Cátedra de Econometría I

Facultad de Ciencias Económicas

Universidad Nacional de La Plata

Notas de clase

Esta versión: 2020 *

Profesor Titular
Walter Sosa Escudero

Profesora Adjunta
Mariana Marchionni

Jefa de Trabajos Prácticos
Jessica Bracco

Ayudantes y adscriptos
Ivana Benzaquén
Leonardo Peñaloza
Milagros Cejas
Malena Dolcet
Azul Menduiña
Gastón García Zavaleta

*Versión revisada y traducida, basada en las notas de clase "Applied Econometrics. Class Material" del profesor Walter Sosa Escudero. Todos los integrantes de la cátedra han colaborado en la revisión de estas notas. Los comentarios y sugerencias son bienvenidos.

Econometría

Literalmente, *econometría* significa "medición económica" (derivada de econo, economía y metría, medición). Sin embargo, aunque es cierto que la medición es una parte importante de la econometría, el alcance de esta disciplina es mucho más amplio, como puede deducirse de las siguientes citas:

- La econometría, resultado de cierta perspectiva sobre el papel que juega la economía, consiste en la aplicación de la estadística matemática a la información económica para dar soporte empírico a los modelos construidos por la economía matemática y obtener resultados numéricos.[1]
- ...la econometría puede ser definida como el análisis cuantitativo de fenómenos económicos reales, basado en el desarrollo simultaneo de teoría y observaciones, y relacionado por métodos apropiados de inferencia.[2]
- La econometría puede ser definida como la ciencia social en la cual las herramientas de la teoría económica, las matemáticas y la inferencia estadística son aplicadas al análisis de los fenómenos económicos.[3]
- La econometría tiene que ver con la determinación empírica de las leyes económicas.[4]
- El arte del econometrista consiste en encontrar el conjunto de supuestos que sean suficientemente específicos y realistas, de tal forma que le permitan aprovechar de la mejor manera los datos que tiene a su disposición.[5]
- Los econometristas ...son una ayuda en el esfuerzo por disipar la mala imagen pública de la economía (cuantitativa o de otro tipo) considerada como una materia

[1] Tintner, G., (1968). *Methodology of Mathematical Economics and Econometrics*, The University of Chicago Press, Chicago, p. 74.

[2] Samuelson, P. A., T. C. Koopmans and J. R. N. Stone, (1954). Report of the evaluative committee for Econometrica, *Econometrica* 22, p. 141-6.

[3] Goldberger, A. S., (1964). *Econometric Theory*, John Wiley & Sons, Nueva York, p. 1.

[4] Theil, H., (1971). *Principles of Econometrics*. John Wiley & Sons, Nueva York, p. 1.

[5] Malinvaud, E., (1966). *Statistical Methods of Econometrics*, Rand McNally, Chicago, p. 514.

en la cual se abren cajas vacías, suponiendo la existencia de abrelatas, para revelar un contenido que será interpretado por cada diez economistas de 11 maneras diferentes.[6]

- El método de la investigación econométrica busca esencialmente una conjunción entre la teoría económica y la medición real, utilizando como puente la teoría y la técnica de la inferencia estadística.[7]

La econometría propone un desarrollo conjunto de las ideas y los datos económicos.

Objetivos:

- Descubrir relaciones relevantes y sugerir teorías.
- Cuantificar fenómenos económicos.
- Aislar fenómenos causales, suplir la falta de experimentos.
- Evaluar teorías e ideas económicas.
- Predecir.

La econometría debe lidiar con la naturaleza específica de los fenómenos económicos:

- Relaciones no exactas.
- Fenómenos complejos, alta interacción entre los fenómenos.
- Datos no experimentales.
- Fenómenos no observables (inteligencia, suerte, preferencias, etc.).

Ejemplo: Retornos a la educación

- $Ingreso = f$ (educación,experiencia, inteligencia,etc.)
- Factores idiosincráticos inciden en esta relación.
- Inteligencia inobservable.
- Experimento: asignar aleatoriamente educación a individios y ver sus salarios.
- Datos disponibles: encuestas de hogares (datos observacionales)

[6] Darnell A. C., and J. L. Evans, (1990). *The Limits of Econometrics*, Edward Elgar Publishing, Hants, p. 54.
[7] Haavelmo, T., (1944). *The Probability Approach in Econometrics*, *Econometrica*, vol. 12. 1944. preface p. *iii*.

La econometría incorpora todas estas características de los fenómenos económicos (y ello la distingue de la estadística).

Los métodos econométricos y las teorías económicas se desarrollan en forma conjunta, interactuando entre ellas. La econometría es una parte fundamental de la economía, no una disciplina separada.

Este curso discute las características teóricas de los métodos econométricos disponibles, lo cual es de fundamental importancia para elegir optimamente las técnicas a utilizar en el trabajo propio, y para evaluar críticamente el trabajo de otros. Además presenta *herramientas computacionales* recientes para la aplicación de la teoría y de los métodos discutidos en clase.

El curso motiva el *uso de métodos empíricos* en economía cubriendo sus principales aspectos: desarrollo y discusión de ideas básicas, recolección de datos, elección de técnicas econométricas adecuadas, evaluación crítica del trabajo de otros autores, presentación oral y escrita de los resultados obtenidos. Presenta *aplicaciones recientes* en distintas áreas tales como: macroeconomía, economía monetaria y bancaria, economía de los recursos humanos, historia económica, publicidad, finanzas, organización industrial, economía laboral, marketing, economía ambiental, entre otras.

Capítulo 1

Modelo Lineal con Dos Variables

1.1. Relaciones Lineales

En el análisis empírico de los fenómenos económicos usualmente estamos interesados en conocer y describir cómo se relacionan las variables económicas involucradas. Por simplicidad, empezaremos explorando la posibilidad de que dos variables estén relacionadas *linealmente*, con el objetivo de determinar si esta relación efectivamente existe, y de medir la dirección (positiva o negativa) y la fuerza de la misma.

Sean X e Y dos variables aleatorias que representan algún fenómeno económico, como por ejemplo consumo e ingreso de las familias, y sea (X_i, Y_i) con $i = 1, 2, \ldots, n$ una muestra aleatoria de tamaño n de estas variables. Típicamente, nuestro análisis se basará en los datos que surgen de una realización particular de la muestra aleatoria. Resulta útil comenzar examinando una representación gráfica de los datos en el plano (X, Y), conocida como diagrama de dispersión o nube de puntos.

La Figura 1.1 presenta tres diagramas de dispersión alternativos. Cada punto en el plano (X, Y) representa la realización de una observación muestral. Una nube de puntos como la del primer panel sugiere la existencia de una relación negativa entre las variables X e Y. Por el contrario, los datos del segundo panel indican la existencia de una asociación positiva entre las variables y también una mayor intensidad de la relación, que se refleja en una mayor concentración de los puntos. El último diagrama de dispersión sugiere que X e Y no están relacionadas.

A partir de la información muestral, es posible computar ciertos estadísticos descriptivos que nos informan sobre la relación entre X e Y. A continuación definimos los conceptos de covarianza y correlación muestrales, que miden la dirección y, en el caso de la correlación, el grado de asociación lineal entre dos variables.

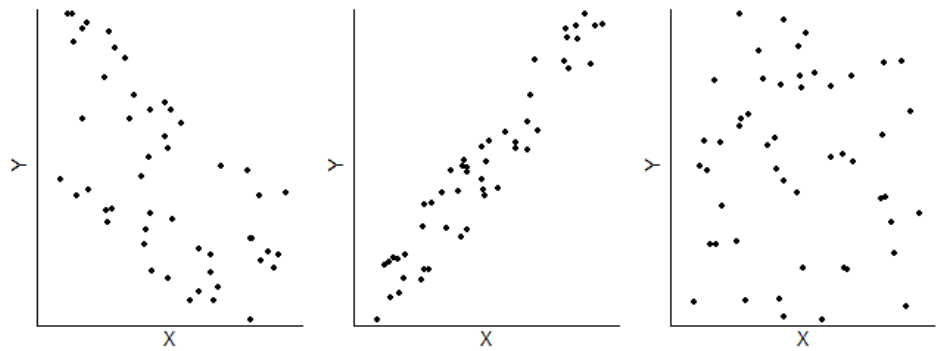


Figura 1.1: Diagramas de dispersión.

a) Covarianza muestral entre X e Y

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \tag{1.1}$$

donde $\bar{X}$ e $\bar{Y}$ son las medias muestrales de X e Y, respectivamente.

Notación: Definamos $z_i = Z_i - \bar{Z}$, es decir, las letras minúsculas denotan a las observaciones como desviaciones respecto de sus propias medias muestrales.

Utilizando esta notación, la covarianza muestral entre X e Y se puede expresar como:

$$Cov(X,Y) = \frac{\sum_{i=1}^{n} x_i y_i}{n-1} \tag{1.2}$$

b) Correlación muestral entre X e Y

$$\begin{aligned} r_{XY} &= \frac{Cov(X,Y)}{S_X S_Y} && (1.3) \\ &= \frac{\sum_{i=1}^{n} x_i y_i/(n-1)}{\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{(n-1)}}\sqrt{\frac{\sum_{i=1}^{n} y_i^2}{(n-1)}}} \\ &= \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} && (1.4) \end{aligned}$$

Notación: La varianza muestral de una variable aleatoria Z (S_Z^2) se puede escribir como:

$$S_Z^2 = \frac{\sum_{i=1}^{n}(Z_i - \bar{Z})^2}{n-1} = \frac{\sum_{i=1}^{n} {z_i}^2}{n-1}$$

Algunas propiedades importantes de la covarianza y la correlación son las siguientes:

1. Ambas medidas son *simétricas:* $Cov(X,Y) = Cov(Y,X)$ y $r_{XY} = r_{YX}$, lo cual se puede comprobar fácilmente a partir de las definiciones.

2. A diferencia de la covarianza, la correlación no depende de las unidades de medida de las variables. Un cambio en las unidades de medida de una variable se obtiene al multiplicar cada observación por una constante. Por ejemplo, el gasto en centavos se obtiene multiplicando por 100 al gasto medido en pesos. Un cambio como este modifica la covarianza pero no la correlación.

   A continuación mostraremos que para cualquier constante positiva a se cumple:

$$Cov(X,Y) \neq Cov(aX,aY)$$

$$r_{X,Y} = r_{aX,aY}$$

   Demostración: Primero notemos que:

$$\overline{aX} = \sum_{i=1}^{n} aX_i/n = a\sum_{i=1}^{n} X_i/n = a\bar{X}$$

   Usando la definición de covarianza muestral (1.1):

$$\begin{aligned} Cov(aX,aY) &= \frac{\sum_{i=1}^{n}(aX_i - \overline{aX})(aY_i - \overline{aY})}{n-1} \\ &= \frac{\sum_{i=1}^{n}(aX_i - a\bar{X})(aY_i - a\bar{Y})}{n-1} \\ &= a^2\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \\ &= a^2 Cov(X,Y) \neq Cov(X,Y) \end{aligned}$$

■

   Ejercicio: Mediante un mecanismo similar, probar el resultado correspondiente al coeficiente de correlación.

3. El coeficiente de correlación es menor o igual a uno en valor absoluto, esto es: $-1 \leq r \leq 1$.

Demostración: Notar que para cualquier constante c se cumple que:

$$\begin{aligned}\sum_{i=1}^{n}(y_i - cx_i)^2 &\geq 0\\ \sum_{i=1}^{n}({y_i}^2 + c^2{x_i}^2 - 2cx_iy_i) &\geq 0\end{aligned}$$

En particular, consideremos $c = \sum_{i=1}^{n} x_iy_i / \sum_{i=1}^{n} x_i^2$. Reemplazando:

$$\begin{aligned}\sum_{i=1}^{n}{y_i}^2 + \left(\frac{\sum_{i=1}^{n} x_iy_i}{\sum_{i=1}^{n} x_i^2}\right)^2 \sum_{i=1}^{n}{x_i}^2 - 2\left(\frac{\sum_{i=1}^{n} x_iy_i}{\sum_{i=1}^{n}{x_i}^2}\right)\sum_{i=1}^{n} x_iy_i &\geq 0\\ \sum_{i=1}^{n}{y_i}^2 + \frac{(\sum_{i=1}^{n} x_iy_i)^2}{\sum_{i=1}^{n}{x_i}^2} - 2\frac{(\sum_{i=1}^{n} x_iy_i)^2}{\sum_{i=1}^{n}{x_i}^2} &\geq 0\\ \sum_{i=1}^{n}{y_i}^2 - \frac{(\sum_{i=1}^{n} x_iy_i)^2}{\sum_{i=1}^{n}{x_i}^2} &\geq 0\end{aligned}$$

$$\begin{aligned}(\sum_{i=1}^{n}{y_i}^2)(\sum_{i=1}^{n}{x_i}^2) &\geq (\sum_{i=1}^{n} x_iy_i)^2\\ \left[\frac{\sum_{i=1}^{n} x_iy_i}{\sqrt{\sum_{i=1}^{n}{y_i}^2}\sqrt{\sum_{i=1}^{n}{x_i}^2}}\right]^2 &\leq 1\\ r^2 &\leq 1\end{aligned}$$

■

4. El coeficiente de correlación es exactamente igual a uno cuando Y es una función lineal exacta de X con pendiente positiva. De manera similar, $r_{XY} = -1$ cuando Y es una función lineal exacta de X con pendiente negativa.

   Demostración: Consideremos primero el caso en que Y es una función lineal exacta de X con pendiente positiva, esto es: $Y_i = a + kX_i$, donde a es cualquier constante y $k > 0$. Partiendo de la definición de media muestral:

$$\begin{aligned}\bar{Y} &= \frac{\sum_{i=1}^{n} Y_i}{n}\\ &= \frac{\sum_{i=1}^{n}(a + kX_i)}{n}\\ &= a + k\frac{\sum_{i=1}^{n} X_i}{n}\\ &= a + k\bar{X}\end{aligned}$$

Consecuentemente:

$$\begin{aligned} y_i = Y_i - \bar{Y} &= a + kX_i - (a + k\bar{X}) \\ &= k(X_i - \bar{X}) \\ &= kx_i \end{aligned}$$

Con estos resultados podemos reexpresar el coeficiente de correlación de (1.4) de la siguiente manera:

$$\begin{aligned} r = \frac{\sum_{i=1}^n y_i x_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}} &= \frac{\sum_{i=1}^n (kx_i) x_i}{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n (kx_i)^2}} \\ &= \frac{\sum_{i=1}^n kx_i^2}{\sqrt{k^2 (\sum_{i=1}^n x_i^2)^2}} \\ &= 1 \end{aligned}$$

■

Ejercicio: Verificar el resultado correspondiente a $k < 0$.

5. Interpretación geométrica de covarianza y correlación. Partamos de la definición de correlación muestral (1.4):

$$r_{XY} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$$

Primero notemos que el signo del coeficiente de correlación depende únicamente del numerador. Además, como ese numerador es n veces la covarianza, el signo de la correlación y de la covarianza coinciden. Si los términos positivos que se suman en el numerador más que compensan a los negativos, la correlación y la covarianza serán positivas. Esto sucede cuando la mayor parte de las observaciones satisface $Y_i > \overline{Y}$ y $X_i > \overline{X}$ ó $Y_i < \overline{Y}$ y $X_i < \overline{X}$. Es decir, cuando la mayor parte de los puntos en un gráfico de dispersión están por encima o por debajo de ambas medias muestrales simultáneamente. En estos casos en que la correlación es positiva, los gráficos de dispersión tienen el aspecto del que se presenta en el segundo panel de la Figura 1.1, indicando cierta relación lineal positiva entre las variables. Por el contrario, cuando la mayor parte de las observaciones satisface $Y_i > \overline{Y}$ y $X_i < \overline{X}$ ó $Y_i < \overline{Y}$ y $X_i > \overline{X}$, la correlación y la covarianza serán negativas, indicando algún grado de relación lineal negativa como el representado por el primer panel de la Figura 1.1.

6. La covarianza y la correlación únicamente miden relaciones lineales. ¿Cómo es la correlación de una nube de puntos como la de la Figura 1.2? Siguiendo el mismo razonamiento empleado

en el punto anterior, concluimos que la correlación y la covarianza serían cercanas a cero. Una correlación (o covarianza) cercana a cero debe interpretarse como evidencia de ausencia de una relación lineal entre las variables, pero de ninguna manera indica ausencia de relación. Es decir, la ausencia de una relación lineal no excluye la posibilidad de otro tipo de relaciones no lineales. Un ejemplo es el caso de la curva de Laffer, que representa la relación entre la recaudación impositiva y las alícuotas impositivas. La curva de Laffer tiene forma de U invertida: aumentos de las alícuotas a partir de niveles bajos primero aumentan la recaudación hasta llegar a un máximo, a partir del cual posteriores aumentos de la alícuota reducen tanto la demanda que la recaudación de impuestos empieza a caer. Este es un ejemplo de una relación no lineal que, empíricamente, produce un coeficiente de correlación muy cercano a cero.

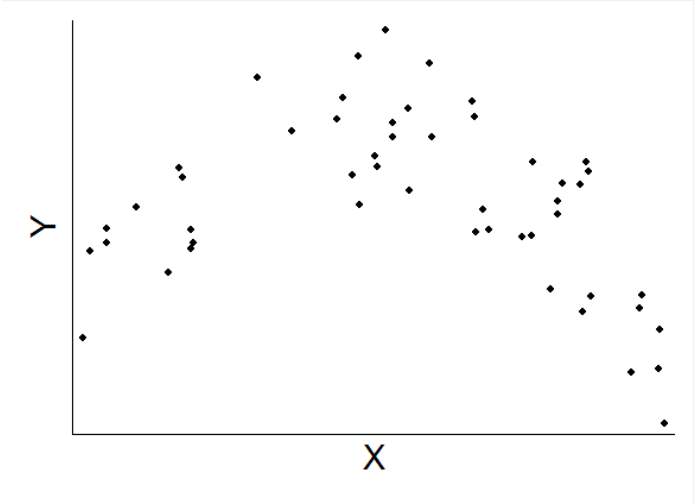


Figura 1.2: Relación no lineal.

7. Correlación no implica causalidad. Un error muy común es pensar que la presencia de correlación implica algún tipo de causalidad entre las variables involucradas. Por ejemplo, consideremos la relación entre inversión en ciencia y tecnología, y crecimiento. Un resultado empírico frecuente es que estas dos variables están positivamente correlacionadas, de modo que el coeficiente de correlación es positivo. ¿Implica esto que para favorecer el crecimiento deberíamos invertir en tecnología? ¿O debemos interpretar que son los países con más crecimiento los que destinan más fondos a la inversión en tecnología?. Lamentablemente, el coeficiente de correlación no dice nada acerca de la dirección de causalidad, sólo confirma que ambas variables se mueven conjuntamente.

En resumen, la covarianza y la correlación son medidas de relación lineal. El signo de ambas

indica la dirección de la asociación: valores positivos evidencian relaciones positivas y viceversa. El coeficiente de correlación tiene una ventaja adicional: indica también el grado o fuerza de la relación lineal. Sus valores están acotados en el intervalo $[-1, 1]$: cuanto mayor es el valor absoluto de la correlación, mayor es el grado de asociación lineal entre las variables; en el extremo, si la relación es perfectamente lineal el coeficiente de correlación es igual a 1 en valor absoluto; finalmente, si la correlación es cercana a cero podemos decir que no hay evidencia de relación lineal, aunque podría existir otro tipo de relación entre las variables.

1.2. El Modelo Lineal

A la luz de la discusión anterior, el objetivo de este capítulo consiste en construir un modelo estimable para la relación lineal *no-exacta* entre dos variables Y y X, vinculadas por alguna teoría económica, como la relación entre el consumo y el ingreso, las cantidades demandadas y el precio, etc.

El modelo propuesto es el siguiente:

$$Y_i = \alpha + \beta X_i + u_i, \qquad i = 1, \ldots, n \tag{1.5}$$

en donde α y β son parámetros desconocidos, objeto de la estimación. u_i es una variable aleatoria no observable, que representa el hecho de que la relación entre Y y X no es exactamente lineal. Es importante observar que si $u_i = 0$ para $i = 1, \ldots, n$, entonces la relación entre Y y X sería lineal y exacta. En este sentido, la presencia de u_i es justamente lo que rompe con esa exactitud. Y es comunmente llamada variable *explicada* o *dependiente*, y X variable *explicativa* o *independiente*. Los "datos" son las n realizaciones (X_i, Y_i), para $i = 1, \ldots, n$.

A la variable aleatoria u_i se la suele llamar "término de error", y representa a todos aquellos factores que afectan a Y y que no son capturados por la variable explicativa X, incluyendo tanto a la "verdadera aleatoriedad" como a factores inobservables. En realidad, la noción de término de error es más adecuada en el contexto de disciplinas experimentales. Por ejemplo, si se deseara conocer el efecto que tiene la aplicación de una dosis de cierta droga (X) sobre la temperatura corporal (Y), u_i podría representar un error de medición asociado al comportamiento impreciso de un instrumento de medición (un termómetro, por ejemplo). En una disciplina social como la economía, más que a errores de medicion, u_i representa a cualquier causa no observada ni medida (por ignorancia u omisión) que afecta a Y más allá de X.

La Figura 1.3 presenta un diagrama de dispersión de las variables X e Y, donde cada punto es una realización observable de las variables explicada y explicativa. Como podemos observar, en este caso, mayores valores de X se corresponden con mayores valores de Y, evidenciando una correlación positiva entre ambas variables. El primer objetivo consiste en encontrar estimaciones razonables para α y β en base a los datos disponibles (X_i, Y_i), $i = 1, \ldots, n$.

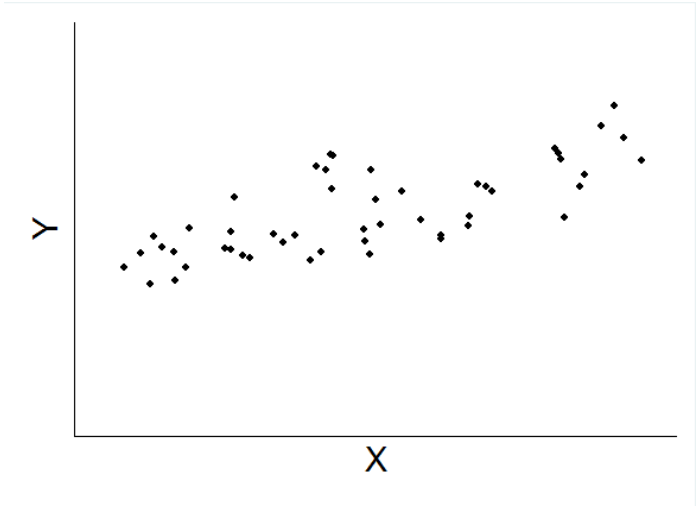


Figura 1.3: Diagrama de dispersión.

1.3. El Método de Mínimos Cuadrados Ordinarios

Denotemos con $\hat{\alpha}$ y $\hat{\beta}$ a los estimadores para α y β en el modelo lineal simple (1.5). Definamos las siguientes magnitudes. La primera es la estimación de Y:

$$\hat{Y}_i \equiv \hat{\alpha} + \hat{\beta} X_i \tag{1.6}$$

Intuitivamente, hemos reemplazado α y β por sus estimaciones, y hemos tratado a u_i como si la relación lineal fuese exacta, esto es, como si u_i fuese cero.

Resulta natural definir al error de estimación como:

$$e_i \equiv Y_i - \hat{Y}_i \tag{1.7}$$

el cual mide la diferencia entre Y_i y su estimación $\hat{Y}_i$.

El objetivo inicial consiste en encontrar $\hat{\alpha}$ y $\hat{\beta}$ en base a la muestra disponible, de modo que los e_i sean lo más pequeños posible, en algún sentido. Es interesante observar cómo funciona este problema desde una perspectiva gráfica. Los datos pueden representarse como n puntos en el plano (X, Y). La relación lineal (1.1) es consistente con una nube de puntos dispersos alrededor de una línea recta imaginaria. De hecho, si todos los u_i fuesen cero, todos los puntos estarían perfectamente alineados sobre una misma recta, en forma consistente con una relación lineal exacta. Como mencionasemos antes, es la presencia de u_i lo que rompe esta exactitud y genera puntos dispersos alrededor de una línea imaginaria.

Es importante notar que para dos valores cualesquiera de $\hat{\alpha}$ y $\hat{\beta}$, los puntos correspondientes al modelo estimado (1.6) se corresponden con una única recta en el plano (X, Y). Consecuentemente,

distintos valores de $\hat{\alpha}$ y $\hat{\beta}$ se corresponden con rectas diferentes, lo que implica que elegir valores particulares para $\hat{\alpha}$ y $\hat{\beta}$ es equivalente a elegir una recta en el plano (X, Y).

Para la i-ésima observación, el error de estimación e_i puede ser representado gráficamente como la distancia *vertical* entre los puntos (X_i, Y_i) y su version estimada, $(X_i, \hat{Y}_i)$, que por construccion cae sobre la recta estimada. Entonces, intuitivamente, queremos valores de $\hat{\alpha}$ y $\hat{\beta}$ de modo que la recta elegida pase lo más cerca posible de los puntos, y entonces los errores sean lo más pequeños posible.

Para hacer más ilustrativa la explicación, en la Figura 1.4 se presentan dos rectas superpuestas a la misma nube de puntos del gráfico anterior. A priori, cualquiera de las dos rectas podria ser la que minimice los errores de estimación. Obviamente, cada una de estas rectas está determinada por una elección diferente de α y β: la recta sólida se corresponde con $\alpha^{'}$ y $\beta^{'}$, mientras que la punteada surge de elegir $\alpha^{''}$ y $\beta^{''}$. Partiendo de cualquier punto (como por ejemplo el punto A), se pueden comparar los errores de estimación que surgen de elegir una recta o la otra. En este caso, la distancia entre el punto A y la recta determinada por $\alpha^{'}$ y $\beta^{'}$ (representada por e') es mayor que la distancia entre ese mismo punto y la recta que surge de elegir $\alpha^{''}$ y $\beta^{''}$ (representada por e'').

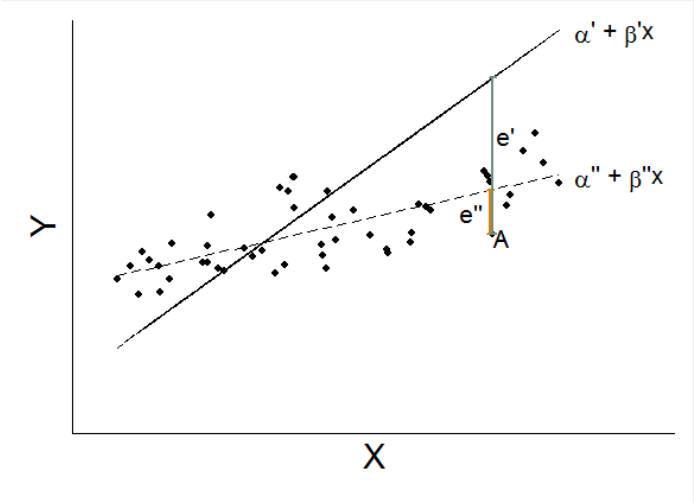


Figura 1.4: Diagrama de dispersión con recta "candidata"

Comencemos con el caso en donde hay sólo dos observaciones distintas. En este caso, nuestro problema tiene una solución trivial, que se reduce a encontrar los valores de $\hat{\alpha}$ y $\hat{\beta}$ que se corresponden con la única recta que pasa por estos dos puntos y que hace que los errores de estimación sean cero. El caso más realista aparece cuando disponemos de más de dos observaciones, no exactamente alineadas sobre una misma recta, tal como sucede en la Figura 1.4. Obviamente, una línea recta no puede pasar por más de dos observaciones no alineadas, lo que sugiere que en estos casos es imposible que los errores sean *todos* iguales a cero. Entonces, parece natural plantear el problema

de encontrar los valores de $\hat{\alpha}$ y $\hat{\beta}$ que determinen una recta que pase lo más cerca posible de todos los puntos, de modo que los errores, en el agregado, sean pequeños. Para ello, es necesario definir qué queremos decir por "cerca". Con este objetivo, definamos una función de penalidad, que consiste en sumar todos los errores de estimación al cuadrado, de modo que los errores positivos y negativos importen por igual. Para cualquier valor de $\hat{\alpha}$ y $\hat{\beta}$, esta función nos dará una idea de cuán grandes son los errores agregados de estimación:

$$SRC(\hat{\alpha}, \hat{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2 \tag{1.8}$$

SRC significa *suma de residuos al cuadrado*. Notar que, dados los datos (X_i,Y_i), SRC es una función que depende de nuestra elección de $\hat{\alpha}$ y $\hat{\beta}$. Esto es, distintos valores de $\hat{\alpha}$ y $\hat{\beta}$ se corresponden con distintas rectas que pasan por la nube de puntos, implicando distintos errores de estimación. Valores altos para SRC se corresponden con rectas que generan errores agregados grandes. En un extremo, SRC es cero si y sólo si todos lo puntos caen en una misma recta. Parece natural, entonces, elegir $\hat{\alpha}$ y $\hat{\beta}$ de modo que SRC sea lo más pequeño posible.

Notación: para facilitar la lectura, desde ahora escribiremos $\sum$ en lugar de $\sum_{i=1}^{n}$, salvo que se indique lo contrario.

Los valores de $\hat{\beta}$ y $\hat{\alpha}$ que minimizan la suma de residuos al cuadrado son:

$$\hat{\beta} = \frac{\sum X_i Y_i - n\bar{Y}\bar{X}}{\sum X_i^2 - n\bar{X}^2}$$

y

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

los cuales son conocidos como los estimadores *mínimo cuadráticos*, o los estimadores de *mínimos cuadrados ordinarios (MCO)* de β y α.

Derivación analítica de los estimadores mínimo cuadráticos

Se puede demostrar que $SRC(\hat{\alpha}, \hat{\beta})$ es globalmente cóncava y diferenciable. Las condiciones de primer orden para un mínimo local son:

$$\begin{aligned} \frac{\partial SRC(\hat{\alpha}, \hat{\beta})}{\partial \hat{\alpha}} &= 0 \\ \frac{\partial SRC(\hat{\alpha}, \hat{\beta})}{\partial \hat{\beta}} &= 0 \end{aligned}$$

La condición de primer orden con respecto a $\hat{\alpha}$ es:

$$\frac{\partial \sum e^2}{\partial \hat{\alpha}} = -2 \sum (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0 \tag{1.9}$$

Dividiendo por menos 2 y distribuyendo las sumatorias:

$$\sum Y_i = n\hat{\alpha} + \hat{\beta} \sum X_i \tag{1.10}$$

Es importante recordar esta última expresión, porque volveremos sobre ella muy seguido.

