ECONOMETRIA
Tema 5: ERRORES DE ESPECIFICACIÓN

César Alonso

UC3M

Curso 2009/2010

Introducción

- Hemos visto que el estimador MCO tiene buenas propiedades bajo los supuestos del Modelo de Regresión Lineal.
- ¿Qué ocurre si por cualquier circunstancia empleamos el marco del Modelo de Regresión Lineal no siendo realmente adecuado?
  ¿Qué propiedades tiene el estimador MCO si cometo algún tipo de error de especificación?
- Los errores de especificación en los que nos vamos a centrar son:
  - Inclusión de variables irrelevantes.
  - Omisión de variables relevantes.
  - Errores de medida en las variables.
- Manteniendo la linealidad en parámetros del modelo, los problemas de incorrecta especificación funcional pueden verse como problemas de omisión de variables (por ejemplo, omisión de X^2, etc.).

Inclusión de variables irrelevantes y Omisión de variables relevantes

Regla de la Variable Omitida

- Cuando veíamos la relación entre el modelo de regresión simple y el modelo de regresión múltiple, confrontábamos la regresión larga y la regresión corta, respectivamente.
- Veamos lo mismo desde otra perspectiva. Sea el modelo de regresión lineal múltiple

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

- Por alguna razón (ignorancia, inobservabilidad de la variable, etc.) construimos un modelo de regresión que no incluye la variable explicativa X_2:

$$Y = \gamma_0 + \gamma_1 X_1 + \varepsilon_1.$$

Inclusión de variables irrelevantes y Omisión de variables relevantes

Regla de la Variable Omitida

- En econometría, se dice que "se ha omitido la variable relevante X_2" (si $\beta_2 \neq 0$), lo que supone un *error de especificación*.
  - **Pregunta:** ¿Qué consecuencias tiene la omisión de X_2 en la relación de Y y X_1?
  - **Respuesta:** Que en lugar de β_1 tendremos γ_1, que se relaciona con β_1 de la forma

$$\gamma_1 = \beta_1 + \beta_2 \frac{C(X_1, X_2)}{V(X_1)}$$

    A esta expresión se le denomina regla de la variable omitida, que muestra que la pendiente de una "regresión corta" es una combinación lineal de pendientes de la "regresión larga".

Inclusión de variables irrelevantes y Omisión de variables relevantes

Regla de la Variable Omitida

- Desde otro punto de vista, al omitir X_2, su efecto formará parte del término de error:

$$\varepsilon_1 = \varepsilon + \beta_2 X_2$$

y por tanto

$$\begin{aligned} E(\varepsilon_1|X_1) &= E(\varepsilon + \beta_2 X_2|X_1) \\ &= E(\varepsilon|X_1) + \beta_2 E(X_2|X_1) \\ &= E\left[E(\varepsilon|X_1, X_2)|X_1\right] + \beta_2 E(X_2|X_1) \\ &= \beta_2 E(X_2|X_1) \neq 0 \end{aligned}$$

- En consecuencia,

$$E\left(Y|X_1\right) = \gamma_0 + \gamma_1 X_1 + \beta_2 E\left(X_2|X_1\right)$$

- En general, en un modelo de regresión lineal múltiple $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K + \varepsilon$, si

$$E(\varepsilon | X_1, X_2, \ldots, X_K) \neq 0$$

se puede entender como que el modelo está **mal especificado**.

Inclusión de variables irrelevantes y Omisión de variables relevantes

Propiedades del estimador MCO

- Vamos a emplear los resultados que obtuvimos cuando vimos la regresión larga frente a la regresión corta tanto en términos poblacionales como muestrales.
- Consideraremos el caso de la regresión múltiple más sencilla posible

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

y su versión estimada

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2,$$

frente a una regresión simple

$$Y = \gamma_0 + \gamma_1 X_1 + \varepsilon_1,$$

y su versión estimada

$$\widehat{Y} = \widehat{\gamma}_0 + \widehat{\gamma}_1 X_1$$

- Sabemos que

$$\widehat{\gamma}_1 = \widehat{\beta}_1 + \widehat{\beta}_2 \widehat{\delta}_1$$

siendo $\widehat{\delta}_1$ el estimador MCO de la pendiente de $L(X_2 | X_1) = \delta_0 + \delta_1 X_1$.

