el criterio notacional que venimos utilizando, designando con mayúsculas las variables genéricas con las que trabajamos: Y, X, y por minúsculas las observaciones numéricas incluidas en las muestras: $y_1, y_2, ..., y_n, x_1, x_2, ..., x_n$. Denotamos el modelo de regresión, como relación entre las variables: $Y = \beta_0 + \beta_1 X$, mientras que

denotamos la relación entre cada par de observaciones por: $y_i = \beta_0 + \beta_1 x_i$. Resulta evidente que es imposible que una relación como (1) se satisfaga para todas y cada una de las observaciones: $i = 1, 2, ..., n$. Si ello ocurriese, podríamos sustituir las dos primeras observaciones muestrales de ambas variables en (1), y determinar exactamente los valores de los coeficientes β_0 y β_1:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 \\ y_2 &= \beta_0 + \beta_1 x_2 \end{aligned}$$

obteniendo las estimaciones de dichos coeficientes con tan sólo estas dos observaciones muestrales. Sin embargo, no debe sorprendernos que al incorporar los valores numéricos de ambos coeficientes, junto con los de las variables Y y X correspondientes a la tercera observación en (1), $y_3 = \beta_0 + \beta_1 x_3$, la relación no se cumpla, salvo por una enorme casualidad.

Queda claro, por tanto, que no es obvio cómo obtener estimaciones de los coeficientes del modelo lineal simple a partir de una determinada muestra de T observaciones temporales, o n observaciones de sección cruzada. A ello dedicaremos algunas de las siguientes secciones. En cualquier caso, nos enfrentamos a una aparente paradoja: el modelo (1) no se satisfará para todas las observaciones muestrales, no importa qué valores numéricos asignemos a sus coeficientes β_0 y β_1. Por ello, no consideramos exactamente el modelo (1), sino una variante del mismo:

$$y_i = \beta_0 + \beta_1 x_i + u_i, \ i = 1, 2, 3, ..., n$$

donde la última variable, u_i, denominada perturbación estructural o término de error del modelo de regresión no es observable, y permite explicar las diferencias entre los dos miembros de la igualdad en (1). El problema de interés estriba en la estimación de los dos coeficientes en el modelo (2), cuando se dispone de una muestra de observaciones para las variables Y_i y X_i, aunque sin disponer de observaciones de la variable u_i.

La variable cuyo comportamiento se pretende explicar, Y_i, recibe el nombre de variable dependiente, mientras que la variable X_i recibe el nombre de variable independiente. En ocasiones, también se denomina a Y_i variable endógena o variable a explicar, mientras que a X_i se le denomina variable exógena o explicativa. Los coeficientes β_0 y β_1 se denominan término constante y pendiente del modelo de regresión simple, respectivamente.

La perturbación aleatoria o término de error del modelo econométrico es una variable no observable para la que, en consecuencia, no dispondremos nunca de observaciones muestrales. Suponemos que su distribución de probabilidad, que puede ser diferente para distintas observaciones muestrales, es independiente de los valores tomados por la variable X_i. Su interpretación es diversa:

a) en primer lugar, puede contener otras variables explicativas que, aun siendo relevantes, no acertamos a especificar;

b) también pudiera ser que, aun siendo conscientes de la existencia de tales variables, no dispusiéramos de observaciones muestrales para las mismas;

c) por último, el término de error puede estar reflejando errores de medida en la variable dependiente Y_i, que suelen surgir porque las variables que utilizamos en la estimación reflejan aproximadamente, pero no exactamente, los conceptos que querríamos incorporar en el modelo.

En el caso de la función de consumo anterior, es difícil en la práctica disponer de datos precisos acerca de los gastos en consumo de una determinada familia: en primer lugar, el consumo es un flujo, y la recogida de datos en un determinado instante de tiempo puede producir todo tipo de distorsiones en dicha variable. Para evitar este tipo de dificultades, en ocasiones, se utiliza como

variable de consumo el resultado de sustraer de los ingresos declarados por la familia, el ahorro realizado durante el período.

Una vez estimados los coeficientes β_0 y β_1 en (2), tendríamos una ecuación lineal, una recta, entre el gasto en consumo y la renta de un conjunto de familias, denominada recta de regresión.

La recta de regresión proporciona la mejor relación existente entre las variables Y y X, en el caso de una regresión simple, o entre la variable dependiente, Y y el conjunto de variables explicativas, en una regresión lineal múltiple. Es tentador interpretar la recta de regresión como si nos proporcionase el valor *esperado* de Y *condicional* en los valores que pueda tomar la variable X. El concepto de *esperanza condicional* es, desde luego, muy importante en el análisis estadístico de datos económicos. Por ejemplo, un banco central puede estar interesado en un determinado momento en estimar la trayectoria que seguiría la tasa de inflación condicional a que dicho banco siga una política monetaria restrictiva. Querría asimismo caracterizar la trayectoria esperada de la inflación condicional a que se ponga en práctica una política monetaria expansiva, y así comparar ambas trayectorias esperadas, y escoger la política monetria acorde a la senda de inflación preferible. De modo simple, este es un ejemplo del importante problema de diseño de política monetaria.

Los modelos econométricos pueden ayudar en este tipo de situaciones. Una vez estimados los coeficientes β, disponemos de valores numéricos para ellos, y fijando una senda numérica para X (tasa de crecimiento monetario) podemos calcular una senda numérica para Y (tasa de inflación). Este ejercicio también se conoce como *predicción por escenarios*. Se trata de establecer sendas o escenarios alternativos para X, cuyos efectos se quieren comparar entre sí, estimar la senda de Y bajo cada uno de dichos escenarios, y calcular el resultado económico o de cualquier otro tipo.

El mismo esquema aplica a la gestión de la empresa, o en muchos contextos financieros. Por ejempo, una empresa se está planteando la conveniencia de dos políticas de publicidad alternativa, una de bajo y otra de alto coste. Si, utilizando datos históricos, estima un modelo de regresión que explique las cifras de ventas (Y) utilizando el gasto en publicidad (X) durante los últimos 40 años, puede utilizar el modelo estimado para calcular aproximadamente las ventas que puede esperar bajo cada una de las dos políticas de publicidad. A continuación, un sencillo cálculo, aplicando los márgenes con que opera a las cifras de ventas estimadas y sustrayendo el coste de la campaña publicitaria, podrá decidir la preferencia por una u otra de las dos campañas.

Existe una limitación, sin embargo, y es que si recordamos el concepto de esperanza condicional, sabemos que dicha esperanza condicional es¡, en general, una función no lineal. Es decir, para calcular el valor esperado de Y para un determinado valor numérico de X, deberíamos utilizar la esperanza de la distribución de Y condicional en X , y ésta es, en general, una función no lineal. Cuando ambas variables, Y y X, tienen una distribución conjunta Normal, entonces, la esperanza condicional es una función lineal, pero no lo es en cualquier caso. Si no aceptamos la Normalidad de la distribución conjunta, entonces la regresión sólo se puede entender como una *aproximación* a la esperanza condicional de Y, dado X.

Por tanto, en este capítulo imponemos una forma funcional lineal para la dependencia de Y respecto de X y no hay ningún razón para pensar que la recta de regresión es una esperanza condicional. Para cada nivel de renta concreto como $Y_i = y^*$, la recta estimada nos proporciona una estimación o predicción de gasto en consumo, $C_i = c^*$. Si hay alguna familia en la muestra con dicha renta, su gasto en consumo observado no coincidirá, salvo por casualidad, con el nivel *previsto* por la recta estimada. La diferencia:

$$\hat{u}_i = C_i - (\hat{\beta}_0 - \hat{\beta}_1 Y_i),$$

que puede ser positiva, si el gasto en consumo excede del estimado por la recta, o negativa, si el gasto observado es inferior al estimado, se conoce como residuo de dicha observación muestral, denotado por $\hat{u}_i$ y, como veremos en la sección 2, juega un papel fundamental en la estimación del modelo de regresión. Es importante observar que la recta de regresión estimada proporciona el nivel de consumo que deberíamos prever para cualquier nivel de renta, incluso si y^* no coincide con el de ninguna familia en la muestra. En tal caso tenemos un verdadero ejercicio de predicción.

En resumen, cuando se lleva a cabo un ejercicio empírico como la estimación del modelo de consumo (2), se tiene en mente un argumento del siguiente tipo: con el modelo (2) no se pretende explicar el comportamiento de la renta disponible de las familias, sino de su nivel de gastos en consumo. Para ello, consideramos las observaciones de la variable explicativa, la renta Y_i, como fijas: es decir, creemos que si hubiésemos entrevistado a otras n familias, hubiéramos generado los mismos datos para dicha variable. Sin embargo, las observaciones muestrales de la variable dependiente, el consumo C_i, habrían sido diferentes, como consecuencia de: a) aspectos específicos, no observables, de las familias encuestadas, b) errores de medida de diferente cuantía a aquellos en los que hemos incurrido en la muestra actualmente disponible, etc., y que aparecen recogidos en la perturbación aleatoria. El término de error es una variable aleatoria, diferente para cada observación muestral, y su realización no es observable. Por el contrario, el residuo es observable, puesto que se construye a partir de las estimaciones y de los datos de las variables dependiente e independiente. Término de error y residuo son entes de diferentes naturaleza.

Desde el punto de vista puramente estadístico, el modelo de regresión no tiene necesariamente una connotación de causalidad en la relación entre variables. Del mismo modo que podemos estimar una regresión de una variable Y sobre otra variable X, podemos estimar una regresión en el orden inverso. Sin embargo, el análisis de este modelo elemental no trata a ambas variables de igual modo: las variables explicativas se consideran deterministas, mientras que la variable dependiente se considera de naturaleza aleatoria. El papel que juega cada una de las variables debe decidirse en función del aspecto teórico que está siendo objeto de estudio. En el ejemplo de consumo y renta, es evidente que queremos explicar los gastos en consumo en función de la renta, y no al revés; el consumo es la variable dependiente, y la renta es la variable independiente. Por eso, el investigador debe decidir de antemano el papel que juega cada una de estas dos variables, porque el tratamiento estadístico del modelo de regresión no concluye nada a este respecto. Sin embargo, su utilización en Econometría se efectúa condicional en una determinada hipótesis acerca de la dirección de la relación, y no al revés.

El modelo de regresión presupone que los valores numéricos de la variable dependiente gastos de consumo, C_i, se generan, en la realidad, a partir de los valores tomados por la variable renta Y_i y precisamente a través de la relación (2). En general, creemos que los procesos económicos son algo más complejos, y que se precisa más de una causa para explicar adecuadamente el comportamiento de una variable como el consumo, C_i, o bien formas funcionales más complicadas que la lineal. Sin embargo, el modelo de regresión simple es también una herramienta útil, al menos en una primera aproximación, desde la que no es muy complejo pasar al análisis del modelo de regresión lineal múltiple, cuyo estudio en profundidad dejamos para temas posteriores.

Comentemos un poco más en detalle estos aspectos:

2.3 Supuestos del modelo de regresión lineal

1. Linealidad en las variables: en algunos casos, el supuesto de que la determinación de los valores del gasto en consumo, C_i, a partir de los de la renta, Y_i, se produce a través de un modelo lineal

es excesivamente restrictiva, pues creemos que el modelo de relación es más bien no lineal. Examinaremos en el próximo capítulo una variedad de modelos alternativos al lineal que aquí analizamos. Sin embargo, en la mayoría de estos casos, el modelo lineal es nuevamente una buena aproximación al verdadero modelo, no lineal, de relación entre variable dependiente e independiente. El caso quizá más paradigmático de no linealidad, surge cuando se cree que el porcentaje de aumento en renta disponible que se transmite a consumo, no es constante, sino que decrece con el nivel de renta. Nótese que el modelo lineal tiene la propiedad de que el cociente de incrementos consumo/renta disponible o, si se prefiere, la derivada del consumo con respecto a la renta disponible, es 1, constante y, por ello, independiente del nivel de renta. Se tendría una relación muy distinta con un modelo del tipo:

$$C_i = \beta_0 + \beta_1 Y_i - \beta_2 Y^2 + u_i,\ i = 1, 2, ..., n$$

Este tipo de no linealidad en las variables puede incorporarse al análisis sin gran dificultad, del modo que veremos en el próximo capítulo,

2. Linealidad en los parámetros: muy diferente es la situación en que los parámetros entran en la relación entre variable dependiente e independientes de modo no lineal. El tratamiento que requieren tales modelos, con excepción de algunos casos sencillos, es sustancialmente más complejo, por lo que no es discutido en este texto,

3. Esperanza matemática nula: suponemos que la esperanza matemática del término de error u_i del modelo es cero: $E(u_i) = 0, i = 1, 2, ..., n$. Si, por el contrario, tuviésemos: $E(u_i) = a \neq 0$, éste sería un efecto constante sobre Y_i y, por ello, determinista, y debería incluirse como parte del término constante β_0 en (1). Una situación en que este supuesto no se cumpliría es cuando el investigador, por error, omite del modelo una variable explicativa relevante. Así, supongamos que en vez de especificar el modelo:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{2t} + u_t,\ t = 1, 2, 3, ..., T$$

se especifica el modelo:

$$y_t = \beta_0 + \beta_1 x_t + v_t,\ t = 1, 2, 3, ..., T$$

en el que, inadvertidamente, se ha omitido la variable explicativa X_2. En este último modelo, erróneamente especificado, el término de error v_t sería igual a: $v_t = \beta_2 x_{2t} + u_t$, y su esperanza matemática: $E(v_t) = E(\beta_2 x_{2t} + u_t) = E(\beta_2 x_{2t}) + E(u_t) = \beta_2 E(x_{2t}) + 0$, donde $E(X_2)$ denota la esperanza matemática de los valores que toma la variable omitida, X_2, que suponemos constante a través del tiempo. Como consecuencia, $E(v_t)$será distinta de cero en general,

4. Varianza constante del término de error (Homocedasticidad): suponemos que la varianza del término de error, que denotamos por $Var(u_i) = \sigma_u^2$ para todo $i = 1, 2, ..., n$, es la misma para todas las observaciones muestrales, ya sean éstas de naturaleza temporal o de sección cruzada,

5. Ausencia de autocorrelación: además, suponemos que los términos de error correspondientes a dos observaciones muestrales cualesquiera, que son dos variables aleatorias diferentes, son estadísticamente incorrelacionadas (autocorrelación espacial en un corte transversal de datos ordenados geográficamente).

6. Estabilidad temporal: otro supuesto incorporado en el modelo es que sus coeficientes, β_0 y β_1, son constantes en el tiempo; igualmente, creemos que el modelo es el mismo para todas las observaciones muestrales. Si disponemos de datos de series temporales, no hay submuestras de tiempo en las cuales los modelos sean diferentes; si estamos explicando los hábitos de consumo de las familias españolas, creemos que la dependencia consumo/renta es igual para familias de renta alta y renta baja, o para familias que habitan en un medio rural y para las que viven en un medio urbano,

7. Causalidad unidireccional: también suponemos que existe una relación causal desde la variable explicativa X hacia la variable endógena Y, es decir, cambios en X influyen sobre cambios en Y, pero no al revés. Ello debe basarse en la naturaleza de la cuestión conceptual que se está analizando, y el investigador siempre debe tener buenos argumentos al respecto, pues ésta no es una cuestión empírica, sino teórica. De aquí surge la denominación de variable exógena para X, es decir, determinada fuera del modelo, y variable endógena, es decir, determinada dentro del modelo, para Y.

   En el ejemplo de relación entre inflación y crecimiento monetario, si durante el período muestral se ha seguido una política monetaria consistente en fijar un determinado crecimiento anual para la cantidad de dinero y seguirlo estrictamente, el crecimiento monetario será una variable exógena en el modelo que pretende explicar la tasa de inflación. Si, por el contrario, se ha seguido una política monetaria en la que el crecimiento monetario se ha decidido en cada período como función de las tasas de inflación que hasta entonces se han registrado, entonces, no estaría justificado calificar de exógeno al crecimiento monetario a la inflación de endógena; quizá ambas deberían ser consideradas variables endógenas, para cuyo necesitaríamos otro tipo de modelos con varias ecuaciones.

8. Variables explicativas deterministas: el modelo incorpora el supuesto, claramente restrictivo, acerca de que la variable explicativa X es determinista. La variable endógena Y no lo es, pues depende de la evolución de una variable aleatoria: el término de error del modelo, u.

   En el ejemplo de relación entre expansión monetaria e inflación, este supuesto significa la creencia de que, si pudiésemos volver al año inicial en las mismas condiciones económicas entonces existentes, y recoger otra muestra para el mismo período, obtendríamos los mismos valores del crecimiento monetario. Desde este punto de vista, las tasas de crecimiento de la oferta monetaria que se han observado en este período son las únicas que pudieron haber ocurrido, con independencia de la información de que dispuso la autoridad monetaria, y de los objetivos de política económica que se trazaron. Sin embargo, nótese que, en esta hipotética situación, las tasas de inflación observadas para el período serían diferentes entre distintas muestras, debido a su componente estocástica u_t.

   En un análisis más general (y más realista) del modelo de regresión, que precisa de un instrumental técnico más complejo que el que presentamos en este texto, se considera que las variables explicativas son también estocásticas, como sin duda queremos creer en la realidad. En estas condiciones más generales, el modelo de regresión lineal simple está plenamente justificado bajo el supuesto de que las dos variables que en él aparecen, X e Y, tienen una distribución de probabilidad conjunta de carácter Normal o Gaussiano. En efecto, ya vimos al estudiar esta familia de distribuciones que la esperanza de la variable Y condicional en la variable X, es una expresión del tipo (1), donde las constantes β_0 y β_1 están relacionadas con los momentos de primer y segundo orden de la distribución bivariante Normal. De hecho, en

tal caso, trabajamos generalmente bajo el supuesto de distribución Normal conjunta de todas las variables que aparecen en el modelo de regresión, e interpretamos éste como la esperanza condicional ya mencionada, lo cual puede extenderse al caso de varias variables explicativas.

3 El estimador de Mínimos Cuadrados Ordinarios

Supongamos que queremos estimar el modelo:

$$y_i = \beta_0 + \beta_1 x_i + u_i,\ i = 1, 2, 3, ..., n$$

donde suponemos que: 1) u_i es una variable aleatoria con $E(u_i) = 0$ y $Var(u_i) = \sigma_u^2$ para todo i, 2) los valores x_i son fijos, 3) β_0 y β_1 son constantes desconocidas. Esta es la especificación del modelo de regresión lineal simple. Para ello, el investigador dispone de una muestra de 16 observaciones acerca de dos variables X e Y, la última de las cuales queremos explicar por medio de la primera:

Cuadro 1

					Valor explicado	Residuo		Residuo cuadrado	Desviaciones en X al cuadrado	Producto de Desviaciones en X e Y respecto de sus medias
n	Y	X	X2	XY	Y-ajustada	u	Xu	u2	(X-mX)2	(Y-mY)(X-mX)
1	16	15	225	240	16.3	-0.33	-5.0	0.11	20.8	15.1
2	18	13	169	234	14.7	3.26	42.4	10.66	6.6	13.6
3	8	11	121	88	13.1	-5.14	-56.5	26.39	0.3	-2.6
4	9	8	64	72	10.7	-1.74	-13.9	3.03	5.9	9.0
5	9	6	36	54	9.1	-0.14	-0.9	0.02	19.7	16.4
6	10	8	64	80	10.7	-0.74	-5.9	0.55	5.9	6.6
7	12	9	81	108	11.5	0.46	4.1	0.21	2.1	1.0
8	14	12	144	168	13.9	0.06	0.8	0.00	2.4	2.1
9	13	10	100	130	12.3	0.66	6.6	0.44	0.2	-0.1
10	10	5	25	50	8.3	1.66	8.3	2.75	29.6	14.6
11	7	9	81	63	11.5	-4.54	-40.9	20.60	2.1	8.2
12	15	12	144	180	13.9	1.06	12.8	1.13	2.4	3.6
13	16	13	169	208	14.7	1.26	16.4	1.60	6.6	8.5
14	18	18	324	324	18.7	-0.73	-13.1	0.53	57.2	40.2
15	15	10	100	150	12.3	2.66	26.6	7.09	0.2	-1.0
16	13	8	64	104	10.7	2.26	18.1	5.11	5.9	-0.8
Sumas :	203	167	1911	2253	203.00	0.00	0.00	80.22	167.94	134.19
Medias :	12.69	10.44	119.44	140.81	12.69	0.00	0.00	5.01	10.50	8.39
Varianzas:	11.71	10.50			6.70	5.01				

Así, tenemos un sistema de ecuaciones:

$$
\begin{aligned}
16 &= \hat{\beta}_0 + 15\hat{\beta}_1 + \hat{u}_1, \\
18 &= \hat{\beta}_0 + 13\hat{\beta}_1 + \hat{u}_2, \\
8 &= \hat{\beta}_0 + 11\hat{\beta}_1 + \hat{u}_3, \\
&\ldots \\
13 &= \hat{\beta}_0 + 8\hat{\beta}_1 + \hat{u}_{16}
\end{aligned}
$$

que no puede resolverse, pues contiene 18 incógnitas, β_0 y β_1, junto con los 16 residuos $\hat{u}_i$ pero sólo 16 ecuaciones. Podríamos fijar los residuos igual a cero en dos ecuaciones y utilizarlas para obtener estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$. Pero dichas estimaciones dependerán del par de ecuaciones seleccionadas, por lo que tal procedimiento no es adecuado. El método apropiado consiste en obtener valores numéricos para β_0 y β_1 que satisfagan de la manera más aproximada posible, simultáneamente, las 16 ecuaciones del sistema anterior.

Una vez estimados los coeficientes, se puede calcular para cada observación i:

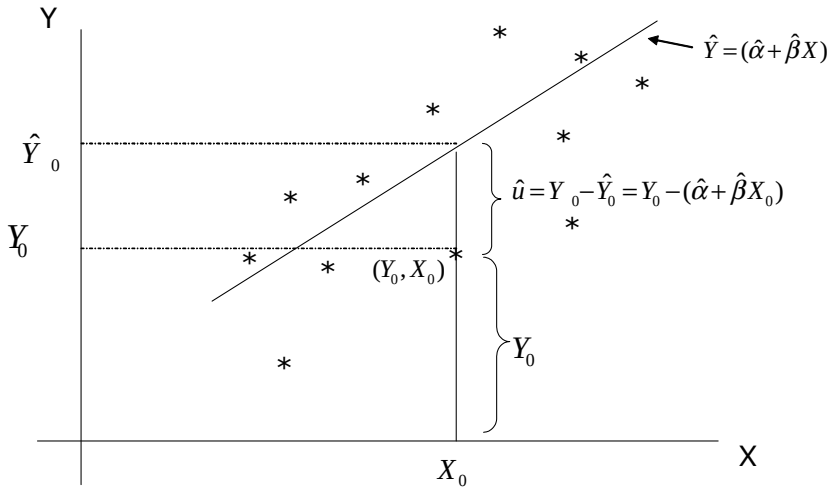
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{3}$$

en el que las estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ han sustituido a los verdaderos valores, desconocidos. La expresión (3) representa la estimación, de acuerdo con el modelo econométrico, del valor que debía haber tomado la variable dependiente Y. Habrá siempre una discrepancia entre el valor realmente observado y_i y la estimación anterior, el residuo correspondiente a dicha observación muestral:

$$\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i),$$

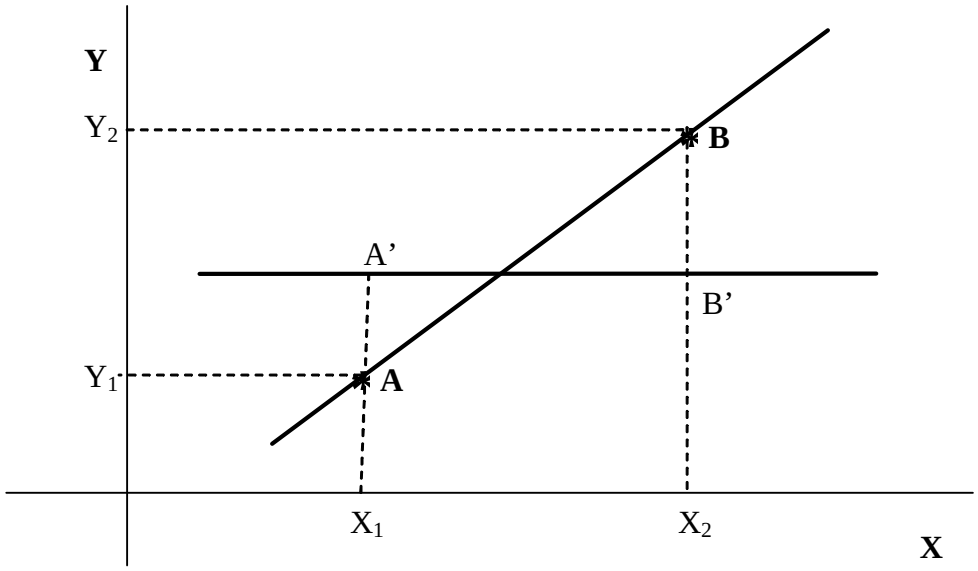
Gráfico 1

Nube de puntos, recta de regresión, valores ajustados, residuos



Hay muchos modos posible sde utilizar la información muestral para asignar valores numéricos a los dos parámetros desconocidos del modelo, β_0 y β_1. Parece razonable que queramos utilizar un procedimiento que genere unos residuos o errores de ajuste que sean reducidos, en algún sentido que hemos de definir. Pero tenemos tantos residuos como observaciones muestrales, y no cabe pensar que podamos conseguir que todos y cada uno de ellos sea pequeño. Generalmente, preferiremos no tener ningún residuo muy grande, sea positivo o negativo, por lo que supondría de tener un importante error de ajuste para la observación muestral correspondiente. Entonces, la posibilidad más inmediata sería utilizar un estimador que consiga que la suma de los residuos sea igual a cero. Pero enseguida vemos que esta idea no funciona:

Gráfico 2



un posible criterio que defina a un estimador sea la minimización de la magnitud de los residuos que dicho estimador genera. Tal idea es correcta, pero hay varias dificultades para hacerla práctica: en primer lugar, tenemos no un residuo, sino un conjunto de n residuos, por lo que no se trata de minimizar un residuo determinado, sino una medida conjunta del tamaño global de todos ellos.

Una vez obtenidas unas estimaciones numéricas de los coeficientes, podría pensarse en sumar los n residuos generados: $\sum_{i=1}^{n} \hat{u}_i$, y escoger como estimación el par de valores $\hat{\beta}_0$ y $\hat{\beta}_1$ que produce la menor suma de residuos. Una dificultad con tal procedimiento es la cancelación de residuos negativos con residuos positivos. Además, si realmente se pretendiese minimizar la suma de residuos, bastaría generar residuos de tamaño muy grande, pero negativos, lo cual no es adecuado.

3.1 Estimador de Mínimos Cuadrados

El estimador de mínimos cuadrados que introducimos en esta sección utiliza como criterio la minimización de la Suma de los Cuadrados de los Residuos (SCR), o también Suma Residual, aunque hay que recordar que es una suma de cuadrados. Se trata, por tanto, de seleccionar valores de los coeficientes β_0 y β_1 que resuelvan el problema:

$$\underset{\hat{\beta}_0,\hat{\beta}_1}{Minimizar}\ SCR = \sum_{i=1}^{n} \hat{u}_i^2$$

Nótese que el residuo asociado a cada observación $i, i = 1, 2, ..., n$, depende de los valores de los coeficientes escogidos, porque:

$$\hat{u}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

de modo que el problema anterior puede escribirse:

$$\underset{\hat{\beta}_0,\hat{\beta}_1}{Minimizar}\ SCR = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

La solución a este problema de optimización se denota por: $\hat{\beta}_0$, $\hat{\beta}_1$, y se denomina estimador de Mínimos Cuadrados Ordinarios (que abreviaremos como MCO) de los coeficientes del modelo de regresión lineal simple. El estimador MCO escoge, de entre todas las posibles, la recta que minimiza la suma de los cuadrados de las distancias entre cada punto de la nube generada por las observaciones muestrales y el asignado por la recta.

Derivando SCR con respecto a ambas variables (β_0 y β_1) e igualando dichas derivadas a cero, tenemos:

$$\frac{\partial SCR}{\partial \beta_0} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \tag{4}$$

$$\frac{\partial SCR}{\partial \beta_1} = -2\sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) x_i = 0 \tag{5}$$

con matriz de derivadas segundas:

$$\frac{\partial^2 SCR}{\partial \beta_0 \partial \beta_1} = \begin{matrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{matrix}$$

que tiene por determinante:

$$DET = 4\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right) = n^2\left(\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2\right) = n^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} = n^2 S_x^2$$

siendo S_x^2 la varianza muestral de X : $S_x^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / n$. Como el primer menor, el elemento (1,1) de esta matriz, que es $2n$, es también positivo, podemos afirmar que la solución al sistema de ecuaciones (4) y (5) serán, los valores numéricos de los coeficientes β_0 y β_1 que, efectivamente, alcanzan un mínimo de la Suma Residual.

3.1.1 Ecuaciones normales

Si resolvemos dicho sistema, obtenemos:

$$\sum_{i=1}^{n} y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i \tag{6}$$

$$\sum_{i=1}^{n} y_i x_i = \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \tag{7}$$

que constituyen un par de ecuaciones simultáneas en las incógnitas, $\hat{\beta}_0$, $\hat{\beta}_1$. Este sistema se conoce como sistema de *ecuaciones normales*.

Utilizando los estadísticos que aparecen en la última fila del Cuadro 1, tendríamos:

$$\begin{aligned} 203 &= 16\beta_0 + 167\beta_1 \\ 2253 &= 167\beta_0 + 1911\beta_1 \end{aligned}$$

que resuelto, proporciona las estimaciones MCO:

$$\hat{\beta}_0 = 4,348;\ \hat{\beta}_1 = 0,799$$

con dichos datos. La sexta columna del cuadro presenta los valores previstos por el modelo para la variable dependiente. La columna siguiente muestra los residuos, es decir, la diferencia entre los valores de Y y los valores previstos por el modelo.

3.1.2 Expresiones para el estimador de Mínimos Cuadrados

En general, si primero despejamos $\hat{\beta}_0$ en (6), tenemos:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n y_i - \hat{\beta}_1 \sum_{i=1}^n x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \tag{8}$$

que podremos utilizar para obtener el estimador MCO de β_0, una vez que tengamos el estimador de 1. Substituyendo en (7), tenemos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{nS_{xy}}{nS_x^2} = \rho_{xy}\frac{S_y}{S_x} \tag{9}$$

donde S_{xy}, S_x^2, S_y^2, S_x, S_y, denotan, respectivamente, la covarianza, varianzas y desviaciones típicas muestrales de X e Y. Las expresiones (8) y (9) son útiles, pues proporcionan directamente las estimaciones MCO como función de estadísticos muestrales, sin necesidad de resolver el sistema de ecuaciones normales. Primero se calcula $\hat{\beta}_1$ y, luego, se obtiene: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Ello demuestra una propiedad del estimador MCO: la recta estimada pasa por el punto $(\bar{y}, \bar{x})$.

3.1.3 Regresión inversa

Supongamos ahora que estimamos la regresión lineal inversa de la anterior, es decir, la regresión que tiene a Y como variable explicativa, y a X como variable dependiente:

$$x_i = \alpha_0 + \alpha_1 y_i + v_i$$

El estimador de mínimos cuadrados de la pendiente en este modelo es:

$$\hat{\alpha}_1 = \frac{S_{xy}}{S_y^2}$$

que es distinta de la que estimamos en la primera regresión. Sin embargo, el estimador de la pendiente de esta regresión no es el inverso del estimador de la pendiente en el modelo de

regresión original. Sin embargo, existe una relación entre ambos. En efecto, si multiplicamos ambos estimadores, tenemos:

$$\hat{\beta}_1\hat{\alpha}_1 = \frac{S_{xy}}{S_x^2}\frac{S_{xy}}{S_y^2} = \frac{(S_{xy})^2}{S_x^2 S_y^2} = (\rho_{xy})^2$$

luego el producto de ambos estimadores es igual al cuadrado del coeficiente de correlación lineal entre ambas variables.

3.1.4 Interpretación del estimador de Mínimos Cuadrados

Podemos ahora deducir la relación que existe entre el estimador MCO de la pendiente β_1 del modelo de regresión lineal simple y el coeficiente de correlación de X e Y :

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y}\frac{S_y}{S_x} = \rho_{xy}\frac{S_y}{S_x}$$

Es decir, el estimador de β_1, el coeficiente que proporciona la relación entre la variables X e Y del modelo de regresión lineal simple, está muy relacionado con el coeficiente de correlación entre ambas variables, siendo una modificación del mismo por el cociente de las desviaciones típicas.

Como la desviación típica es una medida del tamaño de la fluctuación que experimenta una variable a lo largo de la muestra, si X e Y tienen igual desviación típica, experimentan fluctuaciones de igual tamaño. En tal caso, una variación de una unidad en X se traducirá, de acuerdo con $\hat{\beta}_1$ en una variación en Y igual a ρ_{xy}. Si, a lo largo de la muestra, las fluctuaciones en Y son de un tamaño aproximadamente doble del de las fluctuaciones en X, entonces el modelo de regresión implicará que una variación unitaria en esta última variable se traducirá en una variación de dos veces ρ_{xy} en Y. Por ejemplo, si ambas variables tienen igual varianza, y $\rho_{xy} = 0,50$,entonces una elevación de 2 unidades en X vendrá acompañada, en media, de una elevación de 1 unidad en Y. Si la varianza (o volatilidad) de X es doble de la volatilidad de Y, entonces la elevación de dos unidades en X sólo generará, en media, una elevación de $0,5$ unidades en Y. Evidentemente, si el signo de la correlación fuese negativo, entonces una elevación en una de las dos variables vendría acompañada de un desenso (no una elevación) en la otra variable.

3.2 Ejemplo: Peso de bebés recién nacidos[5]

Analizamos en este ejemplo datos tomados de Wooldridge, Introducción a la Econometría: un enfoque moderno, primera edición. Partiendo de un modelo de regresión estimado en dicho texto,

- discutimos el modo de llevar a cabo un análisis descriptivo, tanto de tipo gráfico como de tipo estadístico, acerca de la capacidad explicativa que un conjunto de variables tiene sobre una determinada variable dependiente, y
- describimos cómo el habitual uso mecánico de los estadísticos tipo t de Student y F puede conducir a conclusiones erróneas sobre la capacidad explicativa de una variable o de un conjunto de ellas.

[5] Fichero de trabajo de EVIEWS: Bwght.wf1

3.2.1 Descripción del ejemplo

Consideramos en este ejemplo la especificación de un modelo de regresión para tratar de caracterizar factores que pueden afectar al peso de bebés al nacer. La base de datos[6], tomada de Wooldridge (2001), contiene información sobre el peso de los bebés, recogido en 1.388 nacimientos, la renta de la familia en la que se produce el nacimiento ($renta_i$), el número medio de cigarrillos fumados diariamente por la madre durante el embarazo ($cigarrillos_i$), el número medio de cajetillas de tabaco fumados diariamente por la madre durante el embarazo, el número de orden que ocupa el recién nacido dentro de los hijos de la familia ($ordenac_i$), los años de educación del padre ($educp_i$) y de la madre ($educm_i$), el sexo del bebé y si éste es blanco o de otra raza. Estas dos últimas variables son ficticias, y aparecen en el archivo como variables dicotómicas, es decir, tomando dos valores únicamente. La variable sexo ha sido definida mediante $Sexo = 1$ si el recién nacido es varón, $Sexo = 0$ en caso contrario, mientras que la variable que recoge el grupo étnico se ha definido $Raza = 1$ si el bebé es de raza blanza, $Raza = 0$ en caso contrario. Falta información acerca del nivel educativo del padre del recié nacido en 196 nacimientos, faltando información acerca del nivel educativo de la madre en un caso más, por lo que las regresiones que incluyen estas variables como explicativas utilizan un máximo de 1191 observaciones.

En Wooldridge (2001) se estima el modelo de regresión,

$$\begin{aligned} Peso_i &= \underset{(3,73)}{114,52} - \underset{(0,110)}{0,596}cigarillos_i + \underset{(0,037)}{0,056}renta_i + \underset{(0,659)}{1,788}ordenac_i + \\ &+\underset{(0,283)}{0,472}educp_i - \underset{(0,320)}{0,370}educm_i + u_i,\ i = 1,2,...,N \\ \bar{R}^2 &= 0,035,\ \hat{\sigma}_u = 19,789 \end{aligned}$$

donde se muestran entre paréntesis las desviaciones típicas estimadas de cada uno de los coeficientes. El autor contrasta la significación conjunta de los niveles educativos de ambos padres mediante el estadístico F, en la forma del R^2, no rechazando la hipótesis nula de ausencia de capacidad explicativa de ambas variables, conjuntamente consideradas. Por tanto, el nivel educativo de los padres no parece ser un condicionante significativo del peso de los bebés al nacer.

La discusión que llevamos a cabo en la Sección XX ya sugiere que, en cualquier caso, la interpretación de este resultado no debe hacerse con carácter absoluto. El investigador debería decir que "una vez incluidas considerados como posibles factores explicativas del peso del recién nacido la renta de la familia, el número de cigarrillos fumados por la madre durante el embarazo y el número de orden del recién nacido entre sus hermanos, los indicadores educativos de los padres no aportan *información adicional* relevante".

El segundo matiz que hemos de hacer es que podría darse el caso de que los indicadores educativos contengan información relevante incluso una vez que ya se ha tenido en cuenta la información aportada por las variables mencionadas pero, por alguna razón, la información muestral disponible no permite medir con suficiente precisión el impacto que sobre el peso del bebé tiene el nivel educativo de los padres. Una reducida precisión podría conducir a un estadístico-t reducido y, con ello, a no rechazar la hipótesis nula de ausencia de relación entre nivel educativo de los padres y peso del bebé.

[6] El archivo Bwght.des contiene la descripción de las variables incluidas en el archivo Bwght.raw, algunas de las cuales se han utilizado en el ejemplo.

La tercera consideración a efectuar es que el contraste tipo F efectuado para analizar conjuntamente la información proporcionada por las dos variables educativas descansa sobre el supuesto de Normalidad del término de error del modelo de regresión, cuestión que habríamos de analizar.

Comenzamos nuestro análisis indagando la información que cada una de las potenciales variables explicativas contiene sobre el peso del recién nacido. Al hacerlo individualmente, estamos ignorando el hecho de que distintas variables pueden contener información común; debemos interpretar que se trata de un análisis que trata de detectar la ausencia de capacidad explicativa en alguna variable. Si, como es habitual, nos limitásemos al análisis de los estadísticos tipo t, diríamos que, entre las variables consideradas, el número de cigarrillos fumados por la madre afecta al peso del recién nacido, habiendo asimismo un efecto estadísticamente significativo en relación con el número de orden que el recién nacido ocupa entre los hijos de la familia. Los indicadores de educación no parecen aportar información relevante, al igual que tampoco parece haber relación con la renta de la familia en la que se produce el nacimiento.

3.2.2 Características muestrales de las variables (archivo bwght.wf1)

Los histogramas[7] de las variables revelan características interesantes (ver *HIS_ nombre variable en el fichero* bwght.wf1): la variable dependiente *peso* es una variable continua, cuyo exceso de curtosis genera un comportamiento no Normal en la muestra, rechazándose claramente dicha hipótesis mediante el test de Jarque-Bera. Este resultado despierta dudas acerca del uso de las distribuciones habituales tipo t de Student y F de Fisher-Sendecor para los estadísticos utilizados en la contrastación de hipótesis.

Las variables *cigarrillos* y *paquetes* tienen una correlación exactamente igual a 1,0. Esto significa que se han construido una a partir de la otra, pues si se hubiera encuestado sobre ambas existiría una relación algo menos que perfecta entre ellas. Examinando sus valores, vemos que la primera es igual a 20 veces el valor numérico de la segunda en todos los casos, por lo que utilizaremos únicamente la variable *cigarrillos*. Esta es una variable discreta, con un valor mínimo de 0 y un valor máximo de 50; la mediana es 0, reflejando el hecho de que en casi un 85% de los 1.388 nacimientos recogidos en la muestra, la madre declaró no haber fumado durante el embarazo[8]. Sólo en 212 casos, la madre del recién nacido declaró haber fumado un número medio de cigarrillos por día mayor que cero. Esto sugiere que disponemos de una información relativamente reducida para estimar la contribución al peso del bebé de un cigarrillo adicional, lo que podría hacer que dicha estimación se obtenga con una precisión no muy alta, salvo si la diferencia entre el peso de los bebés de madres fumadoras y no fumadoras es muy sistemática.

La educación de la madre toma valores entre 2 y 18 años, con una mediana de 12 años; ésta es también la moda, recogiendo el 40,5% de las observaciones muestrales. La educación del padre toma valores entre 1 y 18 años, también con una mediana y moda igual a 12 años; valor que aparece en un 37,2% de los nacimientos. El elevado número de observaciones en el nivel educativo correspondiente a 12 años segmenta la muestra de padres y madres entre los que alcanzan el grado medio y los que continúan con estudios superiores.

La información numérica sobre la renta familiar, en miles de dólares, tiene el aspecto de haber sido redondeada, apareciendo únicamente valores numéricos entre 0,5 y 19,5, además de 22,5, 27,5, 32,5, 37,5, 42,5, 47,5, 65,0. Por tanto, la variable renta tiene naturaleza discreta, tomando un número relativamente alto de valores igualmente espaciados en el primer rango mencionado, para

[7] Los nombres en cursivas, entre paréntesis, denotan elementos del archivo de trabajo Bwght.wf1.

[8] Por tanto, la moda de esta variable es cero.

pasar a tomar valores más dispersos posteriormente. Un 38% de las observaciones están en el rango $(0,5; 19,5)$ de renta, estando el 62% restante en niveles de renta superiores, por lo que el proceso de redondeo afecta a un alto número de observaciones. Si hubiera una relación continua entre la renta de la familia y el peso del recién nacido, tal proceso de simplificación numérica podría dificultar notablemente su estimación. Aunque ignoramos el modo en que la concentración de valores numéricos ha sido hecha, imaginemos que se ha asignado un dato de renta de 65,0 a las familias con renta en $(56, 75, 65, 0)$, asignando renta de 47,5 a aquellas familias con renta en $(47,5; 56,75)$. El peso del recén nacido podría crecer suavemente con la renta, pero ésta se ha colapsado en los dos extremos del intervalo, generando una importante cantidad de errores en cualquier relación lineal entre peso y renta. Por tanto, tenderíamos apensar que dicha relación no existe.

La variable $ordenaci_i$, que recoge el orden del recién nacido entre los hijos de la familia, toma valores entre 1 y 6, siendo la moda igual a 1, con una frecuencia relativa de 57,3%. Por tanto, la mediana de esta variable es asimismo igual a 1.