La condición de primer orden con respecto a $\hat{\beta}$ es:

$$\frac{\partial \sum e^2}{\partial \hat{\beta}} = -2 \sum X_i (Y_i - \hat{\alpha} - \hat{\beta} X_i) = 0 \tag{1.11}$$

Dividiendo por -2 y distribuyendo las sumatorias:

$$\sum X_i Y_i = \hat{\alpha} \sum X_i + \hat{\beta} \sum X_i^2 \tag{1.12}$$

(1.10) y (1.12) conforman un sistema de dos ecuaciones lineales con dos incógnitas ($\hat{\alpha}$ y $\hat{\beta}$) conocidas como las *ecuaciones normales*.

Dividiendo (1.10) por n y resolviendo para $\hat{\alpha}$ se obtiene:

$$\hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X} \tag{1.13}$$

el estimador para α. Reemplazando en (1.12) obtenemos:

$$\begin{aligned}
\sum X_i Y_i &= (\bar{Y} - \hat{\beta}\bar{X}) \sum X_i + \hat{\beta} \sum {X_i}^2 \\
\sum X_i Y_i &= \bar{Y} \sum X_i - \hat{\beta}\bar{X} \sum X_i + \hat{\beta} \sum {X_i}^2 \\
\sum X_i Y_i - \bar{Y} \sum X_i &= \hat{\beta} \left(\sum {X_i}^2 - \bar{X} \sum X_i \right)
\end{aligned}$$

$$\hat{\beta} = \frac{\sum X_i Y_i - \bar{Y} \sum X_i}{\sum X_i^2 - \bar{X} \sum X_i}$$

Notar que: $\bar{X} = \sum X_i / n$, luego $\sum X_i = \bar{X} n$. Reemplazando, obtenemos:

$$\hat{\beta} = \frac{\sum X_i Y_i - n\bar{Y}\bar{X}}{\sum X_i^2 - n\bar{X}^2} \tag{1.14}$$

que es el resultado deseado para el estimador de β.

Usando la notación de desviaciones con respecto a las medias muestrales:

$$\begin{aligned}
\sum x_i y_i &= \sum (X_i - \bar{X})(Y_i - \bar{Y}) \\
&= \sum (X_i Y_i - X_i \bar{Y} - \bar{X} Y_i + \bar{X}\bar{Y}) \\
&= \sum X_i Y_i - \bar{Y} \sum X_i - \bar{X} \sum Y_i + n\bar{X}\bar{Y} \\
&= \sum X_i Y_i - n\bar{Y}\bar{X} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} \\
&= \sum X_i Y_i - n\bar{Y}\bar{X}
\end{aligned}$$

que corresponde al numerador de la expresion para $\hat{\beta}$ obtenida mas arriba. Realizando una operación similar en el denominador de dicha expresion obtenemos la siguiente formulación alternativa para el estimador mínimo cuadrático de β:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} \tag{1.15}$$

La Figura 1.5 ilustra la recta (1.6), es decir:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$$

denominada *recta de regresión estimada* y que, como consecuencia de usar el metodo de mínimos cuadrados ordinarios, tiene la propiedad de ser la que pasa más cerca de los puntos, en el sentido de que minimiza la suma de errores al cuadrado. La siguiente sección estudia varias propiedades de esta recta.

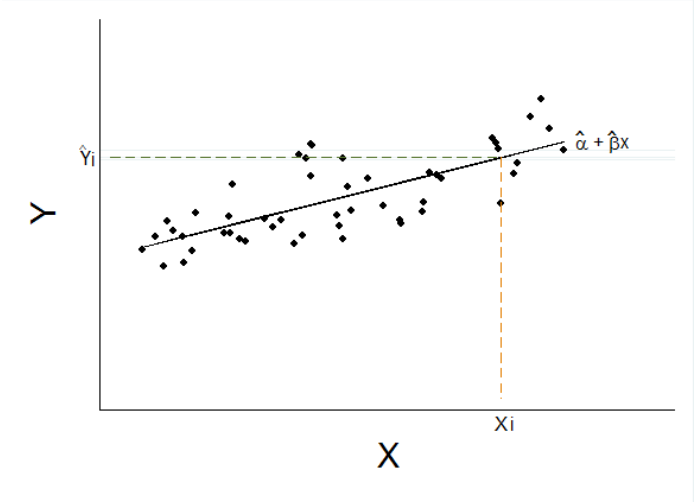


Figura 1.5: Diagrama de dispersión y recta MCO.

1.4. Propiedades Algebraicas de los Estimadores Mínimo Cuadráticos

Entendemos por propiedades *algebraicas* de los estimadores MCO a aquellas que surgen como consecuencia directa del proceso de minimización, destacando la diferencia con las propiedades *estadísticas*, que serán estudiadas en la sección siguiente.

- *Propiedad 1:* $\sum e_i = 0$. Este resultado surge directamente de la derivación analítica de los estimadores mínimo cuadráticos. Dividiendo la ecuación (1.9), por menos 2 y reemplazando por la definición de e_i de la ecuación (1.7), fácilmente se verifica que como consecuencia de minimizar la suma de cuadrados residuales, la suma de los residuos, y consecuentemente su promedio muestral, son iguales a cero.

- *Propiedad 2:* $\sum X_i e_i = 0$. Esta propiedad puede verificarse dividiendo por menos 2 la ecuación (1.11). La covarianza muestral entre X y e viene dada por:

$$\begin{aligned} Cov(X,e) &= \frac{1}{n-1}\sum(X_i-\bar{X})(e_i-\bar{e}) \\ &= \frac{1}{n-1}\left[\sum X_i e_i - \bar{e}\sum X_i - \bar{X}\sum e_i + \sum \bar{X}\bar{e}\right] \\ &= \frac{1}{n-1}\sum X_i e_i \end{aligned}$$

  dado que, por la propiedad anterior, $\sum e_i$ y por lo tanto $\bar{e}$ son iguales a cero. Entonces, esta propiedad dice que, como consecuencia de usar el método de mínimos cuadrados, la covarianza muestral entre la variable explicativa X y el término de error e es cero, o, lo que es lo mismo, los residuos no están linealmente relacionados con la variable explicativa.

- *Propiedad 3: La línea de regresión pasa por el punto de las medias muestrales.* La línea de regresión estimada corresponde a la función $\hat{Y}(X_i) = \hat{\alpha} + \hat{\beta}X_i$ donde se toma a $\hat{\alpha}$ y $\hat{\beta}$ como parámetros, de forma que $\hat{Y}$ es una función que depende de X. Veamos qué ocurre cuando evaluamos esta función en $\bar{X}$, la media de X:

$$\hat{Y}(\bar{X}) = \hat{\alpha} + \hat{\beta}\bar{X}$$

  Pero de (1.13):

$$\hat{\alpha} + \hat{\beta}\bar{X} = \bar{Y}$$

  Luego $\hat{Y}(\bar{X}) = \bar{Y}$, esto es, la línea de regresión estimada por el método de mínimos cuadrados pasa por el punto de las medias muestrales.

- *Propiedad 4: Relación entre regresión y correlación.* Se puede demostrar que:

$$\hat{\beta} = r_{xy}\frac{S_Y}{S_X}$$

  Demostración: Recordemos primero que en (1.3) definimos al coeficiente de correlación muestral entre X e Y para una muestra de n observaciones (X_i, Y_i), $i = 1, 2, \ldots, n$ como:

$$r_{xy} = \frac{Cov(X,Y)}{S_X S_Y}$$

De (1.15):

$$\begin{aligned}
\hat{\beta} &= \frac{\sum x_i y_i}{\sum x_i^2} \\
&= \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum x_i^2}} \\
&= \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum x_i^2}} \frac{\sqrt{\sum y_i^2}}{\sqrt{\sum y_i^2}} \\
&= \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}} \frac{\sqrt{\sum y_i^2}/\sqrt{n-1}}{\sqrt{\sum x_i^2}/\sqrt{n-1}} \\
\hat{\beta} &= r_{xy}\frac{S_Y}{S_X}
\end{aligned}$$

■

Notar que si $r_{xy} = 0$, entonces $\hat{\beta} = 0$. Notar también que si ambas variables tienen la misma varianza muestral, el coeficiente de correlación es igual al coeficiente $\hat{\beta}$ de la regresión. Además podemos ver que, a diferencia del coeficiente de correlación, $\hat{\beta}$ *no es* invariante a cambios en la escala o la unidad de medida.

- *Propiedad 5: Y_i e $\hat{Y}_i$ tienen la misma media muestral.*

  Demostración: Por (1.7), $Y_i = \hat{Y}_i + e_i$, $i = 1, \ldots, n$. Luego, sumando para cada i:

$$\sum Y_i = \sum \hat{Y}_i + \sum e_i$$

  y dividiendo por n:

$$\frac{\sum Y_i}{n} = \frac{\sum \hat{Y}_i}{n}$$

  dado que $\sum e_i = 0$ por la condición de primer orden (1.9). Entonces:

$$\bar{Y} = \bar{\hat{Y}}$$

■

- *Propiedad 6: $\hat{\beta}$ es una función lineal de los Y_i.* Es decir, $\hat{\beta}$ puede escribirse como $\hat{\beta} = \sum w_i Y_i$, donde los w_i son números reales no aleatorios, que dependen exclusivamente de X_i, y no todos son iguales a cero.

Demostración: Comenzaremos escribiendo a $\hat{\beta}$ como sigue:

$$\hat{\beta} = \sum \left(\frac{x_i}{\sum x_i^2} \right) y_i$$

y definamos $w_i = x_i / \sum x_i^2$. Notar que:

$$\sum x_i = \sum (X_i - \bar{X}) = \sum X_i - n\bar{X} = 0$$

lo que implica $\sum w_i = 0$. Del resultado anterior:

$$\begin{aligned} \hat{\beta} &= \sum w_i y_i \\ &= \sum w_i (Y_i - \bar{Y}) \\ &= \sum w_i Y_i - \bar{Y} \sum w_i \\ &= \sum w_i Y_i \end{aligned}$$

■

Aunque esta propiedad no tenga un significado intuitivo claro, será útil para obtener resultados posteriores. También puede demostrarse que el estimador mínimo cuadrático de α es un estimador lineal de Y.

1.5. El Modelo Lineal con Dos Variables bajo los Supuestos Clásicos

Nuestro modelo lineal (1.5) tiene la siguiente forma:

$$Y_i = \alpha + \beta X_i + u_i, \qquad i = 1, \ldots, n$$

En forma adicional, introduciremos los siguientes supuestos:

1. $E(u_i) = 0, \;\; i = 1, 2, \ldots, n$. Este supuesto implica que "en promedio" la relación entre Y y X es exactamente lineal, aunque las realizaciones particulares de los u_i pueden ser distintas de cero.

2. $Var(u_i) = E[(u_i - E(u_i))^2] = E(u_i^2) = \sigma^2, \;\; i = 1, 2, \ldots, n$. La varianza del término aleatorio es constante para todas las observaciones. Esto se conoce como supuesto de *homocedasticidad* del término de error.

3. $Cov(u_i, u_j) = 0, \ \ \forall i \neq j$. El término de error para una observación i no está linealmente relacionado con el término de error de cualquier observación j distinta de i. Para el caso de una variable medida a lo largo del tiempo (por ejemplo, $i = 1980, 1981 \ldots, 1997$), nos referiremos a este supuesto como "ausencia de *autocorrelación*". En términos generales, diremos que no hay *correlación serial*. Notar que, dado $E(u_i) = 0$, suponer $Cov(u_i, u_j) = 0$ es equivalente a decir que $E(u_i u_j) = 0$.

4. Los valores de X_i son *no estocásticos*, es decir, que no son variables aleatorias. Este supuesto se conoce como de *regresores fijos*.

5. Los valores de X_i no son todos iguales, lo que se conoce como *no multicolinealidad perfecta*. Más adelante discutiremos con más detalles las consecuencias algebraicas de este supuesto.

Estos supuestos son conocidos como *supuestos clásicos*, y proporcionan una estructura probabilística básica para estudiar modelos lineales. Algunos de ellos tienen un sentido pedagógico, y estudiaremos luego cómo dichos supuestos pueden ser levantados y cuáles son las consecuencias de hacerlo. Sin embargo, proveen un marco simple sobre el cual poder analizar la naturaleza de los estimadores mínimos cuadráticos. Nuestro problema será encontrar estimaciones de α, β y σ^2 basándonos en una muestra (X_i, Y_i), $i = 1, \ldots, n$, sin poder observar los u_i.

1.6. Propiedades Estadísticas de los Estimadores Mínimos Cuadráticos

En realidad, el problema en cuestión consiste en encontrar *buenos* estimadores de α, β y σ^2. La sección previa presentó estimadores de los primeros dos coeficientes, basados en el principio de mínimos cuadrados, por lo que trivialmente estos estimadores son "buenos" en el sentido de que minimizan cierta noción de distancia: ellos hacen la suma de cuadrados residuales lo más pequeña posible. Es importante remarcar que para obtener los estimadores mínimos cuadráticos no se hizo uso de los supuestos clásicos descriptos anteriormente. Por lo tanto, el paso natural sería analizar si se pueden deducir propiedades adicionales que sean satisfechas por los estimadores mínimos cuadráticos, para que podamos decir que son "buenos" en un sentido que va más allá del implícito en el criterio mínimo cuadrático. Las siguientes son llamadas *propiedades estadísticas* dado que se derivan de la estructura estadística del modelo, es decir, como consecuencia de los supuestos clásicos.

Usaremos repetidamente las expresiones (1.15) y (1.13) para los estimadores mínimos cuadráticos (MCO), por lo que vale la pena recordarlas:

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

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

A continuación analizaremos en detalle las principales propiedades de $\hat{\beta}$, y dejaremos el análisis de $\hat{\alpha}$ como ejercicio para el lector. El punto conceptual inicial es ver que $\hat{\beta}$ depende explícitamente de las Y_i las cuales, a su vez, dependen de los u_i que son, por construcción, variables aleatorias. En el marco de estas notas, todas las funciones de variables aleatorias son de por sí variables aleatorias. Entonces, $\hat{\beta}$ es una variable aleatoria y por lo tanto tiene sentido hablar de sus momentos, como la media o la varianza por ejemplo, o de su distribución.

Empecemos con el modelo lineal (1.5). Sumando todas las observaciones a ambos lados y dividiendo por n se obtiene:

$$\bar{Y} = \alpha + \beta\bar{X} + \bar{u}$$

Sustrayendo en (1.5), es decir, tomando $Y_i - \bar{Y}$, obtenemos:

$$y_i = x_i\beta + u_i^*$$

donde $u_i^* = u_i - \bar{u}$. De acuerdo a los supuestos clásicos, es inmediatamente verificable que $E(u_i^*) = 0$ y, por lo tanto, $E(y_i) = x_i\beta$. Ahora estamos listos para establecer algunas propiedades básicas del estimador de mínimos cuadrados $\hat{\beta}$.

- *$\hat{\beta}$ es un estimador insesgado*, esto es: $E(\hat{\beta}) = \beta$

  Demostración:

$$\begin{aligned}
\hat{\beta} &= \sum w_i y_i \\
E(\hat{\beta}) &= \sum w_i E(y_i) \qquad \text{(los } w_i \text{ son no estocásticos)} \\
&= \sum w_i x_i \beta \\
&= \beta \sum w_i x_i \\
&= \beta \sum x_i^2 / (\sum x_i^2) \\
&= \beta
\end{aligned}$$

■

- *La varianza de $\hat{\beta}$ es $\sigma^2 / \sum x_i^2$*

  Demostración: De la propiedad de linealidad se tiene que $\hat{\beta} = \sum w_i Y_i$, luego

$$V(\hat{\beta}) = V\left(\sum w_i Y_i\right)$$

  Observemos ahora dos cosas. Primero:

$$V(Y_i) = V(\alpha + \beta X_i + u_i) = V(u_i) = \sigma^2$$

dado que X_i es no aleatoria. Segundo, notemos que $E(Y_i) = \alpha + \beta X_i$, por lo tanto

$$\begin{aligned} Cov(Y_i, Y_j) &= E\left[(Y_i - E(Y_i))(Y_j - E(Y_j))\right] \\ &= E(u_i u_j) = 0 \end{aligned}$$

por el supuesto de no correlación serial. Entonces $V\left(\sum w_i Y_i\right)$ es la varianza de una suma (ponderada) de términos no correlacionados. Por lo tanto:

$$\begin{aligned} V(\hat{\beta}) &= V\left(\sum w_i Y_i\right) \\ &= \sum w_i^2 V(Y_i) + 0 \quad \text{(por no autocorrelación)} \\ &= \sigma^2 \sum w_i^2 \quad \text{(por homocedasticidad)} \\ &= \sigma^2 \sum (x_i^2) / \left[\sum x_i^2\right]^2 \\ &= \sigma^2 / \sum x_i^2 \end{aligned}$$

■

- *Teorema de Gauss-Markov*: bajo los supuestos clásicos, $\hat{\beta}$, el estimador mínimo cuadrático de β, tiene la menor varianza (es el más eficiente) dentro del grupo de estimadores lineales e insesgados. Formalmente, si β^* es cualquier estimador lineal e insesgado de β y se cumplen todos los supuestos clasicos, entonces:

$$V(\beta^*) \geq V(\hat{\beta})$$

La demostración de una versión más general de este resultado será pospuesta, y se hará en un capítulo posterior.

Por el Teorema de Gauss-Markov, el estimador MCO es MELI (mejor estimador lineal e insesgado). Es importante notar que *mejor* no es lo mismo que *bueno*, puesto que la primera es una noción relativa, y la segunda, absoluta. Sería deseable obtener un estimador insesgado de varianza mínima, sin restringirlo al conjunto de estimadores lineales, ya que la linealidad no es una propiedad demasiado interesante per se. Además, si levantamos cualquiera de los supuestos clásicos, el estimador MCO ya no será MELI. Este hecho justifica el uso de MCO cuando todos los supuestos clásicos se satisfacen.

Estimación de σ^2

Hasta ahora nos hemos concentrado en el análisis de α y β. A continuación, como estimador de σ^2 proponemos:

$$S^2 = \frac{\sum e_i^2}{n-2}$$

Luego demostraremos que S^2 proporciona un estimador insesgado de σ^2.

1.7. Bondad del ajuste

Resulta interesante prover una medida de cuan lejos se encuentra la recta estimada con respecto a los datos. A fines de obtener tal medida de la *bondad del ajuste*, comencemos por la definición (1.7), luego despejemos Y_i y restemos a ambos lados la media de Y_i para obtener:

$$\begin{aligned} Y_i - \bar{Y} &= \hat{Y}_i - \bar{Y} + e_i \\ y_i &= \hat{y}_i + e_i \end{aligned}$$

usando la notación definida anteriormente y notando que por la propiedad 5, $\bar{Y} = \bar{\hat{Y}}$. Elevando al cuadrado a ambos lados y sumando todas las observaciones:

$$\begin{aligned} y_i^2 &= (\hat{y}_i + e_i)^2 \\ &= \hat{y}_i^2 + e_i^2 + 2\hat{y}_i e_i \\ \sum y_i^2 &= \sum \hat{y}_i^2 + \sum e_i^2 + 2\sum \hat{y}_i e_i \end{aligned}$$

El siguiente paso consiste en mostrar que $\sum \hat{y}_i e_i = 0$:

$$\begin{aligned} \sum \hat{y}_i e_i &= \sum (\hat{Y}_i - \bar{\hat{Y}}) e_i \\ &= \sum (\hat{\alpha} + \hat{\beta} X_i - \bar{\hat{Y}}) e_i \\ &= \hat{\alpha} \sum e_i + \hat{\beta} \sum X_i e_i - \bar{\hat{Y}} \sum e_i \\ &= 0 + 0 - 0 \end{aligned}$$

porque se cumplen las condiciones de primer orden (1.9) y (1.11). Luego, podemos obtener la siguiente descomposición:

$$\begin{aligned} \sum y_i^2 &= \sum \hat{y}_i{}^2 + \sum e_i^2 \\ SCT &= SCE + SCR \end{aligned}$$

Este es un resultado clave que indica que cuando usamos el método de mínimos cuadrados, la variabilidad total de la variable dependiente alrededor de su media muestral (SCT: suma de cuadrados totales) puede descomponerse como la suma de dos terminos. El primero corresponde a la variabilidad de $\hat{Y}$ (SCE: suma de cuadrados explicados), que representa la variabilidad explicada por el modelo estimado. El segundo término representa la variabilidad no explicada por el modelo (SCR: suma de cuadrados residuales), asociada al término de error.

Para un modelo dado, la mejor situación se presenta cuando los errores son todos iguales a cero, caso en el cual la variabilidad total (SCT) coincide con la variabilidad explicada (SCE). La peor situación corresponde al caso en el cual el modelo estimado no explica nada de la variabilidad total, caso en el cual la SCT coincide con SCR. De esta observación, es natural sugerir la siguiente medida de bondad del ajuste, conocida como R^2 o *coeficiente de determinación*:

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

Puede mostrarse (se deja como ejercicio para el lector) que $R^2 = r^2$. Consecuentemente, $0 \leq R^2 \leq 1$. Cuando $R^2 = 1$, $|r| = 1$, que corresponde al caso en el cual la relación entre Y y X es exactamente lineal. Por otro lado, $R^2 = 0$ es equivalente a $r = 0$, que corresponde al caso en el que Y y X no están linealmente relacionadas. Es interesante observar que SCT no depende del modelo estimado, es decir, no depende de $\hat{\beta}$ ni de $\hat{\alpha}$. Entonces, si $\hat{\beta}$ y $\hat{\alpha}$ son elegidos de tal forma que minimicen la SCR estarán automáticamente maximizando R^2. Esto implica que, para un modelo dado, la estimación por mínimos cuadrados maximiza R^2.

Podría decirse que el R^2 es la medida de la calidad de estimación de un modelo más usada y abusada. En capítulos posteriores discutiremos en detalle hasta qué punto puede usarse el R^2 para determinar si un modelo estimado es bueno o malo.

1.8. Inferencia en el modelo lineal con dos variables

Los métodos discutidos hasta ahora proporcionan estimadores razonablemente buenos de los parámetros de interés α, β y σ^2, pero usualmente estaremos interesados en evaluar hipótesis vinculadas a dichos parámetros, o en construir intervalos de confianza. Por ejemplo, tomemos el caso de una simple función de consumo, especificada como una función lineal del ingreso. Entonces, podríamos estar interesados en evaluar si la propensión marginal a consumir es igual a, digamos, 0.75, o si el consumo autónomo es igual a cero.

Una hipótesis sobre un parámetro del modelo es una conjetura sobre el mismo, que podrá ser tanto falsa como verdadera. El problema central radica en el hecho de que, para saber si dicha hipótesis es verdadera o falsa, no tenemos la posibilidad de observar los parámetros. En su lugar, tenemos una estimación de los mismos basada en los datos disponibles.

Como ejemplo, supongamos que estamos interesados en evaluar la hipótesis nula de que el ingreso no es un factor explicativo del consumo, contra la hipótesis alternativa que dicha variable es relevante. En nuestro esquema simplificado, esto corresponde a evaluar $H_0 : \beta = 0$ contra $H_A : \beta \neq 0$. La lógica que utilizaremos será la siguiente: si la hipótesis nula (H_0) fuera de hecho verdadera, β sería exactamente igual a cero. Las realizaciones de $\hat{\beta}$ pueden tomar potencialmente cualquier valor, dado que $\hat{\beta}$ es por construcción una variable aleatoria. Pero si $\hat{\beta}$ es un buen estimador de β, cuando la hipótesis nula es verdadera, $\hat{\beta}$ debería tomar valores *cercanos* a cero. Por otro lado, si la hipótesis nula fuera falsa, las realizaciones de β deberían ser significativamente distintas de cero. Luego, el procedimiento consiste en computar $\hat{\beta}$ a partir de los datos disponibles, y rechazar la hipótesis nula si el valor obtenido es significativamente diferente de cero, o no rechazarla en caso contrario.

Desde luego, la cuestión central detrás de este procedimiento es la de especificar qué es lo que queremos decir con "muy cerca", dado que $\hat{\beta}$ es una variable aleatoria. Más específicamente, necesitamos saber la distribución de $\hat{\beta}$ bajo la hipótesis nula, de tal forma de poder definir precisamente la noción de "*significativamente diferente de cero*". En este contexto, tal afirmación es necesariamente probabilística, es decir, tomamos como región de rechazo un conjunto de valores que caen lejos de cero, o un grupo de valores que bajo la hipótesis nula aparecerían con muy baja probabilidad.

Las propiedades discutidas en secciones anteriores son informativas sobre ciertos momentos de $\hat{\beta}$ o $\hat{\alpha}$ (por ejemplo, sus medias y varianzas), pero no son suficientes a los fines de conocer sus distribuciones. Por ello, necesitamos introducir un supuesto adicional. Supondremos que u_i está normalmente distribuido, para $i = 1, \ldots, n$. Dado que ya hemos supuesto que u_i tiene media cero y una varianza constante igual a σ^2, tenemos:

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

Dado que $Y_i = \alpha + \beta X_i + u_i$ y que los X_i son no aleatorios, podemos observar inmediatamente que los Y_i están también normalmente distribuidos, dado que una transformación lineal de una variable aleatoria normal es también normal. En particular, dado que la distribución normal puede ser caracterizada por su media y varianza solamente, tendremos:

$$Y_i \sim N(\alpha + \beta X_i, \sigma^2), \qquad i = 1 \ldots n.$$

De igual modo, $\hat{\beta}$ está normalmente distribuida, por ser una combinación lineal de los Y_i, esto es:

$$\hat{\beta} \sim N(\beta, \sigma^2 / \sum x_i^2)$$

Si σ^2 fuera conocido, podríamos usar este resultado para implementar un test para las hipótesis:

$$H_o : \beta = \beta_o \text{ vs. } H_A : \beta \neq \beta_o$$

donde β_o es cualquier valor.

Sustrayendo de $\hat{\beta}$ su valor esperado y dividiendo por su desvío estándar obtenemos:

$$z = \frac{\hat{\beta} - \beta_o}{\sigma / \sqrt{\sum x_i^2}} \sim N(0, 1)$$

Por lo tanto, si la hipótesis nula es cierta, z debería tomar valores pequeños (en valor absoluto), y relativamente grandes en caso contrario. Como podemos recordar de algún curso básico de estadística, el test puede implementarse definiendo una región de rechazo y otra de aceptación, como sigue. La región de aceptación incluye valores que caen "cerca" al propuesto por la hipótesis nula. Sean $c < 1$ y z_c un número tal que:

$$Pr(-z_c \leq z \leq z_c) = 1 - c$$

Reemplazando z por su definición:

$$Pr\left[\beta_o - z_c\left(\sigma/\sqrt{\sum x_i^2}\right) \leq \hat{\beta} \leq \beta_o + z_c\left(\sigma/\sqrt{\sum x_i^2}\right)\right] = 1 - c$$

Luego, la región de aceptación está dada por el siguiente intervalo:

$$\beta_o \pm z_c(\sigma/\sqrt{\sum x_i^2})$$

por lo que aceptaremos la hipótesis nula si la realización observada de $\hat{\beta}$ cae dentro de dicho intervalo, y la rechazamos en caso contrario. El número c es especificado previamente y usualmente es un número pequeño. Se lo llama *nivel de significatividad* del test. Notar que c representa la probabilidad de rechazar la hipótesis nula cuando ésta es en verdad correcta (es decir, de cometer un *error tipo I*). Bajo el supuesto de normalidad, el valor de z_c puede obtenerser fácilmente de la tabla de percentiles de la distribución normal estándar.

Una lógica similar puede aplicarse para construir *intervalos de confianza* para β_o. Notar que:

$$Pr\left[\hat{\beta} - z_c(\sigma/\sqrt{\sum x_i^2}) \leq \beta_o \leq \hat{\beta} + z_c(\sigma/\sqrt{\sum x_i^2})\right] = 1 - c$$

Luego, un intervalo de confianza $(1 - c) * 100\,\%$ para β_0 estará dado por:

$$\hat{\beta} \pm z_c\left(\sigma/\sqrt{\sum x_i^2}\right)$$

El problema práctico con los procedimientos previos es que requieren que conozcamos σ^2, algo que habitualmente no ocurre. En su lugar, podemos computar su versión estimada S^2. Definimos t como:

$$t = \frac{\hat{\beta} - \beta}{S\Big/\sqrt{\sum x_i^2}}$$

t es simplemente z para el cual hemos reemplazado σ^2 por S^2. Un resultado importante es que haciendo este reemplazo tendremos:

$$t \sim t_{n-2}$$

Esto es, el "estadístico t" tiene distribución *t de Student* con $n - 2$ grados de libertad. Por lo tanto, cuando usamos la versión estimada de la varianza, la distribución será diferente a la del estadístico que usamos para hacer simples test de hipótesis o construir intervalos de confianza.

Aplicando una vez más la misma lógica, para poder testear la hipótesis nula $H_o : \beta = \beta_o$ contra $H_A : \beta \neq \beta_o$ usamos el estadístico t:

$$t = \frac{\hat{\beta} - \beta_o}{S/\sqrt{\sum x_i^2}} \sim t_{n-2}$$

y un intervalo de confianza $(1 - c) * 100\,\%$ para β_0 estará dado por:

$$\hat{\beta} \pm t_c(S/\sqrt{\sum x_i^2})$$

donde ahora t_c es un percentil de la distrubución "t" con $n - 2$ grados de libertad, que usualmente son tabulados en libros básicos de estadística o econometría.

Un caso particular importante es la *hipótesis de no significatividad*, esto es $H_o : \beta_o = 0$ contra $H_A : \beta_0 \neq 0$. Bajo la hipótesis nula, X no contribuye a explicar Y, y bajo la alternativa, X está linealmente relacionado con Y. Reemplazando β_o por 0 , obtenemos:

$$t_I = \frac{\hat{\beta}}{S/\sqrt{\sum x_i^2}} \sim t_{n-2}$$

el cual viene habitualmente incorporado como un resultado estándar en la mayoría de los programas estadísticos.