- Además:

$$E(\widehat{\beta}_1) = \beta_1$$
$$p\lim \widehat{\beta}_1 = \beta_1$$

y

$$E(\widehat{\gamma}_1) = \gamma_1$$
$$p\lim \widehat{\gamma}_1 = \gamma_1$$

- Otra cosa distinta es que estemos interesados en hacer inferencia sobre γ_1 o sobre β_1, que son diferentes (si queremos conocer el efecto causal de X_1 sobre Y estaremos generalmente interesados en β_1).
- Por tanto, tenemos que:

$$E\left(\widehat{\gamma}_1 \,|X_1, X_2\right) = E\left(\widehat{\beta}_1 \,|X_1, X_2\right) + E\left(\widehat{\beta}_2\widehat{\delta}_1 \,|X_1, X_2\right) = \beta_1 + \beta_2\widehat{\delta}_1$$

y que

$$E(\widehat{\gamma}_1) = \beta_1 + \beta_2 E(\widehat{\delta}_1)$$
$$p\lim(\widehat{\gamma}_1) = \beta_1 + \beta_2\delta_1$$

- En general:
  - $\widehat{\gamma}_1$ no será apropiado si queremos hacer inferencia sobre β_1.
  - Además, es fácil demostrar que $V(\widehat{\gamma}_1) \leq V(\widehat{\beta}_1)$

Inclusión de variables irrelevantes y Omisión de variables relevantes

Propiedades del estimador MCO: Omisión de variables relevantes

- Siempre que X_2 sea una variable "relevante" (es decir, $\beta_2 \neq 0$),
  - $\widehat{\gamma}_1$ será un estimador inconsistente (y sesgado) de β_1 (aunque tendrá menor varianza que $\widehat{\beta}_1$).
- Es decir: la "**omisión de variables relevantes**" genera inconsistencia (y sesgos) en la estimación de los efectos de las variables, (aunque suponga una reducción en la varianza del estimador).

- En otras palabras: el coeficiente de X_1 en la regresión que (incorrectamente) omite X_2:
  - no recoge el efecto *ceteris paribus* sobre Y de un cambio en X_1 (puesto que cuando varía X_1 también lo hace X_2, al estar X_1 y X_2 correlacionadas).
  - recoge el efecto sobre Y de un cambio en X_1 más el efecto indirecto de X_1 sobre X_2 (que, al existir correlación, termina afectando a Y).

- Podemos resumir el sesgo en la estimación de β_1 cuando se omite (incorrectamente) X_2 como sigue:

	$C(X_1, X_2) > 0$	$C(X_1, X_2) < 0$
$\beta_2 > 0$	$+$	$-$
$\beta_2 < 0$	$-$	$+$

- Si X_2 es una variable "irrelevante" (es decir, $\beta_2 = 0$), $\widehat{\gamma}_1$ será un estimador consistente (e insesgado) de β_1, y tendrá menor varianza que $\widehat{\beta}_1$, que también será insesgado y consistente.
- Es decir: la "**inclusión de variables irrelevantes**" en el análisis, no afecta a la consistencia de la estimación de los efectos de las variables
  - **Intuición:** En la población, el coeficiente de una variable irrelevante será igual a 0, de manera que al estimar el modelo que incorrectamente incluye dicha variable los estimadores de los coeficientes de las restantes variables no se verán afectados en el límite (ni en promedio).