El 48% de los recién nacidos (665) son mujeres y el 52% (723) varones, por lo que la muestra está bastante equilibrada en este sentido; por el contrario, el 78% son de raza blanca y el 22% restante de otras razas. Los posibles efectos del sexo y la raza del recién nacido sobre su peso no han sido considerados en la regresión anterior, pero los consideraremos más adelante. Es asimismo interesante observar que de las 212 madres que declararon haber fumado durante el embarazo, 165 eran de raza blanca, mientras que de las 1089 madres que declararon no haber fumado durante el embarazo, 924 eran de raza blanca.[9] Como se muestra en Bwght.xls, haber fumado durante el embarazo es independiente de la raza de la madre.

3.2.3 Asociación con la variable dependiente, peso del recién nacido.

Los coeficientes de correlación habituales son reducidos (Tabla *correlaciones*), siendo el más elevado numéricamente (-0,16) el del número de cigarrillos fumados, que es de signo negativo, como esperaríamos. Recuérdese que una desviación típica aproximada del coeficiente de correlación es el inverso de la raíz cuadrada del tamaño muestral, que estaría en torno a 0,027. Ello haría que la correlación mencionada, aun siendo reducida, fuese estadísticamente significativa. Sin embargo, el resto de las correlaciones recogidas en la tabla sugiere que la búsqueda de capacidad explicativa del peso del recién nacido en las variables disponibles puede resultar poco fructífera. Entre las variables explicativas, la renta de la familia tiene coeficientes de correlación superiores a 0,40 con los niveles educativos del padre y la madre que, a su vez, muestran una correlación de 0,64 entre ellos.

Sin embargo, las variables explicativas tienen naturaleza discreta, por lo que los coeficientes de correlación habituales no están plenamente justificados. Esto mismo hace que las nubes de puntos con la variable dependiente no sean tan informativas como en otros casos; como muestra, recogemos en el fichero de trabajo la nube de puntos entre el peso y el orden que el recíen nacido ocupa entre los hijos. Un efecto negativo, por ejemplo, vendría dado por una reducción del peso al aumentar el valor de la variable $ordenac_i$. La nube de puntos nos da un intervalo de pesos observados entre los recién nacidos que comparten un mismo valor de la variable $ordenac_i$, y se trataría de ver si el valor representativo de cada intervalo de pesos es decreciente al aumentar $ordenac_i$.

[9] Esto se muestra en Bwght.xls, multiplicando las columnas de variables dicotómicas {0,1} "Fuma" y "Blanco", y hallando la suma de dicho producto, y repitiendo el cálculo con "Blanco" y 1-"Fuma". Suponemos aquí que la raza de la madre y del recién nacido son las mismas. De modo análogo, puede verse que de las 212 madres que declararon haber fumado, 100 tuvieron un hijo varón. Esta división aproximada entre hijos varones y mujeres es, por supuesto, muy razonable.

Esto nos dirige a estimar la asociación entre variables mediante tablas de clasificación de sus valores, así como contrastando la igualdad de medias y medianas entre clases. Por ejemplo, para analizar la posible asociación entre el peso del bebé y la educación de la madre, calculamos la mediana del peso de los bebés para cada uno de los posibles niveles educativos de la madre, contrastando la igualdad de dichos valores mediana. Si estas dos variables no estuvieran relacionadas, las medidas de posición central (mediana o media) de la variable *peso* serían similares para los distintos niveles educativos; si existe una asociación positiva entre ambas variables, esperaríamos que la media o mediana de *peso* fuese creciente con el nivel educativo, y lo contrario ocurriría si existiera una relación negativa entre ambas. En ambos casos se rechazaría la hipótesis nula de igualdad de medias así como la de igualdad de medianas. Para ello, debe calcularse la media o mediana de la variable dependiente para cada uno de los distintos rangos de valores numéricos de la variable explicativa que se considera. Nos centramos en las medianas y no en las medias debido a la fuerte desviación que muestran las distribuciones de estas variables respecto de la Normalidad, tanto por razón de la muy elevada frecuencia observada en el valor modal, como de su asimetría. El lector interesado puede reproducir nuestro análisis contrastando la igualdad de medias muestrales del peso para los distintos niveles educativos de la madre o el padre.

Al comparar las variables *peso* y *educm*, los contrastes Kruskal-Wallis y van der Waerden de igualdad de medianas rechazan la igualdad de medianas, sugiriendo asociación entre ambas variables ($MEDN\_PESO\_EDUCM$). Repetimos el contraste llevando a cabo cierta agrupación de los niveles educativos, para eliminar el problema de que algunos niveles educativos recogen un número muy reducido de observaciones: para algunos niveles educativos hay una sóla observación muestral. La agrupación proporciona indicios aún más claros en contra de la igualdad de medianas. Los valores numéricos de las medianas por clases de niveles educativos[10] después de la agregación, recogidas en ($MEDN\_PESO\_EDUCM2$) *sugiere cierta asociación positiva* entre ambas variables, puesto que la mediana del peso parece ser creciente con el nivel educativo de la madre. Así lo sugieren asimismo los valores p de los contrastes de la chi-cuadrado, de Kruskal-Wallis y de van der Waerden que aparecen en la tabla. Tal asociación podría reflejarse en un gráfico de barras que mostrase los pesos medianas que aparecen debajo del rótulo *Category Statistics* en la tabla $MEDN\_PESO\_EDUCM2$ como función de los valores centrales de los intervalos que aparecen para la variable $educm_i$. Sin embargo, tal como muestra el gráfico de barras de *Med_peso_educm2*, la asociación, si existe, es débil.

También en la relación con el nivel educativo del padre, hemos efectuado dos veces el contraste de igualdad de medianas: una, sin agrupar los niveles educativos ($MEDN\_PESO\_EDUCP$), y otra, agrupándolos ($MEDN\_PESO\_EDUCP2$); la segunda es preferible, a pesar de que el nivel de agrupación es relativamente arbitrario. En casos como los que estamos analizando, 15 clases parece un número razonable, pues permite que aflore cierta disparidad entre medianas, a la vez que permite recoger una mínima frecuencia dentro de cada clase. Si juzgamos por los valores p de los contrastes, la evidencia contraria a la hipótesis nula de igualdad de medianas, lo que sugeriría una posible asociación entre las variables *peso* y *educp*, es claramente menor que en el caso del nivel

[10] Para obtener una clasificación de la variable *Peso* utilizando como clasificador los niveles educativos de la madre, seleccionar *Peso* y entrar en *"Descriptive Statistics/Stats by Classification"* escribiendo EDUCM en la ventana *"Series/Group for Classify"*. Para contrastar la igualdad de medianas entre grupos a la vez que se lleva a cabo la clasificación, entrar en *"Tests for Descriptive Statistics/ Equality Tests by Classification"*, escribiendo EDUCM en *"Series/Group for Classify"*, y marcando *"Mediana"*, en vez de *"Media"* bajo *"Test Equality of"*. Para obtener una clasificación con agrupación de niveles educativos, a la derecha, donde aparece *"Group into Bins if "* marcar un número reducido (por ej., 10) en la ventana *"# of values"*, que se refiere al número de rangos de valores que se quieren utilizar para la variable que se utiliza como clasificador, en este caso, EDUCM.

educativo de la madre, sugiriendo que el nivel educativo del padre podría no ser muy relevante para explicar el peso del bebé. Sin embargo, no hemos de olvidar que estamos comparando únicamente una medida de posición central de la variable *peso* para los distintos grupos definidos para *educm* o *educp;* no examinamos el conjunto de todos los valores de *peso* observados dentro de cada nivel educativo, lo que podría arrojar ciertas diferencias entre distintos niveles de $educm_i$. Por ejemplo, podríamos observar que los rangos observados para $peso_i$ se amplían o se estrechan al aumentar $educm_i$, sugiriendo que la varianza de la variable $peso_i$ es función del nivel educativo de la madre. Una evolución creciente de los pesos mínimo y máximo sugeriría asimismo una relación positiva, siendo negativa si si se observase la evolución contraria; esto podría ocurrir sin observar variaciones significativas en los valores mediana.

La evidencia a favor de asociación es bastante más clara en la comparación de *peso* y *renta* ($MEDN\_PESO\_RENTA$), y todavía más clara en el caso de *peso* y *cigarrillos* ($MEDN\_PESO\_CIGS2$). Un diagrama de barras de las medianas de *peso* por clases de *renta* sugiere una asociación positiva (MED_PESO_RENTA), mientras que un diagrama de medianas de *peso* por clases de valores de *cigarrillos* sugiere una asociación negativa (MED_PESO_CIGS2), si bien esta última clasificación está contaminada por el elevado porcentaje muestral con un valor cero de la variable *cigarrillos*. En el fichero de trabajo se incluye asimismo la variable $FUMA$, que hemos definido de modo que el valor 0 si la madre no fumó durante el embarazo, y el valor 1 si lo hizo. El valor mediana de los pesos de los bebés fue de 111 y 120 onzas, respectivamente, en cada caso, lo que sugiere cierta dependencia negativa entre el peso y el hábito de fumar. Los valores p de los contrastes en MED_PESO_FUMA son bastante concluyentes respecto a la existencia de tal dependencia.

La igualdad de medianas no se rechaza cuando se clasifica la variable *peso* de acuerdo con los valores de la variable *ordenac,* sugiriendo que el orden del recién nacido entre sus hermanos podría no ser información relevante para explicar su peso. Este análisis descriptivo es preliminar, habiendo relacionado, alternativamente, cada una de las variables explicativas, con la variable dependiente. No hemos considerado, por tanto, la posible colinealidad entre variables explicativas, es decir, que éstas puedan proporcionar información común. A título preliminar, podríamos concluir con una ordenación de variables por niveles de capacidad explicativa, comenzando con el número de cigarrillos y la renta familiar, junto con una posible dependencia débil respecto del nivel educativo de la madre, mientras que el orden del recién nacido dentro de los hijos de la familia parece no aportar información relevante acerca de su peso. Esta evidencia es coherente con la obtenida en la regresión mostrada al inicio en lo relativo al efecto del número de cigarrillos fumados, pero no en cuanto a los posibles efectos de las variables $renta_i, ordenac_i$, o $educm_i$.

3.2.4 Análisis de regresión

Nuevamente hay que hacer notar que aunque esta sección debería comenzar presentando las nubes de puntos de las variables de la regresión pero, debido a la naturaleza de las variables explicativas, no lo hacemos. Si lo desea, el lector puede utilizar el fichero de trabajo para construir dichos gráficos. Estimamos regresiones individuales sobre las dos variables aparentemente más relevantes, *cigarrillos* y *renta*, obteniendo,

$$\begin{aligned} Peso_i &= \underset{\substack{(0,57)\\(209.3)}}{119.77} - \underset{\substack{(0,090)\\(-5,68)}}{0,514}\, cigarrillos_i + \hat{u}_i, && (10)\\ \bar{R}^2 &= 0,022,\ \hat{\sigma}_u = 20,13,\ Ratio = 0,011 && (11)\\ Peso_i &= \underset{\substack{(1,00)\\(115.0)}}{115,27} + \underset{\substack{(0,029)\\(4,08)}}{0,118} renta_i + \hat{u}_i, && (12)\\ \bar{R}^2 &= 0,011,\ \hat{\sigma}_u = 20,24,\ Ratio = 0,005 && (13) \end{aligned}$$

donde *Ratio* denota el cociente entre la desviación típica muestral de los residuos, y la de la variable peso, que es de 20,35.

Estos modelos de regresión simple puedan estar incorrectamente especificados por omitir algún efecto significativo. Si así fuese, el coeficiente estimado (la pendiente del modelo de regresión) en la primera estaría sesgado, en el sentido de no medir el efecto que sobre el peso tiene la única variable explicativa incluída en la regresión, *cigarrillos*; la estimación de dicho coeficiente estaría recogiendo asimismo los efectos de variables omitidas que no sean independientes de la variable incluida, por ejemplo, la renta de la familia, o la ordenación del recién nacido entre sus hermanos. Sabemos algo más: de acuerdo con la discusión teórica relativa al sesgo por variables omitidas, al omitir una variable explicativa negativamente correlacionada con *cigarrillos*, el coeficiente de ésta se subestimará,sobreestimándose si la variable omitida tiene correlación positiva con *cigarrillos* pues, en ambos casos, asignaríamos a *cigarrillos* el efecto combinado de ambas variables. Esto es precisamente lo que diría nuestra intuición.

El primer paréntesis debajo de cada coeficiente estimado contiene la desviación típica de la estimación, mientras que el segundo contiene el estadístico tipo-t, cociente entre la estimación y su desviación típica. En muestras amplias de sección cruzada es habitual obtener un valor numérico muy reducido para el coeficiente de determinación, si bien desearíamos que fuese algo mayor del obtenido en estas regresiones individuales. En todo caso, los niveles obtenidos del $\bar{R}^2$ en absoluto indican ausencia de relación.

Este es un caso en el que el uso habitual de los estadísticos tipo-t sugeriría que ambas variables tienen capacidad explicativa relevante, siendo *estadísticamente significativas*; de acuerdo con tal criterio, nadie dudaría en incluirlas en un modelo de regresión. Sin embargo, las desviaciones típicas residuales, y los *Ratios* indican que la capacidad explicativa de cada una de estas variables por separado es, verdaderamente, muy reducida. El coeficiente estimado para *cigarrillos*, implica que, para el valor mediana de los cigarillos fumados durante el embarazo (cuando no son cero), que es de 10, la diferencia en peso de bebés de madres fumadoras y madres no fumadoras sería de 5 onzas, menor que la diferencia observada en la muestra, de 112 a 121 onzas, a que antes nos referimos.

Evidencia adicional acerca de la reducida información que *cigarrillos* y *renta* proporcionan sobre *peso* aparece en FIG_RES_CIGS y FIG_RES_RENTA, que representan los valores ajustados y los residuos de ambas regresiones. Este es un tipo de gráficos que siempre hemos de examinar, tras estimar un modelo de regresión. Estos gráficos son la evidencia más clara acerca de la reducidísima capacidad explicativa de las dos variables, ya que la mayor parte de la fluctuación en peso de unos bebés a otros permanece en los esiduos, npo habiendo sido explicada por las variables utilizadas como explicativas en la regresión.

Indicios adicionales acerca de la baja capacidad explicativa aparecen en $CORR\_PESO\_AJUSTE$,

que muestra coeficientes de correlación entre *peso* y los residuos de las dos regresiones, así como de la regresión que incluye ambas variables, *cigarrillos* y *renta*, como variables explicativas, y de otras regresiones que analizaremos posteriormente. Las variables mencionadas son las que han sido incluidas como explicativas en cada regresión. Todas las correlaciones son muy elevadas, lo que significa que la parte de la variable *Peso* que queda sin explicar por las variables *renta* y *cigarrillos* es muy similar a la propia variable *Peso*, es decir, que las regresiones apenas explican las diferencias en *peso* entre bebés. Es interesante que la correlación sea algo menor cuando se utilizan ambas variables, lo que sugiere que la información que contienen no es exactamente común, si bien es reducida en ambos casos.

Correlaciones tan elevadas pueden interpretarse asimismo en el sentido de que, si utilizásemos las regresiones estimadas para predecir el peso de un recién nacido utilizando las variables *cigarrillos* y *renta* como predictores, la correlación entre la previsión resultante y el peso observado del bebé sería muy pequeña o, lo que es equivalente, la calidad de la predicción sería muy baja. Por ejemplo, para el nivel mediana de renta, 27,5, el modelo (12) predice un peso de 118,52 onzas. En la muestra se observa[11], para dicho nivel de renta, un rango de pesos entre 80 y 167 onzas; demasiada dispersión para poder prever con precisión, lo que explica el bajo ajuste del modelo.

3.3 Ejemplo: Discriminación salarial[12]

Este ejemplo tiene como objetivo describir la utilización de variables ficticias para contrastar la estabilidad del modelo de regresión entre submuestras. En la primera parte del ejercicio, utilizamos un modelo de determinación de salarios en función del nivel educativo y la experiencia laboral del trabajador, y se examinan las posibles diferencias en el modelo estimado entre las submuestras de hombres y mujeres. Tras detectar evidencia consistente con la existencia de discriminación salarial en contra de las mujeres, se profundiza en analizar si la dscriminación se debe a una menor valoración del nivel educativo, la experiencia laboral, o de ambos factores. En la segunda parte explicamos diversas maneras en que el uso de variables ficiticias, convenientemente definidas e introducidas en el modelo econométrico, permite contrastar la homogeneidad salarial entre trabajadores de distintas características. En esta segunda parte nos centramos en caracterizar el posible impacto diferencial que sobre el salario tengan la experiencia laboral y el nivel educativo.

3.3.1 Descripcion de los datos

El archivo Bwages.wf1 contiene datos relativos a 1.472 personas encuestadas en Bélgica en 1994, como parte del European Community Household Panel, a las que se ha preguntado por: *a*) su salario en Bef. (40 Bef. equivalen aproximadamente a 1 euro), *b*) su nivel de educación, indicando uno entre cinco niveles posibles, y *c*) su experiencia laboral, en términos del número de años que el encuestado ha trabajado hasta el momento. El salario es una variable de naturaleza continua, si bien toma únicamente valores positivos; la experiencia profesional es una variable discreta, observable numéricamente, por lo que no se trata de una variable ficticia. Por el contrario, el nivel educativo es una variable cualitativa, para la que hay que definir un variable ficticia. Esto podría hacerse de muchas formas distintas, pero lo más natural es utilizar los cinco primeros números enteros, asignándolos a cada nivel educativo, en orden creciente. Hay que entender, sin embargo, que cualquier otra asignación numérica sería asimismo posible.

[11] Ver Bwght.xls

[12] Fichero de trabajo: Bwages.wf1. La base de datos Bwages.txt está tomada de los archivos que acompañan a Kuleuven

A priori, es logico considerar que ambas variables, experiencia y educación, pueden incidir positivamente sobre el nivel salarial del trabajador. Un aspecto a tener en cuenta en la interpretación del modelo es que, a diferencia de la experiencia laboral, cuyos valores numéricos sucesivos están separados siempre por un año más de experiencia, los valores sucesivos de la variable educación recogen distintos niveles educativos, no siendo en absoluto evidente que la diferencia entre dos cualesquiera de dichos niveles sucesivos haya de tener siempre un mismo efecto sobre el salario.

Por último, disponemos asimismo de información acerca de si el encuestado es hombre o mujer; ésta es también una variable cualitativa, para la que el investigador debe construir, por tanto, una variable ficticia. Dicha variable, con el nombre *male*, ya está incluida en el fichero de trabajo. En la descripción del archivo (*Bwages.txt*)se nos dice que el valor $male = 1$ corresponde a varones; en la base de datos, la variable solo toma valores 0 ó 1, por lo que se ha asociado $male = 0$ a las mujeres encuestadas.

3.3.2 Estadísticos descriptivos

El cálculo de estadísticos de la variable *male*, en la tabla de clasificación proporcionada por Eviews,[13] revela que hay en la encuesta 579 individuos con valor cero de esta variable (mujeres), siendo varones el resto, 893. Podemos calcular la media muestral (0,606) y la mediana (1,000) de esta variable pero, evidentemente, estos estadísticos carecen de interés, salvo que quisiéramos calcular, de modo muy indirecto, utilizando la media muestral, el porcentaje de personas de uno y otro sexo dentro de la muestra. La mediana únicamente nos dice en este caso que hay más individuos en la muestra con valor $male = 1$ que con valor $male = 0$; la media nos dice que el 60,6% de los encuestados son varones.

Los estadísticos muestrales, recogidos en $HIST\_WAGES, HIST\_EDUC, HIST\_EXPER$ nos revelan asimismo que los niveles educativos observados en la muestra son $1, 2, 3, 4, 5$, con media muestral de 2,38 y mediana de 3,0, mientras que los niveles de experiencia oscilan entre 0 y 47, siendo únicamente números enteros, con una media de 17,22 años y una mediana de 16,5 años.

Vamos a utilizar estos datos para efectuar un doble análisis: por un lado, caracterizar los determinantes del salario que recibe un trabajador; por otro, contrastar si existe discriminación salarial entre hombres y mujeres en el mercado de trabajo de donde provienen los datos.

Para analizar los determinantes salariales, comenzamos calculando el salario medio por nivel educativo, obteniendo, respectivamente, 340,03; 371,74; 411,60; 461,13; 563,20, para los cinco sucesivos niveles educativos.[14] Que dichos salarios medios sean crecientes sugiere que el nivel educativo es un posible determinante del salario. La evidencia similar para la experiencia profesional es más difícil de obtener, entre otras cosas, porque en este caso hay 48 niveles de experiencia diferentes, en muchos de los cuales hay un número reducido de encuestados, lo que hace que el salario medio de dicho grupo se mida con poca precisión. Estrictamente hablando, el salario medio no es creciente con cada año de experiencia profesional[15], aunque se observa una clara tendencia a aumentar con dicha variable; en todo caso, la evidencia muestral acerca de que el salario aumenta con la experiencia profesional es más tenue que respecto al nivel educativo. El coeficiente de correlación entre salario y experiencia es de 0,307, sugiriendo que existe cierta relación entre estas variables. La

[13] Tras marcar la variable *male*, marcar las pestañas: View_Descriptive Statistics_ Stats by classification, y en la ventana Series/Group for classify, indicar: *male*.

[14] Para ello, utilizamos la variable *educ* como clasificador al calcular los estadísticos descriptivos de la varible *wage,* obteniendo $WAGE\_BY\_EDUC$.

[15] Como se puede ver clasificando la variable $WAGE$ utilizando $EXPER$ como clasificador, lo que da lugar a la taba $WAGE\_BY\_EXPER$.

nube de puntos (*nube_ wage_ exper*) que representa el salario en función de la experiencia arroja asimismo una cierta imagen de dependencia, aunque débil, especialmente para los trabajadores de mayor experiencia.

Para obtener una primera evidencia acerca de una posible discriminación salarial, calculamos el salario medio por sexos, obteniendo 413,95 Bef. para las mujeres, frente a 466,42 Bef. para los hombres, y un salario medio global, para todos ellos, de 445,78 Bef. que está, lógicamente, comprendido entre el promedio de uno y otro grupo. Este dato inicial ya es favorable a la posible existencia de discriminación salarial, por cuanto que el salario medio de los hombres es superior al de las mujeres, supuesto que todos ellos desempeñan tareas laborales comparables. Otro aspecto interesante es que el histograma de la variable salarios revela una fuerte curtosis, que conduce al rechazo de la hipótesis de Normalidad de acuerdo con el test de Bera-Jarque. Esto significa que no es muy razonable creer que dicha variable quede determinada mediante la suma de un conjunto de variables deterministas y un término estocástico con distribución Normal, como se supone en el modelo de regresión lineal habitualmente.

Antes de afirmar categóricamente la existencia de discriminación salarial, sin embargo, debemos considerar la posibilidad de que la diferencia entre los salarios medios de ambos grupos pueda obedecer a que las características de los hombres y mujeres encuestados, en términos de educación y experiencia, sean diferentes: que el salario medio de los varones fuera más elevado no indicaría necesariamente discriminación salarial, si su nivel educativo y experiencia profesional fuesen más altos. La experiencia profesional media es de 15,2 años entre mujeres y de 18,5 años entre hombres[16], con una media global de 17,2 años. Los valores mediana son 14 y 18, respectivamente, para ambos grupos. El nivel educativo medio es de 3,59 entre mujeres, y de 3,24 entre hombres, con una media muestral global de 3,38. Los niveles educativos mediana son de 4 y 3, respectivamente, para mujeres y hombres. Así, en esta muestra los hombres tienen, en media, mayor experiencia laboral pero menor nivel educativo que las mujeres. Al no ser ambas variables más elevadas entre hombres que entre mujeres, la diferencia en salarios entre ambos grupos puede estar indicando, efectivamente, discriminación salarial contra las mujeres.

3.3.3 Análisis de regresión

Comenzamos el análisis de regresión utilizando separadamente el nivel educativo y la experiencia laboral como variables explicativas del salario,

$$\begin{array}{rcl} Salario_i = 249,51 + & 58,10 & Educación_i + u_i,\ \bar{R}^2 = 0,151,\ \hat{\sigma}_u = 165,4. \\ & (3,58) & \\ & (16,22) & \end{array} \tag{14}$$

Las desviaciones típicas y los estadísticos t aparecen entre paréntesis, bajo el valor estimado de la pendiente de la recta de regresión. Como la desviación típica muestral de la variable salario es 179,53, el *Ratio* $1 - \frac{\hat{\sigma}_u}{\sigma_y} = 0,08$. El salario parece aumentar, en media, en algo más de 58 Bef. por año de educación; por el modo en que se ha definido la variable *Educación,* la regresión estimada sugiere, además, que el incremento salarial por año de educación adicional es siempre el citado, con independencia de que se trate del paso del nivel educativo 1 al 2, o del nivel 4 al nivel 5. La regresión,

[16] Como se puede ver clasificando la variable *EXPER* utilizando *MALE* como clasificador. View_Descriptive statistics_Stats by classification, escribiendo *male* en la ventana de Series/Group for Classify.

$$Salario_i = 352,36 + \underset{(12,38)}{\underset{(0,44)}{5,43}} Experiencia_i + u_i,\ \bar{R}^2 = 0,094,\ \hat{\sigma}_u = 170,9. \quad (15)$$

proporciona una estimación de que el salario aumenta en unos 5,4 Bef. por año de experiencia laboral. Nuevamente, la regresión estimada restringe a que el incremento salarial por año de experiencia adicional sea el mismo tanto si el aumento se produce a un nivel reducido como a un nivel elevado de experiencia.

4 Medidas de bondad de ajuste del modelo de regresión

Hasta aquí, hemos propuesto un criterio, de entre los muchos posibles, para obtener estimadores de los coeficientes del modelo de regresión lineal simple: minimizar la suma de los cuadrados de los residuos, y hemos obtenido las expresiones analíticas de los estimadores resultantes, así como de sus varianzas y su covarianza. Cada uno de estos estimadores es una función de las observaciones muestrales de ambas variables, X e Y, y son, por tanto, variables aleatorias; por eso hemos calculado sus esperanzas matemáticas y varianzas. Si alguno de ellos fuese función únicamente de las observaciones de la variable X tendría naturaleza determinista, y su valor no cambiaría si en vez de utilizar en la estimación del modelo la muestra de que disponemos, pudiésemos utilizar otra muestra diferente de igual tamaño.

Sin embargo, éste no es el caso: ambos estimadores dependen también de las observaciones de la variable Y, por lo que tienen naturaleza estocástica, es decir, su valor numérico sería distinto con muestras diferentes. Variando la muestra, obtendríamos distintos valores de 0 y 1, todos los cuales nos describirían el histograma de frecuencias correspondiente a su distribución de probabilidad. En los párrafos anteriores hemos demostrado que la esperanza matemática de cada uno de estos estimadores es el verdadero valor, que es desconocido, del parámetro que pretende estimar, y hemos deducido las expresiones analíticas de las varianzas de cada una de sus distribuciones de probabilidad.

El procedimiento MCO que hemos utilizado garantiza que la recta de regresión obtenida es la que proporciona la menor Suma de Cuadrados de Residuos que es posible obtener trazando rectas a través de la nube de puntos. Sin embargo, en unas ocasiones tal mejor ajuste puede ser excelente, en otras, el mejor ajuste puede no ser muy bueno. Necesitamos, en cualquier caso, disponer de criterios que puedan resumir en un indicador el grado de ajuste de la regresión MCO a la nube de puntos de que partimos.

Recordemos que:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Si la perturbación aleatoria sigue una distribución de probabilidad Normal, entonces $y_i = \beta_0 + \beta_1 x_i + u_i$ también sigue una distribución Normal, pues es igual a una constante, más una variable con distribución Normal. Además:

$$\begin{aligned} E(y_i) &= E(\beta_0 + \beta_1 x_i + u_i) = \beta_0 + \beta_1 x_i \\ Var(y_i) &= Var(\beta_0 + \beta_1 x_i + u_i) = Var(u_i) = \sigma_u^2 \end{aligned}$$

de modo que, de acuerdo con el modelo, todas las observaciones de la variable endógena tienen la misma varianza, pero diferente esperanza matemática, pues ésta depende del valor numérico de la variable X, que varía a lo largo de la muestra.

Puede probarse que el residuo correspondiente a cada observación es una combinación lineal de todos los términos de error del modelo y, por tanto, si la perturbación aleatoria del modelo es Normal, el residuo también tiene distribución Normal. Su esperanza matemática es:

$$\begin{aligned} E(\hat{u}_i) &= E(y_i - \hat{y}_i) = E(y_i) - E(\hat{\beta}_0 + \hat{\beta}_1 x_i) = E\left(\beta_0 + \beta_1 x_i + u_i\right) - E(\hat{\beta}_0) - E(\hat{\beta}_1 x_i) = \\ &= \beta_0 + \beta_1 x_i + E(u_i) - \beta_0 - \beta_1 x_i = 0 \end{aligned}$$

Teniendo en cuenta que, entre β_0, β_1, $\hat{\beta}_0, \hat{\beta}_1$y x_i, sólo $\hat{\beta}_0$ y $\hat{\beta}_1$son aleatorios, puede obtenerse la siguiente expresión para la varianza de cada residuo:

$$\begin{aligned} Var(\hat{u}_i) &= Var(y_i - \hat{y}_i) = Var\left[(\beta_0 + \beta_1 x_i + u_i) - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right] = \\ &= \frac{\sigma_u^2}{n} \frac{\sum_{j=1}^{n} (x_j - x_i)^2}{\sum_{j=1}^{n} (x_j - \bar{x})^2} \end{aligned}$$

Al tener esperanza cero, la varianza del residuo es un adecuado indicador de su tamaño. Podemos ver que la varianza es tanto mayor (lo cual no es deseable), cuanto mayor es σ_u^2, pero es menor cuanto mayor sea el tamaño muestral. También es menor cuanto mayor es la varianza muestral de la variable explicativa, lo cual es, por tanto, un aspecto deseable: un apreciable grado de fluctuación en X no es negativo, sino positivo. Por último, nótese que la observación x_i correspondiente al residuo i aparece en el numerador. Cuanto más se separe ésta de la media de todas las x_i, mayor será la varianza del residuo correspondiente a dicha observación muestral.

4.1 Error Estándar de la Regresión (EER)

No sólo es cierto que la esperanza matemática de la distribución de probabilidad de cada uno de los residuos MCO es igual a cero. También se cumple que su media muestral es igual a cero, puesto que la suma de todos ellos lo es, como vimos en las *ecuaciones normales*. Esta es una peculiaridad del método de estimación MCO, que otro procedimiento de estimación no tiene. Si, considerados a lo largo de toda la muestra, los residuos tienen media cero, entonces su desviación típica muestral será un indicador del tamaño promedio de cada uno de ellos. Esto es importante, porque si la recta estimada se ajusta bien a la nube de puntos, entonces los residuos deberían ser pequeños en algún sentido. Utilizar la desviación típica muestral de los residuos parece un criterio razonable de ajuste. Además, sabemos que si utilizamos $n-2$ en el denominador, su cuadrado es un estimador insesgado de σ_u^2. La ausencia de sesgo en este estimador puede demostrarse sin necesidad de obtener previamente los residuos de la regresión, tomando esperanzas en la expresión:

$$\begin{aligned} \hat{\sigma}_u^2 &= \frac{SCR}{n-2} = \sum_{i=1}^{n} \frac{\hat{u}_i^2}{n-2} = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = \\ &= \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i y_i = \frac{1}{n-2} \left(\sum_{i=1}^{n} y_i^2 - \hat{\beta}_0 \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i y_i\right) \end{aligned}$$

Su raíz cuadrada, la desviación típica estimada, recibe el nombre de error estándar de la regresión EER:

$$EER = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}} = \sqrt{\hat{\sigma}_u^2} = \hat{\sigma}_u$$

Es claro que minimizar la varianza residual equivale a minimizar el error estándar de la regresión, EER. Sin embargo, recordemos que la desviación típica tiene, respecto a la varianza, la ventaja de estar medida en las mismas unidades que la variable a la que se refiere, el residuo, que tiene, a su vez, las mismas unidades que la variable endógena y_i. Para valorar si el ajuste obtenido por la recta MCO a la nube muestral de puntos es bueno, es conveniente utilizar el valor numérico del EER en relación con alguna referencia, y la media muestral de la variable endógena es un buen indicador. Ello nos permite presentar el porcentaje que de la media de y_i representa el EER, pudiendo decir, por ejemplo: el modelo estimado es bueno, puesto que el EER es tan sólo un 4% de la media de la variable endógena o, por el contrario: "el ajuste obtenido no es muy bueno, porque el tamaño medio de los residuos, indicado por el EER, es de un 65% de la media de Y".

4.2 El coeficiente de determinación

El interés del EER como indicador del grado de ajuste de un modelo de regresión disminuye cuando queremos comparar la bondad del ajuste de dos modelos que tienen una *variable dependiente diferente*. En tal caso, no es en absoluto cierto que el modelo con menor EER sea el modelo con mejor ajuste; de hecho, no podremos afirmar nada al respecto, salvo que establezcamos alguna medida relativa de grado de ajuste, que es lo que hacemos en esta sección. A diferencia del EER, el *coeficiente de determinación* que ahora definimos, denotado por R^2, es un indicador sin unidades, que no es preciso ni tiene sentido poner en relación con ninguna de las variables del modelo.

En primer lugar, escribamos para cada observación i :

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

que muestra que la distancia entre una observación y_i y su media $\bar{y}$ puede escribirse como la distancia entre su valor ajustado $\hat{y}_i$ y dicha media, más el residuo correspondiente. La distancia a la media del valor ajustado puede ser mayor o menor que la de la observación y_i, por lo que el residuo puede ser negativo o positivo. La regresión estimada por MCO proporciona el valor numérico de $\hat{y}_i - \bar{y}$, que es una aproximación a la distancia $y_i - \bar{y}$. El resto es la parte no explicada, o residuo. Como hemos mencionado, la explicación puede exceder o no de $y_i - \bar{y}$. La igualdad anterior muestra cómo *la desviación total respecto a la media puede escribirse como la suma de la desviación explicada y el residuo.*

Si elevamos al cuadrado ambos miembros, tenemos:

$$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + \hat{u}_i^2 + 2\,(\hat{y}_i - \bar{y})\,\hat{u}_i$$

y sumando a lo largo de toda la muestra:

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} \hat{u}_i^2 + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y})\,\hat{u}_i \tag{16}$$

Pero:

$$\begin{aligned}\sum_{i=1}^{n}\hat{u}_i\left(\hat{y}_i-\bar{y}\right) &= \sum_{i=1}^{n}\hat{u}_i\hat{y}_i-\bar{y}\sum_{i=1}^{n}\hat{u}_i=\sum_{i=1}^{n}\hat{u}_i\hat{y}_i=\sum_{i=1}^{n}\hat{u}_i(\hat{\beta}_0+\hat{\beta}_1x_i)= \\ &= \hat{\beta}_0\sum_{i=1}^{n}\hat{u}_i+\hat{\beta}_1\sum_{i=1}^{n}\hat{u}_ix_i=\hat{\beta}_0(0)+\hat{\beta}_1(0)=0\end{aligned}$$

donde hemos utilizado repetidamente el hecho de que la suma de los residuos MCO es igual a cero, así como que la suma de sus productos por x_i también es igual a cero. Ambas condiciones provienen de las ecuaciones normales.

Finalmente, substituyendo en (16), llegamos a:

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

es decir:

$$nS_y^2=nS_{\hat{y}^2}+nS_{\hat{u}^2}$$

que expresa cómo la *variación muestral* total en la variable Y, que es n veces su varianza, puede descomponerse como la suma explicada por la regresión estimada, $nS_{\hat{y}^2}=\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2$, el primero de los sumandos del miembro derecho, más la suma no explicada, que es la suma de los cuadrados de los residuos. Si dividimos la suma explicada por la variación total en Y, tenemos la definición de *coeficiente de determinación*:

$$R^2=\frac{nS_{\hat{y}^2}}{nS_y^2}=\frac{S_{y^2}-S_{\hat{u}^2}}{S_y^2}=1-\frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}=\frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2}$$

de modo que:

$$R^2=1-\frac{\textit{Variación no explicada en } Y}{\textit{Variación total en } Y}=\frac{\textit{Variación explicada en } Y}{\textit{Variación total en } Y}$$

Proposition 1 *El coeficiente de determinación de todo modelo de regresión toma siempre valores numéricos entre 0 y 1.*

Proof. El miembro derecho de la ecuación es el cociente de dos términos positivos, luego es positivo. Además, hemos visto que el numerador es uno de los dos componentes del denominador, luego su valor numérico es inferior al de éste. En consecuencia, el cociente, que es positivo, es inferior a la unidad. ■

El coeficiente de determinación, a veces denominado $R-cuadrado$, nos indica el porcentaje de la variación total en la variable Y que la regresión estimada es capaz de explicar. La idea es que si la regresión tiene un ajuste suficientemente bueno, será debido a que la variable X explica buena parte de la variación que Y experimenta a lo largo de la muestra, los residuos serán generalmente pequeños, la variación explicada en Y será un porcentaje elevado de su variación muestral total, y el coeficiente de determinación será próximo a la unidad. Lo contrario ocurrirá cuando el ajuste de la recta MCO a la nube de puntos no sea suficientemente bueno, en cuyo caso el coeficiente de determinación será próximo a cero.

Así pues, un coeficiente de determinación próximo a 1 significa que las estimaciones obtenidas para los coeficientes del modelo de regresión hacen a éste capaz de explicar un elevado porcentaje de las variaciones que experimenta la variable endógena. El modelo proporciona en tal caso un buen ajuste a los datos, por lo que puede utilizarse con confianza para efectuar evaluaciones e inferencias acerca de la cuestión conceptual que lo motivó inicialmente. En el extremo contrario, un coeficiente de determinación próximo a cero significa que las estimaciones obtenidas apenas explican las variaciones que experimenta la variable endógena, por lo que el modelo no puede utilizarse con una gran fiabilidad.

Hay que tener bastante cuidado, sin embargo, con la interpretación del coeficiente de determinación de una regresión. En ocasiones, si la muestra consta de pocas observaciones, quizá uno o dos residuos elevados pueden generar un coeficiente de determinación reducido y, por ello, conducir a creer que la regresión estimada es mala, cuando excepto por dichas observaciones, el ajuste puede ser excelente. Por otra parte, si la muestra consta de muy pocas observaciones, y ningún residuo es especialmente alto, se tendrá un coeficiente de determinación muy elevado, sin que deba interpretarse como un excelente ajuste, sino más bien como un indicador de escasa información muestral.

Otro caso delicado se refiere al uso del coeficiente de determinación con muestras de series temporales que muestran una tendencia similar. En tales casos, el coeficiente de determinación se aproxima a la unidad, aunque la relación entre ambas variables, excepción hecha de sus tendencias, pueda ser pobre. Esto viene indicado por dos ejercicios relacionados: a veces, basta estimar y extraer una tendencia determinista de dos series temporales X e Y para que un coeficiente de determinación en torno a 0,90 antes de la extracción de tendencias, se reduzca a 0,3 ó 0,4. El otro ejercicio, casi reverso del anterior, puede efectuarse tomando dos variables con poca relación, y añadiéndoles una tendencia, es decir, el producto de un detemrinado coeficiente, como $\beta = 0,27$, ó $\beta = 3,45$, por una variable de tendencia, que toma valores 1,2,3,... . Pues bien, si el coeficiente de determinación antes de añadir la tendencia estaba en torno a 0,20, por ejemplo, podría pasar a ser de 0,80 tras añadir la misma tendencia a ambas variables. Estos ejercicios son importantes, porque no querríamos decir en ninguno de los dos casos que las dos variables están muy relacionadas y que, en consecuencia, el modelo de regresión estimado es bueno, sólo porque el coeficiente de determinación sea elevado debido a la presencia de la tendencia común a ambas variables. Este aspecto, de suma importancia, es conocido como el problema de regresión espúria, y es estudiado en detalle más adelante.

Todo esto hace que, entre otras cosas, se exija un coeficiente de determinación superior en regresiones estimadas con datos de series temporales que con datos de sección cruzada. En todo caso, es imprescindible acompañar toda estimación de un modelo de regresión, con los estadísticos que permitan evaluar la bondad del ajuste entre modelo y datos. Estos incluirán el coeficiente de determinación R^2, el EER, así como estadísticos que examinaremos en las próximas secciones.

4.3 Correlación en el modelo de regresión lineal

Correlación es el grado de dependencia que existe entre variables. Cuando se trata de sólo dos variables, existe una medida, el coeficiente de correlación, introducido por K.Pearson:

$$\rho_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$$

cuya justificación estamos ahora en condiciones de comprender. Vamos a demostrar que el coeficiente de correlación de Pearson mide el grado de *dependencia lineal* que existe entre dos

variables, X e Y.

Para ello, partimos del coeficiente de determinación de una regresión lineal simple, y extraemos su raíz cuadrada, denotando por r_{xy} al estadístico que así se obtiene:

$$r_{xy} = \sqrt{R^2} = \sqrt{1 - \frac{S_{\hat{u}^2}}{S_y^2}}$$

Ahora bien, puesto que:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) + \hat{\beta}_1 x_i = \bar{y} + \hat{\beta}_1 (x_i - \bar{x})$$

tenemos:

$$\begin{aligned} S_{\hat{u}^2} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \left(\bar{y} + \frac{S_{xy}}{S_x^2}(x_i - \bar{x})\right)\right]^2 = \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[(y_i - \bar{y})^2 + \frac{(S_{xy})^2}{(S_x^2)^2}(x_i - \bar{x})^2 - 2\frac{S_{xy}}{S_x^2}(x_i - \bar{x})(y_i - \bar{y})\right] = \\ &= S_y^2 + \frac{(S_{xy})^2}{(S_x^2)^2}S_x^2 - 2\frac{(S_{xy})^2}{S_x^2} = S_y^2 - \frac{(S_{xy})^2}{S_x^2} \end{aligned}$$

y, en consecuencia:

$$r_{xy} = \sqrt{R^2} = \sqrt{1 - \frac{S_{\hat{u}^2}}{S_y^2}} = \sqrt{1 - \frac{S_y^2 - \frac{(S_{xy})^2}{S_x^2}}{S_y^2}} = \sqrt{\frac{(S_{xy})^2}{S_x^2 S_y^2}} = \frac{S_{xy}}{S_x S_y} = \rho_{xy}$$

obteniendo, precisamente, *el coeficiente de correlación lineal*. Es decir, por haber demostrado que el coeficiente de correlación de Pearson no es sino la raíz cuadrada del coeficiente de determinación en un modelo de regresión lineal, podemos afirmar que el coeficiente de correlación de Pearson mide el grado de relación entre dos variables, X e Y, supuesto que la relación entre ambas sea de tipo lineal. Por tanto, su interpretación sólo está realmente justificada en la medida que la regresión óptima entre ambas variables, es decir, la esperanza condicional de Y dado X, sea lineal, y no en otro caso.

Asimismo, puesto que ya hemos probado que el coeficiente de determinación está comprendido entre 0 y 1, podemos obtener ahora como corolario que *el coeficiente de correlación de Pearson está siempre comprendido entre -1 y +1*, resultado bien conocido de cursos de Estadística.