Otra alternativa para verificar la significatividad de la relación lineal entre dos variables es tratar de saber cuán grande es la suma de cuadrados explicados SCE. Recordemos que si el modelo tiene intercepto tendremos:

$$SCT = SCE + SCR$$

Si no hay relación lineal entre Y y X, SCE debería ser muy cercano a cero. Considere el siguiente estadístico, que es sólo una versión "estandarizada" de la SCE:

$$F = \frac{SCE}{SCR/(n-2)}$$

Puede demostrarse que bajo el supuesto de normalidad, F tiene una distribución F de Fisher con 1 grado de libertad en el numerador, y $n - 2$ grados de libertad en el denominador, que usualmente se denota como $F_{(1,n-2)}$. Notar que si X no contribuye a explicar a Y en un sentido lineal, la SCE debería ser muy pequeña, lo que haría que F sea muy pequeña. Entonces, deberíamos rechazar la hipótesis nula de que X no explica a Y si el estadístico F computado a partir de los datos toma valores relativamente grandes, y no rechazarla en caso contrario.

Notar que, por definición, $R^2 = SCE/SCT = 1 - SCR/SCT$. Dividiendo numerador y denominador del estadístico F por la suma de cuadrados totales SCT, y despejando para SCE y SCR y reemplazando, podremos escribir el estadístico F en términos del coeficiente R^2 como sigue:

$$F = \frac{R^2}{(1 - R^2)/(n-2)}$$

Luego, el estadístico F en realidad analiza si el R^2 es significativamente alto. Como es de esperar, hay una relación muy cercana entre el estadístico F y el estadístico "t" para la hipótesis de no significatividad (t_I). De hecho, cuando no hay relación lineal entre Y y X, la SCE es cero, o $\beta_0 = 0$. Incluso, se puede demostrar fácilmente que:

$$F = t_I^2$$

Ejercicio: Demostrar $F = t_I^2$.

Capítulo 2

Modelo Lineal con Múltiples Variables

2.1. El Modelo de K-Variables bajo los Supuestos Clásicos

En este capítulo vamos a extender el modelo básico para poder utilizar K variables explicativas. Es decir, Y ahora depende de K variables más el termino de error:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_K X_{Ki} + u_i \qquad i = 1, \ldots, n \tag{2.1}$$

La notación utilizada implica que X_{ki} es la i-ésima observación de la k-ésima variable explicativa, con $(k = 2, \ldots, K)$. Por ejemplo, podemos pensar que Y es el ingreso, y verlo como una función de la educación, la experiencia, y el término de error, o puede que Y sea el consumo, y esté en función del ingreso, la riqueza y un término aleatorio.

Es importante aclarar que el modelo posee K variables explicativas y no $K - 1$ como podría parecer a primera vista. Notar que la primera variable puede ser vista como $X_{1i} = 1$ para cada observación, entonces, son K variables explicativas, con la primera de ellas siempre igual a uno. De ahora en más, a menos que se aclare lo contrario, el grupo de las K variables explicativas incluye como primer variable a la constante o intercepto.

La interpretación de los coeficientes sigue siendo la misma. Si sólo dejamos que varíe marginalmente la variable X_{ki}, entonces β_k, k=2,...,K son derivadas parciales que indican el cambio marginal en Y cuando la k-ésima variable explicativa se modifica marginalmente, manteniendo las demás variables constantes. El modelo discutido en el capítulo anterior es un caso particular del modelo general, correspondiente a K=2.

Al igual que en en el caso de dos variables,vamos a introducir, además de la relación lineal entre Y y las variables explicativas, supuestos similares sobre la estructura del modelo:

- $E(u_i)$=0 $\qquad i = 1, \ldots, n.$
- $Var(u_i)$=$E[(u_i - E(u_i))^2]$=$E(u_i)^2$=σ^2 $\qquad i = 1, \ldots, n.$

  La varianza del término de error es constante para todas las observaciones (homocedasticidad).
- $Cov(u_i, u_j) = 0 \qquad \forall i \neq j.$

  De nuevo remarcamos que, como $E(u_i)$=0, asumir $Cov(u_i, u_j)$=0 es equivalente a asumir $E(u_i \cdot u_j) = 0$.
- Las variables explicativas son *no estocásticas* y no existe relación lineal exacta entre ellas. El significado exacto de este supuesto se comprenderá mejor cuando lo expongamos en forma matricial.

Éstos son conocidos como los *supuestos clásicos* y, al igual que antes, nuestro problema es encontrar estimadores para $\beta_1, \beta_2, \ldots, \beta_K$ y σ^2 basados en la muestra $(1, X_{2i}, X_{3i}, \ldots, X_{ki}, Y_i), i = 1, \ldots, n$ sin observar los u_i.

2.2. El modelo en forma matricial

La ecuación (2.1) significa que la relación lineal vale para las n observaciones, es decir, es una forma reducida del siguiente sistema de n ecuaciones lineales, una para cada observación:

$$\begin{aligned}
Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \ldots + \beta_K X_{K1} + u_1 \\
Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \ldots + \beta_K X_{K2} + u_2 \\
\vdots &= \vdots \\
Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \ldots + \beta_K X_{Kn} + u_n
\end{aligned}$$

El propósito de esta sección es expresar el modelo lineal en forma matricial. Esta es una forma elegante y simple de trabajar con el modelo, evitando el uso de sumatorias. El lector debería estar familiarizado con las operaciones básicas de matrices (suma, multiplicación, inversa, etc.). Primero introduciremos las matrices sólo con fines notacionales, luego avanzaremos hacia un uso más sistemático de las mismas.

Consideremos las siguientes matrices y vectores:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & X_{21} & \dots & X_{K1} \\ 1 & X_{22} & & X_{K2} \\ \vdots & & X_{33} & \\ 1 & & & X_{Kn} \end{bmatrix}$$

Entonces, el sistema lineal de (2.1) puede escribirse como:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & \dots & X_{K1} \\ 1 & X_{22} & & X_{K2} \\ \vdots & & X_{33} & \\ 1 & & & X_{Kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$$Y = X\beta + u \tag{2.2}$$

Es crucial comprender las dimensiones de las matrices y vectores utilizados. Y es un vector n x 1. X es una matriz de dimensión n x K cuya primera columna es un vector de unos, β es un vector de K x 1 y u es un vector de n x 1. De esta forma, todos los productos quedan bien definidos. (2.2) es el modelo de K variables en forma matricial.

2.3. Algunos resultados básicos de matrices y vectores aleatorios

Antes de proceder, presentaremos algunos resultados de matrices y vectores:

1. Rango de una matriz: Sea A una matriz de dimensiones m x n. La misma puede pensarse como un matriz formada por n vectores columna con m elementos cada uno. O bien, como una matriz de m vectores fila con n elementos cada uno. El *rango columna* de A se define como el máximo número de columnas linealmente independientes. De forma similar, el *rango fila* es el máximo número de filas linealmente independientes.

2. El rango columna es igual al rango fila. Entonces, definamos el *rango* de una matriz como el número máximo de filas o columnas que son linealmente independientes. Denotaremos el rango de la matriz A como $\rho(A)$.

3. Definamos a A como una matriz cuadrada de dimensiones m x m. A es *no singular* si $|A| \neq 0$. En tal caso, existe una única matriz no singular A^{-1} llamada *matriz inversa* de A, de forma tal que $AA^{-1} = A^{-1}A = I_m$.

4. Este resultado establece la conexión entre el rango de una matriz y su determinante. Sea A una matriz cuadrada de dimensiones m x m.

$$\begin{aligned} Si \quad \rho(A) = m \Rightarrow |A| \neq 0 \\ Si \quad \rho(A) < m \Rightarrow |A| = 0 \end{aligned}$$

5. Definamos a X como una matriz de dimensiones n x k, con rango columna completo $\rho(X) = k$. Entonces:

$$\rho(X) = \rho(X'X) = k$$

Este resultado garantiza la existencia de $(X'X)^{-1}$ en base al rago de la matriz X.

6. Sean b y a dos vectores K x 1. Entonces:

$$\frac{\partial(b'a)}{b} = a$$

7. Sea b un vector de dimensiones k x 1, y definamos A como una matriz simétrica de dimensiones K x K.

$$\frac{\partial(b'Ab)}{b} = 2Ab$$

8. Definamos a Y como un vector de K variables aleatorias:

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_K \end{bmatrix}$$

$$E(Y) = \mu = \begin{bmatrix} E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_K) \end{bmatrix}$$

y:

$$V(Y) = E[(Y-\mu)(Y-\mu)']$$

$$= \begin{bmatrix} E(Y_1-\mu_1)^2 & E(Y_1-\mu_1)(Y_2-\mu_2) & \dots & \dots & E(Y_1-\mu_1)(Y_K-\mu_K) \\ E(Y_2-\mu_2)(Y_1-\mu_1) & E(Y_2-\mu_2)^2 & & & \vdots \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \\ E(Y_K-\mu_K)(Y_1-\mu_1) & \dots & \dots & \dots & E(Y_K-\mu_K)^2 \end{bmatrix}$$

$$= \begin{bmatrix} V(Y_1) & Cov(Y_1Y_2) & \dots & Cov(Y_1Y_K) \\ \vdots & V(Y_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ \dots & \dots & \dots & V(Y_K) \end{bmatrix}$$

Habitualmente, a la varianza de un vector se la llama *matriz de varianzas y covarianzas*, enfatizando que la misma es una matriz y no un número.

9. Si $V(Y) = \sum$ y c es un vector de dimensiones K x 1, entonces $V(c'Y) = c'V(Y)c = c'\sum c$.

10. Sea A una matriz cuadrada de dimensiones m x m. La *traza* de la matriz A (denotada como $tr(A)$) es la suma de todos los elementos en la diagonal principal de la matriz A:

$$tr(A) = \sum_{i=1}^{n} A_{ii}$$

Es fácil verificar las siguientes propiedades de la traza de una matriz:

- Si A es un escalar, trivialmente $tr(A) = A$.
- $tr(AB) = tr(BA)$
- $tr(AB) = tr(A) + tr(B)$

2.4. Los supuestos clásicos en forma matricial

Los supuestos clásicos se pueden expresar en forma de matrices y vectores aleatorios de una manera sencilla y compacta:

1. $E(u) = 0$. Dada la definición previa de esperanza de un vector, esto es equivalente a $E(u_i) = 0, i = 1, \ldots, n$.

2. $V(u) = E[(u - E(u))(u - E(u))'] = E(uu') = \sigma^2 I_n$

$$V(u) = \begin{bmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{bmatrix}$$

   Recordando que el elemento (i, j) de la matriz $V(u)$ es $Cov(u_i, u_j)$, es posible ver que el hecho de poder escribir la varianza de u como un escalar por la matriz identidad es equivalente a asumir:

$$Var(u_i) = \sigma^2 \quad i = 1, \ldots, n \text{ (homocedasticidad) y}$$
$$Cov(u_i, u_j) = 0 \quad \forall j \neq i \text{ (no correlación serial).}$$

3. X es *no estocástica* y $\rho(X) = K$. La primera parte significa, como antes, que cada elemento de X será tomado como un simple número. Para entender la segunda parte, pensemos en X como una matriz formada por K vectores columnas, donde cada uno contiene las n observaciones de cada variable explicativa. De acuerdo con los conceptos introducidos antereriormente, este supuesto implica que es imposible obtener cualquier columna de X como una combinación lineal de las restantes. Por ejemplo, no podemos tener una variable que sea el ingreso en dólares, y otra que sea el ingreso en pesos, ya que una es igual a la otra multiplicada por un número. Tampoco podemos utilizar como variables explicativas ingreso, riqueza, y riqueza más ingreso. Notar que podemos utilizar ingreso e ingreso al cuadrado, ya que el supuesto prohibe relaciones lineales exactas, y elevar al cuadrado no es una operación lineal. Además, como el supuesto no permite relaciones lineales *exactas*, podemos utilizar variables explicativas cuya relación sea tan alta como se desee, sin que llegue a ser exacta. Por ejemplo, podemos utilizar ingreso y riqueza como variables explicativas para el consumo, más allá que en la práctica esten altamente correlacionadas. El supuesto de no correlación lineal exacta es llamado supuesto de *no multicolinealidad*. Aunque no explícitamente bajo este nombre, ya habiamos realizado este supuesto en el capítulo anterior. En ese caso, el supuesto era que las X_i no podían ser todas iguales entre ellas, ya que si así fuera, cada X_i podía obtenerse como el producto entre 1 y un número. En esa situación, X_i sería exactamente *colineal* con el intercepto.

2.5. Estimación mínimo cuadrática

Como en el capítulo anterior, queremos estimar el vector β a partir de las observaciones Y y X. Comencemos definiendo:

$$\hat{Y} \equiv X\hat{\beta}$$

y

$$e \equiv Y - \hat{Y} \equiv Y - X\hat{\beta}$$

El criterio mínimo cuadrático puede ser definido ahora de la misma forma que en el capítulo anterior:

$$SCR(\hat{\beta}) = \sum_{i=1}^{n} e_i^2 = e'e$$

Escribiendo $e'e$ explícitamente en terminos de $\hat{\beta}$:

$$\begin{aligned} e'e &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

En la segunda línea, $-\hat{\beta}'X'Y$ es un escalar y entonces es trivialmente igual a su transpuesto $-Y'X\hat{\beta}$, así es como obtenemos el resultado de la tercer línea.

Se puede demostrar fácilmente que SCR es una función estrictamente convexa y diferenciable en $\hat{\beta}$. Entonces, las condiciones de primer orden para un punto estacionario son suficientes para un mínimo global. Las condiciones de primer orden son:

$$\frac{\partial(e'e)}{\partial\hat{\beta}} = 0$$

Utilizando las reglas de derivación introducidas en la sección anterior:

$$\frac{\partial(e'e)}{\partial\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

que es un sistema de K ecuaciones lineales con K incógnitas $(\hat{\beta})$. Resolviendo para $\hat{\beta}$ obtenemos:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2.3}$$

Es crucial entender que el supuesto de no multicolinealidad, $\rho(X) = K$, es el que garantiza la existencia de $(X'X)^{-1}$ y, por lo tanto, la posibilidad de obtener una solución única para $\hat{\beta}$.

Ejemplo: Volvamos a considerar el modelo de dos variables. Vamos a ver que el resultado obtenido antes se puede derivar fácilmente en forma matricial.

$$
\begin{aligned}
X &= \begin{bmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \\
X'X &= \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \\
X'Y &= \begin{bmatrix} 1 & \dots & 1 \\ X_1 & \dots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum Y_i \\ \vdots \\ \sum X_i Y_i \end{bmatrix} \\
\begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} &= \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix} \\
\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_1 \end{bmatrix} &= \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i \\ \sum X_i Y_i \end{bmatrix}
\end{aligned}
$$

Para obtener la fórmula del capítulo anterior, debemos obtener la inversa y desarrollar el producto. Esto quedará como ejercicio para el lector.

2.6. Propiedades algebraicas del estimador de mínimos cuadrados

- Propiedad 1: $X'e = 0$.

  Al dividir por -2 las condiciones de primer orden y sacando X como factor común, obtenemos:

$$X'(Y - X\hat{\beta}) = 0$$

  Notar que el término entre paréntesis es la definición de e, con lo cual se obtiene el resultado. Para ver lo que esto implica, expresemos e y X explícitamente:

$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{k1} & X_{k2} & \dots & X_{kn} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = 0$$

Notar que el producto $X'e$ es un vector cuyo primer elemento es:

$$\sum e_i = 0$$

Entonces, si el modelo incluye un intercepto, el método de mínimos cuadrados fuerza a la suma de los términos de error, y por lo tanto a su promedio, a ser igual a cero.

Esta propiedad tiene una interpretación adicional. Vamos a demostrar que $X'e = 0$ es equivalente a decir que el término de error no está linealmente relacionado con ninguna variable explicativa, es decir $Cov(X_k, e) = 0, \quad k$=2,...,K.

Demostración: De la definición de $Cov(X_k, e)$:

$$\begin{aligned} Cov(X_k, e) &= \frac{1}{n-1}[\sum(X_{ki} - \bar{X}_k)(e_i - \bar{e})] \\ &= \frac{1}{n-1}[\sum(X_{ki} - \bar{X}_k)e_i] \\ &= \frac{1}{n-1}[\sum X_{ki}e_i - \bar{X}_k \sum e_i] \\ &= \frac{1}{n-1}[\sum X_{ki}e_i] \\ &= \frac{1}{n-1}[X_k'e] = 0 \end{aligned}$$

Si $X_k'e = 0$, entonces $Cov(X_k, e) = 0$, que es el resultado deseado. ■

- Propiedad 2: $\hat{Y}'e = 0$

$$\hat{Y}'e = \hat{\beta}X'e = 0$$

por la propiedad anterior. Como antes, la interpretación de esto es:

$$Cov(\hat{Y}, e) = 0$$

- Propiedad 3: $\hat{\beta}$ es una función lineal de Y. Esto es, existe una matriz de números A, de dimensiones K x n, tal que $\hat{\beta}$ puede escribirse como:

$$\hat{\beta} = AY$$

Demostración: A partir de (2.3), la expresión del estimador mínimo cuadrático, definamos $A \equiv (X'X)^{-1}X'$, obteniendo así la expresión deseada. ■

Vamos a dejar como ejercicio la demostración de las siguientes propiedades:

- Propiedad 4: La línea de regresión pasa por los puntos de las medias muestrales.
- Propiedad 5: Si hay un intercepto en el modelo, las medias muestrales de Y_i e $\hat{Y}_i$ son la misma.

2.7. Propiedades estadísticas del estimador de mínimos cuadrados

- $\hat{\beta}$ es *insesgado* para β, es decir $E(\hat{\beta}) = \beta$.

  Demostración: La prueba es muy elegante:

$$\begin{aligned}
\hat{\beta} &= (X'X)^{-1}X'Y \\
&= (X'X)^{-1}X'(X\beta + u) \\
&= (X'X)^{-1}(X'X)\beta + (X'X)^{-1}X'u \\
&= \beta + (X'X)^{-1}X'u \\
E(\hat{\beta}) &= \beta + E[(X'X)^{-1}X'u] \qquad (2.4) \\
&= \beta + (X'X)^{-1}X'E(u) \\
&= \beta
\end{aligned}$$

  Notar que para obtener el resultado anterior se debe suponer que las X son no estocásticas y que $E(u) = 0$. ■

- La varianza de $\hat{\beta}$ es $\sigma^2(X'X)^{-1}$.

  Demostración: Usando la definición de varianza de un vector:

$$\begin{aligned}
V(\hat{\beta}) &= E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'] \\
&= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)']
\end{aligned}$$

  dado que demostramos que $\hat{\beta}$ es insesgado. De (2.4) obtenemos:

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

  Reemplazando arriba:

$$\begin{aligned}
V(\hat{\beta}) &= E[(X'X)^{-1}X'uu'X(X'X)^{-1}] \\
&= (X'X)^{-1}X'E(uu')X(X'X)^{-1} \\
&= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\
&= \sigma^2(X'X)^{-1}
\end{aligned}$$

  Notar que el paso al segundo renglón se realizó utilizando el supuesto de X no estocásticas, y al tercer renglón utilizando $E(uu') = V(u) = \sigma^2 I_n$ (no correlación serial y homocedasticidad).

Recordemos que en el modelo de dos variables:

$$V(\hat{\beta}) \quad = \quad \sigma^2(X'X)^{-1} = \frac{\sigma^2}{\sum x_i^2}$$

■

- *Teorema de Gauss-Markov*: Bajo los supuestos clásicos, el estimador de mínimos cuadrados $\hat{\beta}$ es el mejor estimador lineal e insesgado de β.

  Específicamente, sea $\hat{\beta}$ el estimador de mínimos cuadrados de β, y sea $\tilde{\beta}$ cualquier otro estimador lineal e insesgado, entonces:

  $$V(\tilde{\beta}) - V(\hat{\beta})$$

  es una matriz semidefinida positiva.

  El teorema dice que, si se cumplen todos los supuestos clásicos, el estimador de mínimos cuadrados es el de mínima varianza dentro del grupo de los lineales e insesgados. Aunque intuitivamente nos gustaría decir que $V(\tilde{\beta}) \geq V(\hat{\beta})$, debemos tener cuidado ya que estamos trabajando con matrices. Para ello, necesitamos una intución de lo que significa que una matriz sea "más grande" que otra o, de forma equivalente, que la diferencia entre ambas sea no negativa. Vamos a decir que la diferencia es una matriz semidefinida positiva.

  De acuerdo a la definición de matriz semidefinida positiva, decir que $V(\tilde{\beta}) - V(\hat{\beta})$ es una matriz semidefinida positiva es equivalente a decir que, para cada vector c de K constantes:

  $$c'[V(\tilde{\beta}) - V(\hat{\beta})]c \quad \geq \quad 0$$

  O, utilizando las reglas vistas anteriormente:

  $$V(c'\tilde{\beta}) - V(c'\hat{\beta}) \quad \geq \quad 0$$

  para cada vector c. Este es el resultado que queríamos probar.

  Demostración: Para que $\tilde{\beta}$ sea lineal, debe existir una matriz A_{Kxn} de rango K tal que $\tilde{\beta} = AY$. Entonces, bajo los supuestos clásicos:

  $$E(\tilde{\beta}) = E(AY) = E(A(X\beta + u)) = AX\beta \tag{2.5}$$

  Para que $\tilde{\beta}$ sea insesgado, debe cumplirse:

  $$E(\tilde{\beta}) = \beta \tag{2.6}$$

Para que $\tilde{\beta}$ sea lineal e insesgado, (2.5) y (2.6) deben cumplirse simultáneamente y para eso se requiere $AX = I_K$.

Trivialmente, $\tilde{\beta} = \hat{\beta} + \tilde{\beta} - \hat{\beta} \equiv \hat{\beta} + \hat{\gamma}$, con $\hat{\gamma} = \tilde{\beta} - \hat{\beta}$. Notar que $V(\tilde{\beta}) = V(\hat{\beta}) + V(\hat{\gamma})$ sólo si $Cov(\hat{\beta}, \hat{\gamma}) = 0$. Entonces, si probamos $Cov(\hat{\beta}, \hat{\gamma}) = 0$, tendremos el resultado deseado (¿por qué?). Para ello, notemos que, trivialmente, $E(\hat{\gamma}) = 0$, entonces:

$$Cov(\hat{\beta}, \hat{\gamma}) = E[(\hat{\beta} - \beta)\hat{\gamma}']$$

Notar que:

$$\begin{aligned}
\hat{\gamma} &= AY - (X'X)^{-1}X'Y \\
&= (A - (X'X)^{-1}X')Y \\
&= (A - (X'X)^{-1}X')(X\beta + u) \\
&= (A - (X'X)^{-1}X')u
\end{aligned}$$

Reemplazando:

$$\begin{aligned}
Cov(\hat{\beta}, \hat{\gamma}) &= E[(\hat{\beta} - \beta)\hat{\gamma}'] \\
&= E[(X'X)^{-1}X'uu'(A - (X'X)^{-1}X')'] \\
&= \sigma^2[(X'X)^{-1}X'(A' - X(X'X)^{-1})] \\
&= \sigma^2[(X'X)^{-1}X'A' - (X'X)^{-1}X'X(X'X)^{-1}] \\
&= 0
\end{aligned}$$

Donde utilizamos $V(u) = E(uu') = \sigma^2 I_n$ y $AX = I$. Por lo tanto, siguiendo el argumento anterior obtenemos:

$$V(\tilde{\beta}) - V(\hat{\beta}) = V(\hat{\gamma})$$

que es semidefinida positiva por definición.

■

Hay varios puntos importantes para discutir.

Primero, el teorema de Gauss-Markov nos da un resultado de optimalidad para mínimos cuadrados. Dice que, bajo los supuestos clásicos, el estimador de mínimos cuadrados es el mejor dentro del grupo de los lineales e insesgados. Es importante notar que esto es bastante restrictivo, ya que limita el resultado a un grupo específico de estimadores (los lineales e insesgados). Por ejemplo, este teorema no puede utilizarse para comparar el estimador de

mínimos cuadrados con otro que sea no-lineal, o tal vez sesgado. Puede que exista un estimador que sea sesgado, pero de menor varianza que el de mínimos cuadrados.

Segundo, el teorema establece una comparación ordinal entre cierta clase de estimadores (los lineales e insesgados). Que dicho estimador sea el 'mejor'de cierta clase, no necesariamente implica que sea 'bueno'. En algunos casos, el estimador de mínimos cuadrados, más allá de que sea el mejor estimador lineal e insesgado (MELI), puede brindar estimaciones bastante pobres. Veremos esto más adelante, al tratar el tema de multicolinealidad. Tercero, es importante observar que se utilizaron todos los supuestos clásicos en la demostración del teorema. Es decir, todos los supuestos son condición necesaria y suficiente para el teorema. Entonces, lógicamente, podemos inferir que si uno de los supuestos no se cumple, el teorema queda invalidado.

2.8. Estimación de σ^2

Hasta este punto nos hemos concentrado en el análisis de β. Pasaremos ahora a estudiar la estimación de σ^2 y para lo que proponemos:

$$S^2 = \frac{\sum e_i^2}{n-K} = \frac{e'e}{n-K}$$

La principal razón para usar tal estimador es que S^2 es un estimador insesgado de σ^2, es decir, $E(S^2) = \sigma^2$. La prueba requiere que previamente definamos la matriz M.

La Matriz M

$$\begin{aligned} e &= Y - \hat{Y} \\ &= Y - X\hat{\beta} \\ &= Y - X(X'X)^{-1}X'Y \\ &= (I_n - X(X'X)^{-1}X')Y \\ &= MY \end{aligned}$$

con $M = I_n - X(X'X)^{-1}X'$. Notemos que e e Y están linealmente relacionados a través de M. Además:

$$\begin{aligned} e &= MY \\ &= M(X\beta + u) \\ &= MX\beta + Mu \\ &= (I_n - X(X'X)^{-1}X')X\beta + Mu \\ &= (X - X(X'X)^{-1}X'X)\beta + Mu \\ &= Mu \end{aligned}$$

Esto último muestra que e y u también están linealmente relacionados a través de M. Se puede demostrar fácilmente que M es simétrica ($M = M'$) e idempotente ($M'M = M$).

Pudiendo hacer uso de la matriz M, volvamos ahora a la demostración de insesgadez de S^2.

Demostración: Comenzando por la definición de $E(S^2)$:

$$\begin{aligned}
E(S^2) &= E(\frac{e'e}{n-K}) \\
&= E(\frac{u'M'Mu}{n-K}) \\
&= \frac{E(u'Mu)}{n-K} \\
&= \frac{E(tr(u'Mu))}{n-K} \\
&= \frac{E(tr(uu'M)}{n-K} \\
&= \frac{tr(E(uu'M))}{n-K} \\
&= \frac{tr(\sigma^2 IM)}{n-K} \\
&= \frac{\sigma^2 tr(M)}{n-K}
\end{aligned}$$

Entonces:

$$\begin{aligned}
tr(M) &= tr(I - X(X'X)^{-1}X') \\
&= tr(I) - tr(X(X'X)^{-1}X') \\
&= n - tr((X'X)^{-1}X'X) \\
&= n - tr(I_K) \\
&= n - K
\end{aligned}$$

Reemplazando:

$$E(S^2) = \sigma^2 \frac{n-K}{n-K} = \sigma^2$$

■

2.9. Bondad de ajuste

Realizamos la misma descomposición que en el capítulo anterior, pero en este caso vale para K variables:

$$\begin{array}{rcccl} \sum y_i^2 & = & \sum \hat{y}_i{}^2 & + & \sum e_i^2 \\ SCT & = & SCE & + & SCR \end{array}$$

Siendo SCT la suma de cuadrados totales, SCE la suma de cuadrados explicados y SCR la suma de cuadrados residuales. Es fácil probar este resultado. De hecho, la demostración es la misma que la realizada en el caso de dos variables. Partiendo de la definición de e_i, elevando al cuadrado y sumando, obtenemos:

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 + 2\sum \hat{y}_i e_i$$

Pero, por la propiedad 2, $\sum \hat{y}_i e_i$ es igual a cero. En consecuencia, el R^2 también será utilizado para medir la bondad de ajuste en el modelo de K variables.

Como mencionamos anteriormente, los investigadores deben hacer un uso cauteloso del coeficiente R^2 como medida de la calidad del modelo. Algunas cuestiones importantes acerca del coeficiente R^2 son las siguientes:

- El estimador de mínimos cuadrados ordinarios maximiza el R^2. Esto se puede ver en la siguiente expresión del coeficiente R^2:

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

  Notar que la suma de cuadrados totales es una magnitud que no depende del modelo elegido, sino que depende solamente de las Y_i. Pero por otro lado, la suma de cuadrados residuales depende del $\hat{\beta}$ elegido. El método de mínimos cuadrados elige un $\hat{\beta}$ de forma tal que minimice la suma de cuadrados residuales, es decir, que maximice el R^2, dado que la suma de cuadrados totales es una constante.

- El R^2 tiende a aumentar con el número de variables explicativas, es una función no decreciente de K. Es decir que si agregamos variables al modelo original, el R^2 no disminuirá, de hecho tenderá a aumentar. Hay que tener cuidado, ya que un modelo con un mayor número de variables, tendrá un R^2 elevado, o "inflado de manera espúrea", sin tener necesariamente un buen poder explicativo. Es decir, si nos guiamos sólo por el R^2 para definir la bondad de un modelo,esto nos puede llevar a descartar modelos con pocas variables en favor de uno con mayor número de variables, aunque el primero tenga un mayor poder explicativo que el segundo.

  Para solucionar este problema vamos a definir al R^2 *ajustado*, una versión del R^2 que penaliza la adición de variables que no aumenten el poder explicativo del modelo. Lo definimos como:

$$R^2 ajustado \quad = \quad 1 - \frac{SCR/(n-K)}{SCT/(n-1)}$$

siendo $(n-K)$ los grados de libertad de la SCR, y $(n-1)$ los grados de libertad de la SCT.

