ECONOMETRIA I.
Departamento de Fundamentos del Análisis Económico
Universidad de Alicante. Curso 2011/12

GUIÓN TEMA 5.

ESPECIFICACIÓN Y PREDICCIÓN EN EL MRL

5.1. Errores de especificación; selección de variables

Bibliografía apartados :
Greene, 6.4.3, 8.4
A.F.Gallastegui: 6.1
J.M. Wooldridge: 3.3 (pp 96-101) y 3.4 (pp 107 y 108)

5.1.1 Inclusión de variables irrelevantes

La inclusión de variables irrelevantes significa que incluimos una o más variables en el modelo que tienen efecto marginal nulo sobre la variable dependiente es decir que su coeficiente es cero en la población. Por ejemplo, consideremos el modelo

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

La variable X_4 es irrelevante si $\beta_4 = 0$. Como desconocemos que $\beta_4 = 0$, estimamos el modelo incluyendo la variable X_4. ¿Cuáles son las consecuencias sobre el estimador MCO de incluir X_4?. El estimador MCO es insesgado aunque se incluya una variable cuyo coeficiente poblacional es cero, ya que lo que estamos haciendo es estimar un modelo que verifica una restricción ($\beta_4 = 0$) sin imponer dicha restricción y sabemos que esto no influye en la insesgadez del estimador. Sin embargo, también sabemos que, si $\beta_4 = 0$, el estimador MCO de la regresión de Y_t sobre una constante, X_{2t}, X_{3t} y X_{4t} es menos eficiente que el estimador MCO de la regresión de Y_t sobre una constante, X_{2t} y X_{3t}, ya que este último estimador es el estimador de mínimos cuadrados restringidos imponiendo la restriccón $\beta_4 = 0$ y si dicha restricción se cumple el estimador del modelo restringido es más eficiente que el estimador del modelo sin restricciones (puesto que la varianza del modelo

restringido es menor, independientemente de si la restricción es cierta o no, y su sesgo es cero).

5.1.2 Omisión de variables relevantes

Supongamos ahora que omitimos una variable relevante del modelo, es decir, omitimos una variable del modelo que tiene coeficiente poblacional distinto de cero. Por ejemplo, supongamos que el modelo poblacional es

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

con $\beta_3 \neq 0$ y que estimamos por MCO la regresión de Y_t sobre una constante y X_{2t}. En este caso estamos imponiendo una restricción en el modelo ($\beta_3 = 0$) que en realidad no se verifica y por tanto el estimador está, en general, sesgado. En este caso sencillo es fácil calcular el sesgo. Sea $\widetilde{\beta}_2$ el estimador MCO del coeficiente de X_2 en la regresión de Y_t sobre una constante y X_{2t}, utilizando los resultados del Tema 1 sobre el estimador MCO del modelo de regresión simple sabemos que

$$\begin{aligned}
\widetilde{\beta}_2 &= \frac{S_{X_2Y}}{S^2_{X_2}} = \frac{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)(Y_t-\overline{Y})}{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)^2} = \frac{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)Y_t}{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)^2} \\
&= \frac{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)(\beta_1+\beta_2 X_{2t}+\beta_3 X_{3t}+u_t)}{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)^2} \\
&= \beta_1 \frac{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)}{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)^2} + \beta_2 \frac{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)X_{2t}}{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)^2} + \beta_3 \frac{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)X_{3t}}{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)^2} + \frac{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)u_t}{\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)^2}
\end{aligned}$$

y puesto que

$$\begin{aligned}
&\sum_{t=1}^{T}(X_{2t}-\overline{X}_2) = 0, \quad \sum_{t=1}^{T}(X_{2t}-\overline{X}_2)X_{2t} = \sum_{t=1}^{T}(X_{2t}-\overline{X}_2)^2, \\
&\sum_{t=1}^{T}(X_{2t}-\overline{X}_2)X_{3t} = \sum_{t=1}^{T}(X_{2t}-\overline{X}_2)(X_{3t}-\overline{X}_3)
\end{aligned}$$

tenemos que

$$\widetilde{\beta}_2 = \beta_2 + \beta_3 \frac{S_{X_2X_3}}{S^2_{X_2}} + \frac{\sum_{t=1}^{T}(X_{2t} - \overline{X}_2)u_t}{\sum_{t=1}^{T}(X_{2t} - \overline{X}_2)^2}$$

donde $S_{X_2X_3}$ es la covarianza muestral entre X_2 y X_3 y $S^2_{X_2}$ es la varianza muestral de X_2. Si calculamos la esperanza de $\widetilde{\beta}_2$, como los errores tienen media cero y los regresores no son aleatorios, tenemos

$$E(\widetilde{\beta}_2) = \beta_2 + \beta_3 \frac{S_{X_2X_3}}{S^2_{X_2}}$$

y $\widetilde{\beta}_2$ está, en general, sesgado. $\widetilde{\beta}_2$ será un estimador insesgado de β_2 solamente si $\beta_3 = 0$ (lo que querría decir que la variable X_3 no es relevante) o si $S_{X_2X_3} = 0$, es decir si X_2 y X_3 no están correlacionadas entre si en la muestra.