Es importante destacar que si la verdadera relación entre dos variables no es lineal, y utilizamos el coeficiente de correlación de Pearson como un indicador del grado en que ambas están relacionadas, podemos cometer todo tipo de errores. En tal situación, habría que tratar de identificar qué forma funcional adopta el mejor modelo de relación entre ambas variables con el objeto de proceder a su estimación y posterior evaluación de los correspondientes residuos. No es difícil encontrar ejemplos de *relación no lineal exacta* entre dos variables a pesar de que ambas presentan un coeficiente de correlación igual a cero.

Como sabemos, si dos variables son independientes, entonces su covarianza es igual a cero. Pero el coeficiente de Pearson es el cociente entre ésta y el producto de las desviaciones típicas de X e Y,

por lo que, si dos variables son independientes, entonces su coeficiente de correlación lineal es igual a cero. Ello no puede sorprendernos en modo alguno: estamos afirmando que si dos variables X e Y son independientes, y ajustamos una recta de regresión, es decir, un modelo lineal, a un conjunto de observaciones muestrales de ambas variables, entonces detectaremos un grado de asociación nulo entre ambas.

También podríamos ajustar modelos de otro tipo, con funciones no lineales; aunque no los hemos examinado aquí, existen procedimientos de estimación de tales modelos. Hecho tal ejercicio, volveríamos a detectar una capacidad nula del modelo no lineal, para relacionar X e Y, si bien es cierto que deberíamos utilizar algún estadístico adecuado, que relacionase la suma de cuadrados de los residuos con la suma de cuadrados de la variable Y. En resumen, si dos variables son independientes, no podemos estimar ninguna forma funcional de relación entre ellas que genere capacidad explicativa alguna; en particular, una recta no explicará ninguna asociación.

Por el contrario, si el coeficiente de correlación de Pearson es nulo, sólo podremos afirmar que la relación lineal entre ambas variables no es muy buena, pues no se detecta un grado apreciable de asociación entre ambas, supuesto que la forma funcional de tal hipotética relación sea lineal. Sin embargo, ello no excluye la posibilidad de que otra forma funcional, no lineal, reflejase un grado de asociación notable entre ambas variables que, en tal caso, serían dependientes. Por tanto, ausencia de correlación lineal entre dos variables, o incorrelación, que es lo que mide el coeficiente de correlación de Pearson, no implica en modo alguno su independencia.

Ahora que conocemos la estrecha relación entre coeficiente de correlación de Pearson y coeficiente de determinación, podemos apreciar que el primero nos proporciona una información acerca de la relación entre las variables que el coeficiente de determinación no consigue transmitirnos. Ello se debe a que el coeficiente de determinación es el cuadrado del coeficiente de correlación, por lo que pierde la información concerniente a su signo; ésta es relevante, excepto en algunas situaciones en que es perfectamente conocido a priori, dada la naturaleza de las variables X e Y. Por ejemplo, si estimamos una regresión de la cantidad vendida de un producto en un mercado con cierto poder de monopolio, sobre su precio, sabemos a priori que ésta será una relación de signo negativo: un coeficiente β_1 negativo implicará que variaciones positivas, es decir, aumentos en el precio del producto, se transmiten en variaciones negativas, es decir, descensos, en la cantidad vendida, y viceversa. En este ejemplo, nos interesará tan sólo tratar de estimar el grado en que el precio explica la cantidad vendida: si lo hace en gran medida o si, por el contrario, la capacidad explicativa no es muy elevada y debemos encontrar otros factores explicativos (quizá precios de otros productos con cierto grado de sustitución del nuestro, la renta de las familias, etc.) que añadir al modelo de regresión.

Cuando no contamos con esta información, queremos estimar no sólo la capacidad que X tiene para explicar las variaciones que experimenta Y, sino también el signo de su relación. Para ello, observemos que el signo del coeficiente de correlación es el mismo que el de la covarianza, de modo que si ésta es positiva, la relación entre ambas variables es positiva o creciente, siendo negativa o decreciente en el caso alternativo. Por otra parte, los valores numéricos absolutos del coeficiente de correlación de Pearson evolucionan muy en relación con los que toma el coeficiente de determinación: si uno es cero, lo es el otro, mientras que si el valor absoluto del coeficiente de correlación es uno, también es igual a uno el coeficiente de determinación. Además, puesto que el coeficiente de determinación sólo toma valores numéricos entre 0 y 1, necesariamente el coeficiente de correlación toma valores numéricos entre -1 y +1.

Así, decimos que cuando el coeficiente de correlación lineal es próximo a +1, la relación entre ambas variables es estrecha y directa, o de signo positivo, es decir, cuando una aumenta, también

lo hace la otra, y también tienden a disminuir simultáneamente. Cuando una de las variables está por encima de su media, la otra variable tiende a estar también por encima de su media, y cuando una está por debajo, también tiende a estarlo la otra. Si fuese exactamente igual a +1, lo que es prácticamente imposible cuando se trabaja con datos reales, diríamos que la relación entre ambas variables es *perfecta, y positiva o directa*. Cuando el coeficiente de correlación es próximo a -1, entonces la relación es muy estrecha, pero inversa, o de signo negativo, es decir, cuando una variable aumenta la otra tiende a disminuir, y viceversa. Cuando una variable está por encima de su media, la otra avriable tiende a estar por debajo de su media. Si fuese exactamente igual a -1, diríamos que la relación entre las variables es *perfecta y negativa, o inversa*. Cuando el coeficiente de correlación es próximo a cero, también lo es el coeficiente de determinación, por lo que decimos que la relación lineal entre las variables X e Y es prácticamente inexistente.

No debe olvidarse, sin embargo que, a diferencia del coeficiente de determinación, el coeficiente de correlación no es estrictamente cuantitativo: si tenemos dos modelos de regresión para una misma variable dependiente, con coeficientes de correlación de .35 y .70, no podemos decir que el segundo tiene un ajuste doblemente mejor que el primero, si bien podemos afirmar que muestra un ajuste claramente mejor. Tales afirmaciones acerca de comparaciones estrictamente cuantitativas sólo pueden hacerse para el coeficiente de determinación, por su significado como porcentaje de la variación en la variable dependiente que el modelo es capaz de explicar. Si los anteriores valores numéricos correspondiesen a los coeficientes de determinación de ambos modelos, entonces sí que podríamos afirmar que el segundo muestra un ajuste doblemente superior al primero.

En definitiva, los análisis de correlación y de regresión proporcionan respuestas similares acerca de la evolución conjunta de dos variables (o más de 2 variables, en el caso de la regresión múltiple). El análisis de correlación, basado estrictamente en el cálculo del coeficiente de correlación de Pearson, facilita el grado y signo de la asociación, pero no proporciona una idea acerca de la forma funcional de dicha relación, ni tampoco su dirección. Esta, que sí se obtiene con el análisis de regresión, es una ventaja del mismo, pero está condicionada a que se satisfagan las hipótesis del modelo de regresión lineal, que condicionan la validez del método MCO para la estimación del modelo lineal de regresión: así, si a) la verdadera función de relación entre variables, que el analista desconoce, es realmente lineal, b) no se omiten variables explicativas relevantes, c) el término de error del modelo no tiene media significativa, d) ni sus valores para distintas observaciones están correlacionados entre sí, e) si su varianza es la misma para todas las observaciones, y f) si no existe una relación causal de Y hacia X, entonces el análisis de regresión mediante la estimación MCO está plenamente justificada y será conveniente utilizarlo, por cuanto que nos proporciona más información que el mero análisis de correlación.

Además, el uso del estimador MCO en el modelo de regresión lineal simple está justificado por sus propiedades de eficiencia: es el estimador lineal de mínima varianza y si, además de las condiciones anteriores, las perturbaciones tienen distribución Normal, entonces es eficiente, pues su varianza alcanza la cota de Cramer-Rao.

Por el contrario, si tenemos razones para creer que una o más de tales hipótesis dejan de cumplirse en un grado apreciable, podemos perder confianza en los resultados que el análisis de regresión pueda facilitarnos, prefiriendo efectuar un análisis de correlación, cuya validez no descansa sobre tantas hipótesis, si bien precisa del supuesto acerca de que la verdadera función de relación entre X e Y sea lineal.

4.3.1 Propiedades de los residuos de Mínimos Cuadrados

Nótese que las ecuaciones anteriores pueden escribirse también:

$$\begin{aligned}\sum_{i=1}^{n}\hat{u}_i &= 0\\ \sum_{i=1}^{n}x_i\hat{u}_i &= 0\end{aligned}$$

que son dos propiedades del estimador de mínimos cuadrados:

1) la suma de los residuos que genera el estimador de mínimos cuadrados es igual a cero, lo que no necesariamente ocurre con otro procedimiento de estimación [ver suma de la columna 7 del Cuadro 1], y

2) los residuos de mínimos cuadrados están incorrelacionados con la variable explicativa del modelo. Cuando se considera un modelo de regresión lineal general o múltiple, que incluye no una, sino k variables explicativas, los residuos de mínimos cuadrados están incorrelacionados con todas las variables explicativas del modelo [ver suma de la columna 8 del Cuadro 1].

4.4 Esperanza matemática

La expresión del estimador MCO de la pendiente del modelo de regresión lineal simple puede escribirse:

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)y_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right)y_i=\sum_{i=1}^{n}\alpha_i y_i \qquad (17)$$

como una combinación lineal ponderada de las observaciones de la variable endógena, con ponderaciones:

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

En esta cadena de igualdades hemos utilizado el hecho de que la suma de las desviaciones de una variable con respecto a su media muestral, es siempre igual a cero. Las ponderaciones en esta expresión suman cero:

$$\sum_{i=1}^{n}\alpha_i=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right)=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=\frac{0}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=0$$

Además:

$$\begin{aligned}\sum_{i=1}^{n}\alpha_i x_i &= \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right)x_i=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)x_i}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=\frac{\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}\left(x_i^2-2\bar{x}x_i+\bar{x}^2\right)}=\\ &= \frac{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}{\sum_{i=1}^{n}x_i^2-2\bar{x}\left(n\bar{x}\right)+n\bar{x}^2}=\frac{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}=1\end{aligned}$$

Y la suma de los cuadrados de las ponderaciones:

$$\sum_{i=1}^{n}\alpha_i^2=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right)^2=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]^2}=\frac{1}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}=\frac{1}{nS_x^2}$$

Recordemos que estamos suponiendo que los valores x_1,x_2,... tomados por la variable X son fijos, es decir, no están sujetos a ninguna incertidumbre, de modo que, si volviésemos a tomar otra muestra de igual tamaño, tendríamos para dicha variable las mismas observaciones numéricas, una por una, que las que ya disponemos. Tan sólo las observaciones y_1, y_2,... de la variable endógena Y diferirían de las actualmente disponibles, debido a que las realizaciones muestrales de la perturbación aleatoria u_i, el único componente aleatorio de Y, serían diferentes de las actuales. Vamos a utilizar ahora repetidamente el carácter determinista no aleatorio, de la variable X.

Si sustituimos en (17) y_i por su expresión a través del modelo de regresión, tenemos:

$$\begin{aligned}\hat{\beta}_1 &= \sum_{i=1}^{n}\alpha_i\left(\beta_0+\beta_1x_i+u_i\right)=\sum_{i=1}^{n}\alpha_i\beta_0+\sum_{i=1}^{n}\alpha_i\beta_1x_i+\sum_{i=1}^{n}\alpha_iu_i= \qquad (18)\\ &= \beta_0\sum_{i=1}^{n}\alpha_i+\beta_1\sum_{i=1}^{n}\alpha_ix_i+\sum_{i=1}^{n}\alpha_iu_i=\beta_0.0+\beta_1.1+\sum_{i=1}^{n}\alpha_iu_i=\\ &= \beta_1+\sum_{i=1}^{n}\alpha_iu_i\end{aligned}$$

donde hemos utilizado las dos propiedades antes demostradas. Esta es una representación muy útil, que presenta el estimador de mínimos cuadrados de la pendiente como una combinación lineal de las perturbaciones del modelo, con coeficientes α_i, más una constante desconocida, el verdadero valor de dicha pendiente. Los coeficientes α_i en dicha combinación lineal varían de una muestra a otra con los valores de la variable explicativa, X, por lo que el valor numérico del estimador de mínimos cuadrados también variaría si dispusiéramos de distintas muestras recogidas en distintos períodos de tiempo, por ejemplo.

Es importante recordar que suponemos que la variable explicativa es determinista. Es decir, que los valores numéricos observados en la muestra para dicha variable son los únicos posibles, dadas las unidades de observación muestral, sean individuos, empresas, familias, o un conjunto de observaciones de detemrinada frecuencia (diaria, mensual, trimestral anual) a lo largo de un determinado intervalo de tiempo. Recordemos que de una muestra a otra, cambiarían los valores observados de la variable dependiente, y_i porque cambiaría la realización numérica de las perturbaciones u_i, pero no porque cambiasen los valores de la variable explicativa x_i, que serían los mismos entre distintas muestras extraídas de las mismas unidades de observación.

A continuación, vamos a obtener la esperanza matemática y la varianza de los estimadores de mínimos cuadrados de $\hat{\beta}_0$ y $\hat{\beta}_1$. Esto es necesario para poder proceder a contrastar hipótesis acerca de sus verdaderos valores que, recordemos, son desconocidos. Disponemos de una estimación numérica, obtenida con la muestra disponible, que sería diferente si pudiésemos calcularla con otra muestra distinta.

4.4.1 Ausencia de sesgo del estimador de mínimos cuadrados

Tomando esperanzas, y notando que:

$$E\left(\alpha_i u_i\right) = \alpha_i 0 = 0$$

tenemos:

$$E\left(\hat{\beta}_1\right) = \beta_1 + E\left(\sum_{i=1}^{n} \alpha_i u_i\right) = \beta_1 + \sum_{i=1}^{n} E\left(\alpha_i u_i\right) = \beta_1 + \sum_{i=1}^{n} \alpha_i E\left(u_i\right) = \beta_1$$

lo que prueba que el estimador MCO del parámetro β_1 es *insesgado*, puesto que su esperanza matemática coincide con el verdadero valor del parámetro que se pretende estimar, que es desconocido.

Por tanto, el supuesto $E(u_i) = 0, i = 1, 2, ..., N$ es suficiente para garantizar la ausencia de sesgo del estimador de mínimos cuadrados de la pendiente: $\hat{\beta}_1$. Notemos que el supuesto de que la variable explicativa no es aleatoria es crucial para probar la ausencia de sesgo del estimador de mínimos cuadrados. En las expresiones anteriores nos hemos encontrado con $E\left(\alpha_i u_i\right)$, y cada α_i depende de todas las observaciones $x_j, j = 1, 2, ..., n$. Si fuese aleatoria, no sabríamos decir nada acerca de la esperanza matemática $E(x_i u_i)$, salvo haciendo supuestos específicos acerca de la covarianza entre ambas variables aleatorias, x_i y u_i, pero mucho menos acerca de la esperanza $E\left(\alpha_i u_i\right)$.

Recordando que la expresión del estimador MCO del término independiente β_0 es:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

notemos que:

$$E\left(\bar{y}\right) = \beta_0 + E\left(\hat{\beta}_1 \bar{x}\right) + E\left(\bar{u}\right) = \beta_0 + \beta_1 \bar{x}$$

por lo que:

$$E\left(\hat{\beta}_0\right) = E(\bar{y}) - E(\hat{\beta}_1 \bar{x}) = (\beta_0 + \beta_1 \bar{x}) - E(\hat{\beta}_1).\bar{x} = (\beta_0 + \beta_1 \bar{x}) - \beta_1 \bar{x} = \beta_0$$

de modo que, al igual que ocurría con la estimación de β_1, el estimador MCO de β_0 es también *insesgado.*

La recta de regresión estimada pasa por el punto $(\bar{x}, \bar{y})$. Es decir, el valor numérico que la recta de regresión estimada asocia a la variable dependiente Y cuando $X = \bar{x}$ es, precisamente, $Y = \bar{y}$. En efecto:

$$y = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \left(\bar{y} - \hat{\beta}_1 \bar{x}\right) + \hat{\beta}_1 \bar{x} = \bar{y}$$

El punto $(\bar{x}, \bar{y})$ se conoce en ocasiones como el *centro de gravedad* de la nube de puntos $(x_i, y_i), i = 1, 2, ..., N$.

4.5 Matriz de covarianzas

Todo estimador puntual debe ir siempre acompañado de una medida de dispersión del mismo, generalmente su varianza, de modo que podamos juzgar el grado en que se aproxima al verdadero valor del parámetro que pretendemos estimar. Pero además, para poder llevar a cabo un análisis de inferencia estadística, es decir, para poder contrastar si alguno de los coeficientes β_0 ó β_1, o ambos, toman determinados valores teóricos, es preciso disponer de desviaciones típicas de sus estimaciones.

Estos no son sino un caso particular de los problemas de estimación e inferencia estadísticos, y los resolvemos de modo similar, mediante la construcción de intervalos de confianza, al nivel deseado, alrededor del valor hipotético que se pretende contrastar.

4.5.1 Varianza del estimador de mínimos cuadrados de la pendiente del modelo de regresión lineal simple

Recordemos el supuesto de que las perturbaciones aleatorias del modelo correspondientes a todas las unidades muestrales tienen la misma varianza, σ_u^2. Por tanto, si partimos de la expresión (18) que antes obtuvimos para el estimador de β_1, tenemos:

$$Var(\alpha_i u_i) = \alpha_i^2 Var(u_i) = \alpha_i^2 \sigma_u^2$$

para cualquier $i = 1, 2, \ldots$. Entonces, puesto que la covarianza entre u_i y u_j es igual a cero, se tiene:

$$\begin{aligned} E\left(\sum_{i=1}^{n} \alpha_i u_i\right) &= \sum_{i=1}^{n} E\left(\alpha_i u_i\right) = \sum_{i=1}^{n} \alpha_i E\left(u_i\right) = \sum_{i=1}^{n} \alpha_i 0 = 0 \\ Var\left(\sum_{i=1}^{n} \alpha_i u_i\right) &= \sum_{i=1}^{n} Var(\alpha_i u_i) = \sum_{i=1}^{n} \alpha_i^2 Var(u_i) = \sigma_u^2 \left(\sum_{i=1}^{n} \alpha_i^2\right) = \\ &= \frac{1}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} = \frac{1}{n S_x^2} \end{aligned}$$

Como el estimador $\hat{\beta}_1$ es la suma de una constante (el verdadero valor β_1) y una variable aleatoria (la suma ponderada de las perturbaciones) [ver (18)], la varianza de $\hat{\beta}_1$ será igual tan sólo a la varianza de esta última suma:

$$Var\left(\hat{\beta}_1\right) = Var\left(\sum_{i=1}^{n} \alpha_i u_i\right) = \sigma_u^2 \left(\sum_{i=1}^{n} \alpha_i^2\right) = \frac{\sigma_u^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} = \frac{\sigma_u^2}{n S_x^2}$$

Resaltemos que esta expresión es válida bajo los supuestos de que todos los términos de error tienen la misma varainza (es decir, que la varianza es constante a lo largo de la muestra), y de independencia entre dos cualesquiera de los términos de error.

4.6 Estimación de la varianza del término de error o perturbación aleatoria del modelo

Conociendo las expresiones analíticas de las varianzas de ambos estimadores, así como también de su covarianza, podremos contrastar hipótesis acerca de valores teóricos para alguno de los dos coeficientes, y también contrastar hipótesis conjuntas, acerca de ambos simultáneamente. Pero en ellas aparece la varianza del término de error σ_u^2, que es desconocida. Debemos, por tanto, estimar este parámetro, y utilizar su estimación en lugar de su verdadero valor, que es desconocido.

Por similitud, parece razonable utilizar la varianza muestral de los residuos como un estimador de la varianza poblacional σ_u^2. Los residuos de mínimos cuadrados tienen media cero, como muestra

la primera ecuación normal, por lo que su varianza muestral es: $S_{\hat{u}}^2 = \frac{1}{n}\sum_{i=1}^n \hat{u}_i^2 = SCR/n$. Pero estimamos con una pequeña corrección:

$$\hat{\sigma}_u^2 = \frac{1}{n-2}\sum_{i=1}^n \hat{u}_i^2 = \frac{n}{n-2}S_{\hat{u}}^2$$

Tomamos $n-2$ y no simplemente n en el denominador, pero que el estimador $\hat{\sigma}_u^2$ sea insesgado [ver Apéndice]. Una vez que se dispone de una estimación de la varianza, puede utilizarse en las expresiones de la varianza de los estimadores de los coeficientes, de manera que tenemos así estimaciones de las varianzas de los coeficientes estimados, lo que indicaremos con un circunflejo encima de la palabra "Varianza".

Ejemplo.- Con los datos del Cuadro 1, tenemos una Suma Residual, es decir, una suma de cuadrados de residuos, de 80,2. Ello nos lleva a la estimación de la varianza del término de error:

$$\begin{aligned}
\hat{\sigma}_u^2 &= \frac{1}{n-2}\sum_{i=1}^n \hat{u}_i^2 = \frac{80,2}{16-2} = 5,729 \Rightarrow \hat{\sigma}_u = 2,393 \\
R^2 &= 1 - \frac{Suma\ Cuadrados\,\mathrm{Re}siduos}{Suma\,Total} = 1 - \frac{5,014}{11,715} = 1 - 0,428 = 0,572
\end{aligned}$$

Podemos utilizar ahora la estimación de σ_u^2 en las expresiones de las varianzas de los estimadores de Mínimos Cuadrados que aparecen en el Apéndice:

$$\begin{aligned}
Var\left(\hat{\beta}_1\right) &= \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{5,729}{167,9} = 0,03417 \Rightarrow DT\left(\hat{\beta}_1\right) = 0,185 \\
Var\left(\hat{\beta}_0\right) &= \sigma_u^2 \frac{\sum_{i=1}^n {x_i}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{5,729}{16}\frac{1911}{167,9} = 4,075 \Rightarrow DT\left(\hat{\beta}_2\right) = 2,02 \\
Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) &= -\frac{\bar{x}\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = -\frac{5,729}{167,9}(10,4) = -0,354
\end{aligned}$$

Finalmente, el modelo estimado se representa escribiéndolo como la función lineal que es, anotando debajo de los coeficientes estimados sus desviaciones típicas que son, asimismo, estimadas, como acabamos de ver, pues sus verdaderos valores dependen de σ_u^2:

$$y_i = \underset{(2,02)}{4,35} + \underset{(0,185)}{0.799} x_i + u_i, \quad R^2 = 0,572;\ \hat{\sigma}_u = 2,393$$

4.7 El modelo de regresión lineal en desviaciones respecto de la media

Como hemos visto en la sección anterior, a partir del modelo de regresión lineal:

$$y_i = \beta_0 + \beta_1 x_i + u_i,\ i = 1, 2, 3, ..., n$$

se deduce que:

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$$

y, restando, tenemos un modelo en el que cada variable aparece en desviaciones respecto de su media muestral:

$$y_i - \bar{y} = \beta_1 (x_i - \bar{x}) + (u_i - \bar{u}), \ i = 1, 2, 3, ..., n$$

Nótese que la primera y tercera ecuaciones son válidas para cada observación muestral y tenemos, en cada una de ellas, tantas relaciones como observaciones muestrales, mientras que la segunda ecuación aplica sólo a las medias muestrales y constituye, por tanto, una única relación.

En el modelo en desviaciones no hay término independiente, y el término de error es distinto del término de error del modelo original.

Si estimamos este modelo en diferencias por mínimos cuadrados, tendremos el mismo estimador de β_1 que en el modelo original, ya que:

$$\begin{aligned} Var(x_i - \bar{x}) &= Var(x_i) \\ Cov[(x_i - \bar{x}), (y_i - \bar{y})] &= Cov(x_i, y_i) \end{aligned}$$

Aunque no habremos estimado β_0, puesto que dicho parámetro ha desaparecido del modelo, podemos utilizar la relación que obtuvimos antes para calcular $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

La varianza del término de error del modelo en diferencias es ligeramente distinta del modelo original, puesto que:

$$\begin{aligned} Var(u_i - \bar{u}) &= E[u_i (u_i - \bar{u})] = E(u_i^2) - E(u_i \bar{u}) = \\ &= E(u_i^2) - E\left(u_i \sum_{i=1}^{n} \frac{u_i}{n}\right) = \sigma_u^2 - \frac{1}{n}\sigma_u^2 = \frac{n-1}{n}\sigma_u^2 \end{aligned}$$

Los residuos del modelo estimado con las variables en desviaciones respecto de la media son:

$$\hat{v}_i = (y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) = y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 x_i$$

y, por tanto, coinciden numéricamente, con los que se obtienen estimando el modelo con las variables originales.

4.8 El modelo constante

Consideremos un modelo muy sencillo:

$$y_i = \beta_0 + u_i,$$

en el que aparece una constante como única variable explicativa, por lo que se denoomina modelo constante de regresión. El estimador MCO será el estadístico muestral que minimice la suma de los residuos, que en este caso es:

$$SCR = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \beta_0)^2,$$

por lo que se trata de minimizar la suma de las desviaciones al cuadrado entre los valores muestrales de la variable Y y un estadístico. La solución a dicho problema de minimización está dada

por la media muestral, y el valor minimizado es, por tanto, la varianza muestral. En consecuencia, el estimador del modelo constante de regresión es la media muestral. Ello significa que la media muestral es el estimador óptimo, cuando no se dispone de información acerca de ninguna otra variable. En tal situación, lo mejor que podemos hacer es aproximar cada valor potencialmente observable de la variable Y por la media muestral de que dispongamos. Es, desde luego, un estimador algo pobre, pero nos sirve de referencia a la que hay que mejorar; es decir, contando con información muestral acerca de alguna otra variable, hemos de conseguir estimaciones MCO de un modelo de regresión tales que la Suma de Cuadrados de Residuos que generan sea inferior a la varianza muestral de Y. Pero ello va a ocurrir siempre. Cuando se estima el modelo constante, la Suma de Cuadrados de Residuos, que es la varianza de Y, coincide con la Suma Total, por lo que el coeficiente de determinación es igual a cero. Ningún otro modelo tendrá un coeficiente de determinación inferior.

4.9 Eficiencia

En el modelo de regresión, la aleatoriedad proviene del término de error, de quien suponemos que tiene esperanza matemática nula y varianza σ_u^2. La aleatoriedad se transmite a la variable y_i, que tiene esperanza $E(y_i) = \beta_0 + \beta_1 x_i$ y varianza σ_u^2, igual a la de u_i, de quien se diferencia en una constante, $\beta_0 + \beta_1 x_i$. Por otra parte, (17) muestra que el estimador MCO de β_1 depende linealmente de las observaciones de la variable aleatoria Y. También $\hat{\beta}_0$ es una combinación lineal de las observaciones y_i:

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = \bar{y} - \left(\beta_1 + \sum_{i=1}^{n} \alpha_i u_i \right) \bar{x} = \bar{y} - \beta_1 \bar{x} - \bar{x} \left(\sum_{i=1}^{n} \alpha_i (y_i - \beta_0 - \beta_1 x_i \right) = \\
&= \bar{y} - \beta_1 \bar{x} - \bar{x} \sum_{i=1}^{n} \alpha_i y_i + \beta_0 \bar{x} \sum_{i=1}^{n} \alpha_i + \beta_1 \bar{x} \sum_{i=1}^{n} \alpha_i x_i = \bar{y} - \beta_1 \bar{x} - \bar{x} \sum_{i=1}^{n} \alpha_i y_i + \beta_0 \bar{x}.0 + \beta_1 \bar{x}.1 = \\
&= \bar{y} - \bar{x} \sum_{i=1}^{n} \alpha_i y_i = \frac{1}{n} \sum_{i=1}^{n} y_i - \bar{x} \sum_{i=1}^{n} \alpha_i y_i = \sum_{i=1}^{n} \left(\frac{1}{n} - \bar{x} \alpha_i \right) y_i
\end{aligned}$$

Pues bien, el estimador MCO es de mínima varianza dentro de la clase de estimadores lineales:

Theorem 2 *(Teorema de Gauss-Markov).- Bajo los supuestos del modelo llineal de regresión, el estimador MCO es el estimador lineal insesgado de mínima varianza de los coeficientes del modelo de regresión.*

Proof. Consideremos un estimador lineal de la pendiente del modelo de regresión:

$$\tilde{\beta}_1 = \sum_{i=1}^{n} c_i y_i$$

que supondremos distinto del estimador de mínimos cuadrados, es decir, que no todas las constantes c_i son iguales a las α_i. Para que este estimador sea insesgado ha de cumplirse:

$$E\left(\tilde{\beta}_1\right) = E\left(\sum_{i=1}^{n} c_i y_i\right) = E\left(\sum_{i=1}^{n} c_i(\beta_0 + \beta_1 x_i + u_i)\right) = E\left(\beta_0 \sum_{i=1}^{n} c_i\right) + \beta_1 E \sum_{i=1}^{n} c_i x_i + E \sum_{i=1}^{n} c_i u_i =$$
$$= \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} x_i + 0$$

que será igual a β_1 y, con ello, el estimador $\tilde{\beta}_1$ será insesgado sólo si se cumple, simultáneamente:

$$\sum_{i=1}^{n} c_i = 0$$
$$\sum_{i=1}^{n} c_i x_i = 1$$

Suponemos, por tanto, que las constantes c_i satisfacen ambas condiciones. Teniendo en cuenta que tanto $\sum_{i=1}^{n} c_i$ como $\sum_{i=1}^{n} c_i x_i$ son constantes, la varianza de este estimador es:

$$Var\left(\tilde{\beta}_1\right) = Var\left(\sum_{i=1}^{n} c_i u_i\right) = \sum_{i=1}^{n} Var\left(c_i u_i\right) = \sigma_u^2 \sum_{i=1}^{n} c_i^2$$

de modo que, para probar que el estimador de mínimos cuadrados tiene menor varianza que este estimador lineal insesgado genérico, habremos de probar que:

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

con independencia de cuáles sean las constantes $c_i, i = 1, 2, ..., n$.

Para ello, consideremos la expresión:

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

$$= \sum_{i=1}^{n} c_i^2 - 2\frac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} + \sum_{i=1}^{n} \frac{\left(x_i - \bar{x}\right)^2}{\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right]^2} = \sum_{i=1}^{n} c_i^2 - 2\frac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} + \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right]^2} =$$
$$= \sum_{i=1}^{n} c_i^2 - \frac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \geq 0$$

donde la última desigualdad proviene del hecho de que el punto de partida es una suma de cuadrados y por tanto, necesariamente positiva.

Pero esto significa que, como queríamos mostrar:

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

■

El teorema de Gauss-Markov es importante, por cuanto que afirma que la matriz de covarianzas del estimador MCO es inferior a la de cualquier otro estimador lineal e insesgado. Es decir, la diferencia entre ambas matrices, en el orden citado, es semidefinida negativa. Ello tiene implicaciones más útiles: la varianza del estimador MCO de β_0 es inferior a la varianza de cualquier otro estimador lineal e insesgado de dicho coeficiente, y lo mismo ocurre con la varianza del estimador MCO de β_1.

Cuando el término de error del modelo tiene una distribución Normal, tenemos un resultado aún más importante, que afirma que el estimador MCO es eficiente, es decir, tiene la menor varianza posible (la menor matriz de covarianzas), dentro de la clase de los estimadores insesgados, sean estos lineales o no.

Theorem 3 *Teorema de Rao.- Si se cumplen las condiciones de la Sección 13.1 y, además, el término de error del modelo tiene distribución Normal, entonces el estimador MCO es el estimador insesgado de mínima varianza de los coeficientes del modelo de regresión.*

La demostración se basa en probar que, cuando el término de error del modelo de regresión tiene distribución Normal, $u_i \sim N(0, \sigma_u^2)$, entonces el estimador de Mínimos Cuadrados coincide con el estimador de Máxima Verosimilitud. Como este último es siempre (bajo condiciones muy generales y, por tanto, fáciles de satisfacer) el estimador de mínima varianza o eficiente, habremos probado que, en este caso especial, el estimador de mínimos cuadrados también lo es.

Consideremos el modelo de regresión con término de error Normal:

$$y_i = \beta_0 + \beta_1 x_i + u_i \;\; u_i \sim N(0, \sigma_u^2)$$

del que escribimos la función de verosimilitud:

$$L(\beta_0, \beta_1, \sigma_u^2 / y_1, x_1, y_2, x_2, ..., y_n, x_n) = \prod_{i=1}^{n} \frac{1}{\sigma_u \sqrt{2\pi}} e^{-u_i^2/2\sigma_u^2}$$

y su logaritmo:

$$\begin{aligned} \ln L(\beta_0, \beta_1, \sigma_u^2 / y_1, x_1, ..., y_n, x_n) &= -\frac{n}{2} \ln \sigma_u^2 - \frac{n}{2} \ln (2\pi) - \sum_{i=1}^{n} \frac{u_i^2}{2\sigma_u^2} = \\ &= -\frac{n}{2} \ln \sigma_u^2 - \frac{n}{2} \ln (2\pi) - \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma_u^2} \end{aligned}$$

El estimador de Máxima Verosimilitud se obtiene derivando en la expresión anterior con respecto a los parámetros desconocidos: β_0, β_1, σ_u^2, e igualando a cero dichas derivadas.

Pero sin necesidad siquiera de hacer dicho cálculo, ya apreciamos que los valores numéricos de β_0 y $\beta_1$que maximizan ln$L$ son los mismos que minimizan la Suma de Cuadrados de los Residuos, ya que ésta entra con signo menos en la expresión de $\ln L$. Por tanto, los estimadores de Mínimos Cuadrados y de Máxima Verosimilitud de ambos parámetros coinciden, y el teorema quda probado.

Este resultado es importante, porque justifica el uso del estimador de Mínimos Cuadrados, dado que es un estimador eficiente. Pero, como con cualquier teormea, es preciso entender el conjunto de ocndiciones bajo las que puede afirmarse la conclusión que se ha obtenido. En nuestro caso, es

especilamente importante recordar que la eficiencia del estimador de Mínimos Cuadrados se obtiene si el término de error del modelo sigue una distribución Normal, pero no necesariamente en otro caso.

El estimador de Máxima verosimilitud de la varianza del término de error es:

$$\hat{\sigma}^2_{MV} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$$

que es parecido, aunque no idéntico, al estimador MCO de dicho parámetro. De hecho, como sabemos [ver Apéndice] que el estimador MCO de σ_u^2 es insesgado, podemos asegurar que el estimador de máxima verosimilitud es sesgado:

$$E\left(\hat{\sigma}^2_{MV}\right) = E\left(\frac{n-2}{n}\hat{\sigma}^2_{MCO}\right) = \frac{n-2}{n}E\left(\hat{\sigma}^2_{MCO}\right) = \frac{n-2}{n}\sigma_u^2$$

Sin embargo, su sesgo desaparece al aumentar el tamaño muestral por cuanto que el factor $(n-2)/n$ tiende a uno. El estimador MV de la varianza es, por tanto, asintóticamente insesgado.

4.10 Cambios de escala y de origen

4.10.1 Cambios de escala

En ocasiones, es conveniente multiplicar o dividir una variable por una constante. Por ejemplo, esto sucede cuando los valores numericos de una variable son muy elevados, por estar dados en euros, y para facilitar su lectura preferimos utilizar la variable en millones de euros, lo que equivale a dividir sus datos por 1.000.000. En otras ocasiones, podemos estar interesados en multiplicar todos los valores de una variable por una misma constante, por ejemplo, 100. Aunque habitualmente esto afectará a una de las variables del modelo, puede suceder simultaneamente tanto con la variable dependiente como con alguna de las variables explicativas. En el caso de un modelo de regresión simple, se trataría de comparar el modelo:

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{u}_i$$

en el que aparecen las variables originales, con:

$$y_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* x_i + \hat{u}_i^*$$

donde: $y_i^* = \lambda y_i, x_i^* = \mu x_i, i = 1, 2, ..., N$, siendo λ y μ constantes conocidas. Si solo la variable x cambia, estaríamos en una situación con $\lambda = 1$, mientras que si es la variable y la que cambia, entonces $\mu = 1$. Si lo que hacemos es pasar los valores de x de euros a millones de euros, entonces $\lambda = 1$ y $\mu = 1/1000000$. Como se ve, consideramos la posibilidad de que tanto la constante como la pendiente del modelo se vean afectados por este cambio de escala. También los residuos pueden variar y, con ellos, el R^2, la Suma de cuadrados de residuos y la varianza residual, por lo que también el valor numérico de los estadísticos de contraste t y F podrían verse afectados.

Es sencillo analizar los posibles efectos de un cambio de escala, pues todo se basa en propiedades conocidas de la covarianza y varianza. En este caso,

$$
\begin{aligned}
E(x^*) &= \mu E(x);\ E(y^*) = \lambda E(y) \\
Cov(x^*, y^*) &= E\left[(x^* - \bar{x}^*)\,(y^* - \bar{y}^*)\right] = E\left[(\mu x - \mu\bar{x})\,(\lambda y - \lambda\bar{y})\right] = \\
&= E\left[\mu\,(x - \bar{x})\,\lambda\,(y - \bar{y})\right] = \mu\lambda E\left[(x - \bar{x})\,(y - \bar{y})\right] = \mu\lambda Cov(x, y) \\
Var(x^*) &= E\left[(x^* - \bar{x}^*)^2\right] = E\left[(\mu x - \mu\bar{x})^2\right] = E\left[\mu^2\,(x - \bar{x})^2\right] = \mu^2 Var(x)
\end{aligned}
$$

Por tanto,

$$
\begin{aligned}
\hat{\beta}_2^* &= \frac{Cov(x^*, y^*)}{Var(x^*)} = \frac{\mu\lambda Cov(x, y)}{\mu^2 Var(x)} = \frac{\lambda}{\mu}\hat{\beta}_2 \\
\hat{\beta}_1^* &= \bar{y}^* - \hat{\beta}_2^*\bar{x}^* = \lambda\bar{y} - \frac{\lambda}{\mu}\hat{\beta}_2(\mu\bar{x}) = \lambda\left(\bar{y} - \hat{\beta}_2\bar{x}\right) = \lambda\hat{\beta}_1
\end{aligned}
$$

es decir, que la estimación de la pendiente se ve afectada por ambas constantes, mientras que la estimación de la constante sólo se ve afectada por un posible cambio de escala en la variable dependiente. Por ejemplo, si dividimos los datos de x por 1000000, es decir, $\mu = 1/1000000$, y no alteramos los datos de la variable dependiente, la estimación de la pendiente queda multiplicada por 1.000.000, mientras que la constante no cambiará.

También tendremos:

$$
\begin{aligned}
\hat{u}_i^* &= y_i^* - \hat{\beta}_1^* - \hat{\beta}_2^* x_i^* = \lambda y_i - \lambda\hat{\beta}_1 - \frac{\lambda}{\mu}\hat{\beta}_2(\mu x_i) = \lambda y_i - \lambda\hat{\beta}_1 - \lambda\hat{\beta}_2 x_i = \lambda\hat{u}_i \\
SCR^* &= \sum_{i=1}^{N}(\hat{u}_i^*)^2 = \sum_{i=1}^{N}(\lambda\hat{u}_i)^2 = \sum_{i=1}^{N}\lambda^2\hat{u}_i^2 = \lambda^2\sum_{i=1}^{N}\hat{u}_i^2 = \lambda^2 SCR \\
\hat{\sigma}_u^{2^*} &= \frac{SCR^*}{T-k} = \lambda^2\frac{SCR}{T-k} = \lambda^2\hat{\sigma}_u^2 \\
ST^* &= \sum_{i=1}^{N}(y_i^* - \bar{y}^*)^2 = \sum_{i=1}^{N}(\lambda y_i - \lambda\bar{y})^2 = \sum_{i=1}^{N}\left[\lambda\,(y_i - \bar{y})\right]^2 = \sum_{i=1}^{N}\lambda^2\,(y_i - \bar{y})^2 = \lambda^2 ST \\
R^{*2} &= 1 - \frac{SCR^*}{ST^*} = 1 - \frac{\lambda^2 SCR}{\lambda^2 ST} = 1 - \frac{SCR}{ST} = R^2
\end{aligned}
$$

de modo que el ajuste del modelo no se ve afectado por posibles cambios de escala ni en la variable dependiente ni en las variables explicativas. Como hemos visto, los coeficientes en las variables explicativas y la constante del modelos se modifican con los posibles cambios de escala, pero el grado del ajuste del modelo no se ve afectado por los posibles cambios de escala.

¿Cómo se vería afectado el estadístico t para le contraste de significación estadística de la pendiente del modelo? En el contraste de significación, el valor teórico β_2^0 que estamos contrastando para la pendiente es cero, por lo que el estadístico t es igual a:

$$
t^* = \frac{\hat{\beta}_2^* - \beta_2^0}{DT(\hat{\beta}_2^*)} = \frac{\hat{\beta}_2^*}{DT(\hat{\beta}_2^*)} = \frac{\frac{\lambda}{\mu}\hat{\beta}_2}{DT(\frac{\lambda}{\mu}\hat{\beta}_2)} = \frac{\hat{\beta}_2}{DT(\hat{\beta}_2)} = t
$$

por lo que el estadístico t tampoco varía.

4.10.2 Cambios de origen

Supongamos que el modelo transformado es:

$$y_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* x_i + \hat{u}_i^*$$

donde las variables transformadas vienen dadas por: $y_i^* = y_i.\lambda, x_i^* = x_i - \mu, i = 1, 2, ..., N$, siendo λ y μ constantes conocidas.