Vemos que, al aumentar el número de variables, K, la SCR tiende a disminuir y $(n-K)$ también tiende a disminuir, por lo tanto el efecto sobre el R^2 *ajustado* queda indeterminado.

- Es exactamente 1 cuando el número de variables explicativas es igual al número de observaciones. Este es un caso extremo de la discusión anterior, ya que si aumentamos el número de variables explicativas hasta el número de observaciones obtendremos un R^2 igual a uno.
- No puede usarse para comparar modelos con distintas variables explicadas. El R^2 nos dice cuánto de la variabilidad de una variable puede ser explicada a partir de la variabilidad de otras, en un modelo lineal. Luego, para comparar el coeficiente R^2 de distintos modelos, éstos deben tratar de explicar la misma variable explicada, y expresada en las mismas unidades. Por ejemplo, no podemos comparar el R^2 de un modelo cuya la variable explicada esté expresada en niveles con el de otro en el que dicha variable esté medida en logaritmos.

2.10. Inferencia básica en el modelo de K variables

2.10.1. Significatividad individual e hipótesis lineales simples

Consideremos el modelo lineal:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_K X_{Ki} + u_i, \qquad i = 1, \ldots, n$$

Que puede ser expresado en forma matricial como:

$$Y \quad = \quad X\beta + u$$

Consideremos las siguientes hipótesis acerca de los coeficientes β_k, k=1,...,K.

- Caso 1: Significatividad individual: $H_0 : \beta_j = 0$, esto es, bajo la hipótesis nula la j-ésima variable explicativa no es relevante para explicar Y.
- Caso 2: Valores particulares de los coeficientes: $H_0 : \beta_j = r$, esto es, bajo la hipótesis nula el coeficiente asociado a la j-ésima variable explicativa es igual a un valor particular r.
- Caso 3: Igualdad entre dos coeficientes: $H_0 : \beta_j = \beta_i$ con $i \neq j$, o sea, bajo la hipótesis nula los coeficientes asociados a la j-ésima y a la i-ésima variables explicativas son iguales entre sí.

- Caso 4: Restricciones sobre la suma o resta de coeficientes (o combinaciones lineales simples de parámetros): $H_0 : \beta_i + \beta_j = r$, es decir, bajo la hipótesis nula la suma de los coeficientes asociados a la j-ésima y a la i-ésima variables explicativas es igual a un valor particular r.

Todos los casos anteriores pueden expresarse de la siguiente forma:

$$H_0 \quad : \quad c'\beta - r = 0$$

Siendo c' un vector de dimensión 1 x K y r una constante. En detalle:

- El Caso 1 se corresponde con $c' = (0, \ldots, 1, \ldots, 0)$, donde el 1 figura en la j-ésima posición, y $r = 0$.
- El Caso 2 se corresponde con $c' = (0, \ldots, 1, \ldots, 0)$, con el 1 en la j-ésima posición, y r igual a una constante.
- El Caso 3 le corresponde con $c' = (0, \ldots, 1, \ldots, -1, \ldots, 0)$, con el 1 en la j-ésima posición, y el -1 en la i-ésima posición y $r = 0$.
- El Caso 4 le corresponde con $c' = (0, \ldots, 1, \ldots, 1, \ldots, 0)$, con un 1 en la j-ésima posición, el otro en la i-ésima posición y r igual una constante.

Existe otro tipo de hipótesis que responden a esta forma general: propondremos una estrategia para implementar los test de hipótesis.

Al igual que en el capítulo anterior, para realizar un test sobre la hipótesis nula $H_0 : c'\beta - r = 0$ contra la alternativa $H_A : c'\beta - r \neq 0$, vamos a mirar a su contraparte estimada $c'\hat{\beta} - r$ y verificar si es estadísticamente distinta de cero. Para esto, vamos a necesitar la distribución de $c'\hat{\beta} - r$. Como antes, introduciremos un supuesto adicional acerca de la distribución del término de error:

$$u_i \quad \sim \quad N(0, \sigma^2) \qquad\qquad i = 1, \ldots, n$$

Premultiplicando por c' en (2.3), obtenemos:

$$c'\hat{\beta} \quad = \quad c'\beta + c'(X'X)^{-1}X'u$$

Como $c'(X'X)^{-1}X'$ es un vector de 1 x n y $c'\beta$ es un número, $c'\hat{\beta}$ se distribuye en forma normal, ya que es una combinación lineal de los u_i's. Entonces:

$$E(c'\hat{\beta}) \quad = \quad c'\beta$$

y

$$V(c'\hat{\beta}) \quad = \quad c'V(\hat{\beta})c = \sigma^2 c'(X'X)^{-1}c$$

Entonces:

$$c'\hat{\beta} \quad \sim \quad N(c'\beta, \sigma^2 c'(X'X)^{-1}c)$$

De esto obtenemos:

$$c'\hat{\beta} - r \quad \sim \quad N(c'\beta - r, \sigma^2 c'(X'X)^{-1}c)$$

Luego, para una hipótesis $H_0 : c'\beta - r = 0$, se usa un estadístico que, bajo la hipótesis nula, se distribuye:

$$z \quad = \quad \frac{c'\hat{\beta} - r}{\sqrt{V(c'\hat{\beta} - r)}} \sim N(0,1)$$

Con $V(c'\hat{\beta} - r) = \sigma^2 c'(X'X)^{-1}c$. En la práctica, no conocemos σ^2, entonces lo vamos a reemplazar por su estimador S^2 para obtener el siguiente estadístico:

$$t \quad = \quad \frac{c'\hat{\beta} - r}{\sqrt{\hat{V}(c'\hat{\beta} - r)}} \sim t_{n-K}$$

Con $V(c'\hat{\beta} - r) = S^2 c'(X'X)^{-1}c$. Este estadístico tiene, bajo la hipótesis nula, una distribución t con $n - K$ grados de libertad.

Ahora volveremos a los casos vistos anteriormente y explicitaremos los estadísticos de prueba apropiados para cada situación, reemplazando la matriz c' y r según corresponda:

- Caso 1:

$$t \quad = \quad \frac{\hat{\beta}_j}{\sqrt{\hat{V}(\hat{\beta}_j)}}$$

  Con $\hat{V}(\hat{\beta}_j) = S^2 a_{jj}$, siendo a_{jj} el elemento (jj) sobre la diagonal principal de la matriz $(X'X)^{-1}$.

- Caso 2:

$$t \quad = \quad \frac{\hat{\beta}_j - r}{\sqrt{\hat{V}(\hat{\beta}_j)}}$$

  Con $\hat{V}(\hat{\beta}_j) = S^2 a_{jj}$, siendo a_{jj} el elemento (jj) sobre la diagonal principal de la matriz $(X'X)^{-1}$.

- Caso 3:

$$t = \frac{\hat{\beta}_j - \hat{\beta}_i}{\sqrt{\hat{V}(\hat{\beta}_j - \hat{\beta}_i)}}$$

Con $\hat{V}(\hat{\beta}_j - \hat{\beta}_i) = S^2(a_{jj} + a_{ii} - 2a_{ij})$, siendo a_{hs} el elemento (h, s) en la matriz $(X'X)^{-1}$.

- Caso 4:

$$t = \frac{\hat{\beta}_j + \hat{\beta}_i - r}{\sqrt{\hat{V}(\hat{\beta}_j + \hat{\beta}_i - r)}}$$

Con $\hat{V}(\hat{\beta}_j + \hat{\beta}_i - r) = S^2(a_{jj} + a_{ii} + 2a_{ij})$, siendo a_{hs} el elemento (h, s) en la matriz $(X'X)^{-1}$.

2.10.2. Significatividad global

Consideremos ahora la hipótesis nula:

$$H_0 \quad : \quad \beta_2 = 0, \beta_3 = 0, \ldots, \beta_K = 0$$

Contra la hipótesis alternativa:

$$H_A \quad : \quad \beta_2 \neq 0 \vee \beta_3 \neq 0 \vee \ldots \vee \beta_K \neq 0$$

Esto es, bajo la hipótesis nula todos los coeficientes, salvo el intercepto, son iguales a cero, y bajo la hipótesis alternativa al menos uno de ellos es distinto de cero. Esta es la hipótesis de *significatividad global*. Bajo la hipótesis nula, ninguna de las variables del modelo ayuda a explicar Y, y bajo la alternativa, al menos una de ellas ayuda a explicar a Y.

Se puede demostrar que bajo la hipótesis nula, el estadístico de prueba:

$$F = \frac{SCE/(K-1)}{SRC/(n-K)}$$

sigue una distribución F con $K-1$ y $n-K$ grados de libertad en el numerador y en el denominador, respectivamente. Intuitivamente, si la hipótesis nula es correcta, entonces el modelo lineal explica poco (o nada) más allá de la constante. Entonces, la SCE debe ser cercana a cero, y luego F también sera cercano a cero. Bajo la alternativa, al menos una variable ayuda a explicar Y. Por lo tanto, la SCE será mayor que en el caso anterior, y consecuentemente F también lo será. La regla de decisión es rechazar la hipótesis nula si F toma valores relativamente grandes, de acuerdo con la distribución F.

Podemos obtener una representación alternativa del estadístico F, dividiendo numerador y denominador por SCT y recordando que $R^2 = SCE/SCT = 1 - SCR/SCT$:

$$F = \frac{R^2/(K-1)}{(1-R^2)/(n-K)}$$

De una forma u otra, lo que intentamos ver con este test, es si el R^2 es estadísticamente distinto de cero.

2.10.3. Significatividad conjunta de un subgrupo de variables

A diferencia del caso anterior de significatividad global, ahora se quiere evaluar la significatividad conjunta de un subgrupo de variables explicativas. Para ello, consideremos la siguiente hipótesis nula:

$$H_0 \quad : \quad \beta_j = \beta_i = 0$$

Contra la hipótesis alternativa:

$$H_A \quad : \quad \beta_j \neq 0 \vee \beta_i \neq 0$$

Esto es, bajo la hipótesis nula la j ésima y la i ésima variables del modelo no son relevantes mientras que la hipótesis alternativa sostiene que al menos una de ellas sí lo es. En general, bajo la hipótesis nula estamos evaluando si un *subgrupo* particular de variables no son significativas. Otra vez, por la misma razón que antes, este caso no se corresponde con el análisis de los tests t ya que estamos imponiendo *dos* restricciones lineales de manera conjunta.

Para cada una de estas hipótesis, podemos estimar dos modelos, el modelo *restricto* y el *irrestricto*. Por ejemplo, para la hipótesis nula del test de significatividad global, el modelo irrestricto es el modelo lineal original, y el modelo restricto es un modelo que incluye sólo el intercepto como variable explicativa. En el segundo ejemplo, el modelo irrestricto es el modelo original, y el modelo restricto es también el original pero excluyendo la i ésima y j ésima variables explicativas. Vamos a denotar como SSR_U y SSR_R a la suma de cuadrados residuales de los modelos irrestricto y restricto, respectivamente. Es crucial notar que:

$$SSR_R \geq SSR_U$$

El estimador de mínimos cuadrados en el caso irrestricto determina $\hat{\beta}_1$, $\hat{\beta}_2$, ..., $\hat{\beta}_k$ con el fin de minimizar SSR_U. Por lo tanto, la SSR_R puede verse como el resultado de hacer lo mismo, pero forzando a alguna de las estimaciones a cumplir una restricción. Por ejemplo, en el caso de significatividad global, se obliga a todos los coeficientes estimados (excepto el intercepto) a ser cero. Por consiguiente, SSR_R es un mínimo *restricto* y, por definición, este no puede ser menor que el mínimo libre o irrestricto SSR_U. Esta lógica nos lleva al siguiente estadístico para la hipótesis nula de que la restricción es correcta:

$$F = \frac{(SSR_R - SSR_U)/J}{SSR_U/(n-K)} \tag{2.7}$$

Se puede demostrar que, bajo la hipótesis nula, este estadístico tiene una distribución F con J grados de libertad en el numerador y $n - K$ grados de libertad en el denominador. J es el número de restricciones impuestas al modelo original. En el caso de significatividad global, J=$K - 1$, y en el segundo ejemplo, J=2. Para proporcionar una idea intuitiva de cómo funciona este test, consideremos el numerador de (2.7). Hemos visto que $SSR_R - SSR_U$ es un número no negativo. Puesto que el denominador es también un número positivo (SSR_U es la suma de números al cuadrado y $n - K$ es positivo), F es un número no negativo. Recordemos que hemos introducido la función SSR como una función de penalidad, que mide la incapacidad del modelo lineal para explicar Y, la cual es minimizada por el método de mínimos cuadrados. Intuitivamente, el procedimiento de estimación irrestricto tiende a producir estimaciones que son "cercanas" a los verdaderos parámetros, mientras que el procedimiento restricto fuerza a las estimaciones a cumplir la restricción impuesta por la hipótesis nula. Por ejemplo, en el caso de significatividad global, esta establece que todos los coeficientes son iguales a cero. Trivialmente, si la restricción de la hipótesis nula es verdadera, ambos procedimientos generan casi los mismos errores, haciendo F cercano a cero.

Por otro lado, si la hipótesis nula no es correcta, el procedimiento irrestricto todavía sigue haciendo que las estimaciones tiendan hacia los verdaderos valores, pero el procedimiento restricto "insiste" en fijar las estimaciones en los valores incorrectos de la hipótesis nula. Entonces, la discrepancia entre SSR_R y SSR_U debe ser grande. Esto significa que la penalidad captada por SSR es significativamente mayor usando el modelo restricto. Por consiguiente, la lógica implica rechazar H_0 cuando F es "grande" según la distribución F bajo la hipótesis nula.

Capítulo 3

Usos y Extensiones del Modelo Lineal con Varias Variables

3.1. El modelo lineal en parámetros

Consideremos una versión simple del modelo lineal tratado hasta el momento:

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

Este modelo es lineal en el sentido de que la variable explicada está expresada como una combinación lineal de las variables explicativas. En este caso decimos que el modelo es *lineal en variables*. Entonces, es natural preguntarnos cómo proceder cuando tenemos interés en relaciones no lineales. Resulta que el modelo lineal es mucho menos restrictivo que lo que una mirada superficial podría sugerirnos. Notemos que si invertimos los roles y tratamos X_{2i} y X_{3i} como parámetros, y β_1, β_2, β_3 como variables, entonces la relación entre Y_i y estas últimas es también lineal. Desde esta perspectiva, el modelo lineal se dice que es *lineal en parámetros*. En esta sección vamos a mostrar que el uso del modelo lineal junto con el método de mínimos cuadrados se puede aplicar a cualquier modelo que es lineal en parámetros, y no necesariamente lineal en variables.

Tomemos un ejemplo sencillo. Supongamos que estamos interesados en la siguiente ecuación de demanda:

$$Q_i = AP_i^{\beta} e^{u_i} \tag{3.1}$$

donde Q es la cantidad demandada, P es el precio, u es un término aleatorio, y A y β son parámetros desconocidos a estimar. Este modelo es claramente no lineal en variables. Tomemos logaritmo natural a ambos lados para obtener:

$$\ln Q_i = \ln A + \beta \ln P_i + u_i$$

que puede escribirse como:

$$q_i = \alpha + \beta p_i + u_i$$

con $q_i = \ln Q_i$, $p_i = \ln P_i$, y $\alpha = \ln A$. Esta última versión muestra que si usamos q como variable explicada y p como variable explicativa, el modelo tiene la estructura del modelo lineal discutido en el capítulo anterior, y entonces el método de mínimos cuadrados puede ser utilizado para estimar α y β. Hemos transformado el modelo original no lineal en variables (3.1) en un modelo que es lineal en parámetros.

En términos generales, consideremos el siguiente modelo potencialmente no lineal:

$$Y_i = f(X_{1i}, X_{2i}, ..., X_{Ki}, u_i)$$

El modelo es lineal en parámetros si hay funciones $g_1, g_2, ..., g_K$ tal que:

$$g_1(Y_i) = \beta_1 + \beta_2 g_2(X_{1i}, X_{2i}, ..., X_{Ki}) + \cdots + \beta_K g_k(X_{1i}, X_{2i}, ..., X_{Ki}) + u_i$$

y entonces, el modelo puede expresarse como:

$$Y_i^* = \beta_1 + \beta_2 {X_{2i}}^* + \cdots + \beta_K {X_{Ki}}^* + u_i$$

Intuitivamente, cualquier modelo no lineal puede expresarse en términos del modelo lineal de los capítulos anteriores, siempre que podamos escribirlo como un modelo lineal en parámetros a través de una transformación.

A continuación presentamos un catálogo de algunas transformaciones y especificaciones comúnmente usadas:

1. *Logarítmico*: $Y_i = AX_i^{\beta} e^{u_i}$. Esta es la forma utilizada para el ejemplo de demanda (3.1). Tomando logaritmos a ambos lados obtenemos:

   $$y_i = \alpha + \beta x_i + u_i$$

   donde $y_i = \ln Y_i$, $x_i = \ln X_i$ y $\alpha = \ln A$. Es importante interpretar los parámetros correctamente. Notemos que:

   $$\beta = \frac{dy}{dx} = \frac{dlnY}{dlnX} = \epsilon_{YX}$$

   Esto es, β tiene la interpretación de una *elasticidad*, mide en qué proporción cambia Y ante cambios en un uno porciento en X. Esto es consistente con el modelo original y su transformación.

2. *Semilogarítmico*: $Y_i = \exp(\alpha + \beta X_i + u_i)$. Tomando logaritmos a ambos lados:

$$y_i = \alpha + \beta X_i + u_i$$

Notar que:

$$\beta = \frac{dy}{dX} = \frac{dlny}{dX}$$

Entonces, β se interpreta como la *semielasticidad* de Y con respecto a X, esto es, en qué proporción cambia Y ante cambios de una unidad en X. Esta especificación es comúnmente usada en economía laboral para modelar la relación entre salarios (Y) y años de educación (X). En ese caso, β mide en qué porcentaje se incrementan los salarios como consecuencia de obtener un año adicional de educación, un número usualmente llamado *retorno a la educación*.

3. *Recíproco*: $Y_i = \beta_1 + \beta_2(1/X_i) + u_i$. la relación entre Y y X está dada por una hipérbola. Puede ser fácilmente expresada como un modelo lineal en parámetros de la siguiente manera:

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

con $X_i^* = 1/X_i$.

4. *Cuadrático*: $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i$. Aunque trivial, puede expresarse como un modelo lineal en parámetros de la siguiente manera:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + u_i \tag{3.2}$$

con $Z_i = X_i^2$, esto es, la relación cuadrática entre Y y X se modela como una relación lineal entre Y, X y Z. Es importante tener una interpretación consistente del modelo. Notar que:

$$\frac{dY}{dX} = \beta_2 + 2\beta_3 X_i$$

Entonces, β_2 ya no puede ser interpretado como la derivada de Y con respecto a X, ya que hemos agregado un término cuadrático. Este es el "precio" pagado por usar un modelo no lineal "más rico": las derivadas ya no son constantes y, en lugar de ello, dependen de los valores que tome X. Por ejemplo, supongamos que Y es el salario horario y X mide la edad en años. Algunos investigadores sostienen que la relación entre salarios y edad tiene la forma de U invertida, en el sentido que la edad incrementa los salarios hasta cierto punto en el cual los salarios empiezan a decrecer. (3.2) proporciona un modelo para esta situación donde, naturalmente, el efecto de la edad en los salarios varía de acuerdo a la edad del individuo, primero aumentándolos y luego decreciendo a mayor edad.

5. *Interacción*: $Y_i = \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 X_{1i} X_{2i} + u_i$. Este puede expresarse fácilmente como un modelo lineal en parámetros llamando $Z_i = X_{1i} X_{2i}$. Notar que:

$$\frac{dY}{dX_1} = \beta_2 + \beta_4 X_{2i}$$

Entonces, el efecto de X_1 sobre Y depende de X_2, por lo tanto X_1 interactúa con X_2. De nuevo, es importante no interpretar β_2 como la derivada de Y con respecto a X_1. Por ejemplo, consideremos otra vez el modelo de salarios donde Y es el salario mensual, X_1 es educación y X_2 es una medida de inteligencia. Algunos investigadores sostienen que la inteligencia y la educación interactúan, por lo que el efecto de una unidad adicional de educación sobre los salarios es mayor para personas más inteligentes. La variable $X_1 X_2$ estaría capturando este efecto.

Desafortunadamente, no siempre es posible encontrar una transformación para un modelo no lineal en variables, de manera de poder expresarlo en la forma lineal en parámetros. Por ejemplo, el modelo:

$$Y_i = \alpha + \beta_0 {X_{1i}}^{\beta_1} + u_i$$

no puede ser transformado en la forma lineal en parámetros.

3.2. Variables dummy

Supongamos que queremos estudiar la relación entre salarios y género. Para este propósito tenemos una muestra de hombres y mujeres con sus salarios y experiencia, medida en años desde que entraron al mercado laboral. Idealmente, nos gustaría que todos los hombres y mujeres tengan la misma experiencia. En ese caso, podríamos comparar el salario promedio para hombres y para mujeres. Pero si, por ejemplo, los hombres tienen más experiencia que las mujeres, entonces la diferencia de salarios promedio podría estar reflejando diferencias en la experiencia en lugar de una cuestión de género. El modelo de regresión con varias variables proporciona una estrategia elegante para aislar, en este caso, el efecto del género del de la experiencia.

Consideremos el siguiente modelo para salarios:

$$W_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i \tag{3.3}$$

donde para el i ésimo individuo, W_i es el salario medio en dólares, y u_i es el término de error que cumple todos los supuestos clásicos. D_i es una variable que toma el valor 1 si el i ésimo individuo es hombre y 0 si es mujer. D_i es un indicador o variable *dummy*, que indica si el individuo es hombre o mujer. Al igual que en los capítulos anteriores, β_2 puede interpretarse como una derivada

parcial, indicando el cambio esperado en los salarios ante un cambio marginal en la experiencia, manteniendo fijos los restantes factores distintos de la experiencia. Para el caso del género, no podemos hacer el experimento de cambiar marginalmente D_i, ya que esta variable toma solamente dos valores. Por lo tanto, para lograr una interpretación coherente de β_3, vamos a computar el salario esperado para un hombre. En este caso $D_i = 1$, entonces:

$$E(W_i/D_i = 1) = \beta_1 + \beta_2 X_i + \beta_3 \tag{3.4}$$

Ya que para mujeres $D_i = 0$, el salario esperado para mujeres es:

$$E(W_i/D_i = 0) = \beta_1 + \beta_2 X_i \tag{3.5}$$

Entonces, la diferencia entre el salario esperado para un hombre y para una mujer es:

$$E(W_i/D_i = 1) - E(W_i/D_i = 0) = \beta_3$$

Esto proporciona una interpretación natural de β_3 como la diferencia en los salarios esperados entre hombres y mujeres, y también proporciona una solución a nuestro problema de aislar el efecto del género sobre los salarios del efecto de la experiencia. Por lo tanto, β_3 mide el efecto diferencial en salarios que puede atribuirse exclusivamente a diferencias de género. Una intuición gráfica ayuda a entender este punto. Las ecuaciones (3.4) y (3.5) pueden verse como modelos de salarios diferentes para hombres y mujeres. A los fines de interpretar lo anterior, asumamos que β_3 es positivo. Implícitamente, esto dice que el modelo de salarios para las mujeres es exactamente el de los hombres, con un intercepto más bajo.

La estimación de la ecuación (3.3) es sencilla, ya que el modelo es solamente un caso especial del modelo lineal tratado en el capítulo anterior. El único punto particular es la introducción de la variable explicativa D_i, que toma valores 0 ó 1. Notar que, además del supuesto de no multicolinealidad, no hemos impuesto ninguna otra restricción en los valores que pueden tomar las variables explicativas. Así, usando una variable explicativa dummy no hemos alterado de ninguna manera la estructura estadística del modelo lineal con varias variables, por lo que los procedimientos de estimación e inferencia son exactamente igual que antes. El punto clave es cómo interpretar el coeficiente de tales variables, que es lo que hemos enfatizado anteriormente. Es interesante notar que la hipótesis de que no hay efectos de género sobre salarios puede implementarse fácilmente con un test "t" estándar siendo $H_0 : \beta_3 = 0$ en (3.3).

En un marco más general, (3.3) puede ser visto como un modelo donde la variable dummy indica si el individuo pertenece a una determinada clase o no, y el coeficiente que acompaña tal variable es interpretado como la diferencia en el valor esperado de la variable explicada entre las observaciones que pertenecen a una determinada clase con respecto a aquellas que no pertenecen, manteniendo constantes los demás factores.

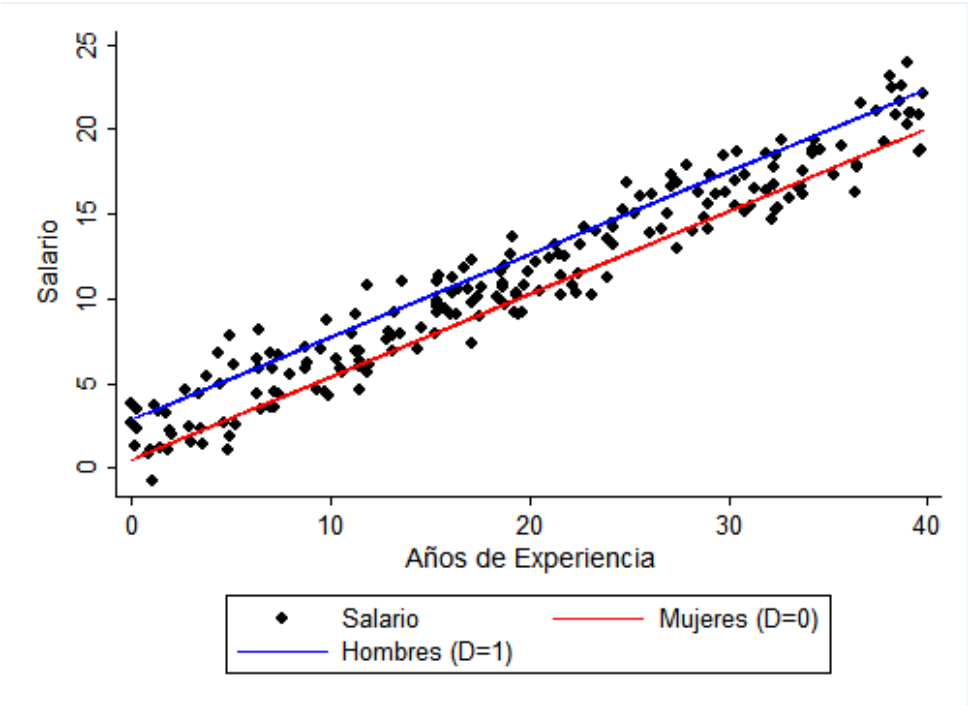


Figura 3.1: Dummy aditiva.

Hay algunos comentarios importantes respecto al uso de variables dummy en esta configuración:

1. Notar que, para distinguir entre individuos que pertenecen a una clase o no (hombres o no hombres) hemos usado solo una variable dummy. Trivialmente, en estos problemas si un individuo pertenece a una cierta clase, automáticamente no pertenece a la otra. Es decir, la variable dummy indica pertenencia a categorías mutuamente excluyentes. Pensemos que pasaría si intentáramos estimar el siguiente modelo:

$$W_i = \beta_1 + \beta_2 X_i + \beta_3 D_{1i} + \beta_4 D_{2i} + u_i$$

   donde D_{1i} es 1 si el individuo es hombre y 0 si el individuo es mujer, y $D_{2i} = 1$ si el individuo es mujer y 0 en caso contrario. En este caso, es sencillo notar que estamos violando el supuesto de no multicolinealidad. Notar que $D_{1i} + D_{2i} = 1$ para todo i, de esta manera, la primer variable explicativa (el número uno acompañando al intercepto), puede obtenerse como la suma de otras dos variables del modelo. Esto determina que necesitamos sólo una variable dummy para distinguir si los individuos pertenecen o no a una categoría. Este es un caso particular de la *trampa de la variable binaria*, que será tratado posteriormente.

2. La interpretación del coeficiente de la variable dummy depende crucialmente de cómo la definamos. Consideremos la siguiente modificación a (3.3):

$$W_i = \beta_1 + \beta_2 X_i + {\beta_3}^* {D_i}^* + u_i$$

   Si en (3.3) definiéramos ${D_i}^* = 1$ si el individuo es mujer y cero de no ser así, entonces ${\beta_3}^*$ mide el efecto de ser mujer sobre los salarios, comparado con el hecho de ser hombre. Sería fácil mostrar que ${\beta_3}^* = -\beta_3$. Esto es, que si en el modelo original $\beta_3 > 0$, lo que se interpreta como que los salarios de los hombres son más altos que los de las mujeres, entonces ${\beta_3}^*$ nos da exactamente la misma información respecto a la diferencia de salarios entre hombres y mujeres. Esto último quiere decir que uno debería elegir libremente qué categoría es denotada con 1 ó 0, mientras que la interpretación sea consistente con esa definición.

3. *Variable dependiente en logaritmos*. Supongamos que ahora el modelo es:

$$w_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i$$

   donde $w_i = \ln W_i$. Vamos a llamar ${w_i}^M$ al logaritmo del salario para hombres, que es ${w_i}^M = \beta_1 + \beta_2 X_i + \beta_3 D_i + u_i$, y ${w_i}^F$ al logaritmo del salario para mujeres, esto es ${w_i}^F = \beta_1 + \beta_2 X_i + u_i$. Entonces:

$$\hat{w}_i{}^M - \hat{w}_i{}^F = \ln \hat{W}_i{}^M - \ln \hat{W}_i{}^F = \ln \frac{\hat{W}_i{}^M}{\hat{W}_i{}^F} = \hat{\beta}_3$$

por consiguiente:

$$\hat{W}_i^M / \hat{W}_i^F - 1 = e^{\hat{\beta}_3} - 1$$

Entonces, $e^{\hat{\beta}_3} - 1$ se interpreta como la proporción en la que el salario de los hombres es mayor al de las mujeres. Por ejemplo, si $e^{\hat{\beta}_3} - 1 = 0{,}12$, significa que los hombres ganan 12 por ciento más que las mujeres. Recordemos de un curso de cálculo básico que para $\hat{\beta}_3$ pequeños:

$$e^{\hat{\beta}_3} - 1 \simeq \hat{\beta}_3$$

entonces, cuando la variable explicada está en logaritmos y $\hat{\beta}_3$ es pequeño, podemos interpretar $\hat{\beta}_3$ *directamente* como la diferencia proporcional entre categorías. El lector recordará que, si $\hat{\beta}_3$ no es suficientemente pequeño, la diferencia entre $e^{\hat{\beta}_3} - 1$ y $\hat{\beta}_3$ puede ser considerable.