El signo del sesgo depende del signo de β_3 y del signo de $S_{X_2X_3}$ como muestra la siguiente tabla:

	$Corr(X_2, X_3) > 0$	$Corr(X_2, X_3) < 0$
$\beta_3 > 0$	Sesgo positivo	Sesgo negativo
$\beta_3 < 0$	Sesgo negativo	Sesgo positivo

El motivo principal por el que se omite una variable relevante en la práctica es porque no se dispone de información sobre la misma.

Ejemplo 1

Supongamos que el salario está determinado por

$$salario_t = \beta_1 + \beta_2 educ_t + \beta_3 habil_t + u_t$$

donde $habil_t$ es la "habilidad innata" que es inobservable. En este modelo, por definición, más habilidad lleva a una mayor productividad y por tanto a un salario más elevado por lo que $\beta_3 > 0$. Además existen razones para creer que educación y habilidad innata están correlacionadas positivamente, por lo que la estimación de la regresión del salario sobre una constante y los años de educación sobreestima, por lo general, el efecto marginal de la educación sobre los salarios.

Consideremos ahora el caso más general de un modelo con k ($k > 3$) variables explicativas

$$Y_t = \beta_1 + \beta_2 X_{2t} + .. + \beta_k X_{kt} + u_t$$

en el que $\beta_k \neq 0$. Si omitimos la variable X_{kt} y estimamos la regresión de Y_t sobre una constante, $X_{2t}, .., X_{(k-1)t}$, el estimador de todos los coeficientes estará, en general, sesgado salvo que las correlaciones entre X_{kt} y las restantes variables del modelo sean todas cero. Basta con que X_{kt} esté correlacionada con una de las otras variables para el estimador MCO de todos los coeficientes esté segado. Además, en el caso más general es difícil establecer el signo del sesgo.

5.2. Predicción

Bibliografía apartados : Greene, 7.11;
A.F.Gallastegui: 5.4
J.M. Wooldridge: 6.4

5.2.1 Intervalo de confianza.

Consideremos el modelo de regresión

$$Y_t = \beta_1 + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t$$

que verifica los supuestos del MRL con errores normales y sea $\widehat{\beta} = (\widehat{\beta}_1, \widehat{\beta}_2, .., \widehat{\beta}_k)'$ el estimador MCO basado en una muestra de tamaño T. Estamos interesados en estimar la media de la variable dependiente para ciertos valores particulares de las variables explicativas $X_2 = x_{2s}, ..., X_k = x_{ks}$ (estos valores de las variables explicativas no tienen porque coincidir con los valores observados en la muestra). Puesto que los errores tienen media cero, el valor esperado para la variable dependiente cuando $X_2 = x_{2s}, ..., X_k = x_{ks}$ es

$$E(y_s) = \beta_1 + \beta_2 x_{2s} + .. + \beta_k x_{ks}$$

y utilizando el estimador MCO obtenemos la estimación de la media de la variable dependiente

$$\widehat{E(y_s)} = \widehat{y}_s = \widehat{\beta}_1 + \widehat{\beta}_2 x_{2s} + .. + \widehat{\beta}_k x_{ks}$$

y como el estimador MCO es insegado, el valor esperado para el estimador de la media es:

$$E(\widehat{y}_s) = \beta_1 + \beta_2 x_{2s} + .. + \beta_k x_{ks} = E(y_s)$$

y por tanto $\widehat{E(y_s)} = \widehat{y}_s$ es un estimador insesgado de $E(y_s)$.