4.11 Apéndice: Varianza del estimador de mínimos cuadrados de la constante del modelo de regresión lineal simple

Para obtener la varianza del estimador MCO de $\hat{\beta}_0$, notemos que:

$$Var\left(\hat{\beta}_0\right) = Var(\bar{y}) + Var(\hat{\beta}_1\bar{x}) - 2Cov(\bar{y}, \hat{\beta}_1\bar{x}) = Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1)$$

donde aparece la varianza de la media muestral de la variable endógena, que podemos calcular, del siguiente modo: si sumamos la expresión (1) del modelo lineal simple para todas las observaciones muestrales, tenemos:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) + \sum_{i=1}^{n} u_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} u_i$$

y, dividimos por el tamaño muestral, n:

$$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$$

lo que puede utilizarse para probar que:

$$\begin{aligned} Var(\bar{y}) &= Var(\beta_0 + \beta_1\bar{x} + \bar{u}) = Var(\beta_0) + Var(\beta_1\bar{x}) + Var(\bar{u}) = 0 + 0 + \frac{\sigma_u^2}{n} = \frac{\sigma_u^2}{n} \\ Cov\left(\bar{y}, \hat{\beta}_1\right) &= Cov\left(\bar{y}, \beta_1 + \sum_{i=1}^{n} \alpha_i u_i\right) = Cov\left(\bar{y}, \beta_1\right) + \sum_{i=1}^{n} \alpha_i Cov\left(\bar{y}, u_i\right) = 0 + \frac{1}{n}\sigma_u^2 \sum_{i=1}^{n} \alpha_i = 0 \end{aligned}$$

por lo que tenemos:

$$\begin{aligned} Var\left(\hat{\beta}_0\right) &= Var(\bar{y}) + \bar{x}^2 Var(\hat{\beta}_1) - 2\bar{x}Cov(\bar{y}, \hat{\beta}_1) = \frac{\sigma_u^2}{n} + \bar{x}^2 \frac{\sigma_u^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \\ &= \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right) = \sigma_u^2 \frac{\sum_{i=1}^{n} {x_i}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{aligned}$$

4.11.1 Covarianza entre los estimadores de mínimos cuadrados de la constante y la pendiente del modelo de regresión lineal simple

Siguiendo un argumento similar, tenemos:

$$\begin{aligned} Cov\left(\bar{y}, u_i\right) &= Cov\left(\beta_0 + \beta_1\bar{x} + \bar{u}, u_i\right) = Cov\left(\beta_0, u_i\right) + \bar{x}Cov\left(\beta_1, u_i\right) + Cov\left(\bar{u}, u_i\right) = \\ &= 0 + 0 + \frac{1}{n}\sum_{j=1}^{n} Cov(u_j, u_i) = \frac{1}{n}\sigma_u^2 \end{aligned}$$

por lo que:

$$\begin{aligned} Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) &= Cov\left(\bar{y} - \hat{\beta}_1\bar{x}, \hat{\beta}_1\right) = \bar{x}Cov\left(\bar{y}, \hat{\beta}_1\right) - \bar{x}Var\left(\hat{\beta}_1\right) = \\ &= 0 - \bar{x}\frac{\sigma_u^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} = -\frac{\bar{x}\sigma_u^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \end{aligned}$$

que indica, entre otras cosas, que el signo de la covarianza entre $\hat{\beta}_0$ y $\hat{\beta}_1$ es el opuesto al signo de la media muestral de la variable X.

Supongamos que dicha media fuese positiva, y también que el error de estimación de β_1 fuese asimismo positivo, es decir, que hubiésemos estimado (sin saberlo), un valor $\hat{\beta}_1$ superior al teórico. Su producto por la media de X generaría, en promedio, una contribución positiva del error de estimación a la explicación de la variable Y :

$$\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{u} = \left[\beta_0 + \beta_1\bar{x}\right] + \left[\left(\hat{\beta}_0 - \beta_0\right) + \left(\hat{\beta}_1 - \beta_1\right)\bar{x}\right]$$

donde en el corchete de la derecha, el segundo sumando está teniendo una contribución positiva. Para compensarlo, la estimación MCO de β_0 estaría por debajo de su valor verdadero: $\beta_0 > \hat{\beta}_0$. Es decir, si el estimador de Mínimos Cuadrados sobreestima β_1, entonces infraestima β_1. Si infraestimamos β_1, entonces sobreestimamos β_0. Lo contrario ocurriría si la media muestral de X fuese negativa.

4.11.2 Argumento alternativo

$$\begin{aligned}
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} = (\beta_0 + \beta_1\bar{x} + \bar{u}) - \hat{\beta}_1\bar{x} = \beta_0 + \left(\beta_1 - \hat{\beta}_1\right)\bar{x} + \bar{u} \\
\hat{\beta}_0 - E\hat{\beta}_0 &= \left(\beta_1 - \hat{\beta}_1\right)\bar{x} + \bar{u} \\
Var\left(\hat{\beta}_0\right) &= E\left[\left(\hat{\beta}_1 - E\hat{\beta}_1\right)\bar{x} + \bar{u}\right]^2 = E\left[\left(\beta_1 - \hat{\beta}_1\right)^2\bar{x}^2\right] + E\left(\bar{u}^2\right) + 2E\left[\left(\hat{\beta}_1 - E\hat{\beta}_1\right)\bar{x}\bar{u}\right] = \\
&= \bar{x}^2 Var\left(\hat{\beta}_1\right) + \frac{\sigma_u^2}{n} - 2\bar{x}E\left[\left(\hat{\beta}_1 - E\hat{\beta}_1\right)\bar{u}\right] \\
\text{Pero} \quad &: \quad E\left[\left(\hat{\beta}_1 - E\hat{\beta}_1\right)\bar{u}\right] = E\left[\left(\sum_{i=1}^{n}\alpha_i u_i\right)\left(\frac{1}{n}\sum_{j=1}^{n}u_j\right)\right] = \frac{\sigma_u^2}{n}\sum_{i=1}^{n}\alpha_i = 0 \\
\text{Luego:} \quad & Var\left(\hat{\beta}_0\right) = \bar{x}^2 Var\left(\hat{\beta}_1\right) + \frac{\sigma_u^2}{n} = \sigma_u^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right) = \sigma_u^2\frac{\sum_{i=1}^{n}{x_i}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \\
Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) &= E\left[\left(\hat{\beta}_0 - \beta_0\right)\left(\hat{\beta}_1 - \beta_1\right)\right] = E\left[\left(\bar{u} - \left(\hat{\beta}_1 - \beta_1\right)\bar{x}\right)\left(\hat{\beta}_1 - \beta_1\right)\right] = \\
&= E\left[\bar{u}\left(\hat{\beta}_1 - \beta_1\right)\right] - \bar{x}E\left[\left(\hat{\beta}_1 - \beta_1\right)^2\right] = 0 - \bar{x}Var\left(\hat{\beta}_1\right) = -\frac{\bar{x}\sigma_u^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}
\end{aligned}$$

5 Contrastación de hipótesis

En esta sección describimos los procedimientos para la contrastación de hipótesis acerca de los valores numéricos que toma uno o varios de los coeficientes del modelo de regresión. Comenzamos prestando atención a los contrastes de hipótesis acerca del valor numérico de un sólo parámetro, para pasar después a considerar hipótesis o restricciones sobre el valor numérico de varios parámetros y, finalmente, hipótesis o restricciones acerca del valor numérico de una o más combinaciones lineales de parámetros. Para todo ello, utilizamos el siguiente resultado:

Una combinación lineal de variables aleatorias con distribución Normal, sigue también una distribución de probabilidad Normal: Si $z_1, z_2, ..., z_n$ *son variables aleatorias independientes, cada una de ellas con una distribución* $N(\mu_i, \sigma_i^2)$*, la combinación lineal:* $w = a_1 z_1 + a_2 z_2 + ... + a_n z_n$*, también sigue una distribución Normal:*

$$w \sim N(a_1\mu_1 + a_2\mu_2 + ... + a_n\mu_n;\ a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + ... + a_n^2\sigma_n^2) = N\left(\sum_{i=1}^{n}\alpha_i\mu_i;\ \sum_{i=1}^{n}a_i^2\sigma_i^2\right)$$

Para este resultado no es necesario que las variables $z_1, z_2, ..., z_n$ sean independientes. Si no lo son, entonces en la expresión de la varianza de w hay que añadir sumandos adicionales correspondientes a la covarianza entre cada dos variables del conjunto: $z_1, z_2, ..., z_n$.

El estimador de mínimos cuadrados de la pendiente del modelo de regresión simple puede escribirse:

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{N}\alpha_i u_i \tag{19}$$

por lo que, si el término de error del modelo, que es una variable aleatoria distinta para cada observación muestral, sigue una distribución Normal, entonces el estimador $\hat{\beta}_1$ también tendrá una distribución Normal.

También el estimador de mínimos cuadrados del término independiente de la regresión, $\hat{\beta}_0$, puede escribirse como combinación lineal de los términos de error del modelo. Para ello, notemos que: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{1}{N}\sum_{i=1}^{N} y_i - \left(\sum_{i=1}^{N} \alpha_i y_i\right)\bar{x} = \sum_{i=1}^{N}\left(\frac{1}{N} - \alpha_i\bar{x}\right) y_i$.Por tanto, también este estimador sigue una distribución Normal, aunque pocas veces estaremos interesados en contrastar hipótesis acerca del verdadero valor de dicho coeficiente.

5.1 Contrastes de hipótesis acerca del valor numérico de un sólo coeficiente

En secciones anteriores obtuvimos las condiciones bajo las cuales el estimador de mínimos cuadrados de la pendiente del modelo de regresión simple es insesgado, y tiene por varianza $Var\left(\hat{\beta}_1\right) = \frac{\sigma_u^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}$. Si, además, los términos de error correspondientes a cada observación siguen una distribución Normal, entonces, vimos que $\hat{\beta}_1$ también seguirá una distribución Normal. Tipificando la variable aleatoria $\hat{\beta}_1$ es decir, restando su esperanza matemática y dividiendo por su desviación típica, tenemos:

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var\left(\hat{\beta}_1\right)}} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma_u^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}}} \sim N(0,1)$$

Este resultado podría utilizarse para llevar a cabo contrastes de hipótesis acerca del verdadero valor numérico de la pendiente β_1, comparando el valor numérico obtenido en la muestra para el estadístico $\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\sigma_u^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}}}$ con las tablas de la distribución $N(0,1)$ al nivel de significación escogido. El problema es que desconocemos el valor numérico del parámetro σ_u^2. Como se muestra en el Apéndice, podemos sustituir el verdadero valor del parámetro σ_u^2, desconocido, por su estimador, y tenemos un estadístico con distribución t_{N-2} :

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\hat{\sigma}_u^2}{\sum_{i=1}^{N}(x_i-\bar{x})^2}}} \equiv \frac{\hat{\beta}_1 - \beta_1}{DT(\hat{\beta}_1)} \sim t_{N-2}$$

Lo importante es que la variable que acabamos de probar que sigue una distribución t_{N-2} es la misma que antes probamos que seguía una distribución $N(0,1)$, sólo que sustituyendo la varianza desconocida del término de error, σ_u^2, que hemos supuesto constante para todas las observaciones muestrales, por su estimación de mínimos cuadrados: $\hat{\sigma}_u^2 = \frac{1}{N-2}\sum_{i=1}^{N}\hat{u}_i^2$. En el denominador de la expresión hemos hecho explícito el hecho de que la raiz cuadrada no es sino la estimación de la desviación típica del estimador de $\hat{\beta}_1$. La propiedad $\frac{\hat{\beta}_1-\beta_1}{DT(\hat{\beta}_1)} \sim t_{N-2}$ es la que utilizaremos para diseñar contrastes de hipótesis en el modelo de regresión lineal simple, puesto que el valor numérico del estadístico puede ahora calcularse con la información muestral.

Introducimos ahora el principio que regirá el modo en que vamos a resolver todo tipo de contrastes de hipótesis:

En cualquier contraste de hipótesis, el estadístico muestral mide el grado de incumplimiento muestral de la hipótesis nula.

Así sucede con el estadístico t anterior, en cuyo numerador aparece la diferencia $\hat{\beta}_1 - \beta_1$, entre la estimación obtenida con nuestra muestra, y el valor hipotético de dicho coeficiente. PEro, como sabemos si una discrepancia, digamos que de 0,40, por ejemplo, entre ambos valor numéricos, el estimado y el teórico, es *suficientemente grande*? Hay que utilizar una unidad de medida, y eso es lo que hacemos al dividir por la desviación típica del estimador. De este modo, evaluamos si la discrepancia $\hat{\beta}_1 - \beta_1$ es igual a una vez, dos veces, o una vez y media la desviación típica del estimador $\hat{\beta}_1$. Una vez calculado, queremos decidir si el grado de incumplimiento muestral es grande o pequeño. En el primer caso, rechazaremos la hipótesis nula, no rechazándola si el grado de incumplimiento muestral es pequeño. Para decidir acerca de esta importante cuestion, comparamos el número resultante de calcular el cociente $\frac{\hat{\beta}_1 - \beta_1}{DT(\hat{\beta}_1)}$, que resume el grado de incumplimiento muestral de la hipótesis nula, con el valor crítico de las tablas de la distribución t_{N-2} al nivel de significación escogido. Si el estadístico muestral es mayor que el valor crítico de las tablas, decimos que el incumplimiento muestral es grande, por lo que rechazamos la hipótesis nula. Si, por el contrario, el estadístico muestral es menor que el valor crítico de las tablas, decimos que el grado de incumplimiento muestral de la hipótesis nula es pequeño, y no rechazamos la hipótesis nula. Por ejemplo, al nivel de significación del 5%, si el tamaño muestral es suficientemente grande (mayor que 120 observaciones), el valor crítico de las tablas es de 1,96 (aproximadamente 2,0). En consecuencia, si la diferencia entre la estimación numérica y el valor teórico contenido en la hipótesis nula es mayor que dos veces la desviación típica del estimador de $\hat{\beta}_1$, decimos que el incumplimiento muestral de la hipótesis nula es sufucientemente grande, y rechazamos la hipótesis nula. Hacemos lo contrario si la diferencia es menor de dos veces la desviación típica del estimador de $\hat{\beta}_1$. En cada caso, para el nivel de significación escogido, y el número de grados de libertad, $N-2$, con que contemos, las tablas nos darán el valor crítico con el que hemos de comparar.

5.1.1 Contrastes de dos colas (bilaterales) acerca del valor numérico de un solo coeficiente

En el caso de un contraste bilateral, por ejemplo, contrastando la hipótesis nula: $H_0: \beta_1 = 1$ frente a la hipótesis alternativa: $H_1: \beta_1 \neq 1$, tomaremos la diferencia $\hat{\beta}_1 - 1,0$ en valor absoluto, porque como en la hipótesis alternativa consideramos tanto la posibilidad de que el verdadero valor del parámetro sea mayor que 1, como de que sea menor que 1. Compararemos con el valor crítico de las tablas de la distribución t_{N-2} que deja en cada cola de la misma una probabilidad igual a $\alpha/2$.

De lo expuesto se deduce otro principio fundamental:

> Ninguna hipótesis nula es completamente correcta en una muestral. Por tanto, se trata de ver si la información muestral aporta evidencia suficientemente contraria a H_0 como para rechazarla o no. Pero la evidencia muestral nunca puede probar que una determinada hipótesis nula es cierta. En consecuencia, las decisiones a considerar son *Rechazar* H_0, o *No rechazar* H_0, pero nunca decidimos *Aceptar* H_0.

Una vez que se toma una decisión, sea de Rechazar o de No rechazar la hipótesis nula, puede ser interesante evaluar por cuánto margen se ha tomado. Para eso, suele utilizarse el valor-p del contraste, definido:

El valor-p de un contraste es la probabilidad de que una muestra de igual tamaño que la utilizada arroje una evidencia más contraria a la hipótesis nula que la que hemos encontrado.

Su cálculo es muy sencillo, puesto que el grado en que la evidencia obtenida de nuestra muestra es contraria a la hipótesis nula queda resumida en el valor numérico del estadístico $\frac{\hat{\beta}_1-\beta_1}{DT(\hat{\beta}_1)}$, en valor absoluto. En consecuencia, el valor-p es la probabilidad que proporcionan las tablas de la distribución t_{N-2} a valores numéricos mayores que $\frac{\hat{\beta}_1-\beta_1}{DT(\hat{\beta}_1)}$, tomados ambos en valor absoluto.

5.1.2 Contrastes de una cola (unilaterales) acerca del valor de un solo coeficiente

Otro principio fundamental de la contrastación de hipótesis es el siguiente:

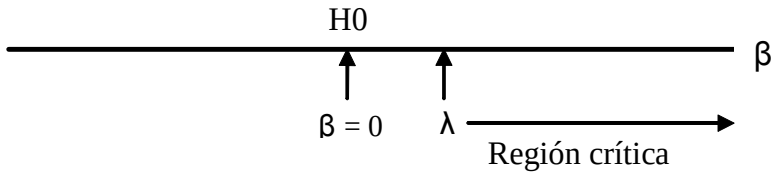
Para rechazar una hipótesis nula es preciso que se cumplan dos condiciones:

1. que la evidencia muestral sea contraria a la hipótesis nula
2. que la evidencia muestral sea favorable a la hipótesis alternativa.

Estas condiciones pueden parecer redundantes, porque en primera lectura parece que significan lo mismo. Esto es cierto en contrastes bilaterales, en los que ambas condiciones son equivalentes y, por tanto, basta con que se cumpla cualquiera de ellas. No sucede lo mismo en contrastes unilaterales.

Supongamos que queremos contrastar la hipótesis nula $H0 : \beta = 0$ frente a la hipótesis alternativa: $H1 : \beta > 0$. La evidencia muestral será contraria a la hipótesis mnula si la estimación numérica de β es *suficientemente diferente de cero*, si bien deberemos aclarar que entendemos por esto último. Pero si $\hat{\beta} = -2,35$ tendremos que si bien la evidencia muestral es contraria a la hipótesis nula, no es favorable a la hipótesis alternativa, que considera únicamente valores positivos de β. En un caso así, no procede llevar a cabo el contraste. Más bien deberemos preguntarnos cómo puede haberse obtenido dicha estimación. Puede ser que la información muestral sea imperfecta, por haber existido algún sesgo en el modo en que se ha recogido. También podría ser que la precisión es muy pequeña, por lo que aunque hemos estimado un valor negativo, su desviación típica estimada es aún mayor, por ejemplo: $DT(\hat{\beta}) = 5,32$. Pero ni siquiera en esta situación procede contratar porque el hecho es que, en tal situación, no rechazaremos nunca la hipótesis nula. También podría ser que el resultado muestral nos hiciese considerar.lo adecuado de la hipótesis alternativa $H1 : \beta > 0$, y prefiriésemos realizar un contraste bilateral. Pero el hecho es que si, volviendo al caso de una reducida precisión, llevamos a cabo el contraste, como no vamos a rechazar la hipótesis nula, concluyésemos que dicha hipótesis es cierta, con una estimación $\hat{\beta} = -2,35$!!!. Esta última interpretación, que se basa en la idea de que si no se rechaza una hipótesis nula pasamos a creer que dicha hipótesis es cierta, es totalmente equivocada, aunque se olbserva con demasiada frecuencia en trabajos de economía aplicada. Es esto lo que nos lleva a recomendar que salvo que se cumplan las dos condiciones antes enunciadas simultáneamente, no se lleve a cabo el contrate de hipótesis unilateral.

Pero, supongamos que hemos obtenido una estimación que se desvía de la hipótesis nula en la dirección correcta, y pasamos a contrastar la hipótesis nula $H0 : \beta = 0$ frente a la hipótesis alternativa: $H1 : \beta > 0$, al nivel de significación del 1% La región crítica está formada por los valores numéricos del estimador positivos y suficientemente grandes, pues constituyen evidencia muestral contraria a $H0$, a la vez que favorable a $H1$. Por tanto, la región crítica es el intervalo de



$$Valor - p = P\left[\hat{\beta} > 0,26 / \beta = 0\right] = P\left[\frac{\hat{\beta} - \beta}{DT(\hat{\beta})} > \frac{0,26 - \beta}{DT(\hat{\beta})} / \beta = 0\right] =$$

$$= P\left[N(0,1) > \frac{0,26}{0,11} = 2,3636\right] = 1 - 0,9910 = 0,009$$

valores a la derecha de un determinado margen de seguridad a la derecha de $\beta = 0$. Concretamente, rechazaremos H_0 en favor de H_1 si la estimación numérica está a la derecha de λ.

Pero ¿cómo se determina este margen de seguridad? Como siempre, tomando un número determinado de veces la desviación típica del estimador. Dicho número de veces se obtiene acudiendo a las tablas de la distribución t_{N-2}. Si suponemos que el número de observaciones muestrales es suficientemente grande como para aproximar la distribución t_{N-2}.por una distribución $N(0,1)$, observamos en las tablas de esta distribución que el valor numérico a la derecha del cual hay una probabilidad de 0,01 es 2,33. Supongamos que hemos estimado $\hat{\beta} = 0,26$, y que la desviación típica de nuestro estimadorr ha sido $DT(\hat{\beta}) = 0,11$. Por tanto, $\lambda = (2,33)(0,11) = 0,2563$. Nuestro estimador cae a la derecha de este valor de λ, luego dentro de la Región Crítica, aunque por poco, por lo que rechazamos la hipótesis nula, si bien por muy poco margen.

Si hubiésemos utilizado un nivel de significación del 5%, el valor numérico de las tablas de la $N(0,1)$sería 1,645, por lo que $\lambda = (1,645)(0,11) = 0,18$, y rechazaríamos $H0$ con relativa claridad. Si hubiésemos utilizado un nivel de significación del 10%, el valor numérico de las tablas de la $N(0,1)$ sería 1,28, por lo que $\lambda = (1,28)(0,11) = 0,14$, y rechazaríamos $H0$ con mayor claridad. Es lógico que sea así.

Vamos a calcular el valor-p de este contraste, es decir, la probabilidad de que la evidencia muestral pudiese ser más desfavorable a $H0$ que la que hemos obtenido. Valores p muy reducidos sugieren que es poco probable encontrar una muestra más contraria a la hipótesis nula que la nuestra. Por lo tanto, la nuestra es bastante contraria a H_0, por lo que rechazaremos dicha hipótesis. Lo contrario sucede si el valor-p es grande. ¿Cómo decidimos si un valor-p es grande o pequeño? Comparando con el nivel de significación. Si el valor-p es menor que el nivel de significación, concluimos que la probabilidad de hallar una muestra más contraria que la nuestra a H_0 es pequeña, y rechazamos la hipótesis nula. Hemos rechazado H_0 a un nivel de significación del 1% por poco, y no hemos rechazado H_0 a niveles de significación del 5% o del 10%. Este análisis sugiere que el valor-p del contraste debe estar por debajo de 0,01, pero muy próximo a dicho nivel. En este caso, una evidencia más contraria a H_0 que la nuestra significa que el estimador fuese aún mayor que el nuestro, que es 0,26, bajo el supuesto de que $H0$ sea cierta. Así:

que está, efectivamente, muy próximo a 1%.

5.2 *Significación estadística versus relevancia económica*:

Consideremos las dos hipotéticas situaciones prácticas:

$$\begin{aligned} \ln(\Pr ecio_i) &= \hat{\beta}_0 + \underset{(0.50)}{0.85} \ln(Lotsize_i) + \hat{u}_i \\ \ln(\Pr ecio_i) &= \hat{\beta}_0 + \underset{(0.02)}{0.05} \ln(Lotsize_i) + \hat{u}_i \end{aligned}$$

Son dos modelos que explican el precio de una vivienda en función únicamente del tamaño de la parcela que ocupa. Estamos interesados en saber si el tamaño de la parcela condiciona el precio o si, por el contrario, éste está determiando exclusivamente por las características de la construcción, pero no por el tamaño de la parcela. Por tanto, queremos contrastar: $H_0 : \beta_1 = 0$, y supongamos que consideramos como hipótesis alternativa:[17] $H_1 : \beta_1 \neq 0$. El estadístico t para el contraste de una hipótesis acerca del valor de un único coeficiente es: $\frac{\hat{\beta}-\beta^0}{DT(\hat{\beta})} \sim t_{N-coefs}$. En este caso concreto, $\beta^0 = 0$, por lo que el estadístico t se obtiene dividiendo la estimación numérica por su desviación típica, que aparece en paréntesis. En el primer modelo, tenemos un valor numérico del estadístico de contraste igual a $1,70$, mientars qeu en el segundo modelo, el estadístico es igual a $2,50$.Con una muestra de 543 viviendas, el número de grados de libertad de la distribución t es superior a 120, por lo que sus valores críticos coinciden con los de la Normal(0,1). A un nivel de significación del 5%, dicho nivel crítco es $1,96$.

Siguiendo las pautas antes reseñadas, no rechazamos la hipótesis nula en el primer modelo, mientras que la rechazamos en el segundo modelo. Hasta aquí, todo es correcto desde el punto de vista estadístico. el problema surge porque, en Economía, es habitual interpretar estos resultados afirmando que el tamaño de la parcela no es un determinante significativo del precio en el primer caso (ya que no rechazamos $\beta_1 = 0$), mientras que sí lo es en el segundo modelo (puesto que rechazamos $\beta_1 = 0$).

Esto es un grave error de interpretación. el primer modelo predice que un aumento del 10% en el tamaño de la parcela genera un incremento del precio de la vivienda del 8,5%, mientras que en segundo caso, el incremento en el precio sería de tan sólo un 0,5%. Imaginemos que una vivienda con una parcela de 2.000 tiene un precio de 40.000. Esto significa que el primer modelo predice que una parcela de 2.200 tendrá un precio de 43.400, mientras que de acuerdo con el segundo modelo, su precio sería de 40.200. El incremento es de 3.400 dólares en el primer modelo, y de solo 200 dólares en el segundo modelo. Es decir, el primer modelo sugiere un efecto cuantitativo del tamaño de la parcela sobre el precio de la vivienda, claramente mayor que el segundo modelo. ¿Tiene sentido eliminar el tamaño de la parcela como determinante del precio en el primer modelo y no en el segundo modelo, como sugieren los contrastes de significación? Evidentemente, no.

El problema surge porque identificamos, sin justificación alguna, la significación estadística de un coeficiente con la relevancia económica de la variable a la que acompaña. Estas son dos propiedades distintas (significación estadística y relevancia económica) de elementos diferentes (un coeficiente, en el primer caso, y una variable, en el segundo caso). No tiene nada que ver una con otra: podemos tener un coeficientes significativo acompañando a una variable poco relevante económicamente (como podría ser el caso del segundo modelo), o un coeficientes no significativamente distinto de

[17] Para ser rigurosos, la hipótesis alternativa debería ser: $H_0 : \beta_1 > 0$,ya que no concebimos la posibilidad de que el tamaño de la parcela afecte *negativamente* al precio.

cero estadísticamente, acompañando a una variable económicamente importante (como podría ser el caso del primer modelo).

Tenemos que distinguir entre estos dos efectos.

Lo que sucede en el primer modelo es que estimamos con poca precisión, como se aprecia en el hecho de que la desviación típica sea elevada, en relación con la estimación numérica del coeficiente. En consecuencia, no tenemos mucha seguridad en que 0,85 sea un valor de referencia muy exacto. Asimismo, el intervalo de confianza que podamos construir será muy amplio. Pero, si bien es cierto que no podemos ser muy precisos acerca de su valor numérico concreto, es dudoso que debamos suponer que puede aproximarse por cero.

En el segundo modelo, por las razones que sea (por ejemplo, esto suele suceder con muestras muy grandes) hemos estimado con mucha precisión. La desviación típica es realmente muy reducida. Pero es dudoso que queramos decir que el efecto antes descrito sea económicamente relevante.

Debemos examinar siempre la situación de la precisión de las estimaciones numéricas que hemos obtenido para los coeficientes del modelo.Y no podemos olvidar que nos interesa pronunciarnos acerca de la relevancia económcia de las variables explicativas, no necesariamente acerca de la significación estadística de los coeficientes asociados a ellas.

- Otra manera de ver este efecto es apreciar que el estadístico t es el producto de dos factores: por un lado, el incumplimiento muestral $\left(\hat{\beta} - \beta^0\right)$; por otro, la precisión en la estimación del parámetro:$\frac{1}{DT(\hat{\beta})}$. Así, el estadístico t puede ser bajo (inferior a $2,0$) bien porque el incumplimiento muestral sea pequeño, o porque, incluso siendo apreciable, la precisión es muy reducida. En este segundo caso, con un incumplimieno alto, querríamos rechazar la hipótesis nula; sin embargo, puede que no lo hagamos porque el producto de los dos factores resulte inferior a $2,0$.Este es el caso del primer modelo del ejemplo anterior.

Tambien podría resultar que el producto de un pequeño incumplimiento de la hipótesis nula y una elevada precisión, produzcan un valor superior a $2,0$ del estadístico t, a pesar de que, en este caso, no querríamos rechazar la hipótesis nula. este es el caso del segundo modelo en el ejemplo anterior.

- Un modo alternativo de contrastar una hipòtesis acerca de un único coeficiente consiste en construir el intervalo de confianza, al nivel de confianza adecuado (recordemos que el nivel de confianza es igual a $1,0$ menos el nivel de significación), y analizar si el valor hipotético, el que aparece en la hipótesis nula, está dentro de dicho intervalo. Si es así, no podremos rechazar dicha hipótesis nula, ya que el valor teórico está dentro del rango de confianza que hemos construido.Lo contrario sucede si en valor teórico cae fuera del intervalo de confianza, en cuyo caso rechzaremos la hipótesis nula. Este procedimiento es completamente equivalente al uso dle estadístico t que antes describimos.

- *Contrastes unilaterales*: Para rechazar una hipótesis nula se deben dar 2 condiciones simultáneamente:

1. que la evidencia muestral sea contraria a la hipótesis nula
2. que la evidencia muestral sea favorable a la hipótesis alternativa.

En contrastes bilaterales como los que hemos visto hasta ahora, que son aquellos cuya hipótesis alternativa es del tipo: $H_0 : \beta_1 \neq 0$, ambas condiciones son equivalentes, por lo que basta

que se cumpla una cualquiera de ellas. Pero no sucede así en el caso de un contraste unilateral. Consideremos el contraste:

$$\begin{aligned} H_0 &: \quad \beta_1 = 1 \\ H_1 &: \quad \beta_1 < 1 \end{aligned}$$

Una estimación numérica $\hat{\beta}_1 = 0,62$ sería evidencia posiblemente contraria a la hipótesis nula, *a la vez* que favorable a la hipótesis alternativa. En esta situación, es posible que rechacemos H_0 en favor de H_1. Que finalmente lo hagamos o no, dependerá de la precisión con que hayamos estimado el coeficiente β_1. Por el contrario, una estimación $\hat{\beta}_1 = 1,12$ puede que satisfaga la primera condición, pero no satisface la segunda, por lo que no rechazaremos H_0. Es importante puntualizar que tampoco rechazaríamos H_0 con una estimación $\hat{\beta}_1 = 2,62$, por mucho que es muy claramente diferente de $1,0$.

En un contraste bilateral, la región fuera del intervalo de confianza se denomina *región crítica*. Es aquella en la que si cae el valor teórico, rechazaremos la hipótesis nula correspondiente. En un contraste unilateral como el anterior, la región crítica se contruye restando a la estimación numérica un número de veces su desviación típica. Dicho número de veces se obtiene de las tablas correspondientes (por ejemplo, $t_{N-ncoefs}$) al nivel de confianza elegido.[18]

Supongamos que hemos estimado $\hat{\beta}_1 = 1,12(0,10)$. Al 5% de significación, las tablas nos dan (para el contraste unilateral) un valor numérico de 1,645. Por tanto, la región crítica se elabora sumando a la estimación numérica, de $1,12$, el producto: $(1,645)(0,10) = 0,165$, lo que nos lleva a: $1,12 + 0,165 = 1,285$. La región crítica es el intervalo abierto a la izquierda de este punto (-$\infty, 1,285$), mientras que la región de aceptación es el intervalo abierto a la derecha de este punto $(1,285, \infty)$. Rechazamos cualquier hipótesis teórica que caiga en la región crítica, mientras que no rechazamos aquellas hipótesis que caigan en la región de aceptación. En concreto, en este caso no rechazamos $H_0 : \ \beta_1 = 1$. En general, es claro que no rechazaremos $H_0 : \ \beta_1 = 1$ en favor de $H_1 : \ \beta_1 < 1$ si la estimación es superior a $1,0$.

La situación interesante surge si obtenemos una estimación: $\hat{\beta}_1 = 0,92(0,10)$. La región crítica es el intervalo a la izquierda de: $0,92 + (1,645)(0,10) = 1,08$, por lo que tampoco rechazaremos $H_0 : \ \beta_1 = 1$ en favor de $H_1 : \ \beta_1 < 1$. Por el contrario, rechazaríamos $H_0 : \ \beta_1 = 1$ en favor de $H_1 : \ \beta_1 < 1$ tanto si estimamos $\hat{\beta}_1 = 0,92(0,03)$, como si estimamos $\hat{\beta}_1 = 0,82(0,10)$, en todos los casos, suponiendo que el número de grados de libertad es superior a 120.

5.3 Apéndice: Contrastación de hipótesis

- Distribución t de Student:

Nos basamos en una propiedad estadística que damos sin demostración:

$$\frac{1}{\sigma_u^2} \sum_{i=1}^{N} \hat{u}_i^2 = N \frac{\hat{\sigma}_u^2}{\sigma_u^2} \sim \chi_{N-2}^2$$

[18] Recordemos que en los contrastes bilaterales, buscamos en las tablas el valor numerico asociado a *la mitad* del nivel de significación.

donde $\hat{\sigma}_u^2$ denota el estimador de máxima verosimilitud de σ_u^2, que es igual a la SCR dividida por el tamaño muestral N, no por $N-2$. Además, la variable χ^2_{N-2} anterior es independiente de $\hat{\beta}_1$.

Recordemos que el cociente entre una $N(0,1)$ y la raiz cuadrada de una variable χ^2 dividida por su número de grados de libertad, se distribuye como una t de Student, si ambas variables, la $N(0,1)$ y la χ^2 son independientes. El número de grados de libertad de la variable t es el mismo número que el de la variable χ^2.

En consecuencia, si dividimos la distribución $N(0,1)$ por esta distribución χ^2_{N-2}, tenemos:

$$\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\sigma_u^2}{\sum_{i=1}^N(x_i-\bar{x})^2}}\sqrt{\frac{1}{N-2}\frac{1}{\sigma_u^2}\sum_{i=1}^N\hat{u}_i^2}}=\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\frac{1}{N-2}\sum_{i=1}^N\hat{u}_i^2}{\sum_{i=1}^N(x_i-\bar{x})^2}}}=\frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{\hat{\sigma}_u^2}{\sum_{i=1}^N(x_i-\bar{x})^2}}}\sim t_{N-2}$$

Para entender (sin demostrar) que $\frac{1}{\sigma_u^2}\sum_{i=1}^N\hat{u}_i^2\sim\chi^2_{N-2}$, utilizamos la descomposición:

$$u_i=\bar{u}+\left(\hat{\beta}_1-\beta_1\right)(x_i-\bar{x})+\hat{u}_i$$

para probar la igualdad:

$$\sum_{i=1}^N\frac{u_i^2}{\sigma_u^2}=N\frac{\bar{u}^2}{\sigma_u^2}+\frac{\left(\hat{\beta}_1-\beta_1\right)^2}{\sigma_u^2}NS_x^2+\frac{1}{\sigma_u^2}\sum_{i=1}^N\hat{u}_i^2$$

El término a la izquierda de la igualdad sigue una distribución χ^2_N, por ser la suma de cuadrados de variables $N(0,1)$ que suponesmo independientes. El primer término de la derecha sigue una distribución χ^2_1, por ser el cuadrado de una variable $N(0,1)$. Lo mismo sucede con el segundo término de la derecha. Por último, podemos utilizar el resultado de que el número de grados de libertad de variables χ^2 independientes es aditivo, es decir, que si χ^2_m y χ^2_n son independientes, entonces: $\chi^2_m+\chi^2_n=\chi^2_{m+n}$. Los términos de la igualdad anterior son independientes, lo que aquí no probamos.

6 El estimador de Mínimos Cuadrados del modelo de regresión múltiple

Consideremos la estimación del modelo de regresión lineal múltiple:

$$y_i=\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+u_i,\ i=1,2,...n$$

Si tratamos de minimizar la Suma de Cuadrados de los Residuos, tendríamos:

$$\underset{\beta_0,\beta_1,\beta_2}{Min}\,SCR(\beta_0,\beta_1,\beta_2)=\underset{\beta_0,\beta_1,\beta_2}{Min}\sum_{i=1}^n\left(y_i-\beta_0-\beta_1x_{1i}-\beta_2x_{2i}\right)^2$$

Derivando en esta función de Suma de Cuadrados de Residuos $SCR(\beta_0,\beta_1,\beta_2)$ respecto de los tres parámetros β_0,β_1,β_2 e igualando a cero, se obtiene el sistema de *ecuaciones normales*:

$$\begin{aligned}
\sum y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum x_{1i} + \hat{\beta}_2 \sum x_{2i} \\
\sum y_i x_{1i} &= \hat{\beta}_0 \sum x_{1i} + \hat{\beta}_1 \sum x_{1i}^2 + \hat{\beta}_2 \sum x_{1i} x_{2i} \\
\sum y_i x_{2i} &= \hat{\beta}_0 \sum x_{2i} + \hat{\beta}_1 \sum x_{1i} x_{2i} + \hat{\beta}_2 \sum x_{2i}^2
\end{aligned}$$

donde los circunflejos denotan que la solución al sistema de ecuaciones será el estimador de Mínimos Cuadrados. Una vez resuelto este sistema, la diferencia: $\hat{u}_i = y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$ será el residuo correspondiente a la observación i-ésima. Por tanto, aun antes de resolver el sistema de ecuaciones normales:

$$\begin{aligned}
\sum \hat{u}_i &= 0 \\
\sum \hat{u}_i x_{1i} &= 0 \\
\sum \hat{u}_i x_{2i} &= 0
\end{aligned}$$

que muestra propiedades similares a las que obtuvimos en el modelo de regresión lineal simple: a) la suma de los residuos de mínimos cuadrados es igual a cero, y b) los residuos de mínimos cuadrados están incorrelacionados con las variables explicativas del modelo.

La segunda propiedad se deduce de la propiedad a), que implica que los residuos tienen media cero. Entonces las sumas del tipo $\sum \hat{u}_i x_{1i}$ no son sino la covarianza entre ambas variables, multiplicada por el tamaño muestral. Por tanto, la covarianza es cero, y también es cero el coeficiente de correlación. Estas propiedades no se cumplirán con otro procedimiento de estimación. Si en el modelo hubiera k variables explicativas ($k > 2$), el razonamiento sería similar, y se tendrían las mismas propiedades.

Para resolver el sistema y hallar el estimador de Mínimos Cuadrados Ordinarios (MCO), primero despejamos $\hat{\beta}_0$ en la primera ecuación:

$$\hat{\beta}_0 = \frac{1}{n} \sum y_i - \hat{\beta}_1 \frac{1}{n} \sum x_{1i} - \hat{\beta}_2 \frac{1}{n} \sum x_{2i}$$

y sustituimos en las otras dos:

$$\begin{aligned}
\sum y_i x_{1i} &= \left(\frac{1}{n} \sum y_i - \hat{\beta}_1 \frac{1}{n} \sum x_{1i} - \hat{\beta}_2 \frac{1}{n} \sum x_{2i} \right) \sum x_{1i} + \hat{\beta}_1 \sum x_{1i}^2 + \hat{\beta}_2 \sum x_{1i} x_{2i} \\
\sum y_i x_{2i} &= \left(\frac{1}{n} \sum y_i - \hat{\beta}_1 \frac{1}{n} \sum x_{1i} - \hat{\beta}_2 \frac{1}{n} \sum x_{2i} \right) \sum x_{2i} + \hat{\beta}_1 \sum x_{1i} x_{2i} + \hat{\beta}_2 \sum x_{2i}^2
\end{aligned}$$

que pueden escribirse:

$$\begin{aligned}
S_{x_1 y} &= \hat{\beta}_1 S_{x_1}^2 + \hat{\beta}_2 S_{x_1 x_2} \\
S_{x_2 y} &= \hat{\beta}_1 S_{x_1 x_2} + \hat{\beta}_2 S_{x_2}^2
\end{aligned} \tag{20}$$

donde:

$$
\begin{aligned}
S_{x_1y} &= \sum (x_{1i}-\bar{x}_1)(y_i-\bar{y});\ S_{x_2y}=\sum (x_{2i}-\bar{x}_2)(y_i-\bar{y});\\
S^2_{x_1} &= \sum (x_{1i}-\bar{x}_1)^2;\ S^2_{x_2}=\sum (x_{2i}-\bar{x}_2)^2;\ S_{x_1x_2}=\sum (x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2);
\end{aligned}
$$

El sistema (20) es lineal con dos ecuaciones en dos incógnitas, sencillo de resolver, que conduce a:

$$
\hat{\beta}_1=\frac{S_{x_1y}S^2_{x_2}-S_{x_2y}S_{x_1x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2};\ \hat{\beta}_2=\frac{S_{x_2y}S^2_{x_1}-S_{x_1y}S_{x_1x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2};\ \hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}_1-\hat{\beta}_2\bar{x}_2; \tag{21}
$$

Obtener las expresiones analíticas de las varianzas y covarianzas de los estimadores de los coeficientes es complejo. Baste dar aquí las expresiones:

$$
\begin{aligned}
Var(\hat{\beta}_1) &= \sigma^2_u\frac{S^2_{x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2};\\
Var(\hat{\beta}_2) &= \sigma^2_u\frac{S^2_{x_1}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2};\\
Cov(\hat{\beta}_1,\hat{\beta}_2) &= -\sigma^2_u\frac{S_{x_1x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2}\\
Var(\hat{\beta}_0) &= \sigma^2_u\left(\frac{1}{n}+\frac{\bar{x}^2_1S^2_{x_2}+\bar{x}^2_2S^2_{x_1}-2\bar{x}_1\bar{x}_2S_{x_1x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2}\right)
\end{aligned}
$$

La expresión de la covarianza será útil para contrastar hipótesis que involucren simultáneamente a ambos parámetros, como veremos en ejercicios.

Proposition 4 El estimador de Mínimos Cuadrados es insesgado

Proof. El estimador MCO puede escribirse: ■

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{S^2_{x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2}\sum (x_{1i}-\bar{x}_1)(y_i-\bar{y})-\frac{S_{x_1x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2}\sum (x_{2i}-\bar{x}_2)(y_i-\bar{y})=\\
&= \frac{S^2_{x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2}\sum (x_{1i}-\bar{x}_1)(\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+u_i-\bar{y})-\\
&\quad -\frac{S_{x_1x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2}\sum (x_{2i}-\bar{x}_2)(\beta_0+\beta_1x_{1i}+\beta_2x_{2i}+u_i-\bar{y})=
\end{aligned}
$$

$$
=\beta_1+\frac{S^2_{x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2}\sum (x_{1i}-\bar{x}_1)(u_i-\bar{u})+\frac{S_{x_1x_2}}{S^2_{x_1}S^2_{x_2}-\left(S_{x_1x_2}\right)^2}\sum (x_{2i}-\bar{x}_2)(u_i-\bar{u})
$$

y los dos últimos sumandos tienen esperanza igual a cero.

Para llevar a cabo contrastes de hipótesis acerca de posibles valores numéricos de los coeficientes del modelo, es preciso disponer de las varianzas de los estimadores, al igual que ocurría en el modelo simple. Sin discutir aquí su deducción analítica, que es compleja, baste decir que las varianzas de los tres coeficientes pueden escribirse:

Además, el estimador MCO es el estimador lineal insesgado de mínima varianza, puesto que el mismo Teorema de Gauss-Markov que enunciamos en el caso del modelo lineal simple continúa siendo válido en este modelo más general. Asimismo, el Teorema de Rao también se cumple, de modo que el estimador de Mínimos Cuadrados es eficiente cuando el término de error sigue una distribución Normal, es decir, tiene la menor varianza posible de entre todos los estimadores insesgados, sean estos lineales o no lineales.