- Sin embargo, **sí se genera una pérdida de eficiencia** en la estimación (tanto mayor cuanto mayor sea el número de variables irrelevantes que se incluyan).
  - **Intuición:** Cuanto más correlacionada está una variable irrelevante con las variables relevantes, más aumentarán las varianzas de los estimadores de los coeficientes de las variables relevantes.
  - Esto supone que incluir variables irrelevantes, aunque no genere inconsistencia del estimador MCO (y por tanto es un error menos grave que el de omisión de variables relevantes), puede generar un serio problema, en la medida en que a la hora de contrastar hipótesis del tipo $H_0 : \beta_j = 0$ perdemos potencia (es decir, aumenta el error de tipo II), de manera que podríamos inferir que no son relevantes variables que sí lo son.

- ¿En la práctica, es posible saber a priori qué modelización es más adecuada?
  - En sentido estricto: **No**.
  - Lo que sí se puede es modelizar lo mejor posible empleando la Teoría Económica como guía y acumular evidencia a favor o en contra de la "relevancia" o "irrelevancia" de una o varias variables mediante los contrastes de hipótesis.

Inclusión de variables irrelevantes y Omisión de variables relevantes

Ejemplo 1: Efecto del tabaco sobre el cáncer

- Supongamos que tenemos un grupo de fumadores y un grupo de no fumadores y observamos la incidencia del cáncer sobre cada individuo.
- Supongamos además que los no fumadores son más propensos a hacer ejercicio y que el ejercicio físico reduce el cáncer (pero no observamos el ejercicio físico).
- Entonces, la incidencia del cáncer entre los fumadores puede estar sobrevalorada debido a que el consumo de tabaco tiende a disminuir el ejercicio físico.
- Formalmente, $C_i = \beta_0 + \beta_1 F_i + \beta_2 EJ_i + \varepsilon$,
  donde, para el individuo i, C_i es una medida de la incidencia del cáncer, F_i es una variable binaria que toma el valor 1 para fumadores y 0 en otro caso, y EJ_i es la cantidad de ejercicio físico. Además, $\beta_1 > 0$, $\beta_2 < 0$.

Inclusión de variables irrelevantes y Omisión de variables relevantes

Ejemplo 1: Efecto del tabaco sobre el cáncer

- Además, $EJ_i = \delta_0 + \delta_1 F_i + v_i$, donde $\delta_1 < 0$.
- Entonces, al realizar la regresión simple de C_i sobre F_i, estamos estimando

$$C_i = \gamma_0 + \gamma_1 F_i + \varepsilon_1,$$

de manera que el coeficiente estimado va a ser

$$\widehat{\gamma}_1 = \widehat{\beta}_1 + \widehat{\beta}_2 \widehat{\delta}_1$$

y por tanto el efecto del tabaco sobre el cáncer estará sobreestimado si como cabe esperar, $\widehat{\beta}_2 < 0$ (el ejercicio, ceteris paribus, reduce la incidencia del cáncer) y $\widehat{\delta}_1 < 0$ (el ejercicio está negativamente correlacionado con el consumo de tabaco).

Inclusión de variables irrelevantes y Omisión de variables relevantes

Ejemplo 2: Efecto del tabaco sobre los salarios

- Además de los consabidos efectos sobre la salud, ¿tiene el tabaquismo consecuencias económicas?
  - Los fumadores podrían recibir salarios más bajos:
    - si fueran menos productivos (en el caso de que se ausentaran en horario de trabajo para fumar);
    - si las dolencias propias de los fumadores provocaran bajas o absentismo laboral;
    - si las empresas discriminaran contra ellos;
    - etc.