En muchas ocasiones estamos interesados en tener una medida de la incertidumbre de $\widehat{E(y_s)}$ y para eso calcularemos el intervalo de confianza para $E(y_s)$. Puesto que $E(y_s)$ es una combinación lineal de $\beta_1, \beta_2, .., \beta_k$, como vimos en el Tema 4, para calcular el intervalo de confianza para $E(y_s)$ necesitamos calcular el error estándar de $\widehat{y}_s$ (que es una combinación lineal de los estimadores de β de MCO). Sea $x_s = (1, x_{2s}, .., x_{ks})'$, entonces

$$\widehat{y}_s = \widehat{\beta}' x_s$$

y

$$SE(\widehat{y}_s) = \sqrt{\widehat{var(\widehat{y}_s)}} = \sqrt{x_s' \widehat{var(\widehat{\beta})} x_s} = \sqrt{\widehat{\sigma}^2 x_s' (X'X)^{-1} x_s}$$

y el intervalo de confianza para $E(y_s)$ es

$$[\widehat{y}_s - t_{T-k,0.025} * SE(\widehat{y}_s), \widehat{y}_s + t_{T-k,0.025} * SE(\widehat{y}_s)]$$

Ejemplo 2

Utilizando la misma muestra de 526 individuos que en el ejemplo 1 del Tema 4, hemos estimado el siguiente modelo:

$$\widehat{salario}_t = \underset{(0.2845)}{4.145} + \underset{(0.3022)}{2.481}\, Hombre_t + \underset{(0.0111)}{0.0269} exp_t$$

y sea

$$\widehat{var(\widehat{\beta})} = \begin{bmatrix} 0.0810 & -0.0452 & -0.0020 \\ -0.0452 & 0.0914 & -0.00014 \\ -0.0020 & -0.00014 & 0.00012 \end{bmatrix}$$

la matriz de varianzas estimada de $\widehat{\beta}$. Vamos a obtener la estimación del salario medio de las mujeres y de los hombres sin experiencia laboral. La estimación del salario medio para las mujeres sin experiencia laboral corresponde a la estimación para el vector que contiene los siguientes valores de las variables explicativas $x_{M0} = (1, 0, 0)'$ y por tanto

$$\widehat{salario}_{M0} = \widehat{\beta}_1 + \widehat{\beta}_2 * 0 + \widehat{\beta}_3 * 0 = \widehat{\beta}_1 = 4.145$$

Como la estimación coincide en este caso con $\widehat{\beta}_1$

$$SE(\widehat{salario}_{M0}) = SE(\widehat{\beta}_1) = 0.2845$$

y el intervalo de confianza es

$$[4.145 - 1.96 * 0.2845, 4.145 + 1.96 * 0.2845] = [3.587, 4.703]$$

La estimación del salario medio para los hombres sin experiencia laboral corresponde a la estimación para $x_{H0} = (1, 1, 0)'$ y por tanto

$$\widehat{salario}_{H0} = \widehat{\beta}_1 + \widehat{\beta}_2 * 1 + \widehat{\beta}_3 * 0 = \widehat{\beta}_1 + \widehat{\beta}_2 = 4.145 + 2.481 = 6.626$$

Como la estimación coincide en este caso con $\widehat{\beta}_1 + \widehat{\beta}_2$

$$\begin{aligned} SE(\widehat{salario}_{H0}) &= SE(\widehat{\beta}_1 + \widehat{\beta}_2) = \sqrt{\widehat{var(\widehat{\beta}_1 + \widehat{\beta}_2)}} = \sqrt{\widehat{var(\widehat{\beta}_1)} + \widehat{var(\widehat{\beta}_2)} + 2\widehat{cov(\widehat{\beta}_1, \widehat{\beta}_2)}} \\ &= \sqrt{0.0810 + 0.0914 - 2 * 0.0452} = 0.2864 \end{aligned}$$

y el intervalo de confianza es

$$[6.626 - 1.96 * 0.2864, 6.626 + 1.96 * 0.2864] = [6.065, 7.187]$$

Nótese que para calcular el intervalo de confianza para los hombres hubiese sido más sencillo utilizar los resultados de la regresión que incluye la dummy de ser mujer en lugar de la dummy de ser hombre

$$\widehat{salario}_t = \underset{(0.2862)}{6.626} - \underset{(0.3022)}{2.481}\, Mujer_t + \underset{(0.0111)}{0.0269} exp_t$$

ya que utilizando este modelo, la estimación para los hombres sin experiencia laboral corresponde a la estimación para $x^*_{H0} = (1, 0, 0)'$ y por tanto

$$\widehat{salario}_{H0} = \widehat{\alpha}_1 = 6.626$$

y como la estimación coincide en este caso con $\widehat{\alpha}_1$

$$SE(\widehat{salario}_{H0}) = SE(\widehat{\alpha}_1) = 0.2862$$

(el error estándar no coincide exactamente con el que calculamos anteriormente por los errores de redondeo).

