Econometria I

Tema 4: Problemas de Especificación y los Datos

Universidad Carlos III

Getafe, Madrid

Octubre-November 2008

Outline

- Mala especificación de la forma funcional
- Utilización de variables proxy para variables explicativas no observadas
- Propiedades de MCO con errores de medida
- Problemas con los datos

Problemas de Especificación y los Datos

- Discutiremos tres problemas que originan que alguna variable explicativa xj sea endógena:
  - Error en la especificación de la forma funcional
  - Utilización de variables proxy
  - Errores de medida.

Mala especificación de la forma funcional

- Ocurre cuando un modelo de regresión no consigue representar adecuadamente la relación funcional entre la variable dependiente y las explicativas:
  - Omisión de términos cuadráticos.
  - Transformación log mal empleada.
  - Omisión de interacciones con variables binarias
- Hay contrastes para comprobar estos problemas, ya que en principio tenemos datos sobre todas las variables relevantes. Un problema diferente es el uso de aproximaciones de ciertas variables porque no disponemos de datos sobre las variables de interés.

RESET (Regression Specification Error Test)

- Es un contraste general de mala especificación de la forma funcional
- Suponemos que

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

satisface RLM.3, por lo que cualquier función no lineal de las variables independientes tiene que ser significativa si la añadimos a la ecuación (por ej. x_1^2 ó $x_1 x_2$).

- Sin embargo si k es grande, comprobar un número alto de casos consumirá muchos grados de libertad, y no todas las clases de errores de especificación pueden ser descubiertas con funciones simples (como las cuadráticas).

RESET

- El contraste RESET se basa en esta regresión:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + \epsilon$$

aunque en principio podríamos añadir más potencias de los valores ajustados $\hat{y}$ de la especificación original: $\hat{y}^4$, etc.

- De esta ecuación no interesan los valores ajustados, si no contrastar:

$$H0 : \delta_1 = \delta_2 = 0$$

- Si se rechaza $H0$ mediante un contraste de la F (con distribución aproximada $F_{2,n-k-3}$) es evidencia de mala especificación del modelo.
- EX:HPRICE:log vs level

Problemas RESET

- No proporciona información sobre cómo proceder si un modelo se rechaza.
- No sirve para contrastar si hay variables omitidas ni heteroscedasticidad.

Contrastes contra Alternativas no Anidadas

- Si queremos contrastar si una variable explicativa debe aparecer en niveles o en logaritmos,

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

  en contra

  $$y = \beta_0 + \beta_1 log(x_1) + \beta_2 log(x_2) + u$$

- no podemos usar un contraste de la F habitual, porque son modelos no anidados.

Especificación: Contraste de Mizon and Richard

- Construir un modelo más general que anide a los dos:
  $y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 log(x_1) + \gamma_4 log(x_2) + \epsilon$

- y entonces contrastar

- $H0 : \gamma_3 = \gamma_4 = 0$ para comprobar el primer modelo o también
- $H0 : \gamma_1 = \gamma_2 = 0$ para comprobar el segundo modelo.

Especificación: Contraste de Davidson & MacKinnon

- Si un modelo es verdadero, los valores ajustados del otro modelo no deberían ser significativos:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

$$\widehat{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1 log(x_1) + \hat{\beta}_2 log(x_2) + \epsilon$$

- De esta forma se comprobaría el primer modelo,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \widehat{\hat{y}} + error$$

- $H0 : \theta_1 = 0$, mediante un contraste de la t (bilateral).
- Equivalentemente se podría contrastar la misma hipótesis en

$$y = \beta_0 + \beta_1 log(x_1) + \beta_2 log(x_2) + \theta_1 \hat{y} + error$$

Problemas de los contrastes no anidados

- No tiene porque aparecer un modelo claramente superior: Se pueden rechazar los dos modelos o ninguno (aunque se podrían comparar con el R^2 ajustado).
- Rechazar un modelo no implica necesariamente que el otro modelo sea el correcto.
- La situación se complica si las variables dependientes son diferentes, y y $log(y)$.

Utilización de variables proxy para variables explicativas no observadas

- Este modelo

  $$ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 abil + u$$

- reconoce que no podemos mantener habilidad constante cuando medimos el retorno de *educ* and *exper*. Si educ está correlado con *abil*, entonces dejar habilidad en el término de error hará que el estimador de β_1 (y de β_2) esté sesgado.
- El objetivo es obtener estimadores insesgados de β_1 y β_2 : en general no podemos esperar poder obtenerlos de β_0 ni tampoco de β_3 porque no observamos *abil* (además no sabríamos interpretarlo porque habilidad es un concepto muy vago).
- Una posible solución para arreglar el problema de variables omitidas es usar una variable proxy en su lugar.