Inclusión de variables irrelevantes y Omisión de variables relevantes

Ejemplo 2: Efecto del tabaco sobre los salarios

- Con datos representativos para EE.UU. de individuos alrededor de 30 años, Levine, Gustafson y Velenchik (1997)[1] estimaron ecuaciones de salarios utilizando las siguientes variables:
  - $Y = \ln(\text{salario})$
  - $F =$ Variable binaria que toma valor 1 si el indiv. es fumador y 0 en caso contrario
  - $ED =$ Años de educación del individuo
- Hay que tener en cuenta que los fumadores tienen en promedio 1 año de educación menos que los no fumadores (es decir, la educación está negativamente correlacionada con el tabaquismo)

[1] Levine, P., T. Gustafson y A. Velenchik (1997), "More Bad News for Smokers? The Effects of Cigarette Smoking on Wages", *Industrial and Labor Relations Review*, 50(3), 493-509.

- Se consideraron 2 especificaciones:
  - Omitiendo educación

$$\widehat{Y}_i = \underset{(0.021)}{-0.176} \quad F_i$$

(como el salario está en logaritmos, supondría que un fumador gana en promedio un 17.6% menos que un no fumador).

Inclusión de variables irrelevantes y Omisión de variables relevantes

Ejemplo 2: Efecto del tabaco sobre los salarios

- Incluyendo educación

$$\begin{array}{lccc} \widehat{Y}_i = & -0.080 & F_i & +0.070 \quad ED_i \\ & (0.021) & & (0.004) \end{array}$$

(Ahora, para un nivel de educación dado, un fumador gana un 8% menos que un no fumador).

- La omisión de la educación exagera (en más del doble) el efecto negativo de fumar sobre el salario.

Errores de Medida

- Algunas veces no tenemos datos sobre la variable económica que realmente nos interesa.
  Ejemplos:
  - Disponemos del ingreso anual *declarado* por un individuo, no del ingreso anual real.
  - De acuerdo con el modelo de *ciclo vital*, el consumo depende de la *renta permanente*, que difiere de la renta disponible en que elimina las variaciones transitorias de la renta. (Sin embargo, nosotros observamos renta disponible pero no observamos la renta permanente).
  - El tipo impositivo marginal puede ser difícil de obtener o resumir en una única cifra para todos los niveles de renta. Sin embargo, podríamos calcular el tipo impositivo medio basado en la renta agregada y en los impuestos recaudados.

Errores de Medida

- **Errores de medida:**
  Aparecen cuando empleamos una medida poco precisa de una variable económica en un modelo de regresión.
- ¿Cómo afectan los errores de medida a la estimación de MCO?
  Dependerá de los casos.

- Consideremos el modelo:

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

con

$$E(\varepsilon | X) = 0 \Rightarrow E(Y^* | X) = L(Y^* | X) = \beta_0 + \beta_1 X$$

y por tanto, β_0 y β_1 verifican:

$$E(\varepsilon) = 0, \qquad C(X, \varepsilon) = 0 \Rightarrow$$

$$\beta_0 = E(Y^*) - \beta_1 E(X) \qquad \beta_1 = C(X, Y^*) / V(X)$$

- Y^* es medida con error, de modo que:

$$v_0 = Y - Y^* = \text{ Error de medida } \Rightarrow Y = Y^* + v_0.$$

  - Si estimamos por MCO un modelo de regresión de Y sobre X,

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X,$$

  - ¿serán $\widehat{\beta}_0$ y $\widehat{\beta}_1$ estimadores consistentes de β_0 y β_1?
  - **Nota**: Si sustituimos en (1) tenemos:

$$Y = \beta_0 + \beta_1 X + u,$$

  con $u = (\varepsilon + v_0)$.

- Nótese que:

$$p\lim \widehat{\beta}_1 = \frac{p\lim\left(\frac{1}{n}\sum_i x_i y_i\right)}{p\lim\left(\frac{1}{n}\sum_i x_i^2\right)} = \frac{C(X,Y)}{V(X)}$$
$$= \frac{C(X,Y^* + v_0)}{V(X)} = \frac{C(X,Y^*) + C(X,v_0)}{V(X)},$$

  - con: $y_i = Y_i - \overline{Y}$, $x_i = X_i - \overline{X}$,
  - y por tanto:

$$p\lim \widehat{\beta}_1 = \frac{C(X,Y^*)}{V(X)} = \beta_1, \text{ si } C(X,v_0) = 0.$$

  - **Por tanto**, si $C(X,v_0) = 0$ (es decir, si el error de medida en la variable dependiente **no está sistemáticamente relacionado con las variables explicativas**), el estimador MCO es consistente.