6.1 Ejemplo: Ventas de un bien en función del precio propio y del gasto en publicidad[19]

Este ejemplo utiliza 10 observaciones anuales sobre las ventas, gastos en publicidad y precio del producto de una empresa. El interés del ejemplo es:

- ilustrar el modo de interpretar los valores numéricos estimados para coeficientes individuales en un contexto de colinealidad,
- mostrar la manera de analizar el contenido informativo de las variables explicativas en un contexto de alta colinealidad,
- proponer un modo de tratar la colinealidad entre variables explicativas,

6.1.1 Algunas características de las variables

El archivo de trabajo contiene información acerca de la cifra de ventas anuales V_t de una empresa, junto con sus gastos en publicidad, Pub_t, ambos en miles de euros, y el precio de venta de su producto, P_t, asimismo en miles de euros por unidad. Son datos artificiales, formados por 10 observaciones de cada variable, pero serán suficiente para ilustrar las cuestiones que nos interesan. Las tres variables muestran, dentro del breve espacio de tiempo cubierto por la muestra, un comportamiento tendencial, que es creciente en el caso de las ventas y los gastos en publicidad, y decreciente en el caso del precio del producto.

Las nubes de puntos representan la relación entre la cifra de ventas anual y cada una de las dos potenciales variables explicativas, precio y gasto en publicidad $(NUBE\_VENT\_PRECIO, NUBE\_VENT\_PUB)$, mostrando claramente una asociación negativa entre V_t y P_t, y positiva entre V_t y Pub_t.

Las covarianzas y coeficientes de correlación entre las variables pueden resumirse en la matriz,

[19] Fichero de trabajo: Ventas.wf1. Fichero de Excel: Ventas.xls.

$$\Sigma = \begin{pmatrix} Correlaciones/Covarianzas & Ventas & Publicidad & \text{Pr}ecio \\ Ventas & S_V^2 & S_{V,Pub} & S_{V,P} \\ Publicidad & \rho_{V,Pub} & S_{Pub}^2 & S_{Pub,P} \\ \text{Pr}ecio & \rho_{V,P} & \rho_{Pub,P} & S_P^2 \end{pmatrix} =$$

$$= \begin{pmatrix} Corr./Cov. & Ventas & Publicidad & \text{Pr}ecio \\ Ventas & 443,5 & 124,1 & -99,0 \\ Publicidad & 0,950 & 38,5 & -26,8 \\ \text{Pr}ecio & -0,901 & -0,829 & 27,2 \end{pmatrix}$$

que muestra en su diagonal las varianzas de las tres variables; sus coeficientes de correlación consigo mismas son igual a uno, por lo que no es preciso mostrarlos. Debajo de la diagonal aparecen los coeficientes de correlación entre cada par de variables, todos ellos entre -1 y +1, mientras que por encima de la diagonal aparecen las covarianzas. Como puede verse, los tres coeficientes de correlación son muy elevados en valor absoluto.

Las desviaciones típicas muestrales de las variables son,

$$D.T.(V_t) = \sqrt{443,5} = 21,06;\ D.T.(Pub_t) = \sqrt{38,5} = 6,20;\ D.T.(P_t) = \sqrt{27,2} = 5,22.$$

Para obtener una medida comparable entre variables, debe utilizarse el coeficiente de variación, definido como cociente entre la desviación típica y media muestral de una variable.[20]

El modelo de ventas estimado utilizando tanto los gastos en publicidad como el precio del bien como variables explicativas es,

$$\underset{}{V_t} = \underset{(67,3)}{247,6} + \underset{(0,545)}{2,204}\ Pub_t - \underset{(0,649)}{1,464}\ P_t, \tag{22}$$

con coeficiente de determinación y varianza residual,

$$R^2_{V.[Pub,P]} = 1 - \frac{SR^2_{V.[Pub,P]}}{T.S_V^2} = 0,943$$

$$\hat{\sigma}_u^2 = \frac{SR^2_{V.[Pub,P]}}{T-3} = \frac{250,6}{10-3} = 35,8 \Rightarrow \hat{\sigma}_u = \sqrt{35,8} = 5,98$$

El coeficiente de determinación es elevado, indicando que más de un 94% de las fluctuaciones en las cifras de ventas anuales está explicada por cambios en el precio del producto y en el gasto en publicidad. El ajuste parece bastante bueno.

Otra manera de ver este hecho consiste en comparar la desviación típica de los residuos, 5,98, que indica el tamaño[21] del componente de las ventas no explicado por el modelo, y la desviación típica de las propias ventas, que es de 21,2. El término de error parece pequeño en relación con el tamaño medio de las fluctuaciones anuales en las cifras de ventas, y el indicador de ajuste asociado toma un valor numérico:

[20] Que en todo caso, no tiene sentido en el caso de variables tendenciales, como ocurre en este ejemplo, por lo que no calculamos dichos coeficientes, renunciando a comparar entre sí el grado de volatilidad de las variables del modelo.

[21] Una vez más, interpretamos la desviación típica de una variable aleatoria de esperanza matemática cero como inidicador de tamaño de dicha variable.

$$Ratio = 1 - \frac{\hat{\sigma}_u}{\sqrt{Var(V_t)}} = 1 - \frac{5,98}{21,1} = 0,717$$

indicando que el 72% del tamaño medio de las fluctuaciones anuales en ventas ha quedado explicada por el modelo anterior.[22]

6.1.2 ¿Qué variable explicativa es más relevante?

¿Qué explica cada variable? En sus estudios sobre los ciclos económicos, Tinbergen (1939) propuso un interesante método para reflejar la información contenida en cada variable explicativa a lo largo de la muestra.[23] Trabajando en desviaciones respecto de la media, Tinbergen sugería mostrar un gráfico representando simultáneamente los valores observados de y y los ajustados por el modelo, un gráfico para cada producto $\hat{\beta}_i x_i$, y un gráfico de residuos. Para ello se utilizan los coeficientes estimados en el modelo de regresión llineal múltiple. Hemos optado por presentar en Ventas_niveles.doc:[24] un gráfico de los valores anuales observados de y, junto con los valores anuales ajustados por el modelo; dos gráficos que confrontan los valores anuales observados para y con los valores explicados por cada una de las variables explicativas por separado, y un último gráfico que representa los valores observados de y frente a los residuos del modelo.

Comparación de coeficientes: limitaciones Varias son las cuestiones que han de tenerse en cuenta al tratar de evaluar, en términos relativos, el contenido informativo que cada variable explicativa tiene sobre la variable dependiente. En primer lugar, podríamos utilizar el hecho de que el coeficiente estimado para Pub en (22) es mayor en valor absoluto que el estimado para P_t para decir que la primera variable es más relevante al explicar las ventas de la empresa. Esto sería incorrecto por dos razones: una de ellas ha sido explicada en la sección anterior, donde hemos visto que en un contexto de colinealidad, un coeficiente individual no puede interpretarse como el impacto que sobre la variable dependiente tiene una variación unitaria en la variable que acompaña a dicho coeficiente estimado. La segunda razón es que, en todo caso, los coeficientes individuales medirían el impacto que sobre las ventas tiene una variación unitaria, positiva o negativa, en cada una de las variables explicativas; el problema es que una variación de una unidad o de 100 unidades puede ser muy grande para una variable, y muy pequeña para otra. Ello dependerá de las variaciones medias que cada una de las variables experimenta a lo largo de la muestra, lo que nos lleva al siguiente epígrafe,

La volatilidad de la variable explicativa Para afirmar que la publicidad es más importante que las ventas porque el valor absoluto del coeficiente estimado para la primera en (22) es mayor que el de la segunda, deberíamos tener en cuenta el tamaño medio de la variación anual media en ambas variables, medido por sus respectivas desviaciones típicas.[25] El efecto promedio de una

[22] Recordemos que $R^2 = 1 - \frac{SR^2}{ST}$, por lo que $SR^2 = (1 - R^2)ST$, y $\hat{\sigma}_u^2 = \frac{SR^2}{T-k} = (1 - R^2)\frac{ST}{T-k} = (1 - R^2)\frac{T}{T-k}Var(y_t)$, por lo que, $Ratio = 1 - \frac{\hat{\sigma}_u}{\sqrt{Var(V_t)}} = 1 - \sqrt{(1 - R^2)\frac{T}{T-k}}$. Para valores grandes de T en relación con k, $Ratio = 1 - \sqrt{(1 - R^2)}$, siendo siempre $Ratio$ inferior a R^2.

[23] Este procedimiento, olvidado por mucho tiempo, ha sido recordado por Johnston y DiNardo (1997).

[24] Recordamos nuevamente que utilizar desviaciones respecto de la media muestral en presencia de tendencias temporales puede no tener mucho significado. Puede conducir, además, a conclusiones erróneas.

[25] De nuevo, esta interpretación es correcta únicamente si las variaciones anuales son independientes entre sí. Sólo en tal caso pueden interpretarse como fluctuaciones alrededor del valor promedio de la variable.

variable explicativa sobre la variable endógena se obtendría multilpicando el coeficiente estimado por la variación media enla variable explicativa. Utilizando los coeficientes estimados en el modelo de regresión múltiple (22), tendríamos,

$$\begin{aligned} Efecto(Pub &\rightarrow Ventas) = 2,204 * 6,20 = 13,665 \\ Efecto(\mathrm{Pr}\, ecio &\rightarrow Ventas) = -1,464 * 5,22 = -7,642 \end{aligned}$$

resultando superior en valor absoluto para los gastos en publicidad que para el precio del producto. Los gastos en publicidad tienen asociado un mayor coeficiente y, además, sus variaciones anuales son de mayor tamaño; ambos efectos se agregan, en este caso, para sugerir que esta variable es la más relevante para explicar las cifras de ventas. Sin embargo, esta impresión está sujeta a la limitación que impone la fuerte colinealidad entre ambas variables explicativas del modelo. [**CHEQUEAR**: Si utilizásemos los coeficientes estimados en las regresiones simples (33) , (32) para estimar el efecto que sobre las Ventas produce una variación de una desviación típica en cada una de las dos variables , explicativas, tendríamos un resultado similar, ya que en dichas regresiones, la estimación numérica del coeficiente incorpora la correlación existente entre la variable incluida y la variable excluida,

$$\begin{aligned} Efecto(Pub &\rightarrow Ventas) = 3,224 * 6,20 = 19,99 \\ Efecto(\mathrm{Pr}\, ecio &\rightarrow Ventas) = -3,637 * 5,22 = -18,98 \end{aligned}$$

que como se ve son, efectivamente, muy similares].

Comparación de residuos También podemos examinar la matriz de correlaciones entre la variable V_t y los residuos procedentes de las dos regresiones simples anteriores. El criterio es que cuanto más se distinga el residuo de la variable dependiente, mayor es la capacidad explicativa del modelo. En la tabla de correlaciones [$CORRELACIONES$] se aprecia que las ventas tienen menor correlación con el residuo de la regresión sobre publicidad (0,312) que con el residuo de la regresión sobre el precio del producto (0,434), lo que sugiere nuevamente el mayor contenido informativo de los gastos en publicidad para explicar las Ventas. En todo caso, la correlación con los residuos de la regresión que utiliza ambas variables como explicativas es sensiblemente inferior (0,238), sugiriendo que ambas variables tienen contenido informativo no trivial.

Es asimismo importante comparar los residuos del modelo de regresión múltiple con los que se derivan de la estimación de cada uno de los modelos de regresión simple. La misma tabla de correlaciones en el fichero de trabajo nos muestra que si excluimos los gastos en publicidad del modelo de regresión múltiple, el coeficiente de correlación entre ambos conjuntos de residuos es de 0,548, mientras que si excluimos la variable precio de la regresión múltiple, el coeficiente de correlación entre residuos es de 0,761. Esto significa que la exclusión de los gastos en publicidad altera más los residuos del modelo que la omisión de la variable Precio, lo que sugiere que la primera es la variable más relevante, igual conclusión a la que alcanzamos en el párrafo anterior.

Las limitaciones de comparar estadísticos tipo t de Student Es muy habitual en el trabajo empírico en Economía juzgar la capacidad explicativa de cada variable en términos relativos, de acuerdo con los valores numéricos de sus respectivos estadísticos tipo-t. Ello parece deberse a que la

comparación entre el valor absoluto de dicho estadístico y el umbral crítico de 2,0 es el procedimiento habitual para analizar si una variable contiene capacidad explicativa sobre la variable endógena. Es decir, la significación estadística de una variable se interpreta directamente como su capacidad de explicar los valores numéricos de la variable dependiente. De dicho procedimiento parece inferirse que cuanto mayor sea el valor absoluto del estadístico tipo-t, mayor es la capacidad explicativa de la variable en cuestión.

Este procedimiento es generalmente inapropiado porque el economista está interesado en el impacto cuantitativo que cambios en una variable explicativa implican sobre la variable dependiente, y la significación estadística de una variable explicativa puede ser simultánea con un efecto cuantitativo muy reducido de dicha variable sobre la variable dependiente. Esta confusión entre significación estadística y relevancia cuantitativa ha sido muy dañina en la ionterpretación de las estimaciones de modelos de regresión en Economía.

Hay diversas razones por las que una variable explicativa que tiene un notable efecto cuantitativo sobre la variable dependiente puede resultar estadísticamente no significativa. Dada la estructura del estadístico tipo-t, éste conjuga la estimación del impacto numérico de cambios en la variable explicativa, con la precisión con que dicho impacto se estima. Así, la ausencia de significación estadística puede surgir bien porque el impacto numérico de dicha variable es muy reducido, o porque, siendo importante, no se mide con suficiente precisión, es decir, con una varianza suficientemente pequeña. De este modo, no puede interpretarse un valor pequeño del estadístico tipo-t como evidencia de una reducida capacidad explicativa; en particular, un valor de dicho estadístico inferior a 2 no necesariamente significa que la variable explicativa en cuestión no tenga contenido informativo sobre la variable dependiente. Ignorar el papel que la precisión en la estimación de un coeficiente del modelo de regresión tiene sobre los contrastes de significación estadística es la segunda fuente tradicional de error en la interpretación de las estimacioens de modelos de relación entre variables económicas.

6.2 Grado de ajuste del modelo de regresión lineal múltiple

Una vez obtenidas las estimaciones numéricas MCO de los coeficientes del modelo, los valores explicados de la variable endógena, $\hat{y}_i$, así como los residuos $\hat{u}_i$, se definen de modo análogo a como hicimos en el modelo lineal simple.

Estimamos la varianza del término de error mediante la varianza residual: $\hat{\sigma}_u^2 = \frac{SCR}{n-k} = \frac{\sum \hat{u}_i^2}{n-k} = \frac{S^2_{y.x_1x_2}}{n-k}$. En el modelo de la sección anterior, $k = 3$. Hemos introducido la notación $S^2_{y.x_1x_2}$ para denotar la Suma de Cuadrados de Residuos en la regresión que tiene a y como variable dependiente, y a x_{1i} y x_{2i} como variables explicativas. El Error Estándar de la Regresión (EER) es la raíz cuadrada de dicha estimación. Al ser una estimación de la desviación típica residual, es un indicador del tamaño medio de los residuos.

La Suma de Cuadrados de los Residuos puede calcularse sin necesidad de obtener los residuos, mediante:

$$\begin{aligned} SCR &= S^2_{y.x_1x_2} = \sum \hat{u}_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i})\hat{u}_i = \\ &= \sum y_i \hat{u}_i - \hat{\beta}_0 \sum \hat{u}_i - \hat{\beta}_1 \sum x_{1i}\hat{u}_i - \hat{\beta}_2 \sum x_{2i}\hat{u}_i = \\ &= \sum y_i \hat{u}_i = \sum y_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}) = \\ &= \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum y_i x_{1i} - \hat{\beta}_2 \sum y_i x_{2i} \end{aligned}$$

análoga a la que obtuvimos en el modelo lineal simple.

Al igual que ocurría en el modelo de regresión simple, el EER es una medida de ajuste del modelo, pero es una medida relativa, que es útil sólo para comparar modelos alternativos que tratan de explicar una misma variable dependiente y, pero no puede utilizarse para comparar el grado de ajuste de modelos elaborados para explicar variables dependientes diferentes.

Siguiendo un argumento totalmente similar al utilizado en el modelo de regresión simple, puede obtenerse la descomposición de la Suma Total:

$$\begin{aligned} ST &= S_y^2 = \sum (y_i - \bar{y})^2 = \sum \left[(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \right]^2 = \\ &= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2 \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \end{aligned}$$

pero el último sumando es igual a cero, porque $y_i - \hat{y}_i = \hat{u}_i$, de modo que:

$$\begin{aligned} \sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum \hat{u}_i \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} - \bar{y} \right) = \\ &= \hat{\beta}_0 \sum \hat{u}_i + \hat{\beta}_1 \sum \hat{u}_i x_{1i} + \hat{\beta}_2 \sum \hat{u}_i x_{2i} - \bar{y} \sum \hat{u}_i \end{aligned}$$

donde todos los sumandos son igual a cero. Por tanto, la Suma Total puede escribirse:

$$S_y^2 = \sum \hat{u}_i^2 + \sum (\hat{y}_i - \bar{y})^2 = S^2_{y.x_1x_2} + \sum (\hat{y}_i - \bar{y})^2$$

y tenemos nuevamente la descomposición:

$$Suma\ Total = Suma\ Cuadrados\ \text{Re}siduos + Suma\ Explicada$$

De este modo, también en este modelo tenemos que la Suma de Cuadrados de Residuos, o Suma Residual, puede interpretarse como el porcentaje de la variación total en la variable dependiente que no viene explicada por las variables independientes. Por tanto, el coeficiente de determinación está comprendido entre 0 y 1, y es tanto más cercano a 1 cuanto mayor sea la capacidad de las variables independientes: 1, x_{1i} y x_{2i}, para explicar las fluctuaciones en la variable dependiente, y.

Por tanto, una medida absoluta de ajuste del modelo es, nuevamente, el coeficiente de determinación múltiple, definido ahora como:

$$R^2_{y.x_1x_2} = 1 - \frac{S^2_{y.x_1x_2}}{S_y^2}$$

El coeficiente de determinación puede también calcularse sin necesidad de obtener previamente los residuos, sino utilizando únicamente momentos muestrales:

$$R^2_{y.x_1x_2} = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = \frac{\hat{\beta}_0 \sum y_i + \hat{\beta}_1 \sum y_i x_{1i} + \hat{\beta}_2 \sum y_i x_{2i} - n\bar{y}^2}{\sum (y_i - \bar{y})^2} = \frac{\hat{\beta}_1 S_{x_1 y} + \hat{\beta}_2 S_{x_2 y}}{S_y^2}$$

a la que se llega utilizando la expresión $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$.

6.3 Coeficiente de determinación ajustado

Al añadir una variable explicativa adicional a un modelo de regresión, la capacidad explicativa de éste non puede disminuir. Si se asigna un coeficiente cero a dicha varaible, la capaciadad explicativa del nuevo modelo sería, evidentemente, idéntica a la del modelo que no incorpora a dicha variable. Como al estimar por mínimos cuadrados buscamos la menor Suma de Cuadrados de residuos posibles, ésta debera ser, como máximo, igual a la del modelo sin la variable explicativa. En general, será siempre inferior, ya sea por poco o por mucho.

En consecuencia, el coefieicnte de determinación nunca disminuye al añadir variables explicativas a un model de regresión. Por tanto, no está justificado comparar los valores del estadístico R^2 de un modelo con el que se obtiene al añadir nuevas variables explicativas, puesto que éste ultimo siempre será superior. Para hacerlos comparables, se introduce una corrección en el cálculo del coeficiente de determinación,

$$1 - \bar{R}^2 = \left(1 - R^2\right) \frac{n-1}{n-k}.$$

Como se ve, al añadir variables explicativas, el primer factor de la derecha siempre se reducirá, mientras que el segundo factor aumentará. Si el aumento en el cofiiciente de determinación es suficientemente importante, predominara este efecto, y preferiremos el modelo ampliado, mientras que lo contrario sucederá si dicho aumento no es muy notable.

Por otra parte, si el número de observaciones, n,es muy elevado, entonces el segundo factor será prácticamente igual a 1,y el efecto sobre él de incrementar en uno el número de variables explicativas será imperceptible, por lo que siempre preferiríamos el modelo con más variables explicativas, lo que no parece muy razonable.

Recordemos que el *coeficiente de determinación corregido* se define:

$$(1 - \bar{R}^2) = (1 - R^2)\frac{N-1}{N-k}$$

Si el coeficiente de determinación estándar es muy pequeño, el coeficiente de determinación corregido puede ser negativo. En efecto, es fácil ver en la expresión que lo define que si $R^2 < k/(n-1)$,entonces $\bar{R}^2 < 0$.

6.3.1 Ejemplo: peso de bebés recién nacidos

Contenido informativo en una variable, adicional al que ya está incorproado en el modelo El lector puede proceder a continuación a comprobar que ninguna de las dos variables de educación aporta capacidad explicativa adicional. Para ello, puede utilizar cada una de ellas por sí sólas como variable explicativa en una regresión y constatar: *a*) el reducido valor del R^2, y del *Ratio* de ajuste, así como *b*) la elevada correlación entre los residuos resultantes y la variable *Peso* original, tal como hemos hecho con *cigarrillos* y *renta; c)* alternativamente, puede añadir uno de los niveles educativos como variable explicativa a una cualquiera de las regresiones anteriores y comprobar

que el coeficiente de correlación entre los residuos de ambas regresiones es muy elevado; algunas de estas correlaciones aparecen en $CORR\_PESO\_AJUSTE$, significando que la inclusión del nivel educativo de los padres no ha añadido capacidad explicativa significativa a las variables *cigarrillos* y *renta*. Las nubes de puntos $COMP\_RES\_1$ a $COMP\_RES\_4$ aportan una evidencia similar: $COMP\_RES\_1$ y $COMP\_RES\_2$ relacionan la variable dependiente, *Peso*, con los residuos de dos regresiones simples, la primera utilizando *cigarrillos* como unica variable explicativa, mientras la segunda utiliza *Renta* en su lugar. $COMP\_RES\_3$ compara los residuos de la regresión simple sobre *cigarrillos*, con los que se obtienen de una regresión que incluye *cigarrillos* y *renta* como variables explicativas; el hecho de que ambos residuos sean tan similares sugiere que la variable *renta* no aporta mucho a la variable *cigarrillos* para explicar el peso del recién nacido.[26] $COMP\_RES\_4$ compara los residuos de la regresión simple sobre *cigarrillos*, con los que se obtienen de una regresión que incluye *cigarrillos, renta,* los niveles educativos del padre y de la madre, y el número de orden del recién nacido entre sus hermanos como variables explicativas; la interpretación es similar, y no parece que el resto de las variables aporte mucha información a la que pueda incorporar el número de cigarrillos.

Para profundizar en la información proporcionada por los niveles educativos, y dada la excesiva concentración de cada una de estas dos variables en el nivel 12 años, definimos una variable ficticia en el caso de las mujeres, *edm*, que es igual a 0 si *educm* es inferior a 12 años, es igual a 1 si *educm* es igual a 12 años, y es igual a 2 si *educm* toma cualquier valor numérico superior a 12 años. En ocasiones, es difícil medir con precisión el efecto de cambios unitarios en una variable como *educm*, pero se mide mejor el efecto que tiene sobre la variable dependiente el paso de un nivel de *educm* a otro. Aunque no incidimos aquí en los resultados, la variable ficticia así construida, que se incluye en el archivo de trabajo, no parece aportar capacidad explicativa significativa. Finalmente, concluimos que los niveles educativos no son relevantes para explicar el peso de los recién nacidos, una vez que se tiene en cuenta la información proporcionada por *cigarrillos*. Algo similar puede decirse del nivel educativo del padre.

Cuando se considera la variable *ordenac*, la escasa contribución informativa es aún más evidente, como ya sugería el análisis descriptivo que antes hicimos, por lo que concluimos que esta variable no aporta información relevante a la ya proporcionada por las variables *renta* y *cigarrillos*. Esto ocurre a pesar de que esta variable aparece con un valor numérico del estadístico t de Student superior a 2 en la regresión que incluye todas las variables explicativas [REG_TODAS], propiedad que se mantiene si excluimos de dicha regresión todas las variables explicativas con estadístico t inferior a 2 en valor absoluto, y volvemos a estimar el modelo. Si siguiéramos este procedimiento, habitual en el análisis empírico, pero en absoluto recomendable, nos quedaríamos con una regresión que utiliza *cigarrillos* y *ordenac* como únicas variables explicativas [$REG\_CIGS\_ORDENAC$]. Sin embargo, la correlación entre los residuos de esta regresión [RES_CIGS_ORDEN] y los que utiliza únicamente la variable *cigarrillos* como explicativa [RES_CIGS] es superior a 0,997, indicando que *ordenac* apenas añade información a la que pueda incluir la variable *cigarrillos*.

Finalmente, si el investigador decidiera utilizar todas las variables simultáneamente, como hicimos en la regresión mostrada en primer lugar, obtendría unos residuos muy altamente correlacionados con los de las regresiones previas, así como con la variable *Peso* original [ver la correlación entre RES_TODAS (el residuo de la regresión con todas las variables explicativas) y RES_CIGS, RES_RENTA, en la tabla $CORR\_PESO\_AJUSTE$, así como la nube de puntos

[26] Nótese que esto no significa en modo alguno que cigarrillos tenga más o menos capacidad explicativa, sino tan sólo que la información proporcionada por la renta familiar no aporta nada a la contenida en el número de cigarillos fumado porla madre durante el embarazo, que podría ser relevante o no serlo.

$COMP\_RES\_4$]. Nuevamente, la interpretación es la misma, en términos de la reducida capacidad explicativa del conjunto de variables considerado, como perfectamente ilustra FIG_RES_TODAS.

En general, el ejemplo que estamos considerando ilustra la necesidad de huir de la aplicación mecánica de los estadísticos tipo t de Student. A pesar del elevado valor numérico de este estadístico, especialmente en las regresiones individuales, la única conclusión razonable en el análisis que hemos presentado, es que ninguna de las variables, tal como aparece recogida en la muestra, explica de manera importante el comportamiento del peso de los recién nacidos[27]. Por ejemplo, la regresión REG_EDUCP_EDUCM, que explica el peso del recién nacido utilizando únicamente por los niveles educativos de los padres únicamente, también genera un estadístico t superior a 2,0 en valor absoluto para la variable $EDUCP$, sin que de ello deba inferirse que esta variable aporta capacidad explicativa alguna, ni siquiera cuando se utiliza por sí sóla, como ya hemos discutido ampliamente.

También es interesante observar que el estadístico tipo F habitual para el contraste de significación global del modelo, es decir, para contrastar la hipótesis nula que afirma que las variables explicativas, consideradas conjuntamente, no aportan capacidad explicativa alguna, arroja un valor numérico de 9,55, con un valor-p igual a 0, por lo que una interpretación estricta del mismo conduciría a admitir la capacidad explicativa conjunta de las variables consideradas acerca del peso de los recién nacidos, contrariamente a las conclusiones que hemos obtenido.

Sin embargo, un investigador todavía debería pronunciarse acerca de la posible evidencia existente en la información muestral sobre la influencia que las distintas variables consideradas pueden tener sobre el peso del recién nacido. En este sentido, si consideramos los cigarillos fumados durante el embarazo, la diferencia entre las medianas que antes mencionamos para los pesos de los bebés nacidos de mujeres no fumadoras y de mujeres fumadoras es notable, siendo menor la mediana del peso para los hijos de mujeres fumadoras, lo que sugiere una relación negativa entre estas dos variables, como quizá cabría esperar. Esta es la única variable recogida en la muestra para la que detectamos un efecto significativo; los datos disponibles sugieren que el consumo de cigarrillos durante el embarazo tiende a disminuir significativamente el peso de los bebés al nacer, lo que ocurre es que la información muestral no nos permite estimar con precisión las variaciones en peso producidas por cada incremento en el número de cigarrillos fumados por la madre durante el embarazo.

Hay otros aspectos, potencialmente relevantes, que no hemos considerado en la discusión previa: los residuos de la regresión más completa, REG_TODAS, tienen una media de -3,70 para los recien nacidos de raza no blanca (186 bebés), y de 0,68 para los de raza blanca (1.005 bebés). Esto está en consonancia con la posibilidad de que los bebés de raza blanca tengan más peso. Dichos residuos tienen media de -1,92 para las mujeres, y de 1,78 para los varones, sugiriendo asimismo que los varones puden tener un peso al nacer mayor que el de las mujeres. Ambos efectos son además acordes a la intuición, por lo que procede analizarlos en algún detalle.

Al incluir ambas variables ficticias junto con las cinco variables antes analizadas, el R^2 de la regresión aumenta apreciablemente, a 0,049, a la vez que la desviación típica residual se reduce a 19,65, y el Ratio de ajuste se eleva a 3,4%. Si restringimos el modelo a incluir las dos variables ficticias, de sexo y raza, junto con *cigarrillos*, la regresión apenas varía, con residuos muy altamente correlacionados con los obtenidos en todas las regresiones consideradas, un R^2 de 0,042, desviación típica de 19,92, y *Ratio* de ajuste de 2,1%.

Esta última [REG_CIG_FIC] es, sin embargo, quizá la regresión más razonable,

[27] Por supuesto, la ilustración en Wooldridge (2001) acerca de la ausencia de capacidad explicativa de las dos variables de educación es cierta. Sin embargo, el resultado es aún más estricto, por cuanto que tampoco las variables *renta, cigarrillos, ordenac*, tienen verdaderamente una capacidad explicativa de gran significación.

$$
\begin{aligned}
Peso_i &= \underset{(1,306)}{113,277} - \underset{(0,090)}{0,506} cigarillos_i + \underset{(1,071)}{3,052} male_i + \underset{(1,301)}{6,230} white_i + u_i \; i = 1, 2, ..., N \\
\bar{R}^2 &= 0,044, \; \hat{\sigma}_u = 19,925
\end{aligned}
$$

cuyos residuos aparecen en $FIG\_RES\_CIG\_FIC$. El coeficiente estimado para la variable ficticia *WHITE* es de 6,23, siendo de 3,05 el coeficiente estimado para *MALE*. El coeficiente estimado para *WHITE* está en línea con la diferencia observada entre el promedio de los residuos correspondientes a bebés de cada grupo racial en la regresión que no incluía esta variable explicativa. De igual modo, el coeficiente estimado para *MALE* es muy similar a la diferencia entre los residuos de varones y mujeres en la regresión que no incluía esta variable explicativa.

Simulación a partir de la regresión estimada Los valores numéricos mencionados sugieren que un bebé de raza blanca pesa, en promedio, 6,23 onzas más que un bebé de otra raza de iguales características en lo relativo a: número de cigarros fumados por la madre durante el embarazo, renta familiar, niveles educativos del padre y la madre, y orden del bebé dentro de los hijos de la familia. La diferencia de peso estimada entre un niño de raza blanca al nacer, y una niña de raza negra, es de 9,28 onzas, comparable a la que obtuvimos antes entre los bebés de madres fumadoras y no fumadoras. En la hoja de cálculo Bwght.xls se muestra cómo la mediana del número de cigarrillos fumados por la madre durante el embarazo, entre el conjunto de las observaciones en que dicho número es no nulo, es de 10 cigarillos. Ello significa que, de acuerdo con la regresión anterior, la diferencia de peso entre distintos bebés sería,

Pesos estimados

Características recién nacido	*Peso*
Madre fumadora, mujer, no blanca	108,217
Madre fumadora, varón, no blanco	111,269
Madre no fumadora, mujer, no blanca	113,277
Madre fumadora, mujer, blanca	114,447
Madre no fumadora, varón, no blanco	116,329
Madre fumadora, varón, blanco	117,499
Madre no fumadora, mujer, blanca	119,507
Madre no fumadora, varón, blanco	122,559

Como puede apreciarse, la diferencia entre los pesos estimados de dos recién nacidos que sólo difieren en que su madre declarase ser fumadora, es siempre de 5 onzas, procedente del producto de 10 cigarrillos, escogido como representativo del nivel de tabaco consumido diariamente, por el coeficiente estimado en la regresión. Esta diferencia estimada es inferior a la diferencia de 9 onzas entre las medianas de los pesos para ambos grupos de madres.

6.4 Ejemplo: Discriminación salarial

6.4.1 Capacidad explicativa adicional

Ninguna de las dos variables consideradas parecen tener, por sí sólas, gran capacidad explicativa sobre el salario. Si estimamos una regresión con ambas, tenemos un R^2 ajustado de 0,344, con $\hat{\sigma}_u = 145,4$, y ratio de ajuste $1 - \frac{\hat{\sigma}_u}{\sigma_y} = 0,19$.

$$\begin{aligned} Salario_i &= 43,31 + \underset{(3,29)}{\underset{(23,7)}{77,87}}\ Educación_i + \underset{(0,39)}{\underset{(20,8)}{8,10}}\ Experiencia_i + u_i, \qquad (23) \\ R^2 &= 0,345,\ \bar{R}^2 = 0,344,\ \hat{\sigma}_u = 145,45,\ SR = 31079583; \end{aligned}$$

Es interesante que los dos coeficientes estimados difieren apreciablemente de los obtenidos en las regresiones individuales, siendo en ambos casos superiores a los estimados en dichas regresiones simples. Ahora, estimamos que el salario aumenta en casi 78 Bef. por año de educación, y en más de 8 Bef. por año de experiencia.

Que los coeficientes hayan aumentado respecto de la regresión simple, sugiere que ambas variables estén negativamente correlacionadas; en efecto, si así fuera, un año más de educación vendría generalmente asociado con una menor experiencia laboral, como por otra parte parece razonable. De hecho, el coeficiente de correlación habitual entre ambas variables es -0,29, reflejando tal correlación negativa si bien, tratándose de variables cualitativas, el uso de dicho estadistico es cuestionable. Los niveles de experiencia medios son de 25,9, 20,7, 17,0, 14,9 y 14,5 años para los niveles educativos de 1 a 5, respectivamente, con una media global de 17,2 años. La ordenación decreciente de dichos promedios sugiere asimismo la correlación negativa entre ambas variables.

6.4.2 ¿Aporta la variable Experiencia información acerca de la determinación salarial, adicional a la que continen el nivel educativo y el sexo del trabajador?

Aunque los resultados parecen claros, un investigador podría dudar de la verdadera relevancia de la información aportada por la variable experiencia sobre los salarios, pues la evidencia proporcionada en este sentido por el modelo de regresión simple, así como por los estadísticos descriptivos podría sugerirlo. Para ello, estimaríamos el modelo omitiendo la experiencia laboral y, utilizando, por tanto, el nivel educativo y la variable ficticia para explicar los salarios,

$$Salario = \beta_0 + \beta_1 Experiencia + \beta_2 Educación + \beta_3 Male$$

y compararíamos los residuos de esta ecuación restringida, con los de (37). Además del claro descenso en los estadísticos de ajuste (*reg_ w_ edu_ male*), la nube de puntos ($SIGNIFICA\_EXPER$) que representa ambos conjuntos de residuos indica claramente las diferencias entre ellos o, lo que es lo mismo, que la experiencia profesional es una variable significativa para explicar la determinación de los salarios en este grupo de trabajadores. El coeficiente de correlación entre los residuos de ambos modelos, incluyendo y excluyendo la variable *Experiencia* es 0,88, claramente por debajo de 1,0, en coherencia con la interpretación de que la Experiencia aporta contenido informativo que no está incorporado en la variable .

6.5 Ejemplo 15.1

6.6 Relación entre estimadores de Mínimos Cuadrados en la regresión simple y la regresión múltiple

En las expresiones del estimador de mínimos cuadrados de los coeficientes del modelo de regresión múltiple:

$$\hat{\beta}_1 = \frac{S_{x_1y}S^2_{x_2} - S_{x_2y}S_{x_1x_2}}{S^2_{x_1}S^2_{x_2} - (S_{x_1x_2})^2};\ \hat{\beta}_2 = \frac{S_{x_2y}S^2_{x_1} - S_{x_1y}S_{x_1x_2}}{S^2_{x_1}S^2_{x_2} - (S_{x_1x_2})^2} \tag{24}$$

observamos que la estimación del efecto que sobre y tiene una variación unitaria en una de las variables explicativas difiere de la estimación que de dicho efecto tendríamos estimando un modelo de regresión simple. En el modelo de regresión simple, el estimador de mínimos cuadrados del coeficiente asociado a la única variable explicativa sería igual, como sabemos, al cociente entre la covarianza de las variables dependiente y explicativa, y la varianza de la variable explicativa de la regresión.

Sin embargo, existe un caso especial en que ambos estimadores coinciden. Supongamos que las dos variables explicativas del modelo, x_{21} y x_{3i} están incorrelacionadas. En tal caso, $S_{x_1x_2} = 0$, y tendríamos en el modelo de regresión múltiple las expresiones:

$$\hat{\beta}_1 = \frac{S_{x_1y}S^2_{x_2}}{S^2_{x_1}S^2_{x_2}} = \frac{S_{x_1y}}{S^2_{x_1}};\ \hat{\beta}_2 = \frac{S_{x_2y}S^2_{x_1}}{S^2_{x_1}S^2_{x_2}} = \frac{S_{x_2y}}{S^2_{x_2}}$$

que coinciden con los estimadores que de los efectos individuales δ_1 y α_1 tendríamos en las regresiones simples:

$$y_i = \delta_0 + \delta_1 x_{1i} + u_i \tag{25}$$
$$y_i = \alpha_0 + \alpha_1 x_{21} + v_i \tag{26}$$

El resultado es generalizable: Supongamos que estamos interesados en una regresión con k variables explicativas, y que podemos dividir las variables explicativas en dos grupos, con q y $k-q$ variables cada uno, teniendo cada variable del primer grupo correlación nula con cada una de las variables del segundo grupo. Dentro de cada grupo, las correlaciones pueden ser arbitrarias. Entonces, las estimaciones de los coeficientes del primer grupo de variables serán las mismas en el modelo completo que en el modelo que utilizase únicamente dichas variables como explicativas, y lo mismo puede decirse de las estimaciones de los coeficientes del segundo grupo de variables explicativas.

Consideremos ahora la situación más habitual, en que las variables explicativas tienen una correlacion no nula. A partir de (24) es fácil obtener, dividiendo por $S^2_{x_1}S^2_{x_2}$ en cada una de ellas:

$$\hat{\beta}_1 = \frac{\hat{\beta}_{y1} - \hat{\beta}_{y2}\hat{\beta}_{21}}{1 - \hat{\beta}_{12}\hat{\beta}_{21}};\ \hat{\beta}_2 = \frac{\hat{\beta}_{y2} - \hat{\beta}_{y1}\hat{\beta}_{12}}{1 - \hat{\beta}_{12}\hat{\beta}_{21}}$$

donde $\hat{\beta}_{y1}$ denota el estimador de mínimos cuadrados de la pendiente en una regresión de y sobre x_1, $\hat{\beta}_{12}$ es el estimador de la pendiente en una regresión de x_1 sobre x_2, y análogamente los demás parámetros. Dado que una correlación nula entre dos variables implica que la estimación de mínimos cuadrados de la pendiente en la regresión entre ambas será igual a cero, es nuevamente sencillo ver la equivalencia entre el estimador del efecto individual y del efecto global sobre y de un cambio en x_1 en ausencia de correlación entre x_1 y x_2.En este caso, $\hat{\beta}_{21} = \hat{\beta}_{12} = 0$, por lo que $\hat{\beta}_1 = \hat{\beta}_{y1}$y $\hat{\beta}_2 = \hat{\beta}_{y2}$.

Una tercera representación de los estimadores de mínimos cuadrados del modelo de regresión múltiple, en función de los coeficientes de correlación entre las distintas variables del modelo y sus volatilidades relativas. Si dividimos por $S^2_{x_1}S^2_{x_2}$ en el numerador y denominador de $\hat{\beta}_1$, tenemos:

$$\hat{\beta}_1 = \frac{\frac{S_{x_1 y}}{S_{x_1}^2} - \frac{S_{x_2 y} S_{x_1 x_2}}{S_{x_1}^2 S_{x_2}^2}}{1 - \frac{\left(S_{x_1 x_2}\right)^2}{S_{x_1}^2 S_{x_2}^2}} = \frac{\frac{S_{x_1 y}}{S_{x_1}^2}\frac{\sqrt{S_y^2}}{\sqrt{S_y^2}} - \frac{S_{x_2 y} S_{x_1 x_2}}{S_{x_1}^2 S_{x_2}^2}\frac{\sqrt{S_y^2}}{\sqrt{S_y^2}}}{1 - \rho_{12}^2} = \frac{\rho_{y1} - \rho_{y2}\rho_{12}}{1 - \rho_{12}^2}\frac{S_y}{S_{x_1}}$$

y de modo similar, pude obtenerse una expresión análoga para $\hat{\beta}_2$:

$$\hat{\beta}_2 = \frac{\rho_{y2} - \rho_{y1}\rho_{12}}{1 - \rho_{12}^2}\frac{S_y}{S_{x_2}}$$

donde los subíndices tienen la misma interpretación que en las expresiones anteriores.

6.7 Coeficientes de correlación (o de determinación) y estadísticos t

El coeficiente de correlación guarda una estrecha relación con los estadísticos t de las variables. Asi, si contrastamos la significación estadística de la única pendiente estimada en un modelo de regresion simple, se tiene:

$$F_{1,n-2} = \frac{R^2}{1 - R^2}\frac{n-2}{1}$$

Pero una distribución F cuyo primer grado d elibertad es igual a 1, es igual al cuadrado de una distribución t de Student con el segundo número de grados de libertad de la distribución F : $F_{1,n-2} = t_{n-2}^2$.

Por tanto,

$$R^2 = \frac{t_{n-2}^2}{t_{n-2}^2 + (n-2)}$$

El resultado se mantiene, por supuesto, cuando la regresion simple que estimamos es una regresion parcial, en la que hemos descontado tanto de y como de x_1 el efecto que sobre ambas variables tiene x_2, y tenemos:

$$\rho_{yx_1.x_2}^2 = \frac{t_1^2}{t_1^2 + (n-k)}$$

donde t_1 es el estadístico t de la variable x_1 en el modelo global, y siendo k el número de coeficientes estimados en dicho modelo, que es igual a 3 en este ejemplo. En la regresion parcial, no sólo el coeficientes estimado para x_1 sino también su desviación típica estimada, son los mismos que en la regresión original. Por eso se cumple la relación anterior.

Si tenemos un alto número de variables explicativas, supongamos $k = 5$, y queremos calcular el coeficiente de correlación parcial $\rho_{yx_2.x_1x_3x_4}$,deberíamos estimar una regresión de y sobre x_1, x_2, x_3, x_4, y utilizar el estadistico t de la variable x_2 en dicha regresión para calcular:

$$\rho_{yx_2.x_1x_5x_6}^2 = \frac{t_2^2}{t_2^2 + (n-5)}$$

6.7.1 Aplicación: Adición de variables a un modelo de regresión

El coeficiente de correlación parcial es una herramienta útil para decidir si añadir una variable explicativa adicional a un modelo estimado. Supongamos que hemos estimado el modelo:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2t} + \beta_3 x_{3t} + u_i$$

y nos planteamos la conveniencia de añadir una variable explicativa x_{4i}. Es claro que en dicha decisión no debemos guiarnos del coeficiente de correlación simple habitual ρ_{yx_4}. Este coeficiente nos mediría toda la información que x_4 contiene sobre y. Pero una parte de dicha información, ya está presente en el modelo, a través de x_1, x_2, x_3. Por tanto, queremos decidir sobre la base del coeficiente de correlación parcial: $\rho_{yx_4.x_1x_2x_3}$ que, como hemos visto en la sección anterior, guarda una estrecha relación con el estadístico t del coeficiente de la variable x_4 en la regresión múltiple.