Otros usos de las variables dummy

Las variables dummy pueden usarse de una manera más sofisticada que en el apartado anterior. Esta sección trata algunos casos útiles y comúnmente utilizados.

- *Variables dummy de pendiente*: En el caso original de variables dummy, hemos dejado que intercepto difiera entre categorías (hombres y mujeres), pero hemos mantenido la misma pendiente para ambas categorías. En nuestro ejemplo, esto significa que el efecto sobre los salarios de la experiencia adicional es el mismo para ambos géneros. Consideremos ahora el siguiente modelo:

$$W_i = \beta_1 + \beta_2 X_i + \beta_3 (D_i X_i) + u_i$$

siendo $D_i = 1$ para hombres y 0 en caso contrario. En este caso, el modelo subyacente para hombres es:

$$W_i = \beta_1 + (\beta_2 + \beta_3) X_i + u_i \tag{3.6}$$

y el modelo para mujeres es:

$$W_i = \beta_1 + \beta_2 X_i + u_i \tag{3.7}$$

Para darle una interpretación a β_2, tomemos la derivada de W_i con respecto a X_i en (3.6) y (3.7) para obtener:

$$\frac{\partial W_i(hombre)}{\partial X_i} = \beta_2 + \beta_3$$

$$\frac{\partial W_i(mujer)}{\partial X_i} = \beta_2$$

entonces:

$$\beta_3 = \frac{\partial W_i(hombre)}{\partial X_i} - \frac{\partial W_i(mujer)}{\partial X_i}$$

Entonces, en este modelo β_3 mide la diferencia en la pendiente entre el modelo de hombres y mujeres. Si β_3 es positivo, significa que la experiencia tiene un efecto mayor incrementando los salarios de los hombres que los de las mujeres. Como antes, un test simple de que ambas pendientes son iguales puede realizarse a partir de $H_0 : \beta_3 = 0$ en (3.6).

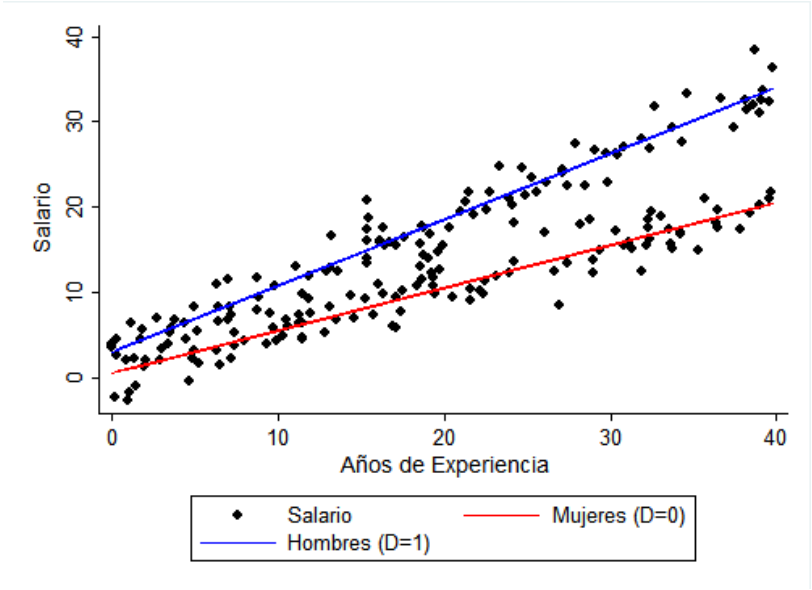


Figura 3.2: Dummy multiplicativa.

Es natural proponer el siguiente modelo, que permite que tanto el intercepto como la pendiente difieran entre categorías:

$$W_i = \beta_1 + \beta_2 X_i + \beta_3 D_i + \beta_4 (D_i X_i) + u_i$$

En este caso, si $\beta_3 = 0$ y $\beta_4 = 0$, entonces el modelo para hombres y mujeres coinciden. Esto equivaldría a una restricción lineal conjunta sobre los coeficientes. De este modo, se puede implementar un test simple con un test F, siendo la hipótesis nula $H_0 : \beta_3 = \beta_4 = 0$.

- *Más de dos categorías*: Supongamos que, además de experiencia, observamos alguna medida de educación. Por ejemplo, vamos a asumir que para cada individuo conocemos si él o

ella pertenecen a alguno de los siguientes grupos de nivel educativo: secundario incompleto, secundario completo o universitario incompleto, y universitario completo. Obviamente, los individuos pertenecen a sólo una de las tres categorías. Consideremos el siguiente modelo:

$$W_i = \beta_1 + \beta_2 X_i + \beta_3 D_{1i} + \beta_4 D_{2i} + u_i \tag{3.8}$$

donde

$$D_{1i} = \begin{cases} 1 & \text{Máxima educación de } i \text{ es secundario completo} \\ 0 & \text{Máxima educación de } i \text{ es secundario incompleto} \end{cases}$$

$$D_{2i} = \begin{cases} 1 & \text{Máxima educación de } i \text{ es universitario completo} \\ 0 & \text{Máxima educación de } i \text{ es universitario incompleto} \end{cases}$$

Podemos verificar fácilmente que un modelo como (3.8) nos proporciona toda la información que necesitamos. Vamos a computar el salario esperado para las tres categorías de educación. Notemos que las dos variables dummy son suficientes para recuperar esta información. Los individuos con secundario incompleto tienen $D_{1i} = 0$ y $D_{2i} = 0$, aquellos con secundario completo o universitario incompleto tienen $D_{1i} = 1$ y $D_{2i} = 0$, y aquellos con universitario completo tienen $D_{1i} = 0$ y $D_{2i} = 1$. Entonces, los salarios esperados para cada categoría son:

$$E(W_i/\text{secundario incompleto}) = \beta_1 + \beta_2 X_i$$
$$E(W_i/\text{secundario completo}) = \beta_1 + \beta_3 + \beta_2 X_i$$
$$E(W_i/\text{universitario completo}) = \beta_1 + \beta_4 + \beta_2 X_i$$

Esto provee una interpretación natural para β_3 y β_4. β_3 mide el impacto en los salarios esperados de terminar la escuela secundaria comparado con no terminarla. β_4 mide el efecto de completar la universidad con respecto a tener secundario incompleto. En general, cada uno de los coeficientes de las variables dummy mide el efecto de pertenecer a la categoría indicada por la variable dummy con respecto a la *categoría base* , en este caso, secundario incompleto. Notemos que el efecto de completar el nivel universitario con respecto a la categoría anterior, secundario completo o universitario incompleto, está dado por $\beta_4 - \beta_3$.

Varias hipótesis interesantes pueden ser testeadas en este marco. Por ejemplo, la hipótesis nula de que la educación no tiene efectos en los salarios corresponde a $H_0 : \beta_3 = \beta_4 = 0$. De nuevo, se puede implementar una restricción lineal conjunta a partir de un test F como los tratados en las secciones anteriores. Otra hipótesis nula interesante es que terminar la escuela secundaria tiene el mismo efecto sobre los salarios que terminar la universidad. Esto corresponde a $H_0 : \beta_3 = \beta_4$.

- *Cambios estructurales*: Consideremos el siguiente modelo simple para observaciones de series de tiempo:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 D_t + \beta_4 D_t X_t + u_t, \qquad i = 1, \ldots, T$$

donde t indica tiempo y los datos son observados desde el período 1 hasta el T. Sea t^* cualquier período entre t y T y definamos D_t como una variable dummy que toma los siguientes valores:

$$D_t = \begin{cases} 0 & \text{si } t < t^* \\ 1 & \text{si } t \geq t^* \end{cases}$$

Esto significa que para todas las observaciones anteriores a t^* el modelo relevante es:

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

mientras que para períodos a partir de t^* el modelo es:

$$Y_t = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) X_t + u_t$$

Entonces, de acuerdo a esta especificación, si β_3 y β_4 son diferentes a cero, en el periodo t^* hay un *cambio estructural*, esto es, la estructura del modelo cambia a partir de ese momento. Hay varios casos particulares interesantes. El caso en el cual no hay cambio estructural corresponde a $\beta_3 = \beta_4 = 0$, y se puede testear fácilmente usando un test F. Cuando $\beta_3 \neq 0$ y $\beta_4 = 0$ sólo el intercepto cambia en t^*, y cuando $\beta_3 = 0$ y $\beta_4 \neq 0$ solo la pendiente cambia en t^*.

- *Efectos estacionales*: Consideremos el siguiente caso. Estamos interesados en un modelo simple de series de tiempo $Y_t = \beta_1 + \beta_2 X_t + u_t$ donde Y son las ventas de helado y X es el precio. Tenemos datos trimestrales desde el primer trimestre de 1970 hasta el último trimestre de 1999. Si tenemos $t = 1, 2, ..., T$ observaciones, la primera corresponde al primer trimestre de 1970, la segunda al segundo trimestre de 1970, la quinta al primer trimestre de 1971, etc. Las ventas de helados están sujetas a fuertes efectos estacionales: independientemente del precio, las ventas aumentan en verano y disminuyen en invierno. Un modelo que se adapta a esta posibilidad es el siguiente:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 D_{2t} + \beta_4 D_{3t} + \beta_5 D_{4t} + u_t$$

donde

$$D_{2t} = \begin{cases} 1 & \text{Si la observación corresponde al segundo trimestre} \\ 0 & \text{En caso contrario} \end{cases}$$

$$D_{3t} = \begin{cases} 1 & \text{Si la observación corresponde al tercer trimestre} \\ 0 & \text{En caso contrario} \end{cases}$$

$$D_{4t} = \begin{cases} 1 & \text{Si la observación corresponde al cuarto trimestre} \\ 0 & \text{En caso contrario} \end{cases}$$

Esta especificación le permite al modelo tener un intercepto diferente para cada trimestre. De acuerdo a la discusión anterior, β_3, β_4 y β_5 se interpretan como la diferencia en el intercepto entre cada trimestre con respecto al primer trimestre (la categoría base). La hipótesis nula de que no hay efectos estacionales se puede evaluar a partir de un test F, siendo $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$.

3.3. Multicolinealidad y micronumerosidad

El supuesto de *no multicolinealidad* requiere que todas las variables del modelo sean linealmente independientes, esto es, que la matriz de observaciones X de las K variables explicativas del modelo esté formada por K vectores columna linealmente independientes o, lo que es exactamente lo mismo, que el rango de X, $\rho(X)$, sea igual a K. Este supuesto juega un rol crucial cuando derivamos los estimadores de mínimos cuadrados. En realidad, la existencia de una única solución al problema de minimizar la suma de cuadrados residuales depende directamente del supuesto de no multicolinealidad. Consecuentemente, relajar este supuesto, es decir, permitir que las variables explicativas sean linealmente dependientes, tiene consecuencias dramáticas, hasta el punto de no ser posible obtener una solución para el problema de mínimos cuadrados.

Un tema completamente diferente se refiere a lo que sucede si la relación lineal entre las variables explicativas es "muy cercana" a ser perfecta. Desde el punto de vista del Teorema de Gauss-Markov, ya que el modelo no impide "muy alta" colinealidad entre las variables explicativas, y por lo tanto no se viola ninguno de los supuestos clásicos, concluimos que "multicolinealidad alta" no tiene efectos en las conclusiones del teorema: el estimador de mínimos cuadrados sigue siendo el mejor estimador lineal e insesgado. Entonces, ¿por qué nos preocupamos por la alta multicolinealidad? El Teorema de Gauss-Markov dice que el estimador de MCO es el mejor dentro de una cierta clase de estimadores, pero no nos dice si el estimador es "bueno" o "malo". El lector recordará la discusión previa sobre las nociones relativas u ordinales, y las absolutas o cardinales. En el caso de

"alta multicolinealidad", las estimaciones de mínimos cuadrados ordinarios son malas a pesar de ser las mejores entre las lineales e insesgadas. Para distinguir explícitamente entre multicolinealidad alta y perfecta, usaremos las siguientes definiciones:

- Hay *multicolinealidad perfecta* o *exacta* cuando $\rho < K$, esto es, cuando se puede obtener al menos una de las variables explicativas como una combinación lineal exacta de las otras.
- Hay *multicolinealidad alta* cuando $\rho = K$ pero la correlación entre al menos dos variables explicativas es muy alta.

Para dar un ejemplo, consideremos el siguiente caso de una función simple de consumo:

$$C_i = \beta_1 + \beta_2 Y_i + \beta_3 W_i + u_i$$

donde C_i es consumo, Y_i es ingreso y W_i es riqueza. Llamemos $\tau_{Y,W}$ al coeficiente de correlación entre Y y W. En este caso, habrá multicolinealidad perfecta si $|\tau_{Y,W}| = 1$, y multicolinealidad alta cuando $|\tau_{Y,W}|$ es cercano a uno. Obviamente, la noción de *multicolinealidad alta* es una cuestión de grado, no existe un umbral más allá del cual hay y debajo del cual no hay multicolinealidad alta. Por lo tanto la pregunta relevante es, en lugar de eso, qué sucede con las estimaciones de MCO cuando la colinealidad entre variables es alta en vez de baja.

Se puede demostrar que, cuando hay multicolinealidad alta:

1. $\hat{\beta}$ sigue siendo MELI, ya que no se relajan ninguno de los supuestos clásicos y, por lo tanto, el Teorema de Gauss-Markov sigue valiendo.
2. La varianza de $\hat{\beta}_k, k = 1, ..., K$ es muy alta.
3. Puede pasar que los estadísticos "t" para la hipótesis nula de no significatividad ($H_0 : \beta_k = 0$) sean todos muy bajos pero que el R^2 sea alto. En consecuencia, el test F de significatividad global podría llevarnos a rechazar la hipótesis nula de que ninguna variable es significativa, mientras que todos los test "t" de significatividad individual nos llevarían a aceptar la hipótesis nula. Más explícitamente, todas las variables parecen irrelevantes individualmente, mientras que conjuntamente son relevantes para el modelo.
4. Las estimaciones son muy inestables ante modificaciones menores del modelo (descartar observaciones, realizar cambios menores en el modelo)

Aunque no sea difícil explorar formalmente estos problemas, eso se encuentra más allá del tratamiento de estas notas. Una intuición gráfica ayudará considerablemente.La figura 3.3 muestra un diagrama de puntos de datos hipotéticos de Y, X y Z, donde las últimas se usan como variables explicativas. Como se puede ver, hay colinealidad alta entre X y Z. Esto es, las coordenadas de los

puntos (X, Z) caen muy cerca de una línea recta. En este caso, el método de mínimos cuadrados intenta pasar un plano a través de los puntos para minimizar la suma de los errores al cuadrado. El gráfico muestra el "plano" de ajuste. Como se puede ver, cualquier rotación menor del plano produce errores muy similares. Esto puede interpretarse de la siguiente manera: el procedimiento tiene problemas en distinguir entre planos rotados (los $\hat{\beta}$ son muy inestables y tienen grandes varianzas) aunque cualquiera de los planos posibles ajusta muy bien a los puntos (el R^2 es alto).

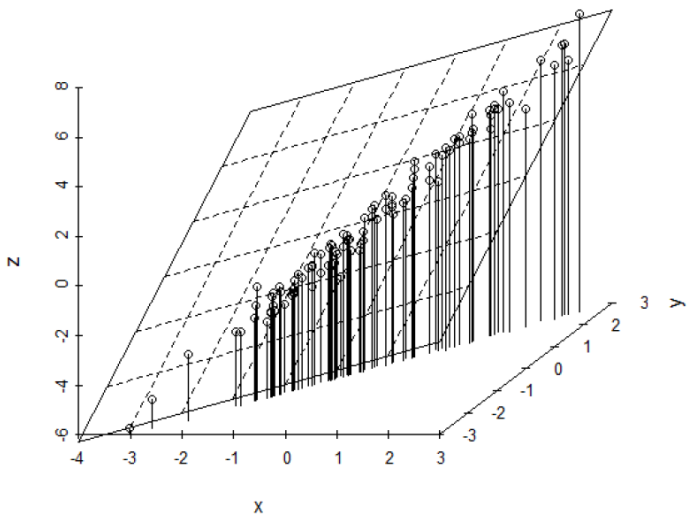


Figura 3.3: Multicolinealidad alta.

Intuitivamente, lo que sucede ante casos de multicolinealidad muy alta es que, dado que las variables están muy estrechamente relacionadas, el procedimiento tiene problemas en separar los efectos de cada variable, aunque puede producir un modelo razonable para el efecto conjunto. En la jerga, y por razones que serán explicitadas luego, decimos que los verdaderos parámetros del modelo son "difíciles de identificar" usando el método de mínimos cuadrados.

Volviendo a la pregunta original (¿por qué nos preocupamos por la multicolinealidad alta?), el problema es, esencialmente, la varianza grande de los estimadores. En un tratamiento muy lúcido e irónico, Goldberger (1991) dice que si éste es el caso, entonces, para ser justos, los investigadores que se preocupan por la multicolinealidad alta deberían preocuparse también por cualquier otra causa que incremente la varianza. Por ejemplo, cuando tenemos un número bajo de observaciones. El supuesto $\rho(X) = K$ requiere que el número de observaciones sea al menos K. Consideremos qué ocurre cuando el número disponible de observaciones es "bajo". Ya que el teorema de Gauss-Markov no requiere que el número de observaciones sea "alto", no se viola ningún supuesto y por lo tanto, el estimador de MCO sigue siendo MELI. En el caso extremo, hemos visto que cuando $n = K$ todos los errores son cero, el modelo ajusta perfectamente a los datos y el R^2 es igual a 1. El

estimador insesgado para la varianza de $\hat{\beta}$ es, como hemos visto:

$$\hat{V}(\hat{\beta}) = S^2(X'X)^{-1} \tag{3.9}$$

con $S^2 = \Sigma e_i^2/(n-K)$. Cuando hay pocas observaciones $n-K$ se hace muy pequeño y la varianza explota, lo que provoca que los estadísticos "t" de significatividad individual sean muy pequeños. Entonces, cuando el número de observaciones es realmente pequeño podemos tener un R^2 alto, varianza de los $\hat{\beta}$ grandes y estadísticos "t" pequeños, es decir, las consecuencias de tener muy pocas observaciones son las mismas que las de tener multicolinealidad. Goldberger dice que, por lo que acabamos de discutir, aquellos investigadores que se preocupan por la alta multicolinealidad deberían prestar igual atención al problema de *micronumerosidad* (bajo número de observaciones).

De (3.9) vemos que hay tres factores que conducen a los estimadores de mínimos cuadrados a tener varianzas muy grandes: 1) pocas observaciones, 2) errores grandes, 3) correlación alta entre variables. Esto sugiere varias acciones a tomar cuando los investigadores se preocupan por las varianzas grandes y, por consiguiente, por la multicolinealidad:

1. *Aumentar el número de observaciones*: Esto remedia los problemas de micronumerosidad. De este modo, el punto es "compensar" las varianzas altas inducidas por la multicolinealidad reduciendo las varianzas vía menor micronumerosidad. En la mayoría de los casos prácticos esta recomendación es trivial (si tuvieramos más observaciones, qué estaríamos esperando para usarlas!) o prácticamente irrelevante (obtener más observaciones puede ser costoso o directamente imposible).

2. *Descartar variables*: una práctica común es descartar una de las variables que esté linealmente relacionada con otra variable del modelo. Por ejemplo, en nuestro caso del consumo, si el ingreso presenta una relación lineal alta con la riqueza, entonces se descartaría una de ellas y se mantendría la otra. Veremos en el siguiente capítulo que si el modelo está correctamente especificado (esto es, si el consumo verdaderamente depende del ingreso y la riqueza), entonces descartar variables puede crear problemas más serios que el que se intenta solucionar. Por otro lado, el modelo podría estar mal especificado (por ejemplo, la riqueza no es un factor explicativo del consumo), en ese caso descartar dicha variable funcionaría. El punto crucial es que, bajo multicolinealidad alta, el modelo lineal estimado por MCO nos proporciona muy poca información respecto a qué variable descartar.

3. *Revisar el modelo*: Podría ser que el modelo esté mal especificado. Esta situación es cercana a la anterior, donde intentamos descartar variables. Quizás una re examinación de la relación teórica entre las variables de interés resulte en un modelo que tenga menor multicolinealidad ex-ante. De nuevo, esto no tiene nada que ver con la especificación estadística del modelo.

4. *Cambiar el método de estimación*: La multicolinealidad alta es un problema del modelo lineal estimado de una forma específica, esto es, usando mínimos cuadrados. Por consiguiente, algunos autores sugieren que una posibilidad es buscar métodos alternativos de estimación que se vean menos afectados por la multicolinealidad. El teorema de Gauss-Markov nos proporciona alguna información: otra alternativa lineal e insesgada debe ser peor que mínimos cuadrados. Existe una amplia literatura de métodos alternativos de estimación al de mínimos cuadrados. Judge et. al (1988) ofrece un tratamiento detallado. Por ejemplo, la regresión *ridge* provee una alternativa que, al precio de un pequeño sesgo, puede potencialmente producir una gran reducción en la varianza.

5. *No hacer nada*: Si el modelo está correctamente especificado y toda la información ha sido usada, una varianza alta es más una característica del grado de dificultad del problema enfrentado que una "patología" o defecto del modelo. Entonces, algunos autores sugieren que, ya que la multicolinealidad alta no viola ninguno de los supuestos clásicos, nada realmente innovador se puede hacer aparte de reconocer explícitamente las dificultades inherentes a obtener estimaciones precisas en presencia de este problema.

Apéndice

Elasticidades y derivadas logarítmicas

Considere una simple función diferenciable $y = f(x)$. La *elasticidad* de y con respecto a x, ϵ_{yx} se define como:

$$\epsilon_{yx} = \frac{dy}{dx}\frac{x}{y}$$

y es interpretada como el cambio porcentual en y como consecuencia de un cambio de x en un 1 por ciento.

Resultado:

$$\epsilon_{yx} = \frac{d\ln y}{d\ln x}$$

Esto es, la elasticidad es igual a una derivada de logaritmos. Para probar este resultado, considere el siguiente resultado general:

$$df(z) = f'(z)dz$$

En el caso de que $f(z) = \ln z$ se tiene que $f'(z) = \frac{1}{z}$.
Por lo tanto,

$$d\ln z = \frac{1}{z}dz$$

Entonces, el cociente entre $d\ln y$ y $d\ln y$ se puede expresar de la siguiente forma:

$$\begin{aligned}\frac{d\ln y}{d\ln x} &= \frac{\frac{1}{y}dy}{\frac{1}{x}dx} \\ &= \frac{dy}{dx}\frac{x}{y} \\ &= \epsilon_{yx}\end{aligned}$$

Capítulo 4

Modelo de Mínimos Cuadrados Generalizados

4.1. El Modelo Lineal Generalizado

4.1.1. Relajación de los supuestos relacionados con la varianza

Recordemos la estructura del modelo lineal general en forma matricial estudiado en capítulos anteriores:

$$Y = X\beta + u$$

- $E(u) = 0$ (exogeneidad)
- $V(u) = \sigma^2 I$ (homocedasticidad y no correlación serial)
- X es una matriz no estocástica con rango completo ($\rho(x) = K$)

La consecuencia inmediata de realizar estos supuestos es la validez del Teorema de Gauss-Markov, del cual aprendimos que el estimador de mínimos cuadrados ordinarios de β es el mejor estimador dentro del grupo de los estimadores *lineales* e *insesgados*. El objetivo de esta sección es relajar los supuestos relacionados con la varianza de u y analizar las consecuencias que se derivan de ello.

Inicialmente seremos lo más ambiciosos que podamos, por lo que desearíamos que la matriz de varianzas tome cualquier forma, esto es $V(u) = \Omega$, donde Ω es cualquier matriz. Pero el hecho de que Ω juegue el rol de una varianza implica la necesidad de que cumpla dos condiciones mínimas:

1. Ω *debe ser una matriz simétrica.* Ya hemos discutido sobre ello. Por definición de la matriz de varianzas, cada elemento (i, j), llamado w_{ij}, es la covarianza entre u_i y u_j. Dado que $Cov(u_i, u_j) = Cov(u_j, u_i)$ entonces $w_{ij} = w_{ji}$, por lo tanto Ω es una matriz simétrica.

2. Ω *debe ser definida positiva.* Recordar que si $V(u) = \Omega$, entonces para cualquier vector c con dimensión n x 1 , $V(c'u) = c'V(u)c = c'\Omega c$. u es un vector de variables aleatorias y por lo tanto su varianza, Ω es una matriz. Pero $c'u$ es una variable aleatoria escalar y por lo tanto su varianza $c'\Omega c > 0$ para cualquier c, luego Ω tiene que ser definida positiva.

En consecuencia, cuando se relaje el supuesto $\Omega = \sigma^2 I$, se dejará que Ω sea cualquier matriz simétrica y definida positiva.

Propiedades importantes

Las matrices simétricas y definidas positivas tienen tres propiedades muy importantes que usaremos repetidamente de aquí en adelante. Si Ω es definida positiva:

1. Existe Ω^{-1}, pues si Ω es definida positiva, entonces $det(\Omega) \neq 0$.

2. Ω^{-1} es simétrica y definida positiva.

3. Existe una matriz $P \in \mathbb{R}^{n \times n}$ no singular, tal que:

$$\Omega^{-1} = P'P$$

La demostración de este resultado puede encontrarse en muchos libros de álgebra. Para tener una intuición aproximada de lo que significan estas propiedades, primero considere a Ω como si fuese una matriz de dimensión 1 x 1, es decir, un escalar. La primer propiedad simplemente dice que los números positivos tienen inversa. La segunda propiedad establece que la inversa de un número positivo sigue siendo positiva. La tercera propiedad dice que para cada número positivo hay una raíz cuadrada que le corresponde. Estas propiedades pueden ser generalizadas para el caso de matrices, entonces P juega el rol de raiz cuadrada de Ω^{-1} en un contexto matricial.

4.1.2. El Modelo Lineal Generalizado y el Estimador de Mínimos Cuadrados Generalizados (MCG)

El *modelo lineal generalizado* será exactamente igual al modelo lineal original pero dejando que $V(u) = \Omega$ sea cualquier matriz simétrica y definida positiva. Esto es, dejando que los errores puedan ser heterocedásticos y/o estar serialmente correlacionados.

La estructura del modelo lineal generalizado es:

$$Y = X\beta + u$$

- $E(u) = 0$
- $V(u) = \Omega$
- X es una matriz no estocástica con rango completo ($\rho(x) = K$)

Dado que todos los supuestos clásicos son utilizados en la demostración del Teorema de Gauss-Markov, relajar el supuesto de la varianza implicará invalidar las conclusiones de dicho teorema: el estimador MCO ya no será el mejor estimador lineal e insesgado. Además, sabemos que, por construcción, el estimador MCO es lineal, y dado que el supuesto de varianza no es utilizado en la prueba de insesgadez, es también insesgado. Por lo tanto, cuando se levantan los supuestos de homocedasticidad o de no correlación serial, el estimador MCO, si bien sigue siendo lineal e insesgado, no será aquel de menor varianza (es decir, no será el más eficiente). Una estrategia común es tratar de encontrar el mejor estimador lineal e insesgado del modelo lineal generalizado. Para hallar un estimador apropiado de β, se debe premultiplicar el modelo por la matriz P que satisface $P'P = \Omega^{-1}$. Que dicha matriz exista es una consecuencia directa del hecho de que Ω sea simétrica y definida positiva.

$$PY = PX\beta + Pu$$

$$Y^* = X^*\beta + u^* \tag{4.1}$$

donde $Y^* = PY$, $X^* = PX$ y $u^* = Pu$. La ecuación 4.1 representa el llamado *modelo lineal transformado*. Notar que Y^* es un vector columna de n elementos, X^* es una matriz $n \times K$ y u^* es un vector columna de n elementos. Es muy importante observar que el modelo original y el transformado comparten el mismo coeficiente β desconocido. Ahora será relevante analizar las propiedades estadísticas del término de error del modelo transformado, u^*:

- $E(u^*) = E(Pu) = PE(u) = 0$ [Por ser P una matriz no estocástica]

- Dado que $E(u^*) = 0$, $V(u^*) = E(u^*u^{*\prime})$,

$$\begin{aligned}
E(u^*u^{*'}) &= E(Puu'P') \\
&= PE(uu')P' \\
&= P\Omega P' \\
&= P[\Omega^{-1}]^{-1}P' \\
&= P[P'P]^{-1}P' && \text{[Usando la propiedad de P]} \\
&= PP^{-1}P'^{-1}P' && \text{[Recordar que } (AB)^{-1} = B^{-1}A^{-1}\text{]} \\
&= I
\end{aligned}$$

¿Tiene sentido que la matriz de varianzas y covarianzas sea la identidad? Me parece más claro si se define V(u)=simgacuadrado*omega.

- El rango de $X^*, \rho(X^*) = \rho(PX)$ es K. Este es un resultado estándar en álgebra lineal, que consiste en que si se premultiplica una matriz de rango K por una matriz no singular, se preserva su rango.

¿Qué es lo que se aprende de estas propiedades? Por construcción, el modelo generalizado no satisface los supuestos clásicos. Sin embargo, el modelo transformado sí lo hace: Sus variables están linealmente relacionadas, su término de error, u^*, tiene media cero y su matriz de varianza puede ser escrita como el producto de un escalar (1, en este caso) por la matriz identidad, y la matriz de variables explicativas, X^*, tiene rango completo. Por lo tanto, el Teorema de Gauss-Markov vale para el modelo transformado, implicando que el mejor estimador lineal e insesgado para β es el estimador de mínimos cuadrados ordinarios que surge de usar las variables transformadas, esto es:

$$\hat{\beta}_{MCG} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$$

Lo llamaremos *estimador mínimos cuadrados generalizados (MCG, o GLS en inglés, por generalized least squares)*. Es importante notar que β es el mismo en el modelo original y en el transformado. Entonces, el procedimiento para derivar un estimador MELI (mejor estimador lineal e insesgado) consistió en transformar el modelo original de tal forma de que el Teorema de Gauss-Markov se siga cumpliendo y luego utilizar el método de mínimos cuadrados ordinarios en el modelo transformado. Por lo tanto, el estimador MCG es el estimador MCO aplicado al modelo transformado. Es interesante ver que no hemos necesitado probar ninguna propiedad de $\hat{\beta}_{MCG}$ otra vez: el modelo del cual surge, el transformado, satisface todos los supuestos clásicos, por lo que $\hat{\beta}_{MCG}$ es lineal, insesgado y de máxima eficiencia (menor varianza) entre el grupo de estimadores lineales e insesgados.