- Nótese que:

$$\begin{aligned} p\lim \widehat{\beta}_0 &= p\lim(\overline{Y} - \widehat{\beta}_1 \overline{X}) = E(Y) - p\lim \widehat{\beta}_1 E(X) \\ &= E(Y^* + v_0) - p\lim \widehat{\beta}_1 E(X), \end{aligned}$$

  - y por tanto:

$$p\lim \widehat{\beta}_0 = E(Y^*) - \beta_1 E(X) = \beta_0,$$

    si $E(v_0) = 0$ y $C(X, v_0) = 0$.

- En **conclusión**:
  - Si estimamos por MCO un modelo de regresión empleando Y en vez de Y^*, los estimadores serán consistentes y la inferencia habitual será válida, si:

$$C(X, v_0) = 0$$
$$E(v_0) = 0$$

    - Si la segunda condición no se cumple, tendríamos un estimador inconsistente de la constante, pero no de la pendiente.
  - Los estimadores de MCO empleando Y son más ineficientes que los obtenidos con Y^* (sin error de medida):

$$V(Y^*|X) = V(\varepsilon|X) = \sigma^2$$

suponiendo que $C(\varepsilon, v_0|X) = 0$, se obtiene:

$$V(Y|X) = V(\varepsilon + v_0|X) = V(\varepsilon|X) + V(v_0|X) = \sigma^2 + \sigma^2_{v_0} > \sigma^2 \Rightarrow$$

mayor ineficiencia en la estimación.

- Se suele considerar un problema más importante que un error de medida en la variable dependiente.
- Consideremos el modelo:

$$Y = \beta_0 + \beta_1 X^* + \varepsilon, \tag{2}$$

con

$$E(\varepsilon | X^*) = 0 \Rightarrow E(Y | X^*) = L(Y | X^*) = \beta_0 + \beta_1 X^*,$$

y por tanto, β_0 y β_1 verifican:

$$E(\varepsilon) = 0, \qquad C(X^*, \varepsilon) = 0 \Rightarrow$$

$$\beta_0 = E(Y) - \beta_1 E(X^*), \qquad \beta_1 = C(X^*, Y) / V(X^*).$$

- Supongamos que X^* se mide con error, de modo que:

$$v_1 = X - X^* = \text{ Error de medida } \Rightarrow X = X^* + v_1$$

y **se cumple que**:

  - $E(v_1) = 0$,
  - $C(X, \varepsilon) = 0$ (ε no está correlacionada con X ni con X^*, ni por tanto con v_1).

- En términos de esperanza condicional:

$$E\left(Y \left| X, X^* \right.\right) = E\left(Y \left| X^* \right.\right).$$

  - Es decir: X no influye en Y si condicionamos en X^*.

- **Qué propiedades tendrá la estimación MCO de un modelo de regresión de Y sobre X?**
  - ¿Serán $\widehat{\beta}_0$ y $\widehat{\beta}_1$ estimadores consistentes de β_0 y β_1?
    - Dependerá de los supuestos que hagamos sobre el error de medida.
- **Nota**: Si sustituimos en (2) tenemos:

$$Y = \beta_0 + \beta_1 X + u,$$

con

$$u = (\varepsilon - \beta_1 v_1).$$

- Supongamos que:
  - $C(X^*, v_1) = 0$ **(Supuesto clásico de errores en variables -CEV)**
  - $C(v_1, \varepsilon) = 0$