Supongamos que hemos estimado un modelo de regresión, con una estimación de la varianza del término de error igual a:

$$\hat{\sigma}_{SR} = \frac{SCR_{SR}}{n-k}$$

y consideramos la posibilidad de excluir del modelo r variables explicativas. Si lo hacemos así, y volvemos a estimar, tendremos:

$$\hat{\sigma}_R = \frac{SCR_R}{n-k+r}$$

El contraste F de significación conjunta de loscoficientes de dichas r variables es:

$$F_{r,n-k} = \frac{SCR_R - SCR_{SR}}{SCR_{SR}} \frac{n-k}{r} = \frac{(n-k+r)\hat{\sigma}_R - (n-k)\hat{\sigma}_{SR}}{(n-k)\hat{\sigma}_{SR}} \frac{n-k}{r}$$

de modo que:

$$\frac{\hat{\sigma}_R}{\hat{\sigma}_{SR}} = \frac{F + \frac{n-k}{r}}{1 + \frac{n-k}{r}}$$

por lo que:

$$\hat{\sigma}_R < \hat{\sigma}_{SR} \Leftrightarrow F < 1$$

pero: $\hat{\sigma}_R = \frac{(1-\bar{R}^2)S_y^2}{n-1}$, de modo que $\hat{\sigma}_R$ disminuye si y sólo si $\bar{R}^2$ aumenta, puesto que S_y^2 y n están dados y no dependen del número de variables explicativas que incluyamos en el modelo de regresión.

6.8 Estimación de efectos individuales en una regresión múltiple

Recordemos la interpretación del residuo de una regresión como el componente de la variable dependiente y no explicado por la variable dependiente x. En el caso de una regresión múltiple, es el componente de y no explicado por el conjunto de las variables explicativas. Por simplicidad, vamos a considerar una regresión múltiple con dos variables explicativas, x_1 y x_2, aunque el argumento puede extenderse al caso más general en que x_1 y x_2 son vectores, es decir, están formados por varias variables.

La primera proposición muestra que el coeficiente de un modelo de regresión múltiple puede obtenerse mediante una regresión entre residuos de regresiones simples. Ya sabemos que salvo en el caso en que x_1 y x_2 estuviesen incorrelacionadas, el coeficiente de x_1 en una regresión de y sobre x_1 y x_2 está afectado por la presencia de x_2 en dicha regresión. Por tanto, va a ser distinto del coeficiente estimado en la regresión simple de y sobre x_1.Sin embargo, el coeficiente de x_1 en la regresión múltiple puede obtenerse en una regresión del componente de y no explicado por x_2 sobre le componente de x_1 no explicado por x_2.

Proposition 5 *El coeficiente de x_{1i} en la regresión múltiple, β_1, puede estimarse mediante la regresión del componente de y no explicado por x_2, $(\hat{\eta}_i)$ sobre el componente de x_1 no explicado por x_2, $(\hat{\xi}_i)$.*

Consideremos nuevamente la regresión:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i,\ i = 1, 2, ...n$$

y estimamos las regresiones auxiliares que explican tanto a y como a x_1 mediante x_2:

$$\begin{aligned} y_i &= \alpha_0 + \alpha_1 x_{2i} + \eta_i \\ x_{1i} &= \delta_0 + \delta_1 x_{2i} + \xi_i \end{aligned}$$

quedándonos con los residuos:

$$\begin{aligned} \hat{\eta}_i &= y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_{2i} \\ \hat{\xi}_i &= x_{1i} - \hat{\delta}_0 - \hat{\delta}_1 x_{2i} \end{aligned}$$

El residuo $\hat{\eta}_i$ es el componente de y_i no explicado por x_{2i}, mientras que el residuo $\hat{\xi}_i$ es el componente de x_{1i} no explicado por x_{2i}. Nótese que:

$$\begin{aligned} \hat{\eta}_i &= (y_i - \bar{y}) - \frac{Cov(y_i, x_{2i})}{Var(x_{2i})}(x_{2i} - \bar{x}_2) = (y_i - \bar{y}) - \frac{S_{x_2 y}}{S^2_{x_2}}(x_{2i} - \bar{x}_2) \\ \hat{\xi}_i &= (x_{1i} - \bar{x}_1) - \frac{Cov(x_{1i}, x_{2i})}{Var(x_{2i})}(x_{2i} - \bar{x}_2) = (x_{1i} - \bar{x}_1) - \frac{S_{x_1 x_2}}{S^2_{x_2}}(x_{2i} - \bar{x}_2) \end{aligned}$$

Ahora, estimamos la regresión de uno sobre el otro:

$$\hat{\eta}_i = \upsilon_0 + \upsilon_1 \hat{\xi}_i + \varepsilon_i$$

obteniendo una estimación de la pendiente:

$$\hat{\upsilon}_1 = \frac{Cov(\hat{\eta}_i, \hat{\xi}_i)}{Var(\hat{\xi}_i)} = \frac{\sum \hat{\eta}_i \hat{\xi}_i}{\sum \hat{\xi}_i^2}$$

debido a que ambos residuos tienen media cero, por lo que sus covarianzas y varianzas se definen mediante productos y cuadrados simples, sin necesidad de restar la media muestral.

Pero:

$$\begin{aligned}\sum \hat{\eta}_i\hat{\xi}_i &= \sum \left((y_i-\bar{y})-\frac{S_{x_2y}}{S^2_{x_2}}(x_{2i}-\bar{x}_2)\right)\left((x_{1i}-\bar{x}_1)-\frac{S_{x_1x_2}}{S^2_{x_2}}(x_{2i}-\bar{x}_2)\right)= \\ &= S_{x_1y}-\frac{S_{x_1x_2}}{S^2_{x_2}}S_{x_2y}-\frac{S_{x_2y}}{S^2_{x_2}}S_{x_2x_1}+\frac{S_{x_2y}}{S^2_{x_2}}\frac{S_{x_1x_2}}{S^2_{x_2}}S^2_{x_2}=S_{x_1y}-\frac{S_{x_1x_2}}{S^2_{x_2}}S_{x_2y}\end{aligned}$$

$$\begin{aligned}\sum \hat{\xi}_i^2 &= \sum \left[(x_{1i}-\bar{x}_1)-\frac{S_{x_1x_2}}{S^2_{x_2}}(x_{2i}-\bar{x}_2)\right]^2=S^2_{x_1}+\left(\frac{S_{x_1x_2}}{S^2_{x_2}}\right)^2S^2_{x_2}-2\frac{S_{x_1x_2}}{S^2_{x_2}}S_{x_1x_2}= \\ &= S^2_{x_1}-\frac{S_{x_1x_2}}{S^2_{x_2}}S_{x_1x_2}\end{aligned}$$

Aunque ahora no lo necesitemos notemos que, por igual razon,

$$\begin{aligned}\sum \hat{\eta}_i^2 &= \sum \left[(y_i-\bar{y})-\frac{S_{yx_2}}{S^2_{x_2}}(x_{2i}-\bar{x}_2)\right]^2=S^2_{y}+\left(\frac{S_{yx_2}}{S^2_{x_2}}\right)^2S^2_{x_2}-2\frac{S_{yx_2}}{S^2_{x_2}}S_{yx_2}= \\ &= S^2_{y}-\frac{(S_{yx_2})^2}{S^2_{x_2}}\end{aligned}$$

Finalmente:

$$\hat{\upsilon}_1=\frac{Cov(\hat{\eta}_i,\hat{\xi}_i)}{Var(\hat{\xi}_i)}=\frac{\sum\hat{\eta}_i\hat{\xi}_i}{\sum\hat{\xi}_i^2}=\frac{S_{x_1y}S^2_{x_2}-S_{x_1x_2}S_{x_2y}}{S^2_{x_1}S^2_{x_2}-S^2_{x_1x_2}}$$

y habremos obtenido el mismo coeficiente que habríamos obtenido en la regresión múltiple.

Como vamos a ver a continuación, en realidad el procedimiento puede simplificarse aún más: ¿Qué pasaría si excluyésemos el efecto de x_{2i} unicamente de x_{1i}, pero no de y?

Proposition 6 *El coeficiente de x_{1i} en la regresión múltiple, β_1, puede estimarse mediante la regresión de y sobre el componente de x_1 no explicado por x_2, $(\hat{\xi}_i)$.*

En ese caso, estimaríamos una regresión de y sobre $\hat{\xi}_i$, el componente de x_{1i} no explicado por x_{2i} :

$$y_i=\gamma_0+\gamma_1\hat{\xi}_i+\varsigma_i$$

obteniendo una estimación de la pendiente:

$$\hat{\gamma}_1=\frac{Cov(y_i,\hat{\xi}_i)}{Var(\hat{\xi}_i)}=\frac{\frac{1}{n}\sum y_i\hat{\xi}_i}{\frac{1}{n}\sum\hat{\xi}_i^2}$$

Pero:

$$\sum y_i\hat{\xi}_i=\sum y_i\left[(x_{1i}-\bar{x}_1)-\frac{Cov(x_{1i},x_{2i})}{Var(x_{2i})}(x_{2i}-\bar{x}_2)\right]=S_{x_1y}-\frac{S_{x_1x_2}}{S^2_{x_2}}S_{x_2y}$$

mientras que, como antes vimos,

$$\sum \hat{\xi}_i^2 = S_{x_1}^2 - \frac{S_{x_1x_2}}{S_{x_2}^2} S_{x_1x_2}$$

por lo que:

$$\hat{\gamma}_1 = \frac{Cov(y_i, \hat{\xi}_i)}{Var(\hat{\xi}_i)} = \frac{S_{x_1y} - \frac{S_{x_1x_2}}{S_{x_2}^2} S_{x_2y}}{S_{x_1}^2 - \frac{S_{x_1x_2}^2}{S_{x_2}^2}} = \frac{S_{x_1y} S_{x_2}^2 - S_{x_1x_2} S_{x_2y}}{S_{x_1}^2 S_{x_2}^2 - S_{x_1x_2}^2}$$

que es el mismo estimador que obtuvimos en la regresión anterior. Sin embargo, las desviaciones típicas del estimador serán diferentes en ambas regresiones.

6.9 Aplicaciones

6.9.1 Extracción de tendencias

Estos resultados se prestan a la siguiente aplicación: Supongamos que tenemos dos variables cuya relación queremos caracterizar, y ambas presentan una clara tendencia temporal. La presencia común de una tendencia generará una impresión de relación entre ambas variables que quiza no se corresponde con la relación que entre ellas existe una vez que descontemos la presencia de dicha tendencia. Si quisiéramos medir esta última correlación, podríamos estimar primero sendas regresiones de y y x_1 sobre una tendencia t, y tomar los residuos de ambas regresiones. Estos podrian interpretarse como el resultado de extraer la tendencia temporal de las variables y y x_1. Por tanto, la correlación entre dichos residuos nos dará la correlación entre las variables y y x_1 que no es debida a la presencia de una tendencia común.

Este procedimiento sería correcto, pero el resultado anterior nos dice que se puede obtener la estimación del coeficiente β con que x_1 influye sobre y de dos maneras más sencillas. Una, estimando una regresión de y sobre el componente que queda en x_1 tras extraer la tendencia de esta variable. Esto evitaría una de las regresiones anteriores. Pero hay un procedimento que en general resultará aún más simple, que consiste en estimar la regresión:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 t + u_i$$

donde, en este caso, no hemos eliminado la tendencia temporal de ninguna de las dos variables. Las dos proposiciones anteriores muestran que la estimación de mínimos cuadrados del parámetro β_1 en esta última regresión sera idéntico al obtenido en cualquiera de las dos regresiones que hemos mencionado al inicio de este párrafo.

6.9.2 Desestacionalización

Algo similar puede aplicarse al caso en que existen componentes estacionales apreciables tanto en y como en x_1. Con datos trimestrales, estos pueden representarse mediante variables ficiticas, una para cada trimestre, que toman el valor 1 en el trimestre correspondiente, y 0 en los restantes. Si estamos interesados en estimar la relación entre ambas variables que no es debida a los factores estacionales, podemos *desestacionalizar* ambas variables, y estimar una regresión entre sus versiones *desestacionalizadas*. ¿Cómo se desestacionaliza? Estimando una regresión de cada variables sobre las 4 variables ficticias trimestrales, y conservando el residuo así obtenido. Alternativamente,

sabemos que bastaría con desestacionalizar x_1. Pero lo que puede ser más sencillo, es que también basta con estimar la regresión:

$$y_i = \beta_1 x_{1i} + \alpha_1 D_{1t} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + u_i$$

donde D_{1t} denota la variable ficticia correspondiente al primer trimestre, y $D_{2t} D_{3t}, D_{4t}$ las correspondientes a las restantes. En esta regresión, hemos eliminado la constante para evitar multicolinealidad exacta por la llamada *trampa de las variables ficticias*. Análogamente, podríamos haber excluido alguna de las variables ficticias, por ejemplo la correspondiente al primer trimestre, estimando,

$$y_i = \beta_0 + \beta_1 x_{1i} + \alpha_2 D_{2t} + \alpha_3 D_{3t} + \alpha_4 D_{4t} + u_i$$

Trabajando con datos mensuales, la estacionalidad se representaría por 12 variables ficticias, una para cada mes. Cada una de ellas tomaria el valor 1 en el mes correspondiente, y el valor 0 en los otros 11 meses del año. Trabajando con datos financieros, tiene gran interés analizar la posible estacionalidad diaria, para detectar posibles pautas sistemáticas que se tengan cada día de la semana. Así, en ocasiones se afirma que hay un *efecto lunes* en las rentabilidades de los mercados de valores, que motiva la toma de posiciones ese día, para vender los viernes. En este caso, definiríamos una variable ficticia para cada día de la semana, de modo similar al anterior.

6.10 Correlación parcial

El coeficiente de correlación parcial entre dos variables, y y x_{1i}, mide el grado de relación entre estas dos variables que no es debida al efecto común que las restantes variables del modelo tienen sobre ambas. Este coeficiente se obtiene como el coeficiente de correlación simple entre y y x_{1i} cuando de cada una de ellas se extrae previamente el efecto de las demás variables. Esta es una manera sencilla de calcular el coeficiente de correlación parcial.

Otro modo de calcular la correlacióm parcial, más complejo, pero ilustrativo, porque utiliza propiedades ya conocidas de las regresiones es el siguiente: Estimamos una regresión de cada una de ellas, y y x_{1i} , sobre las demás variables explicativas del modelo, y reservando los residuos de dichas regresiones. Los residuos serán los componentes de y y x_{1i} que no están explicados por las restantes variables.

Definition 7 *El* coeficiente de correlación parcial *entre y_i y x_{1i} se define como el coeficiente de correlación lineal simple entre ambos residuos $\hat{\eta}_i$ e $\hat{\xi}_i$, los componentes de y_i y x_{1i} no explicados por x_{2i},*

$$\rho_{yx_1.x_2} = \frac{\sum_{i=1}^{n} \hat{\eta}_i \hat{\xi}_i}{\sqrt{\sum_{i=1}^{n} \hat{\eta}_i^2} \sqrt{\sum_{i=1}^{n} \hat{\xi}_i^2}} = \frac{\sum_{i=1}^{n} \left(y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_{2i}\right) \left(x_{1i} - \hat{\delta}_0 - \hat{\delta}_1 x_{2i}\right)}{\sqrt{\sum_{i=1}^{n} \left(y_i - \hat{\alpha}_0 - \hat{\alpha}_1 x_{2i}\right)^2} \sqrt{\sum_{i=1}^{n} \left(x_{1i} - \hat{\delta}_0 - \hat{\delta}_1 x_{2i}\right)^2}}$$

Si todos los coeficientes de correlación entre y, x_1 y x_2 fuesen positivos, entonces la correlación parcial entre y y x_{1i} sería inferior al coeficiente de correlación simple entre ambas variables, al estar descontando de éste el efecto común que sobre ambas variables tiene x_2. Si los coeficientes de correlación son de distinto signo, los resultados numéricos pueden ir en cualquier dirección.

El coeficiente de correlación parcial al cuadrado, $\rho^2_{yx_1.x_2}$, se conoce como coeficiente de determinación parcial, $R^2_{yx_1.x_2}$. El coeficiente de determinación parcial $R^2_{yx_1.x_2}$ mide el porcentaje de la Suma de cuadrados de los residuos $\hat{u}_{y.x_2}$ (porque hemos eliminado de y el componente explicadopor x_2) que está explicada por la variable x_1 (aunque de ella hemos extraído asimismo su componente explicado por x_2). En general, el coeficiente de determinación parcial puede ser mayor, igual o menor que el coeficiente de determinación simple, ya que son proporciones calculadas sobre cantidades distintas.

Es fácil obtener una representación del coeficiente de correlación parcial en términos de los coeficientes de correlación simples:

$$\rho_{yx_1.x_2} = \frac{\sum_{i=1}^n \hat{\eta}_i \hat{\xi}_i}{\sqrt{\sum_{i=1}^n \hat{\eta}_i^2}\sqrt{\sum_{i=1}^n \hat{\xi}_i^2}} = \frac{S_{x_1y} - \frac{S_{x_1x_2}}{S^2_{x_2}} S_{x_2y}}{\sqrt{S^2_y - \frac{\left(S_{yx_2}\right)^2}{S^2_{x_2}}}\sqrt{S^2_{x_1} - \frac{S_{x_1x_2}}{S^2_{x_2}} S_{x_1x_2}}}$$

y dividiendo el numerador y el denominador por el producto $\sqrt{S^2_y S^2_{x_1}}$,tenemos:

$$\rho_{yx_1.x_2} = \frac{\rho_{yx_1} - \rho_{x_1x_2}\rho_{yx_2}}{\sqrt{1-\rho^2_{yx_2}}\sqrt{1-\rho^2_{x_1x_2}}}$$

6.11 Relación entre coeficientes de correlación (y de determinación) simple y parcial

Puede probarse que, en una regresión con dos variables explicativas, además de la constante, se tiene:

$$1 - R^2_{y.x_1x_2} = (1-\rho^2_{yx_1})(1-\rho^2_{yx_2.x_1})$$

mientras que en una regresión con 3 variables explicativas, además del término constante, se tiene:

$$1 - R^2_{y.x_1x_2x_3} = (1-\rho^2_{yx_1})(1-\rho^2_{yx_2.x_1})(1-\rho^2_{yx_3.x_1x_2})$$

Para demostrar el primer resultado, notemos que $1-\rho^2_{yx_2.x_1}$ es igual a 1 menos el coeficiente de determinación de la regresión parcial de y sobre x_2 cuando se ha extraido de ambas variables el efecto de la variable x_1. Los residuos de dicha regresión serán igual al componente de y no explicado ni por x_1, cuyo efecto se eliminó previamente, ni por x_2, que se ha incluido como explicativa (en realidad, se ha incluido como explicativa el componente de $x_2$que no está explicado por x_1). Por tanto, la Suma de cuadrados de residuos será: $SCR_{y.x_1x_2}$. Por otra parte, la Suma Total de dicha regresión es igual a la suma de cuadrados de residuos de la regresión de y sobre x_1, es decir, $SCR_{y.x_1}$. Por tanto,

$$1 - \rho^2_{yx_2.x_1} = \frac{SCR_{y.x_1x_2}}{SCR_{y.x_1}}$$

En definitiva,

$$\left(1-\rho^2_{yx_1}\right)\left(1-\rho^2_{yx_2.x_1}\right) = \frac{SCR_{y.x_1}}{ST_y}\frac{SCR_{y.x_1x_2}}{SCR_{y.x_1}} = \frac{SCR_{y.x_1x_2}}{ST_y} = 1 - R^2_{y.x_1x_2}$$

De estas expresiones se deduce que un coeficiente de correlación parcial puede ser mayor o menor que un coeficiente de correlación simple. Por ejemplo, la variable x_1 puede explicar por sí sóla un 20% de las fluctuaciones en y, pero una vez que se descuenta el efecto de x_2, puede ser que x_1 explique un 40% de la varianza residual, sólo que ahora no son residuos de la variable y, sino del componente de y que no esta explicado por x_2. En este caso, el cuadrado del coeficiente de correlación simple sería $\rho^2_{yx_1} = 0,20$, mientras que el cuadrado del coeficiente de correlación parcial sería: $\rho^2_{yx_1.x_2} = 0,40$. Simplemente, la variabilidad total en y, y la variabilidad en el componente de y no explicado por x_2, no son comparables. En nuestro ejemplo, x_1 explica más del segundo que del primero, lo que perfectamente puede suceder.

6.12 Ejemplo: Ventas de un bien en función del precio y del gasto en publicidad

En este ejemplo podemos definir el coeficiente de correlación parcial entre Ventas y Publicidad, $\rho_{V,Pub.P}$ como el coeficiente de correlación simple entre las variables transformadas que resultan al extraer de ambas variables el efecto común que sobre ellas tiene el Precio del producto. Análogamente, podemos definir el coeficiente de correlación parcial entre Ventas y Precio, $\rho_{V,P.Pub}$ como el coeficiente de correlación simple entre las variables transformadas que resultan al extraer de ambas variables el efecto común que sobre ellas tiene el gasto en Publicidad.

En función de los coeficientes de correlación simples habituales, demostramos en la Sección XX las expresiones,

$$\begin{aligned}
\rho_{V,Pub.P} &= \frac{\rho_{V,Pub} - \rho_{V,P}\rho_{Pub,P}}{\sqrt{\left(1-\rho^2_{V,P}\right)\left(1-\rho^2_{Pub,P}\right)}} = \frac{0,950-(-0,901)(-0,829)}{\sqrt{1-(-0,901)^2}\sqrt{1-(-0,829)^2}} = 0,837 \\
\rho_{V,P.Pub} &= \frac{\rho_{V,P} - \rho_{V,Pub}\rho_{Pub,P}}{\sqrt{\left(1-\rho^2_{V,Pub}\right)\left(1-\rho^2_{Pub,P}\right)}} = \frac{-0,901-0,950(-0,829)}{\sqrt{1-0,950^2}\sqrt{1-(-0,829)^2}} = -0,650
\end{aligned}$$

que nos proporcionan los valores numéricos de ambnos coeficientes de correlación simples.

Vimos asimismo en el texto expresiones para el cálculo de los coeficientes de determinación y correlación parcial en función de Sumas Residuales,

$$\begin{aligned}
R^2_{V/Pub.P} &= 1 - \frac{SR^2_{V./Pub,P]}}{SR^2_{V/P}} \Rightarrow \rho_{V,Pub.P} = \sqrt{R^2_{V/Pub.P}} \\
R^2_{V/P.Pub} &= 1 - \frac{SR^2_{V/[Pub,P]}}{SR^2_{V/Pub}} \Rightarrow \rho_{V,P.Pub} = \sqrt{R^2_{V/P.Pub}}
\end{aligned}$$

Teniendo en cuenta que la relación entre la Suma Residual de un modelo de regresión que tiene como variable dependiente a y y el coeficiente de determinación de la misma es: $SR^2 = T.S^2_y(1-R^2)$, tenemos,[28]

[28] $SR^2_{V/[Pub,P]}$ denota la Suma Residual que resulta cuando Publicidad y Precio explican Ventas, mientras $SR^2_{V/Pub}$ denota la suma de cuadrados de residuos en una regresión de Ventas sobre gastos en publicidad. $SR^2_{V/Pub.P}$ sería

$$
\begin{aligned}
SR^2_{V/[Pub,P]} &= T.S^2_V(1 - R^2_{V/[Pub,P]}) = 10(443,5)\,(1 - 0,943) = 252,80 \\
SR^2_{V/Pub} &= T.S^2_V(1 - R^2_{V/Pub}) = 10(443,5)\,(1 - 0,902) = 434,63 \\
SR^2_{V/P} &= T.S^2_V(1 - R^2_{V/P}) = 10(443,5)\,(1 - 0,812) = 833,78
\end{aligned}
$$

Utilizando estos valores numéricos, tenemos,

$$
\begin{aligned}
R^2_{V/Pub.P} &= 1 - \frac{SR^2_{V/[Pub,P]}}{SR^2_{V/P}} = 1 - \frac{252,80}{833,78} = 0,697 \;\Rightarrow\; \rho_{V,Pub.P} = \sqrt{0.697} = 0,835 \\
R^2_{V/P.Pub} &= 1 - \frac{SR^2_{V/[Pub,P]}}{SR^2_{V/Pub}} = 1 - \frac{252,80}{434,63} = 0,418 \;\Rightarrow\; \rho_{V,P.Pub} = \sqrt{0.418} = -0,647
\end{aligned}
$$

donde hemos asignado un signo negativo a $\rho_{V,P.Pub}$ debido a ser una correlación entre ventas y nivel de precios. La ligera variación observada en los resultados proporcionados por ambos enfoques se debe a la aproximación numérica en las diferentes operaciones realizadas en cada caso.

El coeficiente de correlación simple entre ventas y gastos en publicidad es de 0,950, reduciéndose a 0,835 si excluimos la simultaneidad que ambas variables muestran debido a su correlación común con el nivel de precios. El coeficiente de correlación simple entre ventas y nivel de precios es de -0,901, reduciéndose en valor absoluto al caer el coeficiente a -0,647 cuando excluimos la información común que sobre los valores anuales de ambas variables tiene el gasto en publicidad. Por tanto, el gasto en publicidad explica casi un 30% de la correlación entre venta y nivel de precios, mientras que el nivel de precios es responsable de sólo un 12% de la correlación entre ventas y gastos en publicidad. De ello concluimos que el gasto en publicidad es la variable más importante para explicar la evolución temporal de la cifra de ventas.

Tal conclusión coincide con la que alcanzamos al examinar las correlaciones entre Ventas y los residuos de las regresiones simples, así como al comparar los residuos de la regresión múltiple con los que se obtienen en cada una de las regresiones simples. Estos son los procedimientos que sugerimos utilizar para el análisis de este tipo de cuestiones. En este ejemplo se alcanzaría la misma conclusión por los procedimientos habituales de comparar el valor numérico de los coeficientes individuales estimados en el modelo de regresión múltiple, o el valor absoluto de los estadísticos tipo-t asociados a ambos coeficientes. Sin embargo, ya hemos indicado como ninguna de tales comparaciones está justificada y la coincidencia es casual. Veremos en otros ejemplos que los resultados no son siempre coincidentes[29].

Debe recordarse, de la discusión teórica en la Sección XX, que 0,835 es asimismo el coeficiente de correlación que obtendríamos entre los residuos de regresiones que explican las ventas y los gastos en publicidad, respectivamente, por la variable precio ($RES\_V\_PREC, RES\_PUB\_PRECIO$).

---

la Suma Residual que se tendría en una regresión de ventas sobre gastos en publicidad, después de excluir de ambas variables el componente común que tienen por incorporar información sobre el precio. $SR^2_{V/P.Pub}$ se interpreta análogamente.

[29] W. Kruskal, The American Statisticiam (1987), propuso utilizar el promedio de los cuadrados de los coeficientes de correlación simple y parcial entre Y y cada variable explicativa para evaluar la proporción de la fluctuación en Y que es explicada por cada una de éstas. En nuestro ejemplo tendríamos para los gastos en publicidad: $\frac{0.950^2+0.837^2}{2}$ $=0,801$, y para el nivel de precios: $\frac{(-0.901)^2+(-0.650)^2}{2}$ $=0,617$ alcanzando la misma conclusión [Johnston y DiNardo pág.80].

De modo similar, -0,647 es el coeficiente de correlación entre los residuos de regresiones que explican las ventas y el precio, respectivamente, utilizando los gastos en publicidad como única variable explicativa (RES_V_PUB, RES_PREC_PUB). Ambos resultados pueden comprobarse utilizando las variables descritas, que se contienen en el fichero de trabajo.

Esto nos recuerda el significado de los coeficientes de correlación parcial: al estimar las regresiones de ventas y precios sobre publicidad, estamos extrayendo de estas variables la información común con los gastos en publicidad, y luego correlacionamos los componentes así medidos, obteniendo el grado de asociación entre ventas y precio, excluyendo aquella correlación que pueda estar debida al hecho de que ambas se relacionan con el gasto en publicidad.

7 Colinealidad entre variables explicativas en el modelo de regresión

En la mayoría de los modelos de regresión que nos encontramos, las variables explicativas tienen correlación no nula. Esto se debe a que los datos económicos no proceden de un diseño experimental, como pueda suceder en otro tipo de ciencias. Sus valores proceden de modo complejo de las decisiones de los agentes económicos, que hacen que las distintas variables se influyan mutuamente.

Cuando dicha correlación es elevada, sea de signo positivo o negativo, se hace difícil discriminar entre la relevancia que para explicar la variable dependiente tienen las variables explicativas que presentan correlación elevada.

La implicación de este hecho es que es difícil precisar, a partir de la información muestral acerca de los valores numéricos de los coeficientes asociados a dichas variables. En consecuencia, dichos coeficientes se estiman con una reducida precisión.

Como sabemos, ello se manifiesta en unos valores numéricos reducidos de los estadísticos t, por lo que es probable que para alguno de dichos coeficientes (o quizá para todos ellos) no rechacemos la hipótesis nula de ausencia de significación estadística. Aunque sabemos que tal identificación carece de justificación, esta observación puede conducir a creer que las variables asociadas carecen de relevancia para explicar la variable dependiente.

La reducida precisión en la estimación hace que la varianza de la distribución de probabilidad del estimador de cada coeficiente sea elevada. En consecuencia, el valor numérico del estimador, que no es sino una extracción aleatoria de dicha distribución de probabilidad, puede diferir bastante del verdadero valor numérico del coeficiente, a pesar de tratarse de un estimador insesgado. Esto es debido a que aun siendo insesgado, el estimador tiene una varianza grande.

De hecho, una regresión en que los estadísticos t son todos reducidos (por debajo de 2,0 en valor absoluto), y, sin embargo, el coeficiente de determinación R^2 es claramente mayor que cero, es un indicio claro de colinealidad entre todas las variables explicativas. De hecho, en este tipo de situaciones, el estadístico habitual para el contraste de de la hipótesis nula de ausencia de capacidad explicativa global de la regresión estimada rechazará fácilmente la hipótesis nula, mostrando una capacidad explicativa significativa del modelo, a pesar de tener los coeficientes estimados de todas sus variables explicativas un estadístico t inferior a 2,0 en valor absoluto.

Por el mismo argumento, si se excluyen de una regresión dos o más variables explicativas que tienen un estadístico t muy reducido, y ello reduce apreciablemente el coeficiente de determinación, ello indica que las variables excluidas están altamente correlacionadas, y el reducido valor numérico de los estadísticos t de sus coeficientes se debía a la reducida precisión con que se estiman.

7.1 Consecuencias de la colinealidad

- Las estimaciones numéricas de los coeficientes del modelo obtenidas por mínimos cuadrados, y sus desviaciones típicas pueden variar notablemente al cambiar (añadir o excluir) unas pocas observaciones de la muestra.
- varianzas altas para los estimadores de mínimo cuadrados, lo que conduce a:
- posiblemente, signos sorprendentes en los coeficientes estimados
- intervalos de confianza demasiado amplios, lo que
- dificulta la caracterización del impacto numérico que cambios en una variable explicativa tienen sobre la variable dependiente, y a una
- pérdida de potencia en los contrastes de hipótesis sobre los coeficientes del modelo. En particular, tenderemos a
- mantener demasiado frecuentemente la hipótesis nula de ausencia de significación estadística de cada coeficiente del modelo
- sin que ello implique que el modelo, globalmente considerado, carece de capacidad explicativa sobre la variable dependiente. Esto se reflejaría en un R^2 relativamente alto con pocos (o ningún) estadísticos t superiores a 2,0 en valor absoluto.

7.2 Detección de la colinealidad

- R2 moderadamente alto, con pocos estadísticos t superiores a 2,0 en valor absoluto
- Elevadas correlaciones entre las variables explicativas. Estas correlaciones deben calcularse siempre como parte de la descripción de la información muestral, previamente a la estimación de cualquier modelo de regresión
- Examen de los coeficientes de correlación parcial, para comprobar si un coeficiente de correlación elevado entre dos variables denota verdaderamente una dependencia mutua o, por el contrario, ambas reflejan el efecto común de una tercera variable
- Coeficientes de determinación en regresiones auxiliares que explican cada variable explicativa por todas las demás. De hecho, un coeficiente de determinación no muy elevado entre variables explicativas puede ser síntoma de una colinealidad relativamente importante. Recuérdese, por ejemplo, que en una regresión simple, el R^2 es el cuadrado del coeficiente de correlación lineal. Si estamos explicando Y por las variables X y Z, y estas tienen una correlación de 0,60, el R^2 en la única regresión auxiliar posible (de X sobre Z o de Z sobre X), tiene un R^2 de 0,36, que puede no parecer demasiado elevado.
- Se define como el *factor de influencia* sobre la varianza del estimador de mínimos cuadrados del coeficiente de la variable X_i a $1/(1-R_i^2)$, donde R_i^2 denota el coeficiente de determinación descrito en el apartado anterior. La varianza de dicho estimador es:

$$Var\left(\hat{\beta}_i\right) = \frac{\sigma_u^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2(1-R_i^2)}$$

Esta expresión muestra, por otra parte, que la existencia de colinealidad entre variables explicativas no necesariamente generará una elevada varianza para las estimaciones de mínimos cuadrados de los coeficientes del modelo.

7.3 Qué hacer en presencia de colinealidad?

No existe un tratamiento de la colinealidad que pueda recomendarse en todas las situaciones, por lo que el cuidado del analista de datos en estas situaciones es muy importante.

Exclusión de una variable: si se detecta la presencia de dos variables explicativas muy correlacionadas, ello significa que ambas variables tienen mucha información en común, por lo que excluir una de ellas en el modelo que explica una determinada variable dependiente puede no ser muy grave y, por supuesto, eliminará el problema creado por la alta colinealidad. Sin embargo, hay que tener en cuenta que excluir una variable que contiene información relevante sobre la variable dependiente generará un sesgo en el coeficiente que estimamos para la variable que dejamos en el modelo.

Transformación de variables: en algunos casos, la colinealidad se reduce si se agrupa el efecto de dos variables (por ejemplo, sustituyendo ambas por su suma), sustituyendo dos variables muy relacionadas por su valor relativo (su cociente) o normalizando la regresión dividiendo la variable dependiente y algunas o todas las variables explicativas por una de ellas, habitualmente un factor de escala, como la población, la renta, etc..

Una estrategia de especificación de un modelo que puede funcionar bien en un contexto de colinealidad consiste en comenzar detectando la variable explicativa con mayor contenido informativo sobre la variable dependiente y estimando la regresión simple con dicha variable. El residuo de dicha regresión proporciona el componente de Y que no está explicado por X_1. Correlacionamos dicho componente de Y con las variables explicadas no incluidas en la regresión, y añadimos a la regresión simple anterior la variable explicativa con mayor contenido informativo sobre el componente mencionado de Y. Procedemos de este modo utilizando en cada paso el residuo de la regresión de Y, y calculando el coeficiente de correlación con cada una de las variables explicativas aún no incluidas en el modelo. Nótese que buscamos, en cada etapa, añadir al modelo la mayor información posible que aún no esté incorporada en las variables explicativas que ya están incluidas en el modelo.

7.4 Ejemplo: Ventas de un bien en función del precio y del gasto en publicidad

7.4.1 Regresiones simples cruzadas

Un investigador podría estimar asimismo dos modelos que tratan de recoger la correlación existente entre las variables explicativas,

$$\underset{}{Pub_t} = \underset{()}{107,11} \underset{(0,235)}{-0,986}\ P_t,\ R^2_{Pub.P} = 0,687,\ \hat{\sigma}_u = 3,88 \tag{27}$$

$$\underset{}{P_t} = \underset{()}{103,52} \underset{(0,166)}{-0,697}\ Pub_t,\ R^2_{P.Pub} = 0,687,\ \hat{\sigma}_u = 3,26 \tag{28}$$

en las que:

- el investigador detecta fuerte correlación entre ambas variables, con un coeficiente de determinación que es igual, por supuesto, al cuadrado del coeficiente de correlación simple entre ambas variables. Por eso el coeficiente de determinación es el mismo en ambas regresiones, puesto que el coeficiente de correlación no encierra ninguna idea de causalidad y es, independiente, por tanto, de qué variable tomemos como dependiente y cuál como independiente.

- sin embargo, sería imposible concluir, utilizando estas regresiones estimadas, si la correlación entre precios y gastos en publicidad es fruto o no de una política explícita de comercialización.

- a pesar de la coincidencia entre coeficientes de determinación, las desviaciones típicas del término de error no son iguales, sin embargo, ya que las regresiones explican variables dependientes diferentes. Sin embargo, nuestros ratios habituales coinciden,

$$Ratio\,(Pub_t/P_t) = \frac{3.88}{6.20} = 0,626;\ Ratio\,(P_t/Pub_t) = \frac{3.26}{5.22} = 0,625$$

  donde la leve diferencia se debe exclusivamente a los redondeos a tres decimales.

- las pendientes estimadas en ambas regresiones no son iguales. Su producto es: $(-0,986)(-0,697) = 0,687$ que es, precisamente, el coeficiente de determinación entre ambas variables, gastos en publicidad y nivel de precios. Esta es una propiedad de la regresión lineal simple: si se estiman por mínimos cuadrados regresiones de Y sobre X y de X sobre Y, el producto de las pendientes resultantes es siempre igual al cuadrado del coeficiente de correlación lineal simple entre ambas variables.

7.4.2 Tratamiento de la colinealidad

La regresión auxiliar entre nivel de precios y gastos en publicidad (28), nos permite estimar el componente de la evolución temporal del nivel de precios que no está explicado por las fluctuaciones que anualmente experimenta el gasto en publicidad,

$$P\backslash Pub_t = P_t - 103,52 + 0,697\ Pub_t \tag{29}$$

que no es sino el residuo de la regresión (28). Las propiedades del estimador de mínimos cuadrados garantizan que dicho residuo tiene correlación nula con los gastos en publicidad, por ser ésta la variable explicativa en la regresión a partir de la cual se han generado los residuos. Por tanto, $Corr(Pub_t, P\backslash Pub_t) = 0$.

Si ahora estimamos una regresión que pretende explicar las ventas mediante los gastos en publicidad y el componente de precios no explicado por estos, tenemos,

$$V_t = \underset{(5,26)}{95,99} + \underset{(0,305)}{3,224}\ Pub_t - \underset{(0,649)}{1,464}\ P\backslash Pub_t,\ R^2 = 0,943,\ \bar{R} = 0,927,\ \hat{\sigma}_u = 5,983 \tag{30}$$

donde puede observarse que el coeficiente estimado para $P\backslash Pub_t$ es el mismo que obtuvimos en la regresión inicial (22), y se estima con la misma precisión.

Sin embargo, el coeficiente estimado para los gastos en publicidad es ahora mayor que en (22); la razón es que en (22), al impacto directo sobre las ventas de un aumento en los gastos en publicidad había que añadir el impacto de la reducción en precios que usualmente acompaña el mayor gasto en publicidad. El efecto global es superior al medido por el coeficiente $2,204$ que los gastos en publicidad reciben en (22), y eso aparece claro en (30). De hecho, el coeficiente estimado para Pub_t en (30) es el mismo que obtuvimos en la regresión simple con esta variable, sólo que ahora lo estimamos con una mayor precisión.[30] Aunque numéricamente es mayor, también se estima dicho

[30] Que ambas estimaciones numéricas coincidan no es sino reflejo del resultado teórico que afirma que las estimaciones numéricas del coeficiente de mínimos cuadrados de una variable explicativa no cambia si se excluyen o se incluyen en el modelo variables explicativas incorrelacionados con la primera.

coeficiente con mayor precisión (menor desviación típica) en (30) que en la regresión inicial (22), gracias a que la ausencia de correlación entre las variables explicativas en (30) permite discriminar mejor el efecto de cada variable.

Podría pensarse que una limitación del modelo (30) es el hecho de que en él no aparece el precio del producto, sino tan sólo el componente del mismo que no está explicado por los gastos en publicidad. Esto es, en cierta forma, sólo aparente, pues si combinamos (30) con (29) se recuperan exactamente los mismos coeficientes estimados en la regresión original (22), excepto por el hecho de que el coeficiente de los gastos en publicidad se ha estimado con una precisión superior.

Alternativamente, podríamos utilizar (27) para estimar el componente de los gastos en publicidad no explicado por el nivel de precios,

$$Pub\backslash P_t = Pub_t - 107,11 + 0,986\ P_t,$$

con $Corr(Pub\backslash P_t, P_t) = 0$. A continuación, estimaríamos la regresión de ventas sobre el nivel de precios y $Pub\backslash P_t$, obteniendo,

$$V_t = \underset{(5,26)}{95,99} + \underset{(0,545)}{2,204}\ Pub\backslash P_t - \underset{(0,363)}{3,637}\ P_t,\ R^2 = 0,943,\ \bar{R} = 0,927,\ \hat{\sigma}_u = 5,983 \qquad (31)$$

siendo ahora el coeficiente de la variable auxiliar $Pub\backslash P_t$ el que coincide con el obtenido en el modelo original (22) para Pub_t, mientras que el coeficiente estimado para el nivel de precios es ahora mayor en valor absoluto que en (22), por las mismas razones antes descritas. Coincide con el obtenido en la regresión simple (33), aunque se estima con mayor precisión que en dicha regresión, y también con mayor precisión que en la regresión inicial. El investigador debería quedarse con una de las dos regresiones (30) o (31), dependiendo de la dirección de causalidad en la que interprete la correlación existente entre nivel de precios y gasto en publicidad.

La incorporación del componente del precio no relacionado con los gastos en publicidad eleva el coeficiente de determinación de la regresión de ventas sobre gastos en publicidad desde 0,902 a 0,943. De modo similar, la inclusión del componente del gasto en publicidad no relacionado con el precio en la regresión de ventas sobre precios, eleva el coeficiente de determinación de 0,812 al mismo nivel citado, 0,943. Esto sugiere que el contenido informativo de los gastos en publicidad sobre las ventas es mayor que el que tiene la variable Precio. Sin embargo, la comparación de coeficientes de determinación reduce toda la información muestral relativa a la explicación de las cifras de ventas a una sóla cifra. Preferimos comparar los residuos de modelos que incluyen o excluyen una variable explicativa, pues nos permiten analizar el impacto que dicha variable tiene, observación a observación. Es perfectamente imaginable que tal efecto sea muy notable pero esté concentrado en unas pocas observaciones que tengan alguna característica en común.[31] Ello haría que la comparación de medidas agregadas, como los coeficientes de determinación, no detectase la contribución de la variable explicativa. Si, por el contrario, una comparación detallada de los dos conjuntos de residuos nos detecta variaciones importantes en los residuos correspondientes a ese reducido conjunto de observaciones, podríamos definir una variable ficticia apropiadamente, mejorando con ello la capacidad explicativa del modelo.