Por otra parte:

$$\begin{aligned}
\hat{\beta}_{MCG} &= (X^{*\prime}X^*)^{-1}X^*Y^* \\
&= (X'P'PX)^{-1}X'P'PY \qquad \text{[Recordar que } X^* = PX\text{, y que } X^{*\prime} = (PX)' = X'P'\text{]} \\
&= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y
\end{aligned}$$

Esta es una manera opcional de expresar el estimador MCG. Entonces, tenemos dos posibles formas de computar $\hat{\beta}_{MCG}$. Si conocemos Ω podremos, alternativamente:

- Obtener P (todavía no hemos visto cómo hacerlo!), obtener las variables transformadas $X^* = PX$ y $Y^* = PY$, y computar el estimador de mínimos cuadrados ordinarios usando dichas variables transformadas. Esto es, expresando $\hat{\beta}_{MCG} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$
- Expresando directamente $\hat{\beta}_{MCG} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$

Es interesante observar que cuando $\Omega = \sigma^2 I$ (esto es, cuando no hay heterocedasticidad ni autocorrelación), el estimador MCG no es más que el estimador MCO. Entonces, el estimador de mínimos cuadrados ordinarios es sólo un caso particular dentro del marco de mínimos cuadrados generalizados. La inferencia en el modelo lineal generalizado se deriva, otra vez, directamente del modelo transformado y del hecho de que satisface todos los supuestos clásicos. Dado que el estimador MCG es un estimador MCO para el modelo transformado, su varianza será:

$$V(\hat{\beta}_{MCG}) = (X^{*\prime}X^{*})^{-1}$$

y un estimador insesgado de la misma es:

$$\hat{V}(\hat{\beta}_{MCG}) = S^{*2}(X^{*\prime}X)^{-1}$$

donde $S^{*2} = e^{*\prime}e^{*}/(n-K)$ y e^{*} es un vector de los residuos resultantes de estimar por MCO el modelo transformado.

Si asumimos que los u_i del modelo original están normalmente distribuidos, entonces los $u_i^{*\prime}$ son también normales debido a que surgen como una simple transformación lineal de los primeros. Luego, dado que para el modelo transformado todos los supuestos clásicos se cumplen, la inferencia se realiza procediendo de la forma habitual, esto es, todos los estadísticos 't'y 'F'de los capítulos 2 y 3 son válidos una vez que se utiliza el modelo transformado.

4.1.3. Propiedades del estimador de mínimos cuadrados bajo el modelo lineal generalizado

Expresaremos el estimador de mínimos cuadrados ordinarios como $\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$ para distinguirlo de otros estimadores que aparecerán a lo largo del capítulo. Ya conocemos sus propiedades bajo los supuestos clásicos: es lineal e insesgado, y su varianza $\sigma^2(X'X)^{-1}$ es la menor entre el grupo de estimadores lineales e insesgados. En esta sección analizaremos qué ocurre con estas propiedades cuando se permite la presencia de heterocedasticidad y/o correlación serial, esto es, cuando $V(u) = \Omega$ que no es el producto de un escalar por la matriz identidad.

1. $\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$ sigue siendo, trivialmente, un estimador lineal.

2. $\hat{\beta}_{MCO}$ sigue siendo insesgado. Cuando demostramos la propiedad de insesgadez no se utilizó en absoluto el supuesto $V(u) = \sigma^2 I$, por lo que lógicamente, relajar dicho supuesto no tiene consecuencias sobre esta propiedad.

3. Dado que ya hemos encontrado el MELI para el modelo lineal generalizado (el estimador MCG), $\hat{\beta}_{MCO}$ aunque sigue siendo lineal e insesgado, ya no es más el estimador de mínima varianza entre los estimadores lineales e insesgados.

4. $V(\hat{\beta}_{MCO}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$. Esto es fácil de demostrar:

$$\begin{aligned} V(\hat{\beta}_{MCO}) &= E[(\hat{\beta}_{MCO} - E(\hat{\beta}_{MCO}))(\hat{\beta}_{MCO} - E(\hat{\beta}_{MCO}))'] \\ &= E[(\hat{\beta}_{MCO} - \beta)(\hat{\beta}_{MCO} - \beta)'] && \text{[Dado que } \hat{\beta}_{MCO} \text{ es insesgado]} \\ &= E[(X'X)^{-1}X'uu'X(X'X)^{-1}] \\ &= (X'X)^{-1}X'\Omega X(X'X)^{-1} && \text{[Dado que } X \text{ es no estocástica y } E(uu') = \Omega \text{]} \end{aligned}$$

5. $S^2(X'X)^{-1}$ ya no es más un estimador insesgado de $V(\hat{\beta}_{MCO})$. La demostración de este resultado se encuentra en el Apéndice de este capítulo.

En consecuencia, si insistimos en utilizar $\hat{\beta}_{MCO}$ bajo heterocedasticidad y/o autocorrelación, seguiremos teniendo un estimador insesgado y lineal, pero ya no será el más eficiente, en el sentido que no será el de menor varianza dentro de los estimadores lineales e insesgados. Más problemático resulta el hecho de que si bajo el modelo lineal generalizado, usamos equivocadamente $S^2(X'X)^{-1}$ como un estimador de $V(\hat{\beta}_{MCO})$, entonces estaremos ante una estimación *sesgada* de la verdadera varianza. Esto invalida todos los procedimientos de inferencia, incluídos todos los test 't'y 'F'vistos anteriormente, que explícitamente usan estimaciones de la varianza del estimador.

4.1.4. El Estimador de Mínimos Cuadrados Generalizados Factible

En la sección anterior se supuso que se conocía la matriz Ω. Aunque esto pueda ocurrir en algunos casos, como veremos luego, en muchas situaciones prácticas Ω será completamente desconocida.

Procederemos en dos pasos. Primero, se mostrará cómo se procede si se tiene un estimador válido de Ω; y luego, cuando lidiemos con heterocedasticidad y autocorrelación, analizaremos métodos alternativos para derivar estimadores apropiados de Ω para cada caso en particular.

Supongamos que disponemos de un estimador de Ω llamado $\hat{\Omega}$. Entonces, el estimador que resulta de reemplazar Ω por $\hat{\Omega}$ es el estimador MCG llamado *estimador de mínimos cuadrados generalizados factible (MCGF)*:

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

Hay varias particularidades relacionadas a este estimador que merecen ser mencionadas.

- A menos que seamos más específicos sobre $\hat{\Omega}$, no sabemos nada sobre las propiedades de $\hat{\beta}_{MCGF}$. Por ejemplo, si Y ha sido utilizada en el proceso de obtención de $\hat{\Omega}$, entonces $\hat{\beta}_{MCGF}$ no es ni siquiera un estimados lineal (Y aparece de alguna forma dentro de $\hat{\Omega}$). Entonces, incluso cuando sepamos que el estimador MCG es MELI, esta propiedad no se extiende para el estimador MCGF (el estimador factible no es MELI).

- Entonces, ¿cuál es el punto de usar el estimador MCGF? Si $\hat{\Omega}$ fuera un estimador *consistente* de Ω, esto es, si cuando el tamaño de la muestra es muy grande $\hat{\Omega}$ tiende a acercarse a Ω, entonces es intuitivamente claro que $\hat{\beta}_{MCGF}$ tenderá a acercarse a $\hat{\beta}_{MCG}$. Por lo tanto, para muestras grandes, $\hat{\beta}_{MCGF}$ es muy similar al estimador MELI $\hat{\beta}_{MCG}$. Esto provee una razón válida para utilizar MCGF.

- Si $\hat{\Omega}$ es consistente, puede demostrarse que $\hat{\beta}_{MCGF}$ es *asintóticamente normal*, esto es, cuando el tamaño de la muestra es muy grande, este tiene una distribución muy similar a una distribución normal. Entonces, los procedimientos de inferencia para el MCGF son válidos para muestras grandes.

Es interesante analizar la posibilidad de estimar Ω sin introducir supuestos adicionales. En principio, es fácil de notar que Ω tiene $n + n(n-1)/2$ parámetros diferentes, esto es, de sus $n \times n$ parámetros, el supuesto de simetría implica que los parámetros que son esencialmente diferentes son aquellos en la diagonal (n) y aquellos debajo (o sobre) la diagonal ($(n^2 - n)/2$). Cualquier intento de estimar todos los parámetros diferentes sin imponer ningún supuesto sobre la matriz de varianzas más allá del supuesto de simetría, implica estimar $n + n(n-1)/2$ parámetros con sólo n observaciones, lo cual, en muchas situaciones relevantes, es imposible de resolver. Entonces, la estrategia que seguiremos, que es el tratamiento habitual, será la de analizar heterocedasticidad y correlación serial por separado, para ver si podemos aprender más sobre cada problema específico que eventualmente nos ayudará a imponer supuestos realistas que reduzcan el número de parámetros que deben ser estimados.

Primero trataremos el problema de heterocedasticidad y luego el vinvulado a la correlación serial, dado que el análisis de esta última se simplifica considerablemente una vez que hayamos hablado más explícitamente sobre los modelos básicos de series de tiempo.

4.2. Heterocedasticidad

El supuesto de homocedasticidad implica que la varianza del término de error del modelo lineal es constante para cada observación, esto es:

$$V(u_i) = \sigma^2 \qquad i = 1, \ldots, n$$

La presencia de *heterocedasticidad* significa que se deja que la varianza del término de error difiera entre las distintas observaciones, es decir:

$$V(u_i) = \sigma_i^2 \qquad i = 1, \ldots, n$$

Por razones pedagógicas, mantendremos todos los demás supuestos clásicos, incluyendo el supuesto de no correlación serial. En términos matriciales, a la matriz de varianzas de u se le permitirá tomar la siguiente forma general:

$$V(u) = E(uu') \equiv \Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

La heterocedasticidad es un problema típico cuando se trabaja con datos de corte transversal, aunque una forma particular de heterocedasticidad en el contexto de series de tiempo (el tipo ARCH) ha recibido una considerable atención recientemente.

Consideremos el caso de la relación entre el consumo y el ingreso. Trivialmente, esperaríamos que el consumo promedio crezca con el ingreso, pero también es posible que a medida que el ingreso aumente, el nivel de consumo sea más difícil de predecir, dado que se tiene más discrecionalidad sobre el mismo. Es decir, para niveles de ingresos bajos los niveles de consumo son poco variables ya que todos los individuos podrán mantener un nivel de consumo similar dado el ingreso, mientras que para ingresos altos existe una mayor posibilidad de destinar el ingreso a otros usos diferentes del consumo, por lo que se observaría una mayor variabilidad. Entonces, a medida que el ingreso aumenta, no sólo crece el consumo esperado sino que también se hace más volátil.

Otro ejemplo está relacionado a los errores de tipeo y la experiencia. Suponga que se tiene una muestra de individuos con diferente experiencia (en tiempo) de tipeo. Mayor experiencia conduce a una cantidad menor de errores de tipeo esperada, pero también mayor experiencia implica mayor homogeneidad en el tipeo, es decir, la experiencia hace que los individuos sean más similares en término de los errores que cometen cuando tipean (menor variabilidad para individuos con mayor experiencia).

4.3. Consecuencias de ignorar heterocedasticidad

Si procedemos a estimar el modelo lineal ignorando la posible presencia de heterocedasticidad, utilizaríamos los siguientes estimadores:

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

y

$$\hat{V}(\hat{\beta}_{MCO}) = S^2(X'X)^{-1}$$

Ya hemos discutido anteriormente una situación más general, en la cual permitíamos la posibilidad de heterocedasticidad y correlación serial, y vimos que ignorar cualquiera de ellos afecta las propiedades de optimalidad de los estimadores estándar. A continuación repasemos las consecuencias de ignorar heterocedasticidad:

1. Bajo heterocedasticidad, $\hat{\beta}_{MCO}$ es lineal e insesgado pero no el más eficiente. Sabemos que el estimador $\hat{\beta}_{MCG}$ sí es MELI, y en este caso lo denominaremos estimador de *Mínimos Cuadrados Ponderado*, para el cual $P = 1/\sigma_i$, o alternativamente $P = diag(\sigma_1^{-1}, \ldots, \sigma_n^{-1})$.

2. Bajo heterocedasticidad, $S^2(X'X)^{-1}$ es un estimador sesgado de $V(\hat{\beta}_{MCO})$.

Resumiendo, bajo heterocedasticidad, el estimador MCO de β sigue siendo lineal e insesgado aunque ya no el mejor dentro de dicho grupo (el estimador MCG sí lo es). Lo que es todavía más serio, es que el estimador estándar de la varianza $S^2(X'X)^{-1}$ es sesgado, invalidando todos los test de hipótesis analizados anteriormente, es decir, los test convencionales (t y F) ya no serán válidos.

En consecuencia, la pregunta relevante es la de cómo proceder con la estimación e inferencia bajo heterocedasticidad. Antes de volvernos totalmente sobre esta cuestión, nos concentraremos en el problema de evaluar la presencia de heterocedasticidad, y nos preocuparemos por ella una vez que hayamos encontrado evidencia suficiente que pruebe su existencia.

4.4. Test para detectar presencia de heterocedasticidad

Test de White

La hipótesis nula de este test es la inexistencia de heterocedasticidad, es decir, la varianza es constante para todas las observaciones. La hipótesis alternativa simplemente niega la nula, es decir, bajo la hipótesis alternativa hay *alguna forma* de heterocedasticidad. Por lo tanto:

$$H_o : \sigma_i^2 = \sigma^2 \quad \text{vs.} \quad H_A : \sigma_i^2 \quad \textit{no es constante para todas las observaciones}$$

Para analizar cómo se procede para realizar el test, sin pérdida de generalidad, considere el siguiente modelo lineal con tres variables:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad i = 1, \ldots, n \tag{4.2}$$

Los pasos para realizar el test de White son los siguientes:

1. Estimar el modelo (4.2) por MCO y calcular los errores de estimación para cada una de las observaciones y elevarlos al cuadrado, almacenándolos en un vector llamado $e^2 = (Y_i - \hat{Y}_i)^2$.

2. Realizar una regresión de e^2 en todas las variables explicativas del modelo original, todos los cuadrados de las variables explicativas y todos los productos cruzados no redundantes de las mismas. En nuestro caso, ésto equivale a regresar e^2 en

$$\begin{bmatrix} 1 & X_2 & X_3 & X_2^2 & X_3^2 & X_2X_3 \end{bmatrix}$$

   y obtener el coeficiente de determinación R^2 de esta regresiń auxiliar.

3. Bajo la hipótesis nula, el estadístico nR^2 tiene una distribución que se aproxima asintóticamente a una $\chi^2(p)$, donde p es el número de variables explicativas en la regresión auxiliar sin incluir la constante.

Entonces, la idea es rechazar H_o si nR^2 es significativamente diferente de cero. Para dar una intuición, el modelo auxiliar puede verse como un intento de "modelar"la varianza del término de error. Si el R^2 de esta regresión auxiliar fuera grande, entonces podríamos explicar el comportamiento de los residuos al cuadrado a partir de alguna de las variables incluidas en esa regresión auxiliar, teniendo evidencia de que no son constantes. Lo que hay que tener en cuenta es que el test de White prueba *todas las posibles causas de heterocedasticidad*.

Existen algunas advertencias y comentarios relacionados con la utilización del test de White que se deberían tener en mente:

- Es un test para muestras grandes. Es decir, se comporta correctamente sólo cuando el número de observaciones es muy grande.
- El test parece ser muy informativo bajo la hipótesis nula. En ese caso, podríamos estar seguros que no hay problemas de heterocedasticidad. Pero, es un test que tiene poca potencia (probabilidad de aceptar la hipótesis y que sea cierta), ya que los test que no rechazan la hipótesis nula son informativos en la medida que tenga un gran poder de detectar, en nuestro caso, diferentes patrones de heterocedasticidad. Desafortunadamente, con muestras pequeñas o con un gran número de regresores, el test de White no tiene mucha potencia, proveyendo en muchos casos información limitada cuando no rechaza la hipótesis nula.
- Cuando se rechaza la hipótesis nula, el test sugiere que hay heterocedasticidad, pero no nos provee información sobre la causa ni la forma de dicha heterocedasticidad. Este hecho causará problemas cuando tratemos de construir el estimador MCG, para el cual necesitaremos conocer de manera muy específica qué es lo que causa la heterocedasticidad.

Test de Breusch-Pagan/Godfrey/Koenker

El test de Breusch-Pagan/Godfrey es, mecánicamente, muy similar al test de White. Éste trata de probar heterocedasticidad en un sentido más estrecho y, por lo tanto, ser más informativo sobre

las causas de la heterocedasticidad.

En este caso, evaluaremos si ciertas variables son potenciales causantes de heterocedasticidad. Considere el siguiente modelo que permite la presencia de heterocedasticidad:

$$Y = X\beta + u$$

donde los u_i están normalmente distribuidos con $E(u) = 0$ y

$$V(u_i) = h(\alpha_1 + \alpha_2 Z_{2i} + \alpha_3 Z_{3i} + \ldots + \alpha_p Z_{pi})$$

y $h(.)$ es cualquier función dos veces derivable que toma sólo valores positivos.

Note que cuando $\alpha_2 = \ldots = \alpha_p = 0$, $V(u_i) = h(\alpha_1)$, que es constante. Entonces, la hipótesis de homocedasticidad corresponde a:

$$H_o : \alpha_2 = \ldots = \alpha_p = 0$$

y la hipótesis alternativa es:

$$H_A : \alpha_2 \neq 0 \vee \alpha_3 \neq 0 \vee \ldots \vee \alpha_p \neq 0$$

esto es, al menos una de las variables propuestas explican la varianza. Los pasos para implementar el test son los siguientes:

1. Estimar el modelo original por MCO, ignorando la presencia de heterocedasticidad, y retener los residuos al cuadrado, $e^2 = (Y_i - \hat{Y}_i)^2$.

2. Realizar una regresión de e^2 en las Z_{ki} variables, $k = 2, \ldots, p$ y obtener la suma de cuadrados explicados (SCE) de este modelo auxiliar.

3. El estadístico de prueba es:
$$\frac{1}{2} SCE \sim \chi^2(p-1)$$
es decir, el estadístico del test tiene una distribución asintótica χ^2 con $p-1$ grados de libertad bajo la hipótesis nula.

La intuición es similar a la del test de White. Estamos analizando un modelo que trata de 'explicar'la varianza, en este caso, centrándonos en si las Z_{ik} nos ayudan a explicarla. Cuando la hipótesis nula de homocedasticidad es verdadera, el modelo auxiliar no debería tener ningún poder explicativo, entonces la suma de cuadrados explicados (SCE) debería tomar volores cercanos a cero. Por otro lado, cuando la hipótesis alternativa es correcta, al menos una de las variables Z_{ik} contribuye significativamente a explicar los errores al cuadrado, haciendo que el estadístico de prueba tome

un valor grande. Entonces, rechazaremos H_o si el estadístico es suficientemente grande, de acuerdo a los valores críticos de su distribución bajo la hipótesis nula.

A continuación detallamos algunos cometarios importantes:

- Como en el caso del test de White, éste es un test para muestras grandes.
- Es importante comparar el tipo de información que obtenemos de realizar este test bajo la hipótesis nula y bajo la alternativa. El test de Breusch-Pagan/Godfrey se diferencia del test de White principalmente porque le proveemos más información para hacerlo más operativo, es decir, analizando si la posible heterocedasticidad está relacionada con un grupo específico de variables predeterminadas. Entonces, la hipótesis nula no nos dice que los residuos son homocedásticos, sino que un grupo particular de variables no contribuye a explicar la varianza. O, lo que es lo mismo, bajo la hipótesis nula no hay heterocedasticidad causada por las variables Z propuestas. Desde una perspectiva lógica, el test de Breusch-Pagan/Godfrey es menos informativo que el test de White bajo la hipótesis nula. Por otro lado, bajo la hipótesis alternativa obtendremos información mucho más focalizada en cuanto a los causantes de la heterocedasticidad (las variables Z).
- El supuesto de normalidad puede ser bastante restrictivo para muchas situaciones, y el test muestra un comportamiento incorrecto si este no se cumple. Koenker (1980) propuso usar como estadístico de prueba nR_A^2, donde n es el número de observaciones y R_A^2 es el coeficiente de determinación de la regresión auxiliar. Este estadístico tiene la misma distribución asintótica que el estadístico de Breusch-Pagan/Godfrey, pero es válido bajo no-normalidad.

Test de Goldfeld-Quandt

Este test es útil si creemos que la heterocedasticidad puede ser atribuída a una sola variable. Considere el caso simple $Y_i = \alpha + \beta X_i + u_i$, donde los u_i están normalmente distribuidos con media cero, y se cree que $\sigma_i^2 = V(u_i)$ está relacionada con X_i. Por ejemplo, los datos de consumo-ingreso encuadrarían en esta situación, donde la varianza parece crecer con el ingreso.

El test de Goldfeld-Quandt es muy intuitivo y consiste en los siguientes pasos:

1. Ordenar las observaciones de acuerdo a los valores de X_i.
2. Eliminar las c observaciones centrales, obteniendo dos sub-muestras de tamaño $(n-c)/2$.
3. Correr dos regresiones separadas para cada una de las sub-muestras y computar la suma de cuadrados residuales (SCR) para cada una de estas regresiones. Llamemos SCR_1 a la correspondiente a la primer submuestra, y SCR_2 a la segunda.

4. Dividir cada SCR por $n-k$, para obtener dos números que representen una estimación insesgada de la varianza de los términos de error de cada regresión. Luego, bajo la hipótesis nula de homocedasticidad estos dos números deberían ser iguales. O, alternativamente, SCR_1/SCR_2 debería tomar un valor cercano a 1. Entonces, este test puede basarse en un test F de igualdad de varianza, es decir, utilizaremos:

$$\frac{SCR_1}{SCR_2} \sim F(\frac{n-c-2K}{2}, \frac{n-c-2K}{2})$$

que bajo la hipótesis nula tiene una distribución F con $\frac{n-c-2K}{2}$ grados de libertad tanto en el numerador como en el denominador. Si ponemos la mayor varianza en el numerador, entonces rechazamos H_o si el estadístico de prueba nos da un número muy grande.

4.5. Estimación e inferencia bajo heterocedasticidad

Bajo la presencia de heterocedasticidad (asumiendo la inexistencia de correlación serial), la varianza de u toma la sigueinte forma:

$$Var(u) = \Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

El mejor estimador lineal e insesgado bajo heterocedasticidad será el estimador de mínimos cuadrados generalizados obtenido anteriormente:

$$\hat{\beta}_{MCG} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y^*$$

con $X^* = PX$, $Y^* = PY$, $u^* = Pu$ y P es tal que $P'P = \Omega^{-1}$

Es fácil verificar que para este caso en particular P es:

$$P = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n} \end{pmatrix}$$

Luego, premultiplicar por P tiene el efecto de dividir cada observación por su desvío estándar σ_i, por lo que las variables transformadas son las originales divididas por el desvío estándar correspondiente a cada observación. Por lo tanto, la estimación por MCG consiste en aplicar MCO sobre las variables transformadas.

Alternativamente, si dividimos cada observación del modelo lineal por σ_i:

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{X_{2i}}{\sigma_i} + \ldots + \beta_k \frac{X_{ki}}{\sigma_i} + \frac{u_i}{\sigma_i}$$

$$Y_i^* = \beta_1 X_{1i}^* + \beta_2 X_{2i}^* + \ldots + \beta_k X_{ki}^* + u_i^*$$

Notar que $V(u_i^*) = V(u_i/\sigma_i) = 1$, entonces los residuos de este modelo transformado son homocedásticos.

Luego, si conocemos los errores estándar, se puede abordar una simple estrategia de estimación que consiste en dividir primero todas las observaciones por el error estándar del término de error y entonces aplicar MCO sobre las variables transformadas. En la práctica, raramente se conocen los σ_i, y cualquier intento de estimarlos sin imponer supuestos adicionales requiere de estimar $K + n$ elementos, es decir, los K coeficientes del modelo lineal y los n elementos desconocidos de la varianza, lo que es imposible teniendo solamente n observaciones.

Frente a este problema, existen dos estrategias que se suelen seguir en la práctica:

- Si se tiene información sobre las varianzas, o si se está dispuesto a adoptar algún supuesto simplificador sobre el modo que que opera la heterocedasticidad, se puede buscar un estimador de mínimos cuadrados generalizados. Este procedimiento provee una estimación insesgada y eficiente, e inferencia válida usando los test comunes. Esta es la primer camino que analizaremos.

- Pensemos nuevamente en los efectos de la heterocedasticidad sobre los procedimientos de estimación estándar. El estimador de mínimos cuadrados sigue siendo insesgado aunque no eficiente. Por otro lado, el estimador estándar de la matriz de varianza es sesgado, lo que invalida los procedimientos de inferencia basados en los test 't'y 'F'. Una segunda estrategia consiste en seguir estimando β por MCO, pero buscar un estimador válido para su matriz de varianzas.

Analizaremos estas dos alternativas en las siguientes secciones.

4.5.1. Estructura de varianza conocida: estimación eficiente por MCG

Sin pérdida de generalidad, considere el caso simple de dos variables:

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

para el cual todos los supuestos clásicos se cumplen, excepto el de homocedasticidad. De acuerdo a la sección previa, si conocemos la varianza de u_i, se puede dividir cada observación por la raíz cuadrada de dicha varianza, lo que nos proporcionaría un estimador MCG. En lugar de eso, una estrategia habitual resulta ser la de asumir alguna forma particular de heterocedasticidad.

1. *Varianza proporcional al cuadrado de la variable explicativa*: Cuando se cree que la heterocedasticidad está asociada a una variable explicativa en particular, una práctica habitual es asumir que:

$$V(u_i) = \sigma^2 X_i^2$$

donde σ^2 es una constante desconocida. En este caso, la varianza se incrementa cuadráticamente con X. Esta sería una estrategia razonable para el ejemplo del consumo e ingreso expuesto anteriormente. Para obtener un estimador MCG para el modelo, se dividen todas las observaciones por X_i.

$$\frac{Y_i}{X_i} = \beta_1 \frac{1}{X_i} + \beta_2 + \frac{u_i}{X_i}$$
$$Y_i^* = \beta_1 X_{0i}^* + \beta_2 + u_i^*$$

Notar que:

$$E(u_i^{*2}) = E\left[\left(\frac{u_i}{X_i}\right)^2\right] = \frac{1}{X_i^2} E(u_i^2) = \sigma^2 \frac{X_i^2}{X_i^2} = \sigma^2$$

por lo que los residuos del modelo transformado son homocedásticos y, consecuentemente, aplicar MCO sobre dicho modelo proporciona el estimador MCG.

Se deben tener en cuenta tres comentarios pertinentes sobre esta estrategia:

(a) Notar que para obtener el estimador MCG no necesitamos conocer todos los componentes de la varianza, es decir, no necesitamos conocer σ^2. Lo que hicimos para hacer que el término de error del modelo transformado sea homocedástico es dividir las observaciones por la parte del error estándar que varía entre las observaciones, esto es, por X_i.

(b) Luego, esta estrategia provee un estimador MCG, no un MCGF, dado que, para implementarlo no se necesita estimar partes de la varianza con anticipación. Ésto es consecuencia de la imposición de un supuesto bastante fuerte sobre la varianza.

(c) Se debe tener cuidado cuando se enterpretan los coeficientes en el segundo paso. Notar con atención que el intercepto (ordenada al origen) del modelo transformado es la pendiente del modelo original, y la pendiente del modelo transformado es el intercepto del original.

2. *Varianza proporcional a la variable explicativa*: La especificación para heterocedasticidad en este caso es:

$$V(u_i) = \sigma^2 X_i$$

Desde luego que X_i debe ser positiva para ser consistente con la varianza. Siguiendo la misma estrategia que antes, se divide por la parte del error estándar que varía entre observaciones, esto es, dividir todas las observaciones por $\sqrt{X_i}$, obteniendo:

$$\frac{Y_i}{\sqrt{X_i}} = \beta_1 \frac{1}{\sqrt{X_i}} + \beta_2 \sqrt{X_i} + \frac{u_i}{\sqrt{X_i}}$$
$$Y_i^* = \beta_1 X_{0i}^* + \beta_2 X_{1i}^* + u_i^*$$

Notar nuevamente que $Var(u_i^*) = \frac{\sigma^2 X_i}{X_i} = \sigma^2$, por lo que el término de error del modelo transformado es homocedástico, y se procede a estimar por MCO sobre el modelo transformado usando $1/\sqrt{X_i}$ y $\sqrt{X_i}$ como variables explicativas para obtener el estimador MCG. Este no será MCGF dado que no se necesita una estimación previa para implementar el estimador. Para poder implementarlo e interpretarlo, notar que el modelo transformado no tiene intercepto, y que el coeficiente de la primer variable explicativa corresponde al intercepto del modelo original, y el coeficiente de la segunda variable corresponde a la pendiente del modelo original.

Desde luego que muchas más transformaciones podrían ser analizadas, todas basadas en la misma idea de postular una forma particular de la varianza y luego transformar el modelo dividiendo todas las observaciones por la parte del error estándar que varía entre observaciones. Una estrategia habitual es la de implementar alguna de estas transformaciones y hacer nuevamente un test de heterocedasticidad después de la estimación por MCG. Si la forma de heterocedasticidad asumida es la correcta, entonces los test de heterocedasticidad no deberían rechazar la hipótesis nula de existencia de homocedasticidad.