- Tenemos entonces que:

$$p\lim \widehat{\beta}_1 = \frac{p\lim\left(\frac{1}{n}\sum_i x_i y_i\right)}{p\lim\left(\frac{1}{n}\sum_i x_i^2\right)} = \frac{C(X,Y)}{V(X)} = \frac{C(X^*+v_1,Y)}{V(X^*+v_1)}$$

$$= \frac{C(X^*,Y) + \overbrace{C(Y,v_1)}^{=0}}{V(X^*)+V(v_1)} = \frac{C(X^*,Y)/V(X^*)}{[V(X^*)+V(v_1)]/V(X^*)}$$

$$= \frac{\beta_1}{1+\frac{V(v_1)}{V(X^*)}} \neq \beta_1,$$

y por tanto:

$$\text{sesgo asintótico}\ (\widehat{\beta}_1) = p\lim(\widehat{\beta}_1 - \beta_1) = -\beta_1\ \frac{V(v_1)}{V(X^*)+V(v_1)}.$$

- Nótese que, en presencia de errores de medida, tenderemos a **infraestimar la magnitud (en valor absoluto)** de la pendiente de la variable que se mide con error.
- Si $V(X^*)$ es grande en relación a $V(v_1)$, la inconsistencia puede llegar a ser despreciable.
- Es decir: si la variabilidad del error de medida, relativa a la variabilidad de la variable explicativa original, es pequeña, entonces el efecto del error de medida sobre la consistencia del estimador puede ser insignificante.

Errores de Medida

Error de medida en una variable explicativa

- En un modelo de regresión múltiple, en general el error de medida en una variable explicativa produce inconsistencia de todos los coeficientes estimados $\widehat{\beta}\,'$s. A este respecto, en un modelo de regresión múltiple en el que sólo uno de los regresores se mide con error (y este error no está correlacionado ni con la variable medida con error ni con el resto de las variables explicativas):
  - Se mantiene el resultado de que la pendiente de la variable explicativa que se mide con error tiende a infraestimarse en valor absoluto (se puede demostrar).
  - Las estimaciones de las pendientes asociadas a las restantes variables explicativas serán, en general, inconsistentes, si bien no es fácil saber cuáles serán las direcciones y magnitudes de los sesgos de inconsistencia.
  - Solamente en el caso improbable de que las variables explicativas sean ortogonales a la variable medida con error, los estimadores de sus pendientes serán consistentes.

Errores de Medida

Ejemplo 3: Efecto de los ingresos familiares en el rendimiento universitario

- Queremos ver si la renta familiar tiene algún efecto en las calificaciones medias obtenidas en la universidad.
- No está claro que la renta familiar tenga un efecto directo sobre el rendimiento universitario.
- La estrategia recomendada sería incluir dicha variable como regresor y contrastar si su coeficiente es igual a cero.

$$CAL = \beta_0 + \beta_1 I^* + \beta_2 PRE + \beta_3 SEL + \varepsilon, \quad \text{donde}$$

  - CAL = Nota media en la universidad,
  - I^* = Ingresos familiares,
  - PRE = Nota media del curso previo al acceso a la universidad,
  - SEL = Nota media de selectividad.

Errores de Medida

Ejemplo 3: Efecto de los ingresos familiares en el rendimiento universitario

- Supongamos que los datos se obtienen encuestando directamente a los estudiantes.
  - Es posible que declaren la renta familiar de manera incorrecta, de manera que observamos $I = I^* + v$.
- Aun suponiendo que el error de medida v no está correlacionado con I^* ni con el resto de las variables explicativas (PRE, SEL), los estimadores de los parámetros utilizando I (en vez de I^*, que es inobservable) serían inconsistentes.
  - En particular, tenderíamos a infraestimar β_1.
    - Por tanto, si contrastáramos $H_0 : \beta_1 = 0$, sería más probable que no rechazáramos H_0.
  - En este ejemplo, es difícil dilucidar la magnitud y dirección de los sesgos de inconsistencia de β_2 y β_3.