[31] Por ejemplo, cinco años consecutivos durante los que se produjo una gran elevación en los precios del petróleo, en una muestra de 60 años. En una muestra de sección cruzada correspondiente a un amplio conjunto de países, podrían ser los residuos correspondientes a los países subsaharianos los que experimentan una variación muy notable al incluir una determinada variable epxlicativa en el modelo.

8 Efectos individuales y efectos globales

8.1 Omisión de variables relevantes

La omisión en una regresión de variables relevantes causa sesgos en la estimación de los efectos individuales de las variables incluidas en el modelo. Supongamos que el verdadero modelo es:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

pero estimamos:

$$y_i = \alpha_0 + \alpha_1 x_{1i} + u_i$$

lo que nos genera una estimación de α_1 :

$$\hat{\alpha}_1 = \frac{S_{x_1 y}}{S^2_{x_1}} = \frac{\sum(x_{1i} - \bar{x}_1)(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i)}{\sum(x_{1i} - \bar{x}_1)^2} = \beta_1 + \beta_0 \frac{\sum(x_{1i} - \bar{x}_1)}{S^2_{x_1}} + \beta_2 \frac{S_{x_1 x_2}}{S^2_{x_1}} + \frac{S_{x_1 u}}{S^2_{x_1}}$$

La primera fracción tiene esperanza matemática igual a cero. Por otro lado, la covarianza entre x_{1i} y u_i es también igual a cero. Por tanto, tenemos:

$$E(\hat{\alpha}_1) = \beta_1 + \beta_2 \frac{S_{x_1 x_2}}{S^2_{x_1}} = \beta_1 + \beta_2 \delta_{21}$$

donde δ_{21} denota el coeficiente (pendiente) estimado en la regresión de la varaible excluida de la regresión, x_2, sobre la variable incluida x_{1i}. El sesgo en la estimación del coeficiente de la variable incluida en la regresión es igual, por tanto, al producto de dos factores: 1) el coeficiente que tendría en la verdadera regresión la variable que hemos excluido del modelo, por 2) el coeficiente en una regresión de la vaariable excluida sobre la variable incluida en el modelo.

En el primer caso considerado, si el verdadero modelo es:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

pero estimamos:

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

tendremos una estimación sesgada del efecto individual β_1, pero con menor varianza:

$$Var(\tilde{\beta}_1) = \frac{\sigma^2_u}{S^2_{x_1}}$$

frente a la que tendríamos en el modelo correcto:

$$Var(\hat{\beta}_1) = \frac{\sigma^2_u}{S^2_{x_1}(1 - \rho^2_{x_1 x_2})}$$

por lo que no es claro qué estimador de dicho efecto individual es preferible. Recordemos que un criterio razonable para escoger un estimador frente a otro el es el Error Cuadrático Medio, definido:

$$ECM = (Sesgo(\hat{\beta}))^2 + Varianza(\hat{\beta})$$

Sin embargo, no es claro que la *estimación* de la desviación típica del estimador $\tilde{\beta}_1$sea menor que la estimación de la desviación típica de $\hat{\beta}_1$, debido a que la estimación del parámetro σ_u^2 tampoco será la misma en ambos modelos, siendo mayor en el modelo mal especificado. Como dicha estimación es el cociente entren la Suma de Cuadrados de los Residuos y el número de grados de libertad, se tiene que:

$$\frac{S^2(\hat{\beta}_1)}{S^2(\tilde{\beta}_1)} = \frac{1-\rho_{x_1x_2}^2}{1-\rho_{yx_2.x_1}}$$

la desviación tipica estimada de $\tilde{\beta}_1$será menor que la desviación típica estimada de $\hat{\beta}_1$ si y solo si: $\rho_{x_1x_2}^2 < \rho_{yx_2.x_1}$.

El resultado se generaliza facilmente. Si el verdadero modelo es:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

pero estimamos por error:

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + u_i$$

el sesgo en las estimaciones de β_1 y β_2 es igual al producto del coeficiente que tendría la variable omitida en la regresión, β_3, por los coeficientes de una regresión de la variable excluida, x_3 sobre las dos variables incluidas, (x_{1i}, x_{2i}). Al hacer este producto, es importante conservar el orden de las variables y los coeficientes:

$$Sesgo\begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} = \beta_3 \begin{pmatrix} \delta_{31} \\ \delta_{32} \end{pmatrix} = \begin{pmatrix} \beta_3\delta_{31} \\ \beta_3\delta_{32} \end{pmatrix}$$

Por tanto, el sesgo en β_1 sería igual a $\beta_3\delta_{31}$, mientras que el sesgo en β_2 sería igual a $\beta_3\delta_{32}$.

Si el verdadero modelo es:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

pero estimamos por error:

$$y_i = \alpha_0 + \alpha_1 x_{1i} + u_i$$

entonces, el sesgo en la estimación de β_1 es igual al producto de los coeficientes que tendrían las dos variables omitidas en la regresión, β_2 y β_3, por los coeficientes de una regresión de cada una de ellas sobre la variable incluida, x_{1i}. En esas regresiones:

$$\begin{aligned} x_{2i} &= \delta_{20} + \delta_{21} x_{1i} + u_i \\ x_{3i} &= \delta_{30} + \delta_{31} x_{1i} + u_i \end{aligned}$$

por lo que el sesgo en β_1 sería:

$$Sesgo\,(\hat{\alpha}_1) = (\beta_2, \beta_3)\begin{pmatrix} \delta_{21} \\ \delta_{31} \end{pmatrix} = \beta_2\delta_{21} + \beta_3\delta_{31}$$

$$E\,(\hat{\alpha}_1) = \beta_1 + (\beta_2\delta_{21} + \beta_3\delta_{31})$$

8.2 Inclusión de variables irrelevantes

Este caso es más sencillo: la inclusión de variables irrelevantes no va a introducir ningún sesgo en la estimación de mínimos cuadrados de los coeficientes de las restantes variables del modelo. La estimación del coeficiente de la variable irrelevante, que se ha incluido por error, será próxima a cero, y no se rechazará la hipótesis nula de ausencia de significación. Sin embargo, la inclusión de dicha variable nos lleva a perder precisión en la estimación, por lo que el estimador de mínimos cuadrados de los restantes coeficientes ya no será eficiente, salvo que la variable irrelevante tenga correlacion nula con cada una de las restantes variables explicativas.

8.3 Estimación insesgada de efectos parciales y totales

Pensemos en la interpretación de este resultado: en la regresión múltiple, el estimador β_1 mide el efecto que sobre y tendría una variación unitaria en x_{1i} si la otra variable explicativa x_{2i} no varía. Sin embargo, si x_{1i} y x_{2i} están correlacionadas, entonces tal escenario es poco verosímil pues cuando una de ellas varía, también cambiará la otra variable. En el modelo de regresión simple (25), en el que solo aparece una de las variables, x_{1i}, el coeficiente δ_1 mide el *efecto total* de una variación unitaria en x_{1i} sobre y_i. Dicho efecto es el agregado de un *efecto directo*, y un *efecto indirecto*, que se produce porque cuando x_{1i} varía, x_{2i} también varía. Por el contrario, el estimador del coeficiente en el modelo de regresión múltiple mide tan sólo el efecto directo. Tambien podemos entender estos dos efectos como *efecto total* y *efecto parcial*.

Es evidente que si queremos medir el *efecto parcial o directo* de x_{1i} sobre y, que se produce cuando únicamente x_{1i} varía, permaneciendo inalteradas las restantes variables explicativas, entonces debemos estimar la regresión múltiple y examinar únicamente el coeficiente asociado a esta variable. Basta suponer en la regresión múltiple que las variables distintas de x_1 no varían, mientras que sí lo hace x_1.

Si, por el contrario, queremos medir el *efecto global* (suma de efecto directo y efectos indirectos) que sobre y tiene un cambio unitario en x_{1i}, entonces debemos estimar el modelo de regresión simple. ¿Cómo sabemos esto? La discusión anterior sobre variables omitidas nos dice que, si leverdadero modelo de y incluye a x_1 y x_2 comno variables explicativas, entonces la esperanza matemática del estimador del coeficiente de x_1 en la regresión simple será:

$$E(\hat{\alpha}_1) = \beta_1 + \beta_2\delta_{21}$$

donde δ_{21} denota el coeficiente (pendiente) estimado en la regresión de la variable excluida de la regresión, x_2, sobre la variable incluida x_{1i}. Es decir, el efecto de x_1 sobre y estimado en la regresión simple es igual, en promedio, al efecto de x_1 sobre y que estimaríamos en la regresión múltiple, que llamamos efecto directo o efecto parcial, más el producto $\beta_2\delta_{21}$. Pero δ_{21} nos da la variación esperada en x_2 cuando x_1 aumenta en una unidad. Como β_2 es el efecto de una variación unitaria en x_2 sobre y, el producto $\beta_2\delta_{21}$ nos da el efecto indirecto que sobre y tiene una variación en $x_1$debida al efecto inducido que provoca sobre x_2. Esto e slo que denbominamos efecto indirecto. En consecuencia, la expresión anterior es el agregado del efecto directo más el efecto indirecto, por lo que podemos denominarlo el efecto total que sobre y tiene una variación en x_1.

En la sección 8.1 obtuvimos el sesgo en la estimación de α_1. Nos referíamos al sesgo que cometeríamos si interpretásemos el coeficiente estimado en la regresión simple, α_1, como el efecto directo o parcial de x_1 sobre y. Pero es muy importante apreciar que lo que en la regresión simple estimaríamos con sesgo es el *efecto individual, efecto parcial* o *efecto directo* de x_1 sobre y, pero no el

efecto total de una variación en x_1 sobre y. Este efecto se estima precisamente de manera inesgada en una regresón de y sobre la variable x_1 *exclusivamente*.

Por tanto, cada modelo proporciona una respuesta insesgada a una de las dos preguntas siguientes, y una respuesta sesgada a la otra:

- ¿Cuál es el impacto que sobre y tendría una variación unitaria en x_{1i} si las demás variables explicativas del modelo no variasen? Respuesta: *Efecto parcial* $\Rightarrow$ Estimación insesgada: Regresión múltiple
- ¿Cuál es el impacto total que sobre y tendría una variación unitaria en x_{1i}? Respuesta: *Efecto total* $\Rightarrow$ Estimación insesgada: Regresión simple.

8.4 Ejemplo: Ventas de un bien en función del precio propio y del gasto en publicidad

Una lectura del modelo (22) sugeriría que las cifras de ventas aumentan en 220,4 euros por cada 100 euros de incremento en gastos de publicidad, suponiendo que el precio del producto no variase. Esta es la interpretación *ceteris paribus*, tan habitualmente utilizada, pero también tan poco consisitente con la mayoría de las situaciones a que se enfrenta un analista de datos económicos, con variables epxlicativas correlacionadas entre sí. De modo análogo, las ventas disminuirían en 146,4 euros por cada 100 euros de incremento en el precio unitario del articulo comercializado por la empresa.

Pero, al existir la simultaneidad mencionada entre los niveles de gastos en publicidad y de precios, la interpretación *ceteris paribus* no es rigurosa, puesto que, como indica el elevado coeficiente de correlación negativo entre ambas variables, de -0,829, indicando que la empresa gasta más en publicidad en períodos en que el precio del producto es bajo, y menos cuando el precio del producto es alto, lo cual podría ser espúrio o, por el contrario, fruto de una estrategia deliberada de marketing. Con independencia de las razones que generan dicha correlación, un mayor gasto en publicidad suele venir asociado a una reducción en el precio del producto, siendo el efecto sobre las ventas la conjunción de ambos efectos. En consecuencia, al incrementar el gasto en publicidad en 100 euros, las ventas aumentarían en más de 220,4.

De hecho, las regresiones simples de las ventas anuales sobre cada variable explicativa son,

$$V_t = \underset{(0,375)}{96,0 + 3,224}\ Pub_t,\ R^2_{V.Pub} = 0,902,\ \hat{\sigma}_u = 7,36 \tag{32}$$

$$V_t = \underset{(0,619)}{483,6 - 3,637}\ P_t,\ R^2_{V.P} = 0,812,\ \hat{\sigma}_u = 10,22 \tag{33}$$

cuyos coeficientes de determinación son, por supuesto, el cuadrado de los coeficientes de correlación simples que aparecen en la matriz Σ. En ambos casos estimamos unos coeficientes mayores en valor absoluto a los obtenidos en el modelo de regresión múltiple, por las razones que acabamnos de exponer.

Veamos qué implicaciones tiene el coeficiente de correlación estimado, de -0,829. Puesto que,

$$\rho_{Pub,P} = E\left(\frac{Pub - E(Pub)}{\sigma_{Pub}}\frac{P - E(P)}{\sigma_P}\right)$$

tenemos que $\rho_{Pub,P}$ mide el valor medio que toma el producto de las fluctuaciones en Pub y P alrededor de sus respectivas medias. Supongamos que ambas variables están permanentemente

en torno a sus valores medios, de los cuales se desvían cada período en una cuantía media igual a sus respectivas desviaciones típicas.[32] Esto significa que un incremento nominal de 6,20 euros en el gasto en publicidad, equivale a un aumento de una desviación típica en dicha variable. La expresión anterior, junto con $\rho_{Pub,P} = -0,829$, sugiere que dicho incremento venga asociado con un descenso de 0,829 desviaciones típicas en el precio.[33] Teniendo en cuenta que $\sigma_P = 5,22$, dicho descenso equivale a una reducción en el precio del producto de 4,327 euros. La estimación del modelo de regresión múltiple sugiere que el mayor gasto en publicidad eleva las ventas en 13,665 euros, mientras que el descenso en el precio aumenta las ventas en 6,335 euros, siendo la suma de ambos efectos de 20,00 euros. Este es el efecto que sobre las ventas tiene un incremento de 6,20 euros en el gasto en publicidad, teniendo en cuenta la relación que existe a lo largo de la muestra entre esta variable y el precio unitario del producto. Si consideramos un incremento de 100 euros en el gasto en publicidad, como $100 = (6,20)(16,13)$, tendríamos un efecto estimado sobre las ventas de: $(20,00)(16,13) = 322,6$ euros en ventas, aproximadamente igual a lo obtenido al estimar el modelo (32).[34] Por tanto, los coeficientes estimados en las regresiones simples incorporan el efecto que simultáneamente se produce en la variable omitida cuando cambia el valor numérico de la variable incluida en la regresión simple.

El razonamiento que hemos hecho en el párrafo anterior es, exactamente, la aplicación práctica de las expresiones sobre el sesgo que se produce en el estimador de mínimos cuadrados cuando se omiten del modelo variables relevantes. Consideremos un modelo de regresión múltiple con dos variables explicativas, x_i, x_e, cuyos subíndices denotan que una se incluye en la regresión simple, y otra queda excluida de dicha regresión. Es decir, la regresión múltiple es: $y = \beta_{im}x_i + \beta_{em}x_e + u$, mientras que la regresión simple es, $y = \beta_{is}x_i + v$. En el razonamiento previo hemos descompuesto el efecto β_{is} de una variación unitaria en la variable incluida en una regresión simple x_i, en dos componentes: el efecto directo, medido por el coeficiente de la variable que en el modelo de regresión múltiple tiene la variable incluida β_{im}, y el efecto indirecto. Para estimar éste último, hemos calculado la variación unitaria en términos de desviaciones típicas, $\frac{1}{\sigma_i}$. A continuación, hemos utilizado la definición de coeficiente de correlación para inferir que, en media, esta variación irá acompañada de una variación de $\frac{1}{\sigma_i}\rho_{ie}$ desviaciones típicas en la variable excluida, x_e. Esto equivale a una variación nominal de $\frac{1}{\sigma_i}\rho_{ie}\sigma_e$ en dicha variable. Utilzando las estimaciones de la regresión múltiple, su impacto sobre la variable dependiente será $\frac{1}{\sigma_i}\rho_{ie}\sigma_e\beta_{em}$. Pero esto es igual a $\frac{\sigma_{ei}}{\sigma_i^2}\beta_{em}$, y $\frac{\sigma_{ei}}{\sigma_i^2}$ no es sino la estimación de mínimos cuadrados del coeficiente $\beta_{e/i}$de la regresión simple de la variable omitida x_e, sobre la incluida, x_i. En definitiva, el efecto global de una variación unitaria en la variable incluida en la regresión simple, x_i, sobre la variable dependiente, es,

$$\beta_{im} + \beta_{em}\beta_{e/i}$$

Esta es precisamente la expresión de la esperanza matemática del estimador de mínimos cuadrados del modelo de regresión múltiple $E(\beta_{is})$ que incluye a x_i como única variable explicativa.

[32] Estrictamente hablando, este supuesto es apropiado únicamente en situaciones en que las desviaciones respecto del valor medio en períodos sucesivos son independientes. Esto no ocurre en presencia de comportamientos tendenciales como los de las variables en este ejemplo.

[33] Esta interpretación es estrictamente válida si entendemos que el nivel de precios se fija por la empresa en respuesta al gasto en publicidad acometido, y no al revés; es decir, si interpretamos la alta correlación entre estas variables en el sentido $Publicidad \rightarrow \mathrm{Pr}\, ecio$.

[34] El lector puede repetir el ejercicio partiendo de un descenso de una desviación típica en el precio del producto. Comprobará que el efecto global que obtiene sobre las ventas debido a un descenso de 100 euros en el precio del producto es el que estimaría a partir del modelo de regresión simple (33).

9 Contrastes de restricciones generales

En un modelo de regresión lineal múltiple surgen con frecuencia hipótesis más generales. En unos casos, se trata de contrastar varias hipótesis a la vez acerca de valores numéricos para distintos coeficientes; en otros, contrastamos un valor teórico acerca de una o varias combinaciones lineales de coeficientes. Por ejemplo:

$$\begin{aligned} H_0 &: \quad \beta_1 + \beta_2 = 1 \\ H_1 &: \quad \beta_1 + \beta_2 \neq 1 \end{aligned}$$

$$\begin{aligned} H_0 &: \quad \beta_1 = 0;\ \beta_2 = 1 \\ H_1 &: \quad \beta_1 \neq 0 \text{ ó } \beta_2 \neq 1 \end{aligned}$$

$$\begin{aligned} H_0 &: \quad 2\beta_1 - \beta_2 = 1 \\ H_1 &: \quad 2\beta_1 - \beta_2 \neq 1 \end{aligned}$$

que contrastan una sola restricción en el primer y tercer ejemplos, en cada caso involucrando a dos coeficientes, mientras que el segundo caso es un contraste conjunto de 2 restricciones, cada una de ellas sobre un sólo coeficiente. El número de restricciones es igual al número de condiciones de igualdad, que es 1 en el primer y tercer casos, e igual a 2 en el segundo ejemplo. A diferencia de lo que sucede cuando contrastamos una sóla hipótesis o restricción, cuando se rechaza una hipótesis nula compuesta de dos o más restricciónes, podemos decir que alguna de ellas es falsa, pero no necesariamente todas, y podemos indagar cuáles son falsas y cuáles no lo son.

En estos casos, existen varios enfoques para resolver los contrastes. Quizá el más práctico consiste en estimar dos modelos: el Modelo Sin Restringir (MSR), y el Modelo Restringido (MR), para comparar sus Sumas de Cuadrados de Residuos, $SCRSR$ y $SCRR$, respectivamente. Si las restricciones son ciertas, imponerlas o no imponerlas dará igual, pues aún si no las imponemos, las estimaciones que obtengamos a partir de los datos satisfarán, aproximadamente dichas restricciones, ya que estamos suponiendo que son ciertas. Esto significa que las dos Sumas de Cuadrados de Residuos, serán aproximadamente iguales, y lo contrario sucederá si las restricciones son falsas. En definitiva, parece razonable comparar ambas Sumas de cuadrados de residuos, en términos porcentuales:

$$\frac{SCRR - SCRSR}{SCRSR}$$

Al imponer restricciones a los valores numéricos de los estimadores, el ajuste del modelo nunca puede mejorar. Generalmente, empeorará. La cuestión, es en cuánto se deteriora el ajuste. Si se deteriora en mucho, es que los datos son contrarios a las restricciones y debemos rechazarlas. Por eso es que tomamos el deterioro, es decir, el aumento, que se produce en la Suma de Cuadrados de Residuos al imponer las restricciones, como porcentaje del valor que teníamos antes de imponerlas.

El cociente anterior no tiene una distribución conocida, pero una corrección del mismo:

$$\frac{SCRR - SCRSR}{SCRSR}\frac{N - ncoefs}{q} \sim F_{q,N-ncoefs} \tag{34}$$

donde $ncoefs$ es el número de coeficientes estimados en el Modelo Sin Restringir, incluyendo, como siempre, el término independiente, y q es el número de restricciones que se contrastan.

Ejemplo: Consideremos el modelo: $y_i = \beta_0 + \beta_1 x + \beta_2 z + u_i$, en el que queremos contrastar la hipótesis nula:

$$\begin{aligned} H_0 &: \quad \beta_1 = 0;\ \beta_2 = 1 \\ H_1 &: \quad \beta_1 \neq 0 \text{ ó } \beta_2 \neq 1 \end{aligned}$$

El Modelo Sin Restringir es siempre el modelo original, que en este ejemplo es: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 v_i + u_i$. El Modelo Restringido debe incorporar las dos restricciones $\beta_1 = 0;\ \beta_2 = 1$, por lo que se trata de: $y_i = \beta_0 + 0.x_i + 1.z_i + \beta_3 v_i + u_i = \beta_0 + z_i + \beta_3 v_i + u_i$. A la derecha la variable z_i no tiene coeficiente, por lo que no hay nada que estimar en dicho término. Cuando esto sucede, pasamos dichos términos a la izquierda. Así, tenemos: $y_i - z_i = \beta_0 + \beta_3 v_i + \tilde{u}_i$. Si definimos una nueva variable mediante: $\tilde{y}_i = y_i - z_i$ podemos estimar el Modelo Restringido $\tilde{y}_i = \beta_0 + \beta_3 v_i + \tilde{u}_i$ y comparar las Sumas de Cuadrados de Residuos que generan, mediante el estadístico que hemos presentado. En este caso, $q = 2$ y $ncoefs = 4$.

Ejemplo: Consideremos el modelo: $y_i = \beta_0 + \beta_1 x + \beta_2 z + u_i$, en el que queremos contrastar la hipótesis nula:

$$\begin{aligned} H_0 &: \quad \beta_1 = \beta_2 \\ H_1 &: \quad \beta_1 \neq \beta_2 \end{aligned}$$

Este es una caso interesante, pues se trata de contrastar un valor numérico (cero) para una sóla combinación lineal de coeficientes: $\beta_1 - \beta_2 = 0$, lo que nos permite considerar tres procedimientos distintos, que son numéricamente equivalentes.

Procedimiento 1: Comenzamos por la comparación entre los Modelos Restringido y Sin Restringir. El Modelo Sin Restringir es siempre el modelo original, que en este ejemplo es: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i$. El Modelo Restringido debe incorporar la restricción $\beta_1 = \beta_2$, por lo que se trata de: $y_i = \beta_0 + \beta_2 x_i + \beta_2 z_i + \tilde{u}_i = \beta_0 + \beta_2 (x_i + z_i) + \tilde{u}_i$. Si definimos una nueva variable: $\tilde{x}_i = x_i + z_i, i = 1, 2, ..., N$, el Modelo Restringido es: $y_i = \beta_0 + \beta_2 \tilde{x}_i + \tilde{u}_i$. Se trata de estimar ambos modelos y comparar las Sumas de Cuadrados de Residuos que generan, mediante el estadístico que hemos presentado. En este caso, $q = 1$ y $ncoefs = 3$. En el Modelo Restringido estimamos únicamente β_0 y β_2. La estimación restringida de β_1 es igual a la que obtengamos para β_2.

Procedimiento 2: Una vez estimado el modelo, las estimaciones numéricas no satisfarán exactamente la igualdad: $\hat{\beta}_1 = \hat{\beta}_2$, y el grado de incumplimiento muestral de dicha hipótesis puede evaluarse mediante la discrepancia $\hat{\beta}_1 - \hat{\beta}_2$. De acuerdo con la hipótess nula, esta diferencia debería ser exactamente igual a cero. Para saber si su valor numérico $\hat{\beta}_1 - \hat{\beta}_2$ puede considerarse cero o, por el contrario, debe considerarse significativamente distinta de cero, hemos de compararla con su desviación típica.

Para ello, calculamos la varianza de dicha discrepancia:

$$Var\left(\hat{\beta}_1 - \hat{\beta}_2\right) = Var\left(\hat{\beta}_1\right) + Var\left(\hat{\beta}_2\right) - 2Covar\left(\hat{\beta}_1, \hat{\beta}_2\right)$$

y calculamos el estadístico:

$$\frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{Var\left(\hat{\beta}_1 - \hat{\beta}_2\right)}} \sim t_{N-k}$$

¿Por qué sigue esta distribución t de Student? Recordemos que una combinación lineal de variables Normales es asimismo Normal. Por tanto, la diferencia de dos variables Normales (los estimadores de los dos coeficientes) tienen distribución Normal. Su esperanza matemática es igual a la diferencia de las esperanzas matemáticas de $\hat{\beta}_1$ y $\hat{\beta}_2$. Su varianza es la expresión que hemos calculado arriba. Si restamos a una variable Normal su esperanza matemática (que en este caso es cero, porque bajo H_0, ambos coeficientes son iguales entre sí) y dividimos por su desviación típica obtenemos una variable con distribución $N(0,1)$. Si sustituimos la desviación típica por una estimación de la misma, la distribución pasa a ser una t. Los grados de libertad son siempre igual al número de observaciones menos el número de coeficientes estimados.

Procedimiento 3: Una alternativa sería comparar el cuadrado de la discrepancia con su varianza, y entonces la distribución sería una $F_{1,N-k}$:

$$\frac{\left(\hat{\beta}_1 - \hat{\beta}_2\right)^2}{Var\left(\hat{\beta}_1 - \hat{\beta}_2\right)} \sim F_{1,N-k}$$

Ambos procedimientos son equivalentes, porque los valores de una distribución $F_{1,m}$ son siempre iguales al cuadrado de los valores de una distribución t_m.

Ejemplo: Otro ejemplo: supongamos que en el modelo $y_i = \beta_0 + \beta_1 x + \beta_2 z + u_i$, queremos contrastar: $H_0 : \beta_1 = 0$ frente a la alternativa $H_0 : \beta_1 \neq 0$. El Modelo sin Restringir es el modelo original, mientras que el Modelo Restringido es: $y_i = \beta_0 + \beta_2 z + u_i$. Una vez estimados ambos modelos, utilizamos el estadístico 34 que será en este caso:

$$\frac{SCRR - SCR}{SCR}\frac{N-3}{1} \sim F_{1,N-k}$$

Puede probarse (com veremos en algún ejemplo numérico con ordenador), que el valor numérico de este estadístico es exactamente igual al cuadrado del estadístico t correspondiente al coeficiente β_1. Nuevamente, como $F_{1,N-k} = (t_{N-k})^2$, ambos contrastes son equivalentes.

Ejemplo: Supongamos que en el modelo $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i$, queremos contrastar:

$$\begin{aligned} H_0 &: \quad 2\beta_1 - \beta_2 = 1 \\ H_1 &: \quad 2\beta_1 - \beta_2 \neq 1 \end{aligned}$$

que se trata nuevamente de una sola combinación lineal de coficientes. Para obtener el Modelo Restringido, incorporamos al modelo original (Modelo Sin Restringir) la restricción $2\beta_1 - \beta_2 = 1$ o, lo que es lo mismo: $\beta_2 = 1 + 2\beta_1$, obteniendo: $y_i = \beta_0 + \beta_1 x_i + (1 + 2\beta_1)z_i + \tilde{u}_i = \beta_0 + \beta_1 x_i + z_i + 2\beta_1 z_i + \tilde{u}_i = \beta_0 + z_i + \beta_1(x_i + 2z_i) + \tilde{u}_i$, por lo que si definimos nuevas variables: $\tilde{y}_i = y_i - z_i$, $\tilde{x}_i = x_i + 2z_i$, el Modelo Restringido es: $\tilde{y}_i = \beta_0 + \beta_1 \tilde{x}_i + \tilde{u}_i$, en el que estimamos los coeficientes β_0 y β_1. La estimación de β_2 se obtiene de: $\hat{\beta}_2 = 1 + 2\hat{\beta}_1$. En este caso, $q = 1$ y $ncoefs = 3$.

Podemos utilizar el estadístico t, sin más que observar que $Var(2\hat{\beta}_1 - \hat{\beta}_2 - 1) = 4Var(\hat{\beta}_1) + Var(\hat{\beta}_2) - 2Cov(\hat{\beta}_1, \hat{\beta}_2)$, cuyo valor numérico tendremos que obtener a partir de la matriz de varianzas-covarianzas de los estimadores de mínimos cuadrados.

Ejemplo: Supongamos que en el modelo $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i$, queremos contrastar la restricción:

$$\begin{aligned} H_0 &: \beta_1 + 2\beta_2 = 3 \\ H_1 &: \beta_1 + 2\beta_2 \neq 3 \end{aligned}$$

La restricción puede escribirse: $\beta_1 = -2\beta_2 + 3$. Si sustituimos la restricción en el modelo tenemos, $y_i = \beta_0 + (-2\beta_2 + 3)x_i + \beta_2 z_i + \tilde{u}_i$, por lo que si definimos nuevas variables: $\tilde{y}_i = y_i - 3x_i$, $\tilde{z}_i = z_i - 2x_i$, el Modelo Restringido es: $\tilde{y}_i = \beta_0 + \beta_2 \tilde{z}_i + \tilde{u}_i$, en el que estimamos los coeficientes β_0 y β_2. La estimación de β_1 se obtiene de: $\hat{\beta}_1 = 3 - 2\hat{\beta}_2$. El contrastes de hipótesis se resuelve como en los casos ateriores, mediante el estadístico F que se obtiene comprando las Sumas de Cuadrados de Residuos de ambos modelos, Restringido y Sin Restringir. En este caso, $q = 1$ y $ncoefs = 3$.

9.1 Contraste de significación global del modelo (Análisis ANOVA)

Por ejemplo, en un modelo de regresión múltiple, $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ consideremos el contraste: $H_0 : \beta_1 = \beta_2 = .. = \beta_k = 0$. El número de restricciones es igual al número de condiciones de igualdad, es decir, k.Este es el contraste que conocemos como contraste de *Ausencia de Significación Estadística (o de Capacidad Explicativa o de Contenido Informativo) Global* del modelo.

El Modelo Restringido es: $y_i = \beta_0 + u_i$, que es el modelo constante que antes analizamos. Como sabemos, este es un modelo con $\hat{\beta}_0 = \bar{y}, SCR = N.Var(y) = ST$, y $R^2 = 0$. Por tanto, el estadístico para el contraste de hipótesis es,

$$\frac{ST - SCR}{SCR} \frac{N - (k+1)}{k} \sim F_{k,N-(k+1)}$$

donde hesmo tenido en cuenta que el Modelo sin Restringir es el modelo original y, por tanto, $SCRSR = SCR$, que el número de coeficientes estimados es igual a $k + 1$, y que el número de restricciones es igual al número de variables explicativas, k. Notemos que en el contraste de Ausencia de Significación Global, no se impone que la constante del modelo sea igual a cero. En primer lugar, no es preciso pues, com sabemos, dicha constante no explica nada, como recoge el hecho de que el $R2$ del modelo constante sea igual a cero. En segundo lugar, porque incluso sin ninguna variable explicativa, la variable y tiene media, en general no nula, y necesitamos un término constante para igualar las medias muestrales a ambos lados de la igualdad.

Recordando la definición del coeficiente de determiación, el estadístico anterior puede escribirse:

$$\frac{SE}{SCR} \frac{N - (k+1)}{k} = \frac{R^2}{1 - R^2} \frac{N - (k+1)}{k} \sim F_{k,N-(k+1)}$$

10 Contrastes de cambio estructural

En sentido general, podemos decir que existe cambio estructural en un modelo de regresión cuando los parámetros del mismo no son constantes a lo largo de toda la muestra. El problema es típico de muestras de datos temporales aunque como veremos más adelante, puede plantearse asimismo como contraste de homogenidad en muestras de corte transversal.

La diferencia es únicamente conceptual. A lo largo de la muestra temproal disponibel para una estimación es muy posible que hayan ocurrido sucesos que puedan hacer sospechar al analista que la relación entre variables explicativas y variable dependiente haya cambiado. Puede tratarse de un cambio en el modo en que se pone en práctica la política monetaria, si estamos estimando un modelo de determinación y previsión de la tasa de inflación, o la implantación de una normativa disuasoria del consumo de tabaco si estamos estimando un modelo de demanda de dicho producto. Estos sucesos podrían hacernos pensar que el valor numérico de los coeficientes del modelo, es decir, los efectos que sobre la variable dependiente tienen variaciones en las variables explicativas, hayan cambiado en ese momento. Por tanto, tendríamos inestabilidad paramétrica en algún punto concreto de la muestra.

En muestras de sección cruzada o de corte transversal, hay a veces razones para pensar que los coeficientes del modelo sean distintos entre dos o más submuestras. Es decir, hay heterogeneidad paramétrica entre submuestras. Podríamos estar analizando la rentabilidad de los activos de las empresas industriales en función de su tamaño, la composición de los mismos, su modo de financiación, etc, pero quizá el modo en que estas variables afectan a la rentabilidad no es el mismo para empresas exportadoras que para empresas no exportadoras. El ejemplo que vemos más adelante vuelve a considerar la posibilidad de que el modo en que el mercado de trabajo remunera, por medio de los salarios, el nivel educativo o la experiencia es distinto en hoombres que en mujeres. De este modo, la discriminación salarial se muestra como un caso de heterogeneidad paramétrica.

La hipótesis nula es siempre la hipótesis de homogeneidad paramétrica entre submuestras, o si se quiere, de estabilidad estructural, o de ausencia de cambio estructural. Todos ellos son conceptos análogos.

10.1 Test de estabilidad estructural de Chow

En muchas ocasiones, cuando se habla de cambio estructural en regresión, se tiene en mente una muestra de datos temporales, y el investigador se cuestiona si se ha producido en algún momento una cambio en la relación existente entre variables explicativas y variable dependiente. Un ejemplo sería la estimación de un modelo de determiandción de la tasa de inflación:

$$\pi_t = \beta_0 + \beta_1 r_t + \beta_2 m_t + u_t$$

en función del crecimiento del agregado monetario y del nivel del tipo de interés. Imaginemos el caso de España, y el investigador, que dispone de datos mensuales de los ultimos 40 años, se plantea la posibilidad de que la entrada en el euro haya supuesto un cambio significativo en el modo en que los tipos de interés o el crecimiento monetario afecten a la tasa de inflación. Alternativamente, un investigador que trabajase con datos de EEUU podría plantearse si el cambio a finales de los 70 de una política de control de agregados monetarios a una política de control de tipos de interes ha podido afectar a la relación que pretende estimar.

En estos casos, generalmente se considera que han podido variar todos los coeficientes del modelo. Las restricciones consisten entonces en suponer que los coeficientes son iguales antes que después del hecho que se considera relevante. El Modelo Restringido impone las restricciones y estima, por tanto, un único conjunto de coeficientes, utilizando toda la muestra. La Suma de Cuadrados de Residuos restringida es la que se obtiene de esta estimación.

El Modelo Sin Restringir, al no imponer las restricciones, permite que los coeficientes sean diferentes, y estima el modelo dos veces: una, con los datos hasta la fecha en que se produjo el acontecimiento de interés, y otra, con los datos posteriores a dicha fecha. La Suma de Cuadrados

Sin Restringir es el agregado de las Sumas de Cuadrados de residuos de ambas estimaciones. El estadístico F se forma como antes teniendo en cuenta que el número de grados de libertad que aparece en el mismo es el del Modelo Sin Restringir. Como este se forma estimando dos veces, dicho factor es igual a $n-2k$. Esta es la razón por la que en el ejemplo anterior, era igual a $n-6$, cuando había 3 variables explicativas. Por último, el número de restricciones, que habitualmente denotamos por q es, en este caso, igual al número de variables explicativas,[35] k.

Por tanto, el estadístico F para este contraste suele escribirse:

$$\frac{SCRR-(SCR_1+SCR_2)}{SCR_1+CSR_2}\frac{n-2k}{k}\sim F_{n-2k,k}$$

Un caso especialmente interesante se refiere a la posibilidad de que el investigador dude de que los últimos datos recibidos, en numero reducido, quizá sólo 2 o 3, no respondan al modelo previo, debido a que se ha producido recientemente un cambio estructural.

En este caso, utilizamos el siguiente hecho:

"La estimación de minimos cuadrados de un modelo de regresión con k coeficientes, utilizando una muestra de n observaciones, con $n \leq k$ genera un $R^2=1$ y una Suma de Cuadrados de Residuos igual a 0, ya que todos los residuos son iguales a cero, por poder efectuar un ajuste perfecto del modelo a la muestra."

De hecho, si $n=k$, existe un único estimador de mínimos cuadrados, mientras que si $n<k$, existe todo un continuo de dichos estimadores, todos ellos con las propiedades que acabamos de referir. El lector puede considerar gráficamente la estimación de una regresión simple cuando dispone alternativamente: a) de una única observación para (y,x) ,o b) de dos observaciones para dichas variables.

En tal caso, el estadístico anterior se convierte en:

$$\frac{SCRR-SCR_1}{SCR_1}\frac{n-2k}{k}\sim F_{n-2k,k}$$

Nótese que no hay razón para variar el número de grados de libertad del Modelo Sin Restringir.

10.2 Variables ficticias en la modelización del cambio estructural

El cambio estructural puede tratarse, de modo alternativo al que acabamos de indicar, mediante el uso de una variable ficticia que discrimine entre las dos submuestras. Así, supongamos que disponiendo de una muestra de tamaño T, tenemos la sospecha de que ha podido producirse un cambio estructural a partir del período t_0 ($t_0<T$). definimos una variable ficticia D_t que toma el valor 1 para los años posteriores a t_0, y es igual a 0 en los años previos a dicha fecha. A continuación, si nuestro modelo es:

$$y_t=\beta_0+\beta_1 x_{1t}+\beta_2 x_{2t}+u_t,\ t=1,2,...T$$

estimamos el modelo,

$$y_t=\beta_0+\alpha_0 D_t+\beta_1 x_{1t}+\alpha_1(D_t x_{1t})+\beta_2 x_{2t}+\alpha_2(D_t x_{2t})+u_t,\ t=1,2,...T$$

[35] Salvo que supongamos que algún coeficiente es invariante entre submuestras. En tal caso, hay que realizar el test mediante el uso de variables ficticias, como explicamos más adelante.

El uso de variables ficticias es especialmente útil para examinar la posible heterogeneidad paramétrica en muestras de corte transversal, como ilustra el ejemplo considerado a continuación.

Si estimamos este modelo una sóla vez, con toda la muestra, el modelo se desdobla en dos ecuaciones:

$$\begin{aligned} t &\leq t_0 \Rightarrow y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t,\ t = 1, 2, ...t_0 \\ t &> t_0 \Rightarrow y_t = (\beta_0 + \alpha_0) + (\beta_1 + \alpha_1) x_{1t} + (\beta_2 + \alpha_2) x_{2t} + u_t,\ t = 1, 2, ...t_0 \end{aligned}$$

El contraste de la hipótesis nula de homogenidad es el contraste conjunto de k hipótesis:

$$H_0 : \alpha_0 = \alpha_1 = ... = \alpha_k = 0$$

frente a la hipótesis alternativa de que al menos uno de dichos coeficientes sea distinto de cero.

10.3 Variables ficticias y cambio estructural

Un ejemplo de esta situación sería una posible discriminación salarial en la que la remuneración que recibe un trabajador como salario por su experiencia profesional, es distinta para hombre y para mujeres. Este hecho podría investigarse mediante la consideración de una variable ficticia, $Mujer_i$, que tomase el valor 1 en el caso de las mujeres incluidas en la muestra, y fuese igual a cero para los hombres. Estimaríamos un modelo:

$$Salario_i = \beta_0 + \beta_1 Educación_i + \beta_2 Experiencia_i + \beta_3 (Experiencia_i.Mujer_i)$$

y contrastaríamos la discriminación salarial del tipo citado mediante la hipótesis paramétrica: $H_0 : \beta_3 = 0$, frente a la alternativa unilateral: $H_1 : \beta_3 < 0$. Por supuesto que podrían incorporarse en la misma regresión otros tipos de posible discriminación. Si quisiéramos contrastar la hipótesis global de discriminación salarial de cualquier tipo, ya sea por razón de sexo, o por minusvaloración de la experiencia o del nivel educativo, estimaríamos el modelo:

$$\begin{aligned} Salario_i &= \beta_1 + \alpha_1 Mujer_i + \beta_2 Educación_i + \alpha_2 (Educación_i.Mujer_i) + \\ &\quad + \beta_3 Experiencia_i + \alpha_3 (Experiencia_i.Mujer_i) \end{aligned} \tag{35}$$

y contrastaríamos conjuntamente las 3 hipótesis: $H_0 : \alpha_1 = \alpha_1 = \alpha_2 = 0$.

Si la alternativa no considera un signo concreto, siendo de la forma: $H_1 : \alpha_1 \neq 0$, ó $\alpha_2 \neq 0$, ó $\alpha_3 \neq 0$, podríamos realizar el contraste estimando los Modelos Restringido y in Restringir, y comparando sus Sumas Residuales. El modelo Restringido sería:

$$Salario_i = \beta_1 + \beta_2 Educación_i + \beta_3 Experiencia_i \tag{36}$$

que se estimaría una sóla vez, utilizando todos los datos. El modelo Sin Restringir sería (35) , y formaríamos el estadístico:

$$\frac{SCRR - SCR}{SCR} \frac{n-6}{3} \sim F_{n-6,3}$$

Como es sabido, el Modelo Sin Restringir puede estimarse también mediante dos regresiones como (36) , una estimada con los datos de hombres, y otra estimada con los datos de mujeres. El agregado de las Sumas de Cuadrados de Residuos obtenidas con las dos submuestras sería igual a la Suma de Cuadrados de residuos que obtendríamos estimando con toda la muestra el modelo (35) .

10.4 Estadísticos CUSUM y CUSUMSQ[36]

Se utilizan con datos temporales. Si estimamos la regresión simple:

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

con una muestra hasta T, se conoce como residuo recursivo el error que se comete al ajustar el valor numérico de y_{T+1} con dichas estimaciones, es decir,

$$\hat{e}_t = y_{T+1} - \hat{\beta}_0^T - \hat{\beta}_1^T x_{T+1}$$

Este residuo recursivo puede interpretar como el error cometido al utilizar las estimaciones obtenidas con datos hasta T para predecir el valor de y_{T+1}, suponiendo que x_{T+1} es conocido. Suponemos que el valor futuro de la variable explicativa x_{T+1} es conocido. Dicho error es aleatorio, pues con los datos hasta T ignoramos lo que puede suceder en $T+1$, y puede demostrarse que tiene una varianza:[37]

$$Var\left(\hat{e}_t\right) = \sigma_u^2 \left(1 + \frac{x_{T+1}^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\right)$$

Si normalizamos el residuo recursivo mediante el cociente:

$$\tilde{e}_t = \frac{\hat{e}_t}{\sqrt{1 + \frac{x_{T+1}^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}}}$$

tenemos un residuo recursivo normalizado con varianza constante, σ_u^2.