4.5.2. Estructura de varianza desconocida: estimación consistente

La estrategia anterior proporciona procedimientos para estimar por MCG, que se derivan como consecuencia de imponer alguna estructura a aquello que causa la heterocedasticidad. Pero, en muchas situaciones prácticas, no se conoce la forma exacta de la heterocedasticidad. Una situación típica que aparece cuando se siguen las estrategias anteriores es la de adoptar una forma particular para la varianza, computar la estimación por MCG, y que luego el test de heterocedasticidad en el modelo transformado siga rechazando la hipótesis nula, sugiriendo que la transformación no ha sido exitosa al "remover" el problema.

Por lo tanto, una estrategia alternativa de estimación bajo heterocedasticidad, como mencionamos anteriormente, consiste en retener el estimador de MCO (que todavía es lineal e insesgado aunque no eficiente) y buscar un estimador válido para la matriz de varianzas. Recordar que la matriz de varianzas de $\hat{\beta}_{MCO}$ bajo heterocedasticidad está dada por:

$$V(\hat{\beta}_{MCO}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

donde Ω, en este caso es una matriz diagonal que tiene a σ_i^2 como elementos típicos. Denotaremos a esa matriz como $\Omega = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$.

La intuición parece sugerir que el problema de estimar $V(\hat{\beta}_{MCO})$ requiere la estimación de $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, que implica estimar n parámetros. Pero resulta ser el caso de que una estimación de $X'\Omega X$ satisfacería nuestros propósitos. Después de notar esto, Halbert White ha probado que un estimador consistente de $X'\Omega X$ está dado por $X'DX$, donde D es una matriz diagonal $n \times n$, con elementos típicos iguales al cuadrado de los residuos aplicando MCO en el modelo original, es decir, $D = diag(e_1^2, e_2^2, \ldots, e_n^2)$, donde e_i $(i = 1, \ldots, n)$ son los residuos obtenido de estimar el modelo original por MCO. Es importante remarcar que D no es un estimador consistente de Ω, pero sí lo es de $X'\Omega X$, que no es lo mismo.

Luego, un estimador consistente de la matriz de varianza con heterocedasticidad está dado por:

¿Acá no sobran X' y X?

$$\hat{V}(\hat{\beta}_{MCO}) = (X'X)^{-1}X'DX(X'X)^{-1}$$

Entonces, la estrategia consiste en usar MCO pero reemplazando el estimador estándar $S^2(X'X)^{-1}$ por el estimador consistente de varianza de White. Esto proporciona una estimación insesgada de β, una estimación consistente de $V(\hat{\beta}_{MCO})$ y un marco de inferencia válido para muestras grandes (por consistencia).

Cuando se compara este procedimiento con la estimación por MCG, esta estrategia sacrifica eficiencia, dado que MCG es, por construcción, el más eficiente dentro del grupo de estimadores lineales e insesgados. Pero, por otro lado, no requiere de supuestos sobre la estructura de la heterocedasticidad, y tampoco perdemos insesgamiento ni linealidad.

Apéndice

Sesgo en el estimador de varianza bajo heterocedasticidad

Dijimos que el estimador estándar de la varianza del estimador de mínimos cuadrados es *sesgado* cuando hay heterocedsaticidad. Probaremos un resultado más general.

Resultado: *Si Ω no se puede escribir como un escalar multiplicado por la matriz identidad, entonces $S^2(X'X)^{-1}$ es sesgado para $V(\hat{\beta})$.*

Prueba: En primer lugar, recordar que:

$$V(\hat{\beta}_{MCO}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

Por definición de:

$$\begin{aligned} S^2(X'X)^{-1} &= \frac{e'e}{n-K}(X'X)^{-1} \\ &= \frac{u'Mu}{n-K}(X'X)^{-1} \\ &= \frac{tr(u'Mu)}{n-K}(X'X)^{-1} \\ &= \frac{tr(Muu')}{n-K}(X'X)^{-1} \end{aligned}$$

Tomando esperanzas:

$$\begin{aligned} E(S^2(X'X)^{-1}) &= \frac{tr(ME(uu'))}{n-K}(X'X)^{-1} \\ &= \frac{tr(M\Omega)}{n-K}(X'X)^{-1} \end{aligned}$$

llamemos $\omega = \frac{tr(M\Omega)}{n-K}$. Por contradicción, supongamos que $S^2(X'X)^{-1}$ es insesgado. Entonces, tiene que ser igual a $V(\hat{\beta})$:

$$\omega(X'X)^{-1} = (X'X)^{-1}(X'\Omega X)(X'X)^{-1}$$

lo que equivale a:

$$[(X'X)^{-1}(X'\Omega X) - \omega I](X'X)^{-1} = 0$$

y dado que $(X'X)^{-1}$ tiene rango completo, equivale a:

$$\begin{aligned} (X'X)^{-1}(X'\Omega X) - \omega I &= 0 \\ X'\Omega X &= \omega X'X \\ &= X'DX \end{aligned}$$

en donde D es una matriz diagonal con todos sus elementos iguales a ω, de modo que $\Omega = D$. Pero esto se contradice con el hecho de que Ω permite la presencia de heterocedasticidad y/o autocorrelación.

Capítulo 5

Especificación del modelo

5.1. Especificación del modelo

En los capítulos anteriores hemos estudiado un modelo para la relación lineal entre una variable Y y K variables explicativas más un término de error:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \mu_i, \qquad i = 1, \ldots, n \tag{5.1}$$

De forma implícita, esta estructura lineal fue un supuesto que mantuvimos. Procedimos como si estuviéramos seguros de que las K variables verdaderamente pertenecen a esta relación, es decir, como si el modelo estuviera correctamente especificado. Aunque el concepto es mucho más general, en este capítulo cuando hablamos de *errores de especificación* nos referimos a casos en donde:

- Algunas variables han sido incorrectamente omitidas del modelo.
- Algunas variables han sido incorrectamente incluidas en el modelo.

Para ganar mayor precisión, consideremos el siguiente modelo particular en forma matricial:

$$Y = X_1\beta_1 + X_2\beta_2 + \mu \tag{5.2}$$

Aquí Y es un vector de n observaciones de la variable explicada, X_1 es una matriz de observaciones de K_1 variables explicativas, X_2 es una matriz de K_2 variables explicativas, y β_1 y β_2 son vectores de coeficientes desconocidos. Todos los supuestos clásicos se mantienen. A manera de ejemplo, Y podría ser consumo, X_1 podría ser el ingreso familiar y X_2 la riqueza de la familia. Esto nos ayudará a definir más claramente los dos errores de especificación:

1. *Omisión de variables relevantes*: En este caso erróneamente estimamos un modelo más chico que el verdadero modelo, esto es:

Modelo verdadero: $Y = X_1\beta_1 + X_2\beta_2 + \mu$

Modelo estimado: $Y = X_1\beta_1 + \mu$

En términos de nuestro ejemplo, esto correspondería al caso en donde la riqueza es una variable relevante para explicar el consumo, pero procedemos como si no lo fuera. También se puede ver como un caso en el que procedemos como si $\beta_2 = 0$ cuando en realidad no lo es.

2. *Inclusión de variables irrelevantes*: En este caso erróneamente estimamos un modelo más grande que el verdadero modelo, esto es:

   Modelo verdadero: $Y = X_1\beta_1 + \mu$

   Modelo estimado: $Y = X_1\beta_1 + X_2\beta_2 + \mu$

   En términos de nuestro ejemplo, incluimos a la riqueza como un regresor cuando en realidad no es una variable relevante. Desde una perspectiva diferente, este error de especificación corresponde a proceder como si β_2 fuera distinto de cero, cuando en realidad $\beta_2 = 0$.

Es importante hacer énfasis en que estamos explorando una situación hipotética, ya que nunca podemos conocer cuál es el verdadero modelo. Esto significa que estamos estudiando qué le pasaría a los estimadores cuando estimamos uu modelo incorrecto en lugar del verdadero. Procedemos a estudiar qué le pasa al estimador de MCO en cada caso por separado.

Omisión de variables relevantes

Supongamos que $\beta_2 \neq 0$ pero procedemos asumiendo erróneamente que $\beta_2 = 0$. Entonces, corremos una regresión de Y en X_1, y estimamos β_1 usando MCO:

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$$

El estimador es trivialmente lineal pero, en general, es *sesgado*.

$$\begin{aligned}
\hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'Y \\
&= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \mu) \\
&= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\mu \\
E(\hat{\beta}_1) &= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2
\end{aligned}$$

Por lo tanto, el sesgo del estimador puede ser computado como:

$$Sesgo(\hat{\beta}_1) = E(\hat{\beta}_1) - \beta_1 = (X_1'X_1)^{-1}X_1'X_2\beta_2$$

Entonces, a menos que $X_1'X_2 = 0$ el estimador de MCO para β_1 será sesgado. Por lo tanto, la omisión de variables relevantes puede tener consecuencias drásticas para los estimadores de

mínimos cuadrados ordinarios. Un error común en la práctica es pensar que la omisión de variables relevantes *necesariamente* conduce a estimadores sesgados. Para que el sesgo ocurra debe ser cierto que $X_1'X_2 \neq 0$. Cabe notar que en la práctica es muy raro que $X_1'X_2$ sea exactamente cero, por lo que la manera apropiada de enfocar la discusión previa es que la *fuerza* del sesgo dependerá de cuán cerca esté $X_1'X_2 = 0$ de ser cero.

Con el fin de explorar que más podemos aprender mirando $X_1'X_2$, supongamos por un momento que X_2 es una sola variable ($K_2 = 1$). Consideremos cualquier n-vector Z. Cuando $Z'X_2 = 0$ diremos que Z y X_2 son *ortogonales*. Usando esta terminología, la omisión de variables relevantes lleva a estimaciones sesgadas cuando la variable excluida no es ortogonal respecto a todas las variables incluidas. Para entenderlo mejor, notemos que $(X_1'X_1)^{-1}X_1'X_2$ es el estimador MCO que obtendríamos de regresar X_2 en X_1, por lo tanto, el sesgo será distinto de cero si X_2 puede ser linealmente explicada por las variables incluidas en X_1.

Entonces, en general, la omisión de variables relevantes produce estimaciones sesgadas a menos que las variables omitidas no están relacionadas linealmente con las variables incluidas. Este sesgo es usualmente llamado *sesgo por omisión de variables*.

Inclusión de variables irrelevantes

Supongamos que $\beta_2 = 0$ pero que procedemos como si fuera distinto de cero. En este caso correríamos una regresión de Y en X_1 y X_2. Con el fin de ver que el estimador resultante será insesgado, escribamos al verdadero modelo como:

$$Y = X_1\beta_1 + X_2\beta_2 + \mu$$

donde $\beta_2 = 0$. Para este modelo general, el estimador de mínimos cuadrados ordinarios provee un estimador insesgado para *cualquier* valor que β_1 o β_2 tomen, en particular para $\beta_2 = 0$. Entonces, no habrá necesidad de probar insesgadez ya que es un caso particular de lo que hemos probado en capítulos anteriores. Por lo tanto, la inclusión de variables irrelevantes no atenta contra la insesgadez. Puede argumentarse que la inclusión de variables irrelevantes conduce a estimaciones ineficientes, pero dicho argumento puede causar cierta confusión. Si no conocemos nada acerca de β_1 y β_2 y todos los supuestos clásicos se cumplen, entonces, por el Teorema de Gauss-Markov, el mejor estimador lineal e insesgado de β_1 y β_2 es el estimador de MCO, usando tanto X_1 como X_2 como variables explicativas. Por otro lado, si el verdadero modelo es $Y = X_1\beta_1 + \mu$ -tenemos certeza de que $\beta_2 = 0$-, entonces el mejor estimador lineal e insesgado para β_1 es $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$. En dicho caso, el estimador "extendido" (el que utiliza ambos regresores), dado que sigue siendo lineal e insesgado, de acuerdo al Teorema de Gauss-Markov, deberá tener mayor varianza.

¿Omitir o no omitir?

Si el sesgo es un mayor pecado que la ineficiencia, entonces incluir erróneamente una variable irrelevante parecería ser preferible a omitirla erróneamente. Además, cuando estimamos el modelo más grande, tenemos una estructura para *testear* formalmente si una variable (o grupo de variables) es realmente significativa, usando tests "t" o "F" de significatividad individual o conjunta como hemos visto previamente. Entonces, desde esta perspectiva parecería que es preferible estimar el modelo más "grande".

Sin embargo, debemos tener cuidado con este razonamiento ya que puede conducirnos a importantes pérdidas de precisión por la inclusión de regresores irrelevantes. Sin información adicional, hay un trade-off entre sesgo-varianza que subyace a la mayoría de los trabajos estadísticos. Modelos más grandes son más "seguros" en términos de sesgo, pero tienen el costo de ser estimados de forma más imprecisa. Además, modelos innecesariamente grandes van en contra de la parsimonia principal que subyace cualquier actividad de modelamiento, la cual busca pequeñas estructuras para explicar comportamientos generales.

Cabe hacer un comentario acerca de la práctica de descartar variables irrelevantes del modelo. Hemos distinguido entre dos situaciones: La primera se refiere a descartar del modelo una variable de la cual estamos completamente seguros que es irrelevante en el modelo. Como hemos comentado, si estamos totalmente seguros que una variable explicativa no es relevante, su omisión nos va a llevar a estimaciones con varianzas más chicas.

Una segunda situación se refiere a descartar una variable que es *estadísticamente* no significativa. Una práctica común es estimar un modelo "grande" y a partir de ahí testear la significatividad individual de las variables usando tests "t". Si para alguna variable en particular el test "t" sugiere aceptar la hipótesis nula, algunos investigadores reestiman el modelo descartando dicha variable. Esta es la lógica detrás del enfoque "de lo general a lo particular" respaldada por David Hendry y sus coautores. Aunque atractiva desde una perspectiva lógica, se debe tener cuidado ya que los tests "t" no proveen información exacta acerca de la verdad o falsedad de la hipótesis nula. Por otra parte, bajo el contexto clásico, la probabilidad de rechazar la hipótesis nula cuando es verdadera no es cero (i.e., la probabilidad del error de tipo I). Entonces, con una probabilidad que no es exactamente cero, podríamos estar descartando una variable que es de hecho relevante. El punto crucial es que por construcción los tests no son completamente informativos acerca de la verdad o falsedad de la hipótesis nula, por consiguiente, la práctica de "descartar variables *estadísticamente no significativas*" (en contraposición con descartar variables no significativas en términos teóricos) puede conducirnos a un sesgo por variables omitidas. Este es un asunto muy delicado en el cual la meta de realizar inferencia con un modelo dado interactúa con la construcción o el descubrimiento de un modelo relevante.

5.2. Regresores estocásticos

Uno de los supuestos clásicos indica que las variables explicativas son tratadas como "no estocásticas", es decir, como si fueran números fijos. Este es un supuesto apropiado, por ejemplo, en disciplinas experimentales. A manera de ejemplo, supongamos que podemos asignar a distintos individuos diferentes dosis de una droga (X) y observar un resultado asociado a este efecto (Y), y supongamos que este resultado está linealmente relacionado con X más un término de error. En dicha situación tenemos control sobre la variable X, de hecho, podríamos repetir este experimento tantas veces como quisiéramos para los mismos X's. Esta es la noción de que "X está fijo en muestras repetidas" que aparece en muchos trabajos estadísticos sobre el tema. Pero en economía y en muchas otras ciencias sociales rara vez los datos surgen de tal manera. Tomemos por ejemplo el caso de la educación y los salarios. Un "experimento" del tipo descripto anteriormente asignaría a individuos recientemente nacidos diferentes niveles de educación y eventualmente observaríamos cuáles son sus salarios (¡esperando que los demás factores se mantengan constantes!). En lugar de eso, lo que observamos son pares educación-salario que surgen de elecciones hechas por los individuos. En dicho caso, el investigador no tiene "control" sobre las variables explicativas. Tanto X como Y deberían ser tratadas como aleatorias. Esta sección explora una formulación básica acerca de la naturaleza estocástica de X que es coherente con los modelos vistos en los capítulos anteriores.

El modelo lineal con regresores estocásticos o aleatorios se especifica de la siguiente manera:

$$Y = X\beta + \mu$$

suponiendo:

1. X puede ser una matriz estocástica con $\phi(X) = K$. Ahora estamos dejando que X esté formada por variables aleatorias. Los datos son vistos como realizaciones de dichas variables aleatorias. El requerimiento del rango se refiere a la matriz de datos más que a las variables aleatorias en sí mismas.

2. $E(\mu|X) = 0$. Este es uno de los cambios principales con respecto al modelo básico. Cuando X es estocástica necesitamos que dado X, el valor esperado de μ sea cero. Visto desde otra perspectiva, estamos pidiendo que la información contenida en X no altere el hecho de que el valor esperado de μ es cero.

3. $V(\mu|X) = \sigma^2 I$. En este supuesto requerimos que no haya heterocedasticidad ni correlación serial, condicional a los valores que tome X.

El estimador de mínimos cuadrados ordinarios será:

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

Mostraremos que bajo estos supuestos sigue siendo insesgado:

$$\begin{aligned}
\hat{\beta} &= \beta + (X'X)^{-1}X'\mu \\
E(\hat{\beta}) &= \beta + E[(X'X)^{-1}X'\mu] \\
&= \beta + E[(X'X)^{-1}X'E(\mu|X)] \quad \text{(Usando la Ley de las Esperanzas Iteradas)} \\
&= \beta \quad \text{(Ya que } E(\mu|X) = 0\text{)}
\end{aligned}$$

Es importante clarificar lo que estamos tratando de hacer. No estamos usando una técnica de estimación distinta para el caso de regresores aleatorios. Estamos buscando una manera de definir los supuestos de manera que nos permitan usar la estructura de mínimos cuadrados ordinarios cuando X es estocástica. Esto significa que no necesitamos de una estrategia de estimación e inferencia distinta al permitir que X sea aleatoria. Lo que necesitamos es una manera diferente de expresar el modelo y los supuestos detrás de él, de forma tal que podamos usar las herramientas aprendidas previamente. El "truco" es reemplazar los supuestos de μ por supuestos de μ condicional en X.

Para explorar con mayor detalle lo que el supuesto de $E(\mu|X) = 0$ implica, volvamos al caso de dos variables con intercepto y una sola variable explicativa. En ese caso, la matriz X tiene una primer columna de unos y una segunda columna correspondiente a las observaciones de una variable que llamaremos Z. En este caso $E(\mu|X) = 0$ significa que el valor esperado de μ es igual a 0 condicional en la constante y en Z. Primero, notemos que por la Ley de las Esperanzas Iteradas esto implica que $E(\mu) = 0$. Segundo, consideremos la covarianza entre X y μ:

$$\begin{aligned}
Cov(Z, \mu) &= E[(Z - E(Z))(\mu - E(\mu))] \\
&= E(Z\mu) - E(Z)E(\mu) \\
&= E(Z\mu) \\
&= E(ZE(\mu|Z) \\
&= 0
\end{aligned}$$

Por lo tanto, el supuesto $E(\mu|X)$ implica que el término de error no debe estar correlacionado con las variables del modelo. Alternativamente, si el término de error está correlacionado con las variables explicativas, el estimador MCO será sesgado.

En un contexto diferente hemos encontrado una situación similar a la de omisión de variables relevantes. En dicho caso requeríamos que las variables omitidas no estuvieran linealmente relacionadas con aquellas incluidas en el modelo. Cuando las variables explicativas son tomadas como aleatorias, las conclusiones acerca de la omisión de variables relevantes son más fáciles de ver. Si el verdadero modelo es:

$$Y = X_1\beta_1 + X_2\beta_2 + \mu$$

y erróneamente procedemos como si $\beta_2 = 0$, entonces el modelo estimado puede ser visto como:

$$Y = X_1\beta_1 + \mu^*$$

con $\mu^* = X_2\beta_2 + \mu$. Entonces, el modelo que omite variables relevantes tiene en su término de error parte del verdadero modelo. Este sesgo surge, precisamente, cuando $\mu*$ está linealmente relacionada con X_1, es decir cuando X_2 está relacionada con X_1.

La siguiente sección explora otra situación en la que las variables explicativas son por definición aleatorias, y provee otro ejemplo de variables explicativas correlacionadas con el término de error que conducen a estimaciones sesgadas bajo MCO.

5.3. Errores de medición en las variables

En la sección anterior discutimos la posible aleatoriedad de X desde un terreno "filosófico", es decir, como una consecuencia de la naturaleza no experimental de los datos en las ciencias sociales. Pero en ciertos casos, las variables explicativas estocásticas surgen de una manera más simple y mecánica, por ejemplo, cuando se sabe que una variable está incorrectamente medida. En esta sección vamos a distinguir entre errores de medición en las variables explicativas y en las variables explicadas.

Errores de medición en las variables explicativas

Supongamos que dos variables Y y X^* están linealmente relacionadas como en el modelo simple de dos variables, para una muestra de n observaciones:

$$Y_i = \alpha + \beta X_i^* + \mu_i$$

Si todos los supuestos clásicos se cumplen, obtendríamos un estimador MELI al computar por MCO, regresando Y_i en X_i^*. En cambio, supongamos que X_i^* no es directamente observable, sino que observamos una versión "ruidosa":

$$X_i = X_i^* + \omega_i$$

es decir, la variable explicativa está *medida con un término de error* siendo ω_i el error en la medición. Es crucial darnos cuenta que no observamos X_i^* ni ω_i sino X_i, que es una combinación lineal de ambas. Por simplicidad asumiremos que $E(\omega_i = 0)$, $V(\omega_i) = \sigma_\omega^2$, $Cov(\omega_i, \omega_j) = 0$ para todo $i \neq j$, y $Cov(\omega_i, \mu_j) = 0$ para todo i, j. Es decir, el error de medición tiene valor esperado igual a cero, varianza constante y no está correlacionada serialmente. Además, supondremos que no está linealmente relacionado con el término de error original μ_i.

En caso de ignorar el error de medición, estimaríamos un modelo como:

$$Y_i = \alpha + \beta X_i + \nu_i$$

es decir, procederíamos relacionando a las Y_i con las X_i, la versión observable de X_i^*, para algún término de error ν_i.

Para explorar qué es ν_i, reemplacemos X_i^* en el modelo original:

$$\begin{aligned} Y_i &= \alpha + \beta(X_i - \omega_i) + \mu_i \\ &= \alpha + \beta X_i + (\mu_i - \beta\omega_i) \\ &= \alpha + \beta X_i + \nu_i \end{aligned}$$

Pero notemos que:

$$\begin{aligned} Cov(\nu_i, X_i) &= E[(\nu_i - E(\nu_i))(X_i - E(X_i))] \\ &= E[(\mu_i - \beta\omega_i)\omega_i] \\ &= E(\mu_i\omega_i - \beta\omega_i^2) \\ &= E(-\beta\omega_i^2) \\ &= -\beta\sigma_\omega^2 \neq 0 \end{aligned}$$

En consecuencia, cuando la variable explicativa es medida con un término de error, y usamos la variable "incorrectamente medida", el estimador de MCO conduce a estimaciones sesgadas. Para ver este resultado de forma más intuitiva, supongamos que X_i^* es un regresor aleatorio, no correlacionado con el error de medición ni con el término del error original. Entonces:

$$\begin{aligned} V(X_i) &= V(X_i^*) + V(\omega_i) \\ \sigma_X^2 &= \sigma_{X^*}^2 + \sigma_\omega^2 \end{aligned}$$

Se puede demostrar que para valores grandes de n:

$$\hat{\beta} \simeq \beta[\frac{\sigma_{X^*}^2}{\sigma_\omega^2 + \sigma_{X^*}^2}]$$

Entonces, las discrepancias entre $\hat{\beta}$ y β dependen de qué tan largo es el error en la medición en términos de su varianza σ_ω^2. Cabe notar que el término entre corchetes es un número positivo, por lo tanto, el error de medición hace que $\hat{\beta}$ sea más *pequeño* en términos absolutos que β. Tomando esto en cuenta, el sesgo que surge como consecuencia de usar variables explicativas mal medidas es llamado *sesgo de atenuación*.

Por lo tanto, debemos tener cuidado al interpretar los resultados de una estimación por MCO cuando las variables no son significativas. Bajo esta perspectiva, valores pequeños de $\hat{\beta}$ (en términos absolutos) y por consiguiente valores "t" pequeños, pueden surgir, entre otras causas, porque

la variable verdaderamente no es relevante, o porque está mal medida. El comentario apunta a pensar cuidadosamente si las variables explicativas están sujetas a errores de medición, ya que esto puede erróneamente conducir a la conclusión de que una variable no es significativa en la relación.

Errores de medición en la variable explicada

Ahora vamos a explorar qué es lo que pasa cuando la variable explicada está medida con algún término de error. Supongamos que la verdadera relación es la siguiente:

$$Y_i^* = \alpha + \beta X_i + \mu_i$$

pero Y_i^* no es observable directamente. En cambio, podemos observar:

$$Y_i = Y_i^* + \epsilon_i$$

donde ϵ_i es un error de medición que satisface las siguientes propiedades: $E(\epsilon_i) = 0$, $V(\epsilon_i) = \sigma_\epsilon^2$, $Cov(\epsilon_i, \epsilon_j) = 0$ para todo $i \neq j$, y $Cov(\epsilon_i, \mu_j) = 0$ para todo i, j. Como en el caso anterior, estamos asumiendo que el término de error tiene media cero, varianza constante, y no está correlacionado serialmente y tampoco tiene relación lineal con el término de error original. Reemplazando Y_i^* en el modelo obtenemos:

$$\begin{aligned} Y_i &= \alpha + \beta X_i + \mu_i + \epsilon_i \\ &= \alpha + \beta X_i + \nu_i \end{aligned}$$

que es la versión "estimable".

Cabe notar lo siguiente: Dado que $E(\mu_i) = E(\epsilon_i)$ entonces $E(\nu_i) = 0$, por consiguiente el estimador de MCO es insesgado, ya que el término de error del modelo estimable tiene media cero. Por lo tanto, al contrario del caso en dónde la variable explicativa está medida con un término de error, cuando la variable medida con error es la explicada, el estimador de MCO no es sesgado.

Adicionalmente, notemos que bajo nuestros supuestos:

$$V(\nu_i) = V(\mu_i + \epsilon_i) = \sigma_\mu^2 + \sigma_\epsilon^2$$

Luego:

$$V(\hat{\beta}) = \frac{\sigma_\nu^2}{\sum x_i^2} = \frac{\sigma_\mu^2 + \sigma_\epsilon^2}{\sum x_i^2} > \frac{\sigma_\mu^2}{\sum x_i^2}$$

entonces, la presencia de error en la medición conduce a estimadores más imprecisos, es decir, la varianza del estimador de la pendiente del modelo es más grande que la que obtendríamos

en ausencia del error de medición. Intuitivamente, lo que pasa es que el error de medición de la variable explicada es absorbida en el término de error global, por lo que su efecto es el mismo que el de incrementar la varianza del término de error original.

En resumen, el efecto de los errores de medición depende de si ellos están en la variable explicativa o en la variable explicada. El primer caso conduce a sesgos en el estimador de MCO, y el segundo caso lo hace menos eficiente. Un estimador alternativo para el primer caso será propuesto una vez que introduzcamos el estimador de variables instrumentales.

Apéndice: Ley de las Esperanzas Iteradas

Sea Y una variable aleatoria continua con función de densidad $f(y)$. Usamos Y para denotar a la variable aleatoria e y para denotar los valores que puede tomar, es decir, su *soporte*. Recordemos que el valor esperado se define como:

$$E(Y) = \int yf(y)dy$$

donde la integral corre sobre la recta real. Omitimos los límites de integración por simplicidad. Ahora, sean X e Y dos variables aleatorias continuas conjuntamente distribuidas con densidad $f(x, y)$. Las densidades marginales de Y y X están dadas respectivamente por:

$$f_Y(y) = \int f(x, y)dx$$

y:

$$f_X(x) = \int f(x, y)dy$$

Entonces, el valor esperado de Y será:

$$\begin{aligned} E(Y) &= \int yf_Y(y)dy \\ &= \int\int yf(x, y)dxdy \end{aligned}$$

También definimos la densidad condicional de Y en X como:

$$f(y|x) = \frac{f(x, y)}{f_X(x)}$$

Y la esperanza condicionada de Y dado $X = x$ como:

$$E(Y|x) = \int yf(y|x)dy$$

Recordemos que si esta esperanza condicional existe para cada x, entonces $E(Y|x)$ será una función que depende de x. Entonces, cuando $x = X$, $E(Y|X)$ será una *variable aleatoria* (Para más detalles ver Rice, 1994, pp.135-139).

Dos propiedades de las esperanzas condicionales que resultarán de utilidad son:

- $E(X|X) = \int Xf(y|X)dy = X$. Este es un resultado más bien trivial pero útil que usaremos repetidamente.
- Si $Y = a + bX + cU$, entonces $E(Y|X) = a + bX + cE(U|X)$. Intuitivamente nos dice que cuando Y es una función lineal de X, al computar la esperanza de Y condicional en X, tratamos a X como si fuera una 'constante'.

Ahora estamos listos para explorar el resultado principal de este apéndice, la *ley de esperanzas iteradas*:

$$\begin{aligned} E(Y) &= \int\int yf(x,y)dydx \\ &= \int\int yf(y|x)f_X(x)dydx \\ &= \int[\int yf(y|x)dy]f_X(x)dx \\ &= \int E(Y|x)f_X(x)dx \\ &= E[E(Y|X)] \end{aligned}$$

Esta ley nos provee un camino alternativo para calcular las esperanzas. Intuitivamente nos dice que para computar el valor esperado de Y podemos proceder directamente (calcular E(Y)) o indirectamente, primero condicionando a Y en alguna otra variable, y después calculando la esperanza de su esperanza. Pensemos en el siguiente caso. Supongamos que hay igual número de hombres y de mujeres en una clase y estamos interesados en encontrar la edad promedio. Podríamos calcular el promedio directamente, o primero dividir la clase en hombres y mujeres, computar la edad promedio de cada sexo y posteriormente computar el promedio de los promedios. Cualquiera de los dos caminos nos llevaría al mismo número, la edad promedio.

Capítulo 6

Modelos de Elección Binaria

En esta sección estudiaremos modelos estadísticos para casos en los que la variable dependiente toma solamente dos valores. Por ejemplo, un individuo compra un producto o no, un ente regulador adopta una política regulatoria o no, un alumno es admitido a un programa de posgrado o no, etc.