5.2.2 Intervalo de predicción

El método que acabamos de ver nos permite calcular un intervalo de confianza para la media de la variable dependiente cuando las variables explicativas toman unos valores concretos, es decir para la media de la variable dependiente asociada al subconjunto de la población definido por esos valores de las variables explicativas. Pero un intervalo de confianza para el individuo medio en el subconjunto de la población no es lo mismo que el intervalo de predicción para un individuo cualquiera de ese subconjunto, ya que en el caso de un individuo particular hay otra fuente de variación que viene dada por el error no observable, es decir por todos los factores no observables que afectan a la variable dependiente. Vamos a calcular ahora el intervalo de predicción para un individuo para el que $X_2 = x_{2s}, ..., X_k = x_{ks}$. Utilizando el modelo poblacional

$$y_s = \beta_1 + \beta_2 x_{2s} + .. + \beta_k x_{kt} + u_s$$

Como el error aleatorio, u_s, tiene media cero, la mejor predicción para y_s es

$$\widehat{y}_s = \widehat{\beta}_1 + \widehat{\beta}_2 x_{2s} + .. + \widehat{\beta}_k x_{ks}$$

Nótese que $\widehat{y}_s$, la predicción para y_s, coincide con la estimación para la media de la variable dependiente, $\widehat{E(y_s)}$, no obstante en cada caso se persigue un objetivo distinto. En un caso buscamos predecir una variable aleatoria y en otro buscamos estimar una media desconocida pero constante.

Se define el error de predicción como

$$e_s = y_s - \widehat{y}_s$$

Como $E(\widehat{y}_s) = E(y_s)$

$$E(e_s) = E(y_s - \widehat{y}_s) = E(y_s) - E(\widehat{y}_s) = 0$$

y el error de predicción tiene media cero. Vamos a calcular ahora la varianza del error de predicción

$$var(e_s) = var(y_s) + var(\widehat{y}_s) - 2cov(y_s, \widehat{y}_s) = \sigma^2 + var(\widehat{y}_s)$$

ya que $var(y_s) = var(u_s) = \sigma^2$ y $cov(y_s, \widehat{y}_s) = 0$ ya que $\widehat{y}_s$ es función de los errores asociados a las observaciones de la muestra e y_s es función del error u_s y los errores son independientes. El error estándar del error de predicción e_s es por tanto

$$SE(e_s) = \sqrt{\widehat{\sigma}^2 + \widehat{var(\widehat{y}_s)}} = \sqrt{\widehat{\sigma}^2 + x_s' \widehat{var(\widehat{\beta})} x_s} = \sqrt{\widehat{\sigma}^2 (1 + x_s'(X'X)^{-1} x_s)}$$

y el intervalo de predicción para un individuo para el que $X_2 = x_{2s}, ..., X_k = x_{ks}$ es

$$[\widehat{y}_s - t_{T-k,0.025} * SE(e_s), \widehat{y}_s + t_{T-k,0.025} * SE(e_s)]$$

Nótese que $SE(e_s) > SE(\widehat{y}_s)$, y por tanto, el intervalo de predicción para un individuo cualquiera con ciertos valores para las variables explicativas tiene mayor amplitud que el intervalo de confianza para la media de todos los individuos con esos valores de las variables explicativas. Este hecho refleja que es más difícil predecir para un individuo que estimar la media, ya que el individuo tiene una fuente de heterogeneidad adicional a la media, que viene dada por el término de error, y que refleja todos los factores inobservables que influyen en la variable dependiente.

Ejemplo 2 (cont.)

Siguiendo con el ejemplo anterior y sabiendo $\widehat{\sigma}^2 = 16.85$ podemos calcular el intervalo de predicción para el salario de un hombre sin experiencia laboral

$$SE(\widehat{e}_{H0}) = \widehat{\sigma}^2 + \widehat{var(\widehat{\beta}_1)} = \sqrt{16.85 + 0.0810} = 4.11$$

y el intervalo de predicción es

$$[6.626 - 1.96 * 4.11, 6.626 + 1.96 * 4.11] = [-1.430, 14.682]$$

Nótese que el intervalo de predicción que acabamos de calcular tiene una amplitud muy grande (incluso incluye valores negativos que no tienen sentido para el salario). Este resultado ilustra lo que suele ocurrir en la práctica cuando trabajamos con datos de sección cruzada, ya que la varianza del término de error suele ser grande puesto que suele haber muchos factores no observables que influyen en la variable dependiente. En general cuando trabajamos con datos de sección cruzada estaremos interesados en un intervalo de confianza para el individuo medio con determinadas caraterísticas y no en un intervalo de predicción para un individuo cualquiera con dichas características. Por el contrario, cuando trabajamos con datos de series temporales el objetivo será calcular el intervalo de predicción para la variable dependiente para el siguiente periodo $T + 1$ suponiendo que en dicho periodo las variables explicativas tomarán unos determinados valores.