El estadístico CUSUM se define a partir de sumas de los residuos recursivos normalizados:

$$W_t = \sum_{s=k+1}^{s=t} \frac{\tilde{e}_s}{\hat{\sigma}},\ k+1 \leq t \leq T$$

donde $\hat{\sigma}$ se estima mediante:

$$\hat{\sigma}^2 = \frac{1}{T-k} \sum_{s=k+1}^{s=T} \left(\tilde{e}_s - \overline{\tilde{e}}\right)^2$$

con $\overline{\tilde{e}} = \frac{1}{T-k} \sum_{s=k+1}^{s=T} \tilde{e}_s$.

Bajo la hipótesis nula de estabilidad, W_t tiene esperanza igual a cero, y varianza aproximadamente igual al número de residuos acumulados, $T-k$, de modo que el contraste consiste en superponer los valor numéricos de las sucesión W_t a un par de líneas rectas que delimitan un intervalo de amplitud creciente. Se construyen bandas de confianza mediante líneas rectas que unen los puntos $\left(k, \pm a\sqrt{T-k}\right)$ y $\left(k, \pm 3a\sqrt{T-k}\right)$. Al 95% de confianza, $a = 0,948$, mientras que al 99% de confianza, $a = 1,143$, y se rechaza la hipótesis de estabilidad en los coeficientes del modelo si la secuencia W_t traspasa dichas bandas.

[36] En el caso de una regresión múltiple, las expresiones de la varianza del residuo recursivo que aparecen en esta sección son más complejas. Sin embargo, la construcción de los estadísticos, su interpretación y la resolución de los contrastes de estabilidad son iguales a los que aquí se presentan.

[37] La extensión al caso en que se dispone de varias variables explicativas es inmediato.

El estadístico CUSUMSQ utiliza los cuadrados de los residuos recursivos normalizados:

$$S_t = \frac{\sum_{s=k+1}^{s=t} \tilde{e}_s^2}{\sum_{s=k+1}^{s=T} \tilde{e}_s^2}$$

donde k es el número de variables explicativas. Hemos de comenzar estimando con un número de datos al menos igual a k. Sin embargo, si no comenzamos a partir de un tamaño muestral suficientemente grande, las primeras estiamciones de los coeficientes no serán muy precisas y tenderán a reflejar inestabilidad de manera espúria.

Los residuos recursivos son independientes si los residuos originales también lo eran. En tal caso, cada término de la suma es una variable aleatoria con distribución chi-cuadrado con un grado de libertad, por lo que:

$$E(S_t) = \frac{t-k}{T-k}$$

que comienza en 0 para $t = k$, y converge hacia 1 cuando $t = T$.

El contraste consiste en dibujar la secuencia temporal de S_t así como bandas de confianza de amplitud C (dada por unas tablas para este estadístico) alrededor de $\frac{t-k}{T-k}$. Nuevamente, si S_t se sale de las bandas, se rechaza la hipótesis nula de ausencia de cambio estructural, que podría afectar a los oeficientes del modelo, o a la varianza del término de error.

Por último, bajo la hipótesis nula, la media muestral de los residuos recursivos se distribuye como una Normal (si los errores originales eran Normales), con esperanza cero y varianza $\frac{\sigma_u^2}{T-k}$ por lo que puede construirse un contraste tipo-t :

$$\frac{\sqrt{T-k}}{\hat{\sigma}_u}\bar{\tilde{e}} \sim t_{T-k-1}$$

donde $\hat{\sigma}_u$ se ha estimado como se explicó mas arriba.

10.5 Ejemplo: Discriminación salarial: contraste de discriminación salarial mediante variables ficticias

Los residuos de la regresión combinada tienen, por supuesto, una media muestral igual a cero. Sin embargo, su media es de -31,8 entre las mujeres (observaciones con $male = 0$) y de 20,6 entre hombres (observaciones con $male = 1$), sugiriendo claramente un diferente comportamiento de los salarios entre ambos grupos. Ello significa que dados un mismo nivel de educación y de experiencia, el salario es 52,4 Bef. inferior para las mujeres que para los hombres; esta observación constituye evidencia más clara a favor de discriminación salarial contra las mujeres. Cuando incluimos en la regresión anterior la variable ficticia *MALE*, obtenemos un R^2 ajustado de 0,364, con $\hat{\sigma}_u = 143,1$, y el ratio de ajuste aumenta a 0,20. La regresión estimada es,

$$\underset{\substack{(15,6)\\(0,39)\\(20,1)}}{Salario = 8,62 +} \quad \underset{\substack{(3,25)\\(24,6)}}{7,76\ Experiencia +} \quad \underset{\substack{(7,77)\\(7,0)}}{80,11\ Educación +} \quad 54,30\ Male \tag{37}$$

que sugiere que tanto el nivel educativo como la experiencia profesional explican el salario de un trabajador, y también que existen diferencias entre los salarios que reciben mujeres y hombres que tienen un mismo nivel educativo e igual experiencia laboral.

Puesto que la variable $Male$ toma el valor 0 para las mujeres, y el valor 1 para los hombres, la regresión anterior equivale al par de regresiones,

$$\begin{aligned} Salario &= 62,92 + 7,76 Experiencia + 80,11 Educación, \ para \ los \ \text{hombres} \\ Salario &= 8,62 + 7,76 Experiencia + 80,11 Educación, \ para \ las \ mujeres \end{aligned}$$

Como ejemplo, nuestra estimación sugiere que un hombre de 10 años de experiencia laboral y 4 años de educación recibiría un salario de $62.92 + (7.76)10 + (80.11)4 = 460.96$, mientras que una mujer de igual cualificación recibiría un salario igual a $8.62 + (7.76)10 + (80.11)4 = 406.66$.

Todo ello proporciona evidencia clara acerca de discriminación salarial en el sentido antes descrito: a igualdad de experiencia y educación, un hombre recibe un salario superior en 54,3 unidades monetarias al de una mujer[38]. No parece preciso contrastar explícitamente tal hipótesis. Además, el histograma de frecuencias de la variable salarios se desvia claramente respecto de una distribución Normal por lo que la teoría estadística habitual basada, entre otros, en el supuesto de Normalidad de la perturbación del modelo y el carácter determinista de las variables explicativas, no es estrictamente aplicable en este caso. Afortunadamente, como hemos dicho, tampoco parece necesaria su aplicación.

Las dos regresiones anteriores se diferencian tan sólo en la estimación de la constante, por lo que gráficamente pueden visualizarse como dos líneas de regresión paralelas, con igual pendiente, pero con mayor ordenada en el origen para la regresión de hombres que para la correspondiente a las mujeres. Es decir, la regresión de hombres está sistemáticamente por encima de la de las mujeres, lo que hace que para cada posible combinación de nivel educativo y experiencia, el salario de los hombres (la variable dependiente del modelo) sea mayor que el de las las mujeres[39].

10.5.1 Aspectos concretos de discriminación salarial

Una vez obtenida la evidencia anterior acerca de la existencia de discriminación salarial, podríamos profundizar algo más, en el sentido de preguntarnos si la discriminación salarial en contra de la mujer tiene carácter general, que es lo que hemos supuesto hasta ahora, o alguna forma específica. Concretamente, con la información disponible, podríamos contrastar si la mujer trabajadora es discriminada al reconocer en términos salariales, bien la experiencia profesional, o bien el nivel educativo del trabajador. Para ello necesitamos definir nuevas variables, mediante el producto de la variable ficticia por cada una de las mencionadas. Por ejemplo, en la regresión,

$$Salario_i = \beta_0 + \beta_1 Educación_i + \beta_2 \left(Educación_i . Male_i\right) + u_i \tag{38}$$

la variable producto $Educación_i . Male_i$ toma un valor igual a cero para las mujeres incluidas en la muestra, mientras que coincide con la variable $Educación_i$ en el caso de los hombres. Por tanto el modelo anterior equivale a los dos modelos,

$$\begin{aligned} Salario_i &= \beta_0 + \beta_1 Educación_i + u_i \ para \ las \ mujeres \\ Salario_i &= \beta_0 + (\beta_1 + \beta_2) \, Educación_i + u_i \ para \ los \ \text{hombres} \end{aligned}$$

[38] No es casualidad que esta diferencia coincide con la disparidad antes mencionada entre las medias muestrales de los residuos en ambos grupos de trabajadores: 20,6-(-31,8)=52,4

[39] Si bien este argumento no es estrictamente válido porque con dos variables explicativas, experiencia y educación, no tenemos rectas de regresión, sino planos de regresión. Sin embargo, la idea intuitiva es la misma.

en los que si $\beta_2 = 0$ ambos modelos coinciden, lo que significaría que el salario recoge el nivel de educación del trabajador en igual manera en hombres que en mujeres, no habiendo discriminación salarial en este sentido. Así, el contraste de significación del coeficiente β_2 en el modelo (38) equivale a un contraste de discriminación en el reconocimiento del nivel educativo del trabajador. Este modelo, al igual que el modelo que se obtiene aplicando un tratamiento análogo a la experiencia laboral, aparecen estimados en el fichero de trabajo ($REG\_W\_CROSSEXPER, REG\_W\_CROSSEDU$). Estimar (38) equivale a considerar dos rectas de regresión con igual ordenada en el origen, β_0, pero con una pendiente diferente para hombres y para mujeres. Será mayor la primera si $\beta_2 > 0$, siendo menor si $\beta_2 < 0$. Sin embargo, dada la evidencia ya obtenida acerca de la posible discriminación salarial en contra de las mujeres, en las regresiones citadas se ha incluido asimismo explícitamente la variable *Male*, con el objeto de captar cualquier posible evidencia de discriminación sistemática. De este modo, las regresiones estimadas tienen distinta ordenada en el origen y distinta pendiente para hombres y mujeres.

Es interesante preguntarse en cuál de los aspectos, experiencia o educación, se ve más discriminada la mujer. En ambas regresiones, la variable ficticia $MALE$ y los efectos cruzados, representados por las variables producto, tienen estadísticos t inferiores a 2,0 en valor absoluto. Estas son situaciones que suelen producirse en el análisis aplicado, generando muchas dudas en el investigador, que podría comenzar a cuestionarse si realmente hay diferencias salariales entre hombres y mujeres. Sin embargo, no hay razón para ello: desde que hemos estimado el modelo (37) , sabemos que las dos regresiones que ahora consideramos están mal especificadas, pues falta un indicador en cada una de ellas. En consecuencia, la omisión de variables explicativas relevantes hace que tanto las estimaciones numéricas de los coeficientes, como de sus desviaciones típicas, sean sesgadas.

Si, a pesar de ello, nos atenemos a las estimaciones obtenidas, estas regresiones muestran que cada año de experiencia se valora a los hombres un 37,7% más que a las mujeres $(1,57/4,16)$, mientras que el paso de un nivel educativo al siguiente se valora en los hombres un 9,8% más que en las mujeres $(5,71/58,4)$. Por tanto, parece haber mayor evidencia de discriminación en el reconocimiento de la experiencia profesional que en el reconocimiento del nivel educativo. En los dos casos estimamos una recta con mayor ordenada en el origen y mayor pendiente para los salarios de hombres que para los de mujeres. Esto es evidencia clara sugiriendo discriminación en contra de las mujeres.

Aunque este análisis ha sido ilustrativo, no queremos que la posible detección de evidencia sugiriendo una valoración inferior de la educación en mujeres que en hombres pueda deberse a una mala especificación de los posibles modos de discriminación. Para ello, incluimos ahora los dos indicadores, experiencia y nivel educativo en el modelo de salarios, permitiendo que ambos coeficientes, así como la ordenada en el origen, difieran para hombres y mujeres. Así, necesitamos estimar una regresión,

$$\begin{aligned} Salario_i = {} & 42,48 + \underset{\substack{(5,72)\\(13,2)}}{75,54}\ Educación_i + \underset{\substack{(6,95)\\(0,97)}}{6,77}\ (Educación_i.Male_i) + \\ & + \underset{\substack{(0,64)\\(10,3)}}{6,61}\ Experiencia_i + \underset{\substack{(0,80)\\(2,24)}}{1,80}\ (Experiencia_i.Male_i) + \underset{\substack{(31,5)\\(0,04)}}{1,30}\ Male + u_i \end{aligned} \tag{39}$$

con los resultados que se incluyen en el archivo de trabajo ($REG\_W\_CROSSDOBLE$). El R^2 ajustado de 0,366, con $\hat{\sigma}_u = 142,99$, y ratio de ajuste $1 - \frac{\hat{\sigma}_u}{\sigma_y} = 0,19$. Nuevamente, estimamos

que la experiencia profesional se valora en los hombres un $1,80/6,61 = 27,2\%$ más que en las mujeres, y la educación en un $6,77/75,54 = 8,9\%$ más en hombres que en mujeres. La variable ficticia $MALE$ tiene una contribución reducida, como indica su coeficiente estimado, pero ello es sólo aparente, pues sus posibles efectos están recogidos asimismo a través de la variables de interacción $Experiencia_i.Male_i$ y $Educación_i.Male_i$.

De acuerdo con esta estimación, un varón recibe por cada año de experiencia profesional 8,41 Bef., mientras una mujer recibe tan sólo 6,61 Bef.. Por cada salto en el nivel educativo, un hombre ve incrementado su salario en 82,31 Bef., mientras que dicho incremento es de 75,54 para la mujer. La discriminación salarial estimada entre trabajadores de distinto sexo, pero de igual nivel educativo y experiencia laboral es de 1,30 Bef., más 1,80 Bef. por el número de años de experiencia, más 6,77 Bef. por el número asignado a su nivel educativo común.

Como alternativa, si hubiéramos optado por aceptar la restricción $\beta_2 = 0$ (coeficiente de $Educación_i.Male_i$) como razonable, habríamos estimado el modelo,

$$\begin{aligned} Salario_i &= \underset{(17,26)}{23,78} + \underset{(3,25)}{80,12} Educación_i + \underset{(0,62)}{6,76} Experiencia_i + \\ &\quad + \underset{(0,77)}{1,58} \left(Experiencia_i.Male_i\right) + \underset{(14,85)}{28,41} Male + u_i \end{aligned}$$

que sugiere que hay una evidencia sistemática de discriminación que hace que, a igual nivel educativo, una mujer sin experiencia laboral reciba 28,41 Bef. menos que un trabajador varón que asimismo carezca de experiencia laboral. Además, un varón recibe 8,34 Bef. por cada año de experiencia profesional (la suma de 8,76 y 1,58 Bef.), mientras que una mujer recibe tan sólo 6,76 Bef.. Por tanto, la discriminación entre trabajadores de igual nivel educativo pero de distinto sexo se estima en 28,41 Bef. *más* 1,58 Bef. por el número de años de experiencia profesional de ambos trabajadores. Según este último modelo, el nivel educativo tiene un efecto igual sobre el salario de hombres y de mujeres, por lo que incorpora el supuesto de que no hay discriminación en la remuneración del mismo. Sin embargo, incorpora la idea de que la Experciencia se remunera de manera diferente a hombres y a mujeres.

10.5.2 ¿Existe evidencia de desigual remuneración de la educación entre hombres y mujeres?

El lector debe apreciar las similitudes y diferencias entre las conclusiones numéricas alcanzadas en los dos últimos modelos: la primera, que permite la posibilidad de que la remuneración salarial del nivel educativo sea distinta en hombres y mujeres, y la segunda, que impone la restricción de que dicha remuneración es igual entre ambos grupos de trabajadores. De acuerdo con el primero de los modelos, un aumento de nivel educativo incrementa el salario de hombres en 82,31 Bef., y el de las mujeres en 75,54; según el modelo que acabamos de estimar, el incremento es de 80,12 Bef., común a hombres y mujeres que, por supuesto, está entre los dos valores que estimamos con el modelo que incorporaba este tipo de discriminación. Ambos modelos implican discriminación salarial por razón de sexo, y también que la experiencia laboral se remunera de manera diferenciada en hombres y en mujeres. El primer modelo afirma lo mismo acerca del nivel educativo, mientras que el último modelo impone igual remuneración salarial por nivel educativo en hombres que en mujeres.

Es difícil decidir cuál de los dos modelos es preferible. La última regresión tiene practicamente el mismo R^2 y la misma desviación típica residual que la anterior. En consecuencia, la aplicación de los contrastes estadísticos habituales, basados en Normalidad del término de error, variables

explicativas deterministas, etc., no permiten distinguir entre ambos modelos. En esta situación, parece preferible escoger el modelo más sencillo, y concluir que no hay evidencia en la muestra de trabajadores disponible acerca de diferencias en el reconocimiento salarial del nivel educativo entre trabajadores de ambos sexos.

Dada la similitud de estadísticos, es frecuente que el investigador concluya que ambos modelos son idénticos. Sin embargo, esto no es completamente exacto. Una interpretación alternativa del reducido estadístico t del producto $Educación{*}Male$ es que, aunque el nivel educativo recibe distinta valoración salarial en hombres que en mujeres, las diferencias no se miden con suficiente precisión con los datos disponibles. Esta apreciación se basaría en el hecho de que el efecto discriminatorio estimado en (39) es de un 9%, que no parece que pueda considerarse despreciable. El problema es que la desviación típica con que se estima el coeficiente es prácticamente de igual tamaño que éste, revelando que es un problema de reducida precisión (alta varianza) en las estimación, lo que conduce a un estadístico t reducido, en torno a 1,0. En definitiva, el primero de los dos modelos permite más variedad salarial y puede considerarse, en tal sentido, más informativo.

El peligro es que, por estimar tal parámetro con baja precisión, las inferencias numéricas que se obtengan sobre los salarios estén poco justificadas. Dichas estimaciones son, en algunos casos particulares,

SALARIOS	*Modelo restringido*			*Modelo no restringido*		
	Mujeres	*Hombres*	*Ratio*	*Mujeres*	*Hombres*	*Ratio*
$Educ = 1; Exper = 3$	124,2	157,3	78,9%	137,9	151,3	91,1%
$Educ = 1; Exper = 17$	218,9	274,1	79,8%	230,4	269,1	85,6%
$Educ = 1; Exper = 30$	306,7	382,5	80,2%	316,3	378,4	83,6%
$Educ = 4; Exper = 3$	364,5	397,7	91,7%	364,5	398,3	91,5%
$Educ = 4; Exper = 17$	459,2	514,5	89,3%	457,0	516,0	88,6%
$Educ = 4; Exper = 30$	547,1	622,9	87,8%	542,9	625,3	86,8%

Para niveles educativos bajos, el modelo restringido implica diferencias salariales entre hombres y mujeres bastante mayores que el modelo no restringido. Lo contrario ocurre para niveles educativos altos, en los que el modelo restringido genera menores diferencias salariales entre hombres y mujeres. Es decir, el modelo que incluye explícitamente una valoración diferente para el nivel educativo de hombres y mujeres produce una estimación de la discriminación salarial más uniforme, sin que dependa del nivel educativo de los trabajadores que se comparen. Esta característica podría hacerlo preferible, pero ha de ser en última instancia la creencia del investigador acerca de si el nivel educativo se valora igual en ambos sexos o no, lo que debe llevarle a escoger uno u otro modelo.

Otra forma de analizar esta cuestión se basa en examinar los residuos del modelo restringido, el que estimamos en último lugar. Si la remuneración de la educación fuese sistemáticamente mayor en hombres que en mujeres, esperaríamos ver residuos mayores en hombres que en mujeres, dentro de cada nivel educativo. Ello se debe a que, al no permitir diferencias por sexo, nuestra estimación de la remuneración a la educación estaría comprendida entre los niveles percibidos por hombres y mujeres; de este modo, estaríamos infravalorando la remuneración a la educación percibida por los hombres, y sobrevalorando la que perciben las mujeres. En consecuencia, los residuos correspondientes a los varones deberían ser superiores a los de las mujeres en cada nivel educativo. Si examinamos los residuos del modelo para cada nivel educativo, obtenemos medias aritméticas de 8,6 y 3,1 para hombres y mujeres en el primer nivel educativo, 4,5 y 0,0 en el segundo, -1,6 y 11,3 en el tercero, -12,9 y -5,5 en el cuarto, y 17,9 y -6,3 en el superior. Por tanto, en este

sentido no surge evidencia sistemática de discriminación en la remuneración del nivel educativo, y el modelo restringido parecería suficiente.

10.5.3 Discriminación salarial como cambio estructural

Antes hemos planteado el contraste de discriminación a través del contraste de significación de un determinado coeficiente o conjunto de coeficientes del modelo. Otra manera de plantearlo sería a través de la *estabilidad* del modelo de determinación de salarios entre hombres y mujeres. Al igual que cuando examinamos la estabilidad temporal, se trataría, en definitiva, de dividir la muestra en dos submuestras, y comparar las estimaciones obtenidas en cada submuestra, tanto entre ellas, como con la estimación obtenida con la muestra completa. Si hay alguna variación entre los modelos de salarios estimados para hombres y mujeres, diremos también que hay cambio estructural en el mecanismo de determinación salarial, puesto que las respuestas a los determinantes del salario serían en tal caso distintos en ambos grupos de trabajadores.

Así, limitándonos por simplicidad al análisis de discriminación sistemática, podríamos estimar el modelo utilizando la submuestra de hombres en un caso[40], y la submuestra de mujeres, en otro, obteniendo:

$$\begin{aligned} Salario_i &= 42,48 + 75,54 Educación_i + 6,61 Experiencia_i + u_i,\ para\ mujeres \\ R^2 &= 0,365,\ \bar{R}^2 = 0,363,\ \hat{\sigma}_u = 153,03,\ SR = 20840842; \end{aligned} \tag{40}$$

$$\begin{aligned} Salario_i &= 43,78 + 82,31 Educación_i + 8,41 Experiencia_i + u_i,\ para\ \text{hombres} \\ R^2 &= 0,331,\ \bar{R}^2 = 0,328,\ \hat{\sigma}_u = 125,93,\ SR = 9134157; \end{aligned} \tag{41}$$

Por supuesto, que estas regresiones son comparables a la estimación del modelo (39) . De hecho, el lector debe comprobar que de dicho modelo se deducen dos relaciones, una válida para hombres y otra para mujeres, y que *coinciden exactamente* con las dos regresiones que acabamos de estimar.

El contraste de cambio estructural se basa en la comparación de las Sumas Residuales de los modelos restringido y sin restringir. Las restricciones en este caso consisten en el supuesto de que los coeficientes del modelo de salarios son iguales para hombres y mujeres; en tal caso, el modelo sería *estable* y concluiríamos que no hay evidencia de cambio estructural. El Modelo Sin Restringir está formado por las dos regresiones anteriores, mientras que el Modelo Restringido es (23). El estadístico tipo-F se construye, en este caso,

$$F_{q,gdl_{MSR}} = \frac{(SRR - SRS)/q}{SRS/gdl_{MSR}} = \frac{(31079583 - (20840842 + 9134157))/3}{(20840842 + 9134157)/(N_h + N_m - 6)} = 18,0$$

donde hemos utilizado que el número de restricciones es 3, el número de coeficientes que se supone igual en ambas submuestras. El modelo restringido impone la igualdad de coeficientes para hombres y mujeres, por lo que consiste en estimar una única regresión con todos los datos; es la ecuación (23) y genera, por tanto, una suma residual restringida $SRR = 31079583$. El modelo sin restringir pemite distintos coeficientes para hombres y mujeres; consiste en tratar las observaciones de ambas submuestras como independientes, estimando una regresión para cada una de ellas, como

[40] Introducir "1 1472 IF MALE=0" en la ventana "Sample" para estimar con observaciones de mujeres y "1 1472 IF MALE=1" para estimar con datos de trabajadores varones.

hemos hecho en (40) y (41). La Suma Residual de dicho modelo es el agregado de las Sumas Residuales de cada una de las dos regresiones, para hombres y mujeres. El número de grados de libertad de dicho modelo es igual a la suma de los grados de libertad de las dos regresiones: número de observaciones correspondientes a hombres, menos 3, más el número de observaciones correspondientes a mujeres, menos 3, $N_h + N_m - 6$.

El valor numérico del estadístico F está claramente por encima de los valores críticos de la distribución de probabilidad $F_{3,1466}$ a los niveles de significación habituales, 1%, 5%, 10%, por lo que rechazamos la hipótesis nula a cualquiera de dichos niveles. La hipótesis nula especifica la igualdad de coeficientes entre los modelos de salarios de hombres y mujeres, $H_0 : \beta_h = \beta_m$, por lo que concluiríamos que los modelos de salarios son diferentes. Hay que notar, sin embargo, que el estadístico utilizado sólo tendría distribución F si el término de error del modelo de salarios tuviera distribución Normal, lo que ya hemos comentado que parece altamente improbable, dado el histograma de frecuencias de los salarios.

Por sí sólo, este contraste no dice nada acerca del sentido en que se producen las diferencias, por lo que sería dificil concluir de él nada relativo a la discriminación salarial. Sin embargo, el hecho de que los coeficientes asociados tanto a nivel de educación como a la experiencia laboral sean mayores para los hombres que para las mujeres sugiere que las diferencias son en perjuicio de las mujeres. Como los términos constantes estimados son muy similares, es fácil ver que entre dos trabajadores de distinto sexo, pero de igual nivel educativo y experiencia laboral, el hombre recibe, generalmente, un salario superior al de la mujer.

10.5.4 Especificaciones con variables ficticias: contrastes de homogeneidad salarial entre grupos de trabajadores

En esta segunda parte del ejercicio, vamos a ilustrar el modo en que pueden utilizarse variables ficticias para proponer distintos grados de homogeneidad en el mecanismo de determinación salarial. Trabajando con la misma base de datos, continuamos utilizando el nivel educativo y el grado de experiencia laboral como posibles determinantes salariales. La hipótesis que ahra consideramos es que la experiencia laboral se remunera de igual modo en todos los niveles educativos. Dada la evidencia ya presentada acerca de la existencia de discriminación salarial por razón de sexo, utilizamos inicialmente las observaciones procedentes de trabajadores varones, para centrarnos exclusivamente en analizar las diferencias que puedan provenir de los dos factores citados.

10.5.5 Homogeneidad del modelo de salarios para distintos niveles educativos

Comentábamos al inicio de este ejercicio cómo las diferencias entre cada dos niveles sucesivos de educación pueden ser muy distintas, dependiendo de los niveles educativos que se comparen. Esto no ha sido recogido en nuestro análisis hasta ahora, porque la definición que se ha hecho de la variable educación conduce a que estimemos un incremento salarial con cada cambio de nivel educativo, con independencia de los niveles en los que se produzca. Para analizar esta cuestión en más detalle, estimamos por separado la contribución media de cada nivel educativo a la retribución salarial. Como ya sabemos que existe discriminación salarial por sexos, *vamos a utilizar únicamente las observaciones correspondientes a los hombres.*

Para ello, estimamos cinco regresiones del tipo (15), utilizando en cada caso datos de varones de un mismo nivel educativo[41]. Los resultados son,

[41] Al estimar la regresión, introducir en la ventana "Sample", el mensaje "1 1472 IF MALE=1 AND EDUC=1", e

$$
\begin{aligned}
nivel\ 1 \quad &: \quad Salario_i = 318,03 + 1,67 Experiencia_i + \hat{u}_i, \\
N_i \quad &= \quad 76,\ R^2 = 0,043,\ \ \hat{\sigma}_u = 67,14,\ SR = 333565.4; \\
nivel\ 2 \quad &: \quad Salario_i = 275,19 + 5,48 Experiencia_i + \hat{u}_i, \\
N_i \quad &= \quad 195,\ R^2 = 0,212,\ \ \hat{\sigma}_u = 107,83,14,\ SR = 2244014; \\
nivel\ 3 \quad &: \quad Salario_i = 312,10 + 6,63 Experiencia_i + \hat{u}_i, \\
N_i \quad &= \quad 258,\ R^2 = 0,218,\ \ \hat{\sigma}_u = 125,16,\ SR = 4010484; \\
nivel\ 4 \quad &: \quad Salario_i = 323,86 + 10,56 Experiencia_i + \hat{u}_i, \\
N_i \quad &= \quad 164,\ R^2 = 0,374,\ \ \hat{\sigma}_u = 133,71,\ SR = 2896499; \\
nivel\ 5 \quad &: \quad Salario_i = 389,43 + 13,46 Experiencia_i + \hat{u}_i, \\
N_i \quad &= \quad 200,\ R^2 = 0,257,\ \ \hat{\sigma}_u = 226,56,\ SR = 10162534;
\end{aligned}
\tag{42}
$$

en los que se aprecia un aumento en el coeficiente estimado para la variable Experiencia, según aumenta el nivel educativo. Ello sugiere que el reconocimiento salarial de la experiencia profesional entre varones es mayor cuanto más alto sea su nivel educativo, quizá por ser entonces la experiencia laboral de mayor calidad, un resultado sin duda interesante.

Otro resultado que surge de este modelo estimado es que el salario para trabajadores sin experiencia laboral es creciente con el nivel educativo, excepto entre los dos primeros niveles. Esta comparación no es, sin embargo, la más interesante, por cuanto que apenas hay trabajadores sin experiencia laboral. Otra manera de interpretar el modelo consiste en acudir al promedio de la experiencia laboral, que es de 17,22 años para toda la muestra, pero es[42] de 26,57 años para los hombres de nivel educativo 1, siendo de 20,42 años, 18,28, 16,19, y 15,85 años para los restantes niveles educativos. Por tanto, el salario medio para los trabajadores del primer nivel educativo se estima en $318.03 + (1.67)(26.57) = 362.4$, siendo para los sucesivos niveles: $275.19 + (5.48)(20.42) = 387.09$, $312.10 + (6.63)(18.28) = 433.3$, $323.86 + (10.56)(16.19) = 494.83$, $389.43 + (13.46)(15.85) = 602.77$.

Así, en promedio, un trabajador varón del nivel educativo 2 recibe un salario superior en 24.7 Bef. al del nivel educativo 1. Las remuneraciones promedio asignadas a los cambios sucesivos en nivel educativo son: 46,2 Bef. entre los trabajadores de niveles educativos 2 y 3; 61,5 Bef. para el salto de niveles educativos 3 a 4, y 108 Bef. para el paso de nivel educativo 4 a nivel 5.

Como se ve, estimamos una remuneración creciente para el salto entre cada par de niveles educativos sucesivos, valorándose más un aumento de nivel educativo cuanto más alto sea el nivel educativo de partida. Esto hace que el modelo (41) sea excesivamente restringido; en él, estimábamos en 82,31 Bef. la valoración de cada nivel educativo adicional, con independencia del nivel de partida. Tal estimación debe verse como un promedio de las cuatro remuneraciones que calculamos a partir de (42), pero es inapropiada, dados los resultados de este último modelo.

Un modelo algo menos resringido que (41) sería,

$$
Salario_i = \beta_0 + \beta_1 Educación_i + \beta_2 Experiencia_i + \beta_3 \left(Educación_i . Experiencia_i\right) + u_i,\ para\ \text{hom}bres \tag{43}
$$

ir variando el código asignado a EDUC, de 1 a 5.

[42] Tras marcar la variable Experiencia en el archivo de trabajo, entrar en "View/Descriptive Statistics/Statistics by Classification" y escribir en la ventana "Series/Group for Classify": EDUC*MALE.

Este modelo genera, para los distintos niveles educativos,

$$\begin{aligned}
\text{Nivel educativo 1} &: \quad Salario_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3).Experiencia_i + u_i, \\
\text{Nivel educativo 2} &: \quad Salario_i = (\beta_0 + 2\beta_1) + (\beta_2 + 2\beta_3).Experiencia_i + u_i, \\
\text{Nivel educativo 3} &: \quad Salario_i = (\beta_0 + 3\beta_1) + (\beta_2 + 3\beta_3).Experiencia_i + u_i, \\
\text{Nivel educativo 4} &: \quad Salario_i = (\beta_0 + 4\beta_1) + (\beta_2 + 4\beta_3).Experiencia_i + u_i, \\
\text{Nivel educativo 5} &: \quad Salario_i = (\beta_0 + 5\beta_1) + (\beta_2 + 5\beta_3).Experiencia_i + u_i,
\end{aligned}$$

que impone sobre (42) dos tipos de restricciones: *a*) que la diferencia en la remuneracion que recibe cada año de experiencia en trabajadores de dos niveles educativos sucesivos es la misma, β_3, independientente de los niveles de educación considerados, y *b*) que la diferencia salarial entre trabajadores de igual experiencia y niveles de educación consecutivos es siempre la misma, β_1.

Como consecuencia, este modelo implica que para caracterizar las diferencias salariales entre trabajadores de igual experiencia laboral sólo importa la diferencia que exista entre sus niveles educativos, pero no cuáles sean estos. Si k denota la diferencia entre los niveles educativos de dos trabajadores de igual experiencia, donde k podría ser igual a $0, 1, 2, 3$ ó 4, la diferencia entre sus salarios sería: $k\beta_1 + k\beta_3 Experiencia$, siendo $Experiencia$ el número de años de experiencia de ambos trabajadores.[43] El lector debe asegurarse de que entiende que en (42) no se ha impuesto ninguna de estas dos restricciones.

Por supuesto que estas restricciones pueden contrastarse conjuntamente utilizando los estadísticos habituales, sin más que considerar a (42) como Modelo Sin Restringir, y a (43) como Modelo Restringido.

Un modelo más restrictivo consideraría que la remuneración salarial a cada año de experiencia laboral del trabajador es independiente de su nivel educativo. Dicho modelo sería,

$$Salario_i = \beta_0 + \delta_2 D_{2i} + \delta_3 D_{3i} + \delta_4 D_{4i} + \delta_5 D_{5i} + \beta_1 Experiencia_i + u_i, \tag{44}$$

donde la variable ficticia D_{2i} se define mediante $D_{2i} = 1$ si la observación i-ésima se refiere a un trabajador varón en el segundo nivel educativo, y $D_{2i} = 0$ en todos los demás casos. El resto de las variables ficticias se define de manera análoga. Una vez estimado este modelo tendríamos para los varones del primer nivel educativo, $Salario_i = \hat{\beta}_0 + \hat{\beta}_1 Experiencia_i + \hat{u}_i$, para los del segundo nivel educativo: $Salario_i = (\hat{\beta}_0 + \hat{\delta}_2) + \hat{\beta}_1 Experiencia_i + \hat{u}_i$, y así sucesivamente; por ejemplo, para los trabajadores varones del más alto nivel educativo, tendríamos, $Salario_i = (\hat{\beta}_0 + \hat{\delta}_5) + \hat{\beta}_1 Experiencia_i + \hat{u}_i$. Al estimar δ_2 obtenemos el diferencial salarial que reciben los trabajadores varones del segundo nivel educativo respecto de los del primero, con independencia de su experiencia laboral. Los restantes coeficientes $\delta_3, \delta_4, \delta_5$, se interpretan de manera análoga, por lo que esperaríamos que fueran todos ellos positivos.

La estimación del modelo conduce a,

$$Salario_i = 133,19 + 73,12 D_{2i} + 142,27 D_{3i} + 208,54 D_{4i} + 313,62 D_{5i} + 8,00 Experiencia_i + u_i, \tag{45}$$

A diferencia de los modelos (43) y (42), la diferencia salarial entre trabajadores de distinto nivel educativo pero que tienen igual experiencia, se supone ahora independiente de dicho nivel de

[43] ¿Cuál sería la diferencia en salarios si no tuvieran el mismo grado de experiencia?.

experiencia. Por tanto, (44) es un modelo más restringido que los dos anteriores. En (44) tenemos cinco regresiones paralelas, con distinta ordenada en el origen pero igual pendiente. Por el contrario, (42) genera cinco rectas de regresión con distinta ordenada en el origen y diferente pendiente, eds decir, cinco rectas completamente distintas.

En (43) permitimos que la remuneración a la experiencia varíe con el nivel educativo, lo cual es más general que (44). Es algo más restrictivo en cuando que hace que las diferencias en la ordenada en el origen sean iguales entre niveled educativos. Los modelos (44) y (43) no son directamente comparables, pues uno no puede obtenerse imponiendo restricciones sobre el otro.

El modelo alternativo,

$$Salario_i = \beta_0 + \beta_2 Experiencia_i + \beta_3 \left(Educación_i.Experiencia_i\right) + u_i,$$

no es muy interesante, pues si se piensa que puede haber distinta remuneración salarial a la experiencia dependiendo del nivel educativo, es aún más probable que haya diferencias entre trabajadores de igual experiencia, pero distinto nivel educativo. En consecuencia, los modelos (44) y (43) son generalmente preferibles.

Un modelo aún más restrictivo impondría coeficientes comunes a todos los niveles educativos,

$$\begin{aligned} Todos\ los\ varones \quad &: \quad Salario_i = 360,25 + 5,73 Experiencia_i + \hat{u}_i, \\ N_i \quad &= \quad 893,\ R^2 = 0,093,\ \ \hat{\sigma}_u = 182,65,\ SR = 29723552; \end{aligned} \tag{46}$$

Este modelo equivale a imponer las restricciones $H_0 : \delta_2 = \delta_3 = \delta_4 = \delta_5$=0 en (44) , o bien $H_0 : \beta_1 = \beta_3 = 0$ en (43) , o $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5;\ \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$, en (42) , si denotamos por α_i los términos independientes y por β_i las pendientes de cada ecuación en este último modelo. Este útimo es un conjunto de 8 restricciones, el número de igualdades que se incluyen en la hipótesis nula. Cada uno de estos conjuntos de hipótesis puede contrastarse comparando las Sumas Residuales de los Modelos Restringido y Sin Restringir en cada caso, utilizando en el cálculo del estadístico tipo F el número de restricciones[44] y el número de grados de libertad del Modelo Sin Restringir: número de observaciones utilizadas en la estimación, menos número de coeficientes estimado en dicho modelo.

El modelo (41) queda en un terreno intermedio entre los anteriores: se obtiene a partir de (42) imponiendo las 7 restricciones $H_0 : \alpha_2 - \alpha_1 = \alpha_3 - \alpha_2 = \alpha_4 - \alpha_3 = \alpha_5 - \alpha_4;\ \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$. Se obtiene asimismo a partir de (43) , imponiendo la restricción $H_0 : \beta_3 = 0$, a partir de (44) imponiendo la restricción $H_0 : \delta_2 = \delta_3 = \delta_4 = \delta_5 = 0$. Por otra parte, el modelo (46) se obtiene a partir de (41) imponiendo la restricción $H_0 : \beta_1 = 0$. Por tanto, todas estos conjuntos de restricciones pueden contrastarse comparando las Sumas Residuales apropiadas, mediante el habitual estadístico tipo F en el que habrá que utilizar asimismo la información relativa al número de restricciones que se contrastan y el número de grados de libertad del Modelo Sin Restringir.

A modo de ejemplo, consideremos el contraste de las restricciones $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5;\ \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5$, sobre el modelo (42) , que nos llevaría al modelo (46) como versión restringida del anterior. El coeficiente estimado para la variable Experiencia en este último modelo es un promedio de los obtenidos en las cinco regresiones que configuran el modelo (42) , y lo mismo ocurre con la ordenada en el origen de la recta de regresión. El Modelo Sin Restringir es el conjunto de las cinco regresiones de (42), por lo que la Suma Residual Sin restringir es el agregado de

[44] Es decir, el número de igualdades utilizada para caracterizar la hipótesis nula.

las sumas residuales en ellas, mientras que el Modelo Restringido es el constituido por la última regresión, teniendo, por tanto el estadístico F,

$$\begin{aligned} F_{q,gdl_{MSR}} &= \frac{(SRR - SRS)/q}{SRS/gdl_{MSR}} = \\ &= \frac{(29723552 - (333565.4 + 2244014 + 4010484 + 2896499 + 10162534))/8}{(333565.4 + 2244014 + 4010484 + 2896499 + 10162534)/(893 - 10)} = 56,6 \end{aligned}$$

siendo 10 el número de coeficientes estimados en el Modelo Sin Restringir: dos coeficientes en cada una de las cinco regresiones. Se contrastan 8 restricciones, pues los dos coeficientes del modelo se hacen iguales en cuatro de las regresiones, a lo que ocurra en una de ellas. El estadístico F conduce a un rechazo tan claro de la hipótesis nula de igualdad de la regresión de salarios para los distintos niveles educativos, que, incluso si la distribución de probabilidad del término de error se desvía de la Normal, la evidencia obtenida contra la hipótesis de estabilidad de la regresión ha de juzgarse como muy clara. En consecuencia, el modelo (46) es inapropiado, por ser excesivamente restringido.

10.5.6 Variables ficticias y colinealidad perfecta

Alternativamente al modelo (41), podríamos haber especificado,

$$Salario_i = \delta_1 D_{1i} + \delta_2 D_{2i} + \delta_3 D_{3i} + \delta_4 D_{4i} + \delta_5 D_{5i} + \beta_1 Experiencia_i + u_i, \tag{47}$$

tras definir una variable ficticia D_{1i} del mismo modo que definimos las restantes. En este caso, para los varones del primer nivel educativo tendríamos, $Salario_i = \hat{\delta}_1 + \hat{\beta}_1 Experiencia_i + \hat{u}_i$, para los del segundo nivel educativo: $Salario_i = \hat{\delta}_2 + \hat{\beta}_1 Experiencia_i + \hat{u}_i$, y así sucesivamente; para los trabajadores varones del más alto nivel educativo, tendríamos, $Salario_i = \hat{\delta}_5 + \hat{\beta}_1 Experiencia_i + \hat{u}_i$. Por supuesto, los valores numéricos de los coeficientes δ serían diferentes ahora que en el modelo anterior. Esperaríamos que las estimaciones numéricas de los coeficientes δ fuesen crecientes para los distintos niveles educativos.

La suma de las cinco variables ficticias incluidas en el modelo (47) es igual a uno para todas las observaciones, pues sólo una de ellas es igual a uno en cada observación, siendo las restantes iguales a cero, y esto ocurre para todas las observaciones disponibles. Por tanto, su suma es igual al valor de la variable que acompaña al término constante, por lo que éste no puede incluirse en la regresión, pues tendríamos colinealidad perfecta, no pudiendo estimarse dicho modelo. En el caso de (44) las cuatro variables ficticias suman uno para todas las observaciones, excepto las del primer nivel educativo, para el que suman cero; por tanto, su suma no coincide con el valor numérico de la variable que acompaña al término constante, y el modelo puede estimarse. En dicho modelo, podíamos haber optado por incluir D_{1i} y excluir otra cualquiera de las variables ficticias, y la interpretación de los coeficientes estimados sería análoga a la que propusimos para el modelo (44).