6.1. Motivación

Consideremos el siguiente caso. Estamos interesados en conocer qué factores determinan la admisión de un alumno a un programa de posgrado. A tal efecto, se dispone de una base de datos de 100 personas que solicitaron admisión a un programa. Por cada alumno la información es la siguiente: una variable binaria indicando si la persona en cuestión fue admitida o no al programa (uno si fue admitida, cero si no lo fue), calificación obtenida en los exámenes GRE y TOEFL[1], promedio de notas en la carrera de grado (PROM). Estos datos son ficticios y han sido estandarizados de modo que la nota mínima en cada examen es cero y la máxima diez. Un listado de los primeros 10 alumnos es el siguiente:

[1]El GRE (*Graduate Record Examinations*) es un examen de habilidades de razonamiento verbal y analítico, así como de escritura y pensamiento crítico. El TOEFL (*Test of English as a Foreign Language*) por su parte, es un examen que evalúa el dominio del idioma inglés para no nativos. Ambas pruebas son requeridas en el proceso de admisión de la mayoría de los programas de posgrado.

obs	y	gre	prom	toefl
1	1	5.8043	7.3873	2.8336
2	1	7.0583	0.1907	9.6828
3	0	1.2374	2.5869	1.6566
4	1	4.8138	9.6222	7.4295
5	0	1.2842	0.5603	9.3139
6	1	9.2899	7.0723	7.3215
7	0	4.0955	0.0774	3.0129
8	0	7.0242	1.3122	8.2629
9	1	8.8110	5.5499	8.5857
10	1	4.5807	5.9019	5.2783

A través de los modelos presentados en el capítulo intentaremos responder preguntas como estas:

1. Dado un determinado nivel de comparación en la calificaciones obtenidas en el GRE, TOEFL y en las notas en la carrera de grado, Cuál es la probabilidad de ser admitido al programa?
2. En cuánto mejoran las chances de admisión si mejora la nota en el GRE?
3. Es realmente importante el TOEFL en el proceso de admisión?
4. Cuán confiables son las respuestas a las preguntas anteriores?
5. Aun conociendo las notas de los exámenes y de la carrera de grado, no somos capaces de predecir con exactitud si un alumno será admitido o no. Cuál es el origen de esa aleatoriedad y cómo puede ser tratada e interpretada?

6.2. Modelos de elección binaria

Denotemos con Y a una variable aleatoria que puede tomar solo dos valores, uno o cero y que puede ser asociada a la ocurrencia de un evento. Se dispone de una muestra aleatoria de n observaciones $Y_i \quad \text{con} \quad i = 1, \ldots, n$. Llamemos X_i al vector de k variables explicativas asociado al individuo i, el cual será utilizado para "explicar" la variable Y_i.

Un *modelo de elección binaria* es un modelo de la probabilidad de ocurrencia del evento denotado por Y_i condicional en el conjunto de información X_i.

$$P_i = Pr(Y_i = 1|X_i)$$

Es importante notar que dado que Y_i toma solo los valores cero y uno, esta probabilidad condicional es también la esperanza de Y_i condicional en X_i:

$$E(Y_i|X_i) = 1P_i + 0(1 - P_i) = P_i$$

Supongamos que X_i está constituido por un vector fila de k variables explicativas incluyendo un "uno" para incorporar la ordenada al origen en el modelo ($X_{1i} = 1$). Un primer intento de modelación podría consistir en postular una relación lineal entre Y_i y X_i, por ejemplo:

$$Y_i = X_i\beta + u_i \quad \text{con} \quad E[u_i|X_i] = 0$$

por lo tanto, aplicando esperanza

$$\begin{aligned} E[Y_i|X_i] &= E[X_i\beta + u_i|X_i] \\ E[Y_i|X_i] &= E[X_i\beta|X_i] + E[u_i|X_i] \end{aligned}$$

dado que X_i es no estocástica y β es un parámetro, y que $E[u_i|X_i] = 0$ se tiene:

$$E[Y_i|X_i] = P_i = X_i\beta$$

expresando esto mismo en forma extensiva tenemos que:

$$E[Y_i|X_i] = P_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

En este caso el vector de parámetros β podría ser consistentemente estimado utilizando el método de mínimos cuadrados ordinarios. Este proceso consistirá simplemente en regresar el vector de ceros y unos de las realizaciones de Y en las variables explicativas.

Sin embargo, esta especificación lineal presenta un serio problema: en nuestro caso $E[Y_i|X_i]$ es también una probabilidad condicional, por lo cual debería estar restringida a tomar valores entre cero y uno. El modelo lineal no impone ninguna restricción sobre $X_i\beta$, y en consecuencia podría predecir valores negativos o mayores que uno para una probabilidad. Además, es fácil observar que el término de error de este modelo no es homoscedástico ya que la varianza condicional varía según las observaciones[2].

$$\begin{aligned} Var(u_i|X_i) &= Var(Y_i|X_i) \\ &= E(Y_i^2|X_i) - E(Y_i|X_i)^2 \\ &= P_i - P_i^2 \quad \text{siendo} \quad E(Y_i^2|X_i) = 1^2 P_i + 0^2(1 - P_i) \\ &= P_i(1 - P_i) \\ &= X_i\beta(1 - X_i\beta) \end{aligned}$$

Por lo tanto, depende de X_i y en consecuencia, es heterocedástico.

[2] Así y todo, el modelo lineal de probabilidad ha sido recientemente revitalizado en Heckman y Snyder (1997).

Hay una tercera razón –más intuitiva– acerca de por qué deberíamos abandonar el modelo de probabilidad lineal y es que las derivadas parciales de este tipo de modelos son constantes. Es decir, el efecto marginal de X_k sobre P_i es una constante, lo cual no es una característica deseable para un modelo que evalúa la probabilidad de ocurrencia de un evento dado un conjunto de características.

Consideremos el siguiente ejemplo. Supongamos que Y es una variable que toma el valor 1 si un niño asiste a la escuela y 0 si no lo hace y que estamos interesados en medir el efecto del ingreso familiar sobre la posibilidad de que la familia envíe al niño a la escuela. Bajo un modelo de probabilidad lineal este efecto marginal no depende del ingreso. En consecuencia, un pequeño cambio en el mismo tendría igual efecto para cada familia, independientemente de su ingreso. Pero es natural pensar que para una familia extremadamente rica, un cambio menor en el ingreso generará un efecto pequeño (si es que genera alguno) en la decisión de enviar a sus niños a la escuela. Lo mismo ocurre para una familia extremadamente pobre; un pequeño cambio en el ingreso será destinado probablemente a la satisfacción de necesidades más urgentes como alimentos o vestimenta. Sin embargo, como puede verse en la Figura 6.1 a medida que nos movemos en la escala de ingreso, el efecto marginal de éste tendrá probablemente un efecto más fuerte. Este patrón de menores efectos marginales en los extremos es claramente incompatible con el modelo lineal. Muchos otros ejemplos parecen responder a este patrón (liquidez y quiebra de bancos, la calificación GRE y la admisión al programa de posgrado, etc.).

6.3. Logits y Probits: modelos de índices transformados

A la luz de la discusión anterior, deberíamos adoptar un tipo de especificación bajo la cual los valores de P_i estén restringidos al intervalo $[0, 1]$. Una manera muy conveniente de restringir la forma funcional es la siguiente:

$$P_i = F(X_i\beta)$$

en donde la función $F(.)$ tiene las siguientes propiedades

$$F(-\infty) = 0, \quad F(\infty) = 1, \quad f(x) = dF(x)/dx > 0$$

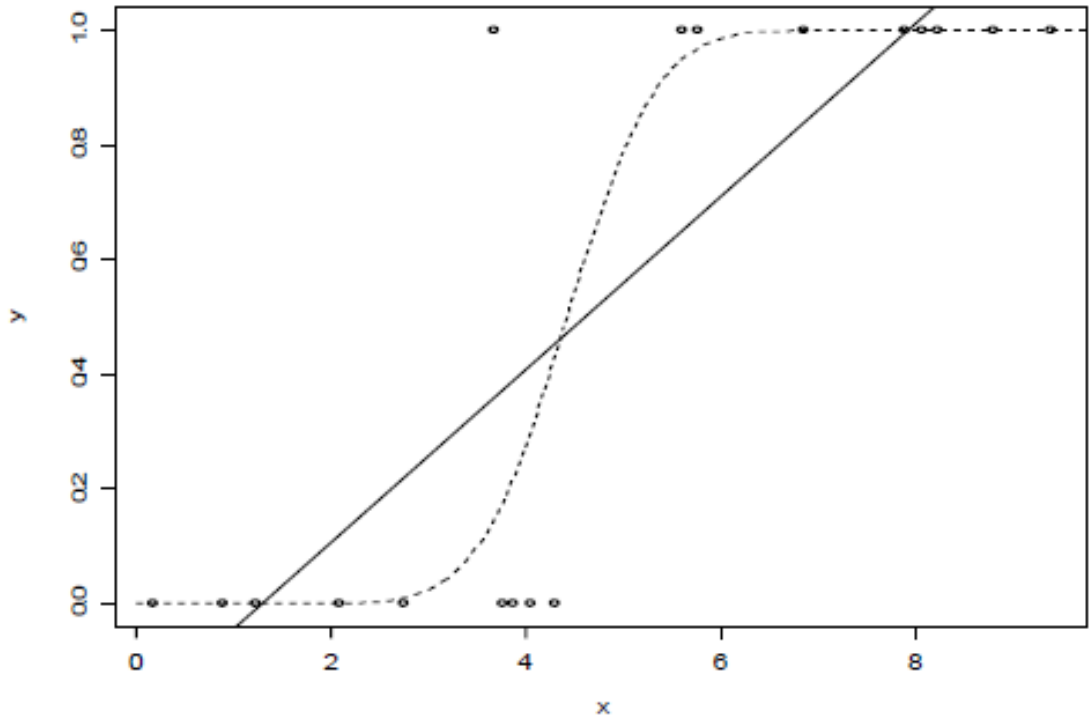
O sea, $F(.)$ es una función diferenciable monótona creciente con dominio real y rango $[0, 1]$. De esta forma nuestro modelo no lineal sería el siguiente:

$$y_i = F(X_i\beta) + u_i \tag{6.1}$$

Observemos más detenidamente algunas características de la función $F(X_i\beta)$:

1. Obviamente se trata de una función no lineal, pero una muy particular, en el sentido de que las variables explicativas afectan a la variable dependiente a través de un *índice lineal* $(X_i\beta)$

Figura 6.1: Modelo de probabilidad lineal vs no lineal



que luego es transformado por la función $F(.)$ de manera tal que los valores de la misma sean consistentes con los de una probabilidad[3].

2. Cómo elegir la función $F(.)$? Nótese que la función de distribución de cualquier variable aleatoria contínua tiene las propiedades de $F(.)$. En esta dirección es que buscaremos una forma funcional para $F(.)$.

Una primer forma funcional que satisface nuestros requisitos es la correspondiente a la función de distribución normal:

$$P_i = F(X_i\beta) = \Phi(X_i\beta) = \int_{-\infty}^{X_i\beta} \phi(s)ds$$

en donde $\phi(.)$ es la función de densidad normal estándar. Esta especificación de $F(.)$ utilizando la función de distribución normal es la que se denomina *probit*. Otra alternativa comúnmente utilizada se basa en la *distribución logística*:

$$P_i = F(X_i\beta) = \Lambda(X_i\beta) = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}}$$

[3]Desde este punto de vista, el modelo binario pertenece a una familia mas general conocida en la literatura como Modelos Lineales Generalizados. La referencia clásica es McCullagh y Nelder (1993).

y corresponde al modelo *logit*. En la siguiente sección discutiremos una interpretación mas intuitiva de estos modelos.

6.4. Cómo se interpretan los parámetros del modelo binario?

Habiendo especificado la función $F(.)$ nos resta estimar sus parámetros, el vector β. Posterguemos por un instante el proceso de estimación de dichos parámetros (supongamos que disponemos de estimaciones de los mismos) y concentrémonos en su interpretación. Un primer objetivo consiste en medir cómo se altera la probabilidad condicional de ocurrencia del evento cuando cambia marginalmente alguna de las variables explicativas. Tomando derivadas parciales en nuestra función de probabilidad condicional[4]:

$$\frac{\partial P_i}{\partial X_k} = \beta_k \, f(X_i\beta) \tag{6.2}$$

De acuerdo a (6.2), el efecto marginal tiene dos componentes multiplicativos: el primero indica cómo un cambio en una variable explicativa afecta al índice lineal $X_i\beta$ y el segundo muestra cómo la variación en el índice se manifiesta en cambios en la probabilidad a través de cambios en la función $F(.)$.

En términos de nuestro ejemplo original, y habiendo encontrado que la calificación en el GRE es una variable significativa en el modelo, de acuerdo a lo anterior estaríamos diciendo que mejorar el rendimiento en el GRE implicaría una mejora sustantiva en la probabilidad de admisión para individuos cerca de la media de la distribución, lo cual tiene bastante sentido. Una mejora marginal en el GRE no debería modificar sustancialmente la situación de un individuo muy malo o demasiado bueno.

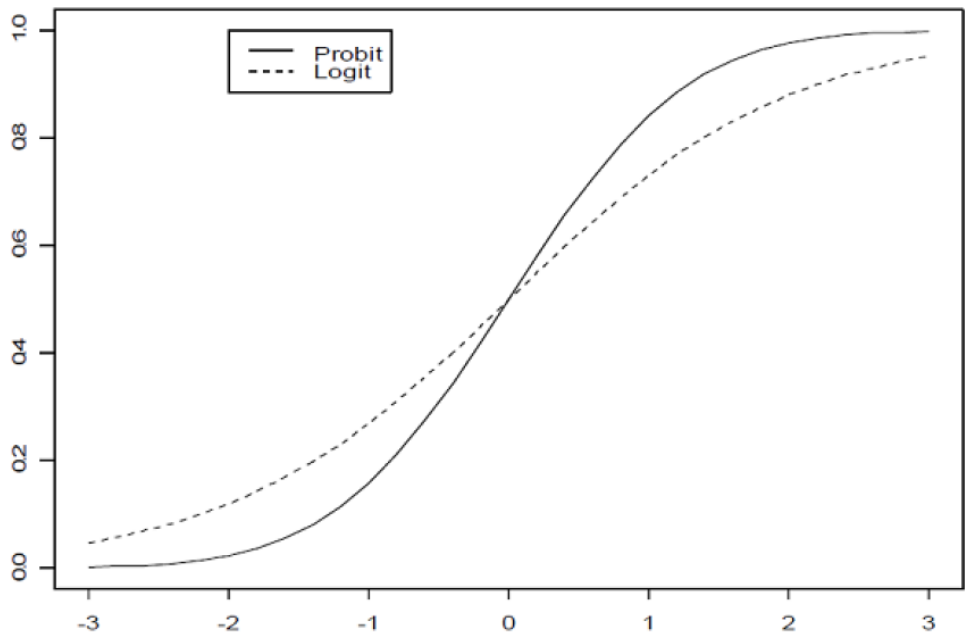
6.5. Logit o Probit?

Una simple inspección visual de la Figura 6.2 permite observar que no existen mayores diferencias entre la distribución logística y la normal, excepto en las colas de la distribución. En la práctica, y con tamaños de muestra no demasiado grandes, los modelos logit y probit tienden a producir resultados muy similares, siendo la única diferencia relevante la forma en la que los coeficientes se encuentran escalados. Esto se debe a que la varianza de una variable aleatoria con distribución logística es $\pi^2/3$ (Anderson, et.al.(1992), p.35) mientras que la normal estándar tiene varianza uno, lo que hace que los coeficientes del modelo logit sean en general mayores que los del probit. En la práctica, usualmente, se multiplican los coeficientes del modelo logit por $\pi/\sqrt{3}$ para poder com-

[4]Para el caso en que x_j sea una variable binaria, el efecto parcial del cambio de 0 a 1 en x_j manteniendo todo lo demás constante será: $F(\beta_0 + \beta_1 x_1 + \ldots + \beta_j + \beta_k x_k) - F(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)$

pararlos con los del modelo probit (ver Figura 6.3). Amemiya(1981), basándose en un método de prueba y error, propone como factor 1/1,6.

Figura 6.2: Logit vs Probit

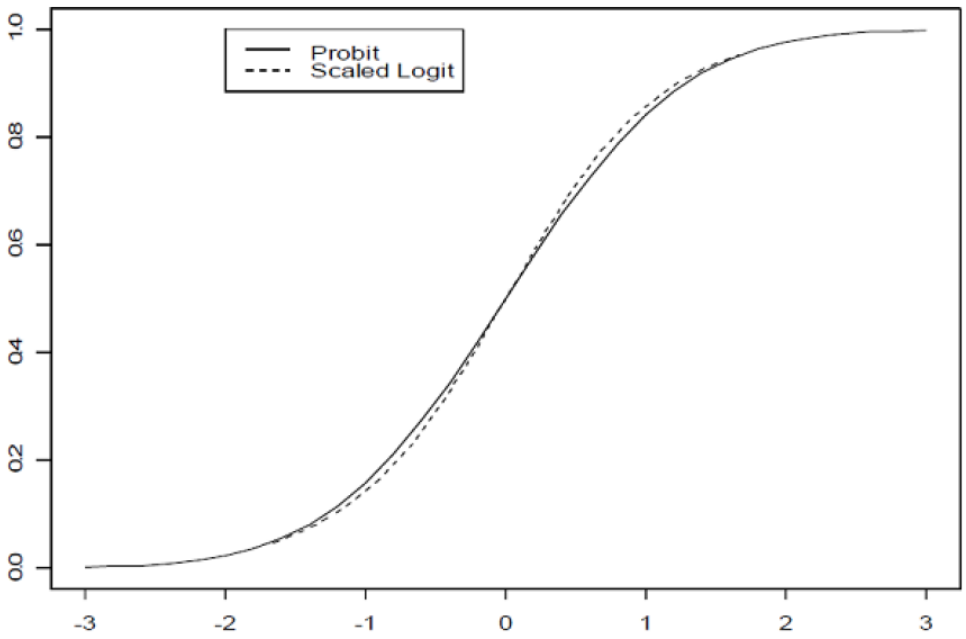


6.6. Estimación e inferencia

El vector de parámetros poblacionales β no es observable y por lo tanto debe ser estimado mediante una muestra (Y_i, X_i), con $i = 1, \ldots, n$. Llamaremos $\hat{\beta}$ al vector que contiene las estimaciones muestrales de β. Por ejemplo, para estimar el modelo lineal de probabilidad se puede utilizar el modelo de MCO utilizando a Y_i como variable dependiente y a las X_i como variables explicativas. Sin embargo, el caso de los modelos probit y logit requiere otro tipo de métodos que no serán contemplados en este curso.

Sin importar cuál sea el método de estimación, el hecho de que se utilice información proveniente de una muestra de la población hace que $\hat{\beta}$ también sea una variable aleatoria, y por lo tanto es necesario conocer su distribución muestral para realizar distintos ejercicios de inferencia estadística sobre los valores de β. Si bien no es el objetivo de este capítulo ahondar sobre las propiedades estadísticas de los estimadores logit y probit, se puede demostrar que cuando el tamaño de la muestra es grande, la distribución normal provee una aproximación adecuada. Es decir, $\hat{\beta}$ es

Figura 6.3: Logit vs Probit reescalado



aproximadamente $N(\beta, V(\hat{\beta}))$. Esto permite utilizar los procedimientos usuales de inferencia.

Un test de la hipótesis nula de que un coeficiente es cero frente a la alternativa de que no lo es, se basa en un (pseudo) estadístico 't':

$$t_k = \frac{\hat{\beta}_k}{\hat{se}(\hat{\beta}_k)} \sim N(0,1)$$

en donde $\hat{\beta}_k$ es el estimador de máxima verosimilitud del coeficiente de la k-ésima variable explicativa y $\hat{se}(\hat{\beta}_k)$ es el la raíz cuadrada del estimador de la varianza de $\hat{\beta}_k$. El resultado es asintótico en el sentido de que la distribución normal provee una buena aproximación para muestras grandes.

6.7. Extensiones

En estas notas hemos presentado los elementos básicos del análisis de datos binarios. La siguiente es una lista (incompleta) de extensiones y referencias básicas que deberían aprender en un curso más avanzado:

1. Es necesario aclarar que si bien es posible estimar el modelo de probabilidad lineal mediante MCO, la naturaleza no lineal de $E(Y|X)$ hace que este método no sea aplicable para probit

y logit. Para estos casos se utiliza el método de Máxima Verosimilitud. Se puede demostrar que bajo los supuestos clásicos el estimador MCO es el estimador de máxima verosimilitud, condicionado en las variables explicativas.

2. El problema de heteroscedasticidad tiene consecuencias un tanto más graves que en el caso del modelo lineal general. Es posible demostrar (Yatchew y Griliches, 1984) que bajo la presencia de heteroscedasticidad el estimador de máxima verosimilitud es inconsistente.

3. Como hemos visto hasta aquí, los modelos logit y probit son más difíciles de estimar y de interpretar que el modelo de probabilidad lineal. A su vez, vimos que este último no induce una estructura muy distinta a estos modelos no lineales y su naturaleza más simple permite trabajar con paneles y con variables instrumentales. Es por ello que no resultan tan claras las ventajas de usar logit y probit.

6.8. Aplicaciones

6.8.1. Modelo para la probabilidad de admisión a un doctorado

De acuerdo al análisis anterior, el modelo econométrico adoptado para la probabilidad de que un individuo sea admitido a un programa de doctorado, condicional en su calificación en el GRE, TOEFL y notas de la carrera (PROM), se puede expresar de la siguiente manera:

$$P_i = F(X_i\beta)$$

con:

$$X_i\beta = \beta_0 + \beta_1 \text{GRE}_i + \beta_2 \text{TOEFL}_i + \beta_3 \text{PROM}_i$$

en donde hemos incluido una constante como variable explicativa adicional. El método de estimación es máxima verosimilitud basado en una muestra de 100 observaciones. La siguiente tabla presenta algunos resultados de la estimación logit y probit.

Resultados para el modelo de admisión

	Logit			Probit			
	Coeff	Std. Err.	t	Coeff	Std.Err.	t	Logit/1.6
INTERC	-10.2431	2.4237	-4.2263	-5.6826	1.2034	-4.7223	-6.4020
GRE	1.0361	0.2568	4.0343	0.5839	0.1308	4.4647	0.6476
PROM	1.0821	0.2425	4.4624	0.5946	0.1206	4.9321	0.6763
TOEFL	-0.0351	0.1252	-0.2806	-0.0169	0.0697	-0.2427	-0.0220

L1 = 38.66597, gl=2, p=4.0159e-009

Las columnas 2-4 presentan coeficientes del modelo logit y las columnas 5-7 los del modelo probit. La columna 8 presenta los coeficientes del modelo logit divididos por 1.6. El primer resultado interesante es que de acuerdo al estadístico 't', la hipótesis nula de que el coeficiente de la variable TOEFL es cero no puede ser rechazada. El resto de los coeficientes es, de acuerdo al mismo test, significativamente distintos de cero y presentan los signos esperados: un aumento en el GRE o en PROM aumentan la probabilidad de admisión. Cuando los coeficientes del modelo logit son reescalados presentan valores similares a los del modelo probit.

En síntesis, de acuerdo con nuestro modelo empírico, el TOEFL no es una variable relevante en el proceso de admisión, mientras que el resultado de GRE y el promedio sí lo son.

En la siguiente tabla calculamos las derivadas reemplazando los resultados obtenidos en la regresión. Estas derivadas son evaluadas en las medias de las variables independientes.

Derivadas calculadas en las medias

	Probit	Logit	Means
GRE	0.1993	0.1977	4.5570
PROM	0.2028	0.2065	4.2760
TOEFL	-0.0058	0.0067	4.8300

Nótese que los modelos logit y probit producen resultados muy similares. En el caso del GRE, nuestro modelo predice que para un individuo con GRE, PROM y TOEFL igual al promedio, un incremento marginal en la nota del GRE aumentará la probabilidad de ser admitido en casi 0,2. Un valor muy similar es obtenido para el caso de PROM. El valor obtenido para el TOEFL no es interpretable dado que el coeficiente asociado con dicha variable no es significativamente distinto de cero, de acuerdo a los resultados obtenidos anteriormente.

6.8.2. Modelo para la probabilidad de desempleo

Consideremos un modelo para la probabilidad de que un individuo caiga en una situación de desempleo (Y vale 1 si está desocupado y 0 si no lo está). Podemos pensar en un modelo simple en donde dicha probabilidad depende de sus características personales: la edad, el género, la situación marital, su educación, el rol en el hogar, el ingreso no laboral y la localización geográfica. En la siguiente tabla mostramos una estimación de Stata con datos de la Encuesta Permanente de Hogares (EPH) para el Gran Buenos Aires:

```
logit desocupado edad mujer soltero educ jefe inla caba
```

```
Logistic regression                               Number of obs   =       6669
                                                  LR chi2(7)      =     189.95
                                                  Prob > chi2     =     0.0000
Log likelihood = -1615.0284                       Pseudo R2       =     0.0555
```

desocupado	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
edad	-.0109965	.0048867	-2.25	0.024	-.0205744	-.0014187
mujer	.2553919	.1050513	2.43	0.015	.0494951	.4612887
soltero	.2554484	.1181695	2.16	0.031	.0238404	.4870564
educ	-.100993	.013732	-7.35	0.000	-.1279073	-.0740787
jefe	-.6682848	.1161409	-5.75	0.000	-.8959168	-.4406529
inla	.000687	.0000901	7.63	0.000	.0005105	.0008635
caba	-.090867	.1227719	-0.74	0.459	-.3314955	.1497614
_cons	-.9996536	.2698181	-3.70	0.000	-1.528487	-.4708199

A diferencia de los modelos de Mínimos Cuadrados Ordinarios (MCO), estos modelos tienen que ser interpretados cuidadosamente. Por empezar, dado que los valores de estos coeficientes no tienen una interpretación cuantitativa (solo es interpretable el signo de los mismos), en el Trabajo Práctico les pedimos que calculen los efectos marginales de cada variable para realizar una interpretación cuantitativa del efecto de cada variable sobre la probabilidad de estar desocupado. Por ejemplo, en el inciso (e) computamos los efectos marginales para un hombre casado que es jefe de familia, con 17 años de educación, edad e ingreso no laboral promedio y que resida en la Ciudad Autónoma de Buenos Aires (CABA).

```
Marginal effects after logit
      y  = Pr(desocupado) (predict)
         =  .02056653
```

variable	dy/dx	Std. Err.	z	P>|z|	[	95% C.I.	]	X
edad	-.0002215	.0001	-2.14	0.033	-.000425		-.000018	40.8832
mujer*	.0058263	.00255	2.28	0.022	.000826		.010826	0
soltero*	.0058277	.00294	1.98	0.048	.000057		.011599	0
educ	-.0020344	.00029	-7.07	0.000	-.002599		-.00147	17
jefe*	-.0187869	.00445	-4.22	0.000	-.027514		-.01006	1
inla	.0000138	.00000	5.63	0.000	9.0e-06		.000019	90.1766
caba*	-.0019124	.00259	-0.74	0.460	-.006986		.003162	1

```
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

A continuación se presenta la interpretación cuantitativa de cada uno de los efectos marginales.

Empecemos con las variables explicativas que son continuas:

- La interpretación del valor `-.0020344`, que corresponde al efecto marginal de la variable años de educación (`educ`) sería que, para una persona con las características consideradas en el vector X (para este caso, hombre casado que es jefe de familia, con 17 años de educación, edad e ingreso no laboral promedio y que resida en la CABA), un aumento en un año de educación, provoca un cambio en la probabilidad predicha de -.0020344, es decir, la probabilidad de estar desocupado se reduciría en **0.203 puntos porcentuales** ($-,0020344 * 100$), dado todo lo demás constante.

- La interpretación para el efecto marginal de la variable `edad` es equivalente. Para una persona con las características consideradas, un aumento en un año de edad reduce la probabilidad predicha de estar desempleado en **0.022 puntos porcentuales** ($-,0002215 * 100$), *ceteris paribus*.

Para el caso del efecto marginal de las variables dummies (como `mujer`, `soltero`, `jefe`, y `caba`) recuerden que se computan de diferente manera (haciendo la diferencia entre las probabilidades predichas para una persona con las mismas características a excepción de la considerada en el efecto marginal), pero se interpreta de manera equivalente. Stata ya hace este cálculo automáticamente (se los comunica en el mensaje `dy/dx is for discrete change of dummy variable from 0 to 1`).

- Un hombre casado que es jefe de hogar, con 17 años de educación, edad e ingreso no laboral promedio y que reside en CABA tiene una probabilidad predicha de estar desempleado de 1.87 puntos porcentuales ($-,0187869 * 100$) menor que un hombre de las mismas características, que no es jefe de hogar.

- De la misma forma, un hombre que reside en CABA y tiene las mismas características que uno que no lo hace, tiene una probabilidad predicha de estar desempleado de 0.19 puntos porcentuales ($-,0019124 * 100$) menor.

Como notarán, se ha hecho énfasis en aclarar que en el caso de los modelos de elección binaria sí se multiplica por 100 al efecto marginal se está midiendo el efecto del cambio en una unidad de X sobre la probabilidad predicha. Ese cambio es en **puntos porcentuales**, y no en tanto por ciento. El primer caso su usa para indicar un cambio marginal, mientras que el segundo se aplica cuando se trata de cambios proporcionales. Por ejemplo, según se muestra en la segunda salida de Stata, la probabilidad de desempleo para un hombre casado que es jefe de familia, con 17 años de educación, edad e ingreso no laboral promedio y que resida en la CABA es de `.02056653` (es decir, 2 por ciento de probabilidad). Dijimos que el efecto marginal de la educación (`educ`) para este caso es de **0.20 puntos porcentuales**, es decir si en lugar de tener 17 años de educación

tuviera 18 (1 año más) entonces la probabilidad pasaría a ser 1.8 % (es decir, el 2 por ciento original menos 0.20 puntos porcentuales). La forma incorrecta de interpretar los modelos probit y logit es si habláramos del cambio de probabilidad como una reducción del 0.2 % (cambio proporcional), porque en ese caso se entiende que la probabilidad predicha para ese caso seria 1.996 por ciento, es decir hacer $2 * (1 - 0{,}002)$, lo cual es incorrecto.