Variables proxy

- Una variable proxy es algo que está relacionado con la variable que nos gustaría controlar pero que no observamos.

- Ejemplo: cociente intelectual (IQ) como un proxy por habilidad. Esto no requiere que IQ sea lo mismo que habilidad, lo que necesitamos es que IQ esté correlado con habilidad.
- Modelo básico:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

- donde tenemos datos de y, x_1 y x_2, pero
  - x_3^* no se observa,
  - aunque se dispone un proxy x_3

Condiciones sobre la variable proxy x_3

- Lo mínimo es que tenga una relación con la variable x_3 : $x_3^* = \delta_0 + \delta_3 x_3 + v_3$
- v_3 es un error debido al hecho de que x_3^* y x_3 no están exactamente relacionados.
- δ_3 mide la relación entre x_3^* y x_3 , típicamente $\delta_3 > 0$. Si $\delta_3 = 0$, entonces x_3 no puede ser un proxy para x_3^* .
- δ_0 permite que x_3^* y x_3 estén medidas en diferentes escalas.

¿Cómo usar variables proxy para obtener estimadores insesgados de β_1 y β_2?

- Solución plug-in: la idea es pensar que x_3^* y x_3 son iguales, y hacer la regresión MCO.

  y sobre $x1$, $x2$, $x3$,

- reemplazando x_3^* por x_3 en su lugar.
- Esto parece sensato, pero hay que estudiar si consigue producir estimadores consistentes de β_1 y β_2.

Supuestos para obtener estimadores MCO consistentes con variables proxy

- Supuestos sobre el término de error u
  - El término de error u está incorrelado con $x1$, $x2$ y x_3^* , lo que es el supuesto habitual en el modelo original:

    $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

  - El término de error u está incorrelado con x_3.
    Es decir, x_3 no es necesario en el modelo poblacional una vez que ya incorporamos $x1$, $x2$ y x_3^* , puesto que sólo x_3^* afecta directamente a y.

Assumptions in order to get consistent OLS estimators with proxy variables

- Assumptions about the error term v_3
  - Hablamos de esto en Clases.

Propiedades de MCO con errores de medida

- No siempre es posible recoger datos sobre la variable que afecta realmente el comportamiento económico.
- Cuando se usa una medida imprecisa de una variable económica en un modelo de regresión, entonces el modelo contiene errores de medida.
- En este caso los EMCO pueden dejar de ser consistentes y es posible calcular su sesgo.
- Las consecuencias son semejantes a los de la utilización de proxies, pero son problemas conceptualmente diferentes.
- En el caso de las variables proxies buscamos una variable que esté asociada con la variable inobservada. Normalmente no nos interesa su efecto parcial, sino el de otras variables.
- En el caso de errores de medida, la variable que no observamos tiene un significado cuantitativo bien definido, pero nuestras medidas pueden contener errores. Además generalmente estamos interesados en el efecto marginal de esta variable.

Errores de Medida en la Variable Dependiente

- Tenemos problemas para medir la variable dependiente y^* en la población

  $$y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

  que satisface todas las condiciones de G-M.

- y es la medida observable de y^* por lo que es esperable que y e y^* difieran, al menos para una parte de la población.
- El error de medida en la población se define como

  $$e_0 = y - y^*.$$

Errores de Medida en la Variable Dependiente

- La clave es como e_0 se relaciona con otros factores:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u + e_0$$

- donde $u + e_0$ es el nuevo término de error y se pueden calcular los EMCO.

¿Cuando los EMCO con y en lugar de y^* son consistentes?

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u + e_0$$

- u tiene media cero y está incorrelado con cada x_j .
- Si e_0 no tuviese media cero, eso sólo afectaría a la estimación de β_0, lo que no suele ser un problema.
- La clave es la relación de e_0 con las x_j .

- Si e_0 está incorrelada con todas las x_j los EMCO son insesgados, consistentes y todos los procedimientos habituales son válidos.

¿Cuando los EMCO con y en lugar de y^* son consistentes?

- Si e_0 y u están incorrelados, entonces

  $$Var(u + e_0) = Var(u) + Var(e_0)$$

  error de medida $\Rightarrow$ menor eficiencia de los EMCO.
- Por tanto, si el error de medición es simplemente error aleatorio independiente de las variables aleatorias, los EMCO son perfectamente apropiados.

Errores de Medida en las Variables Independientes

- Este problema es generalmente más relevante:

  $$y = \beta_0 + \beta_1 x_1^* + u$$

  que satisface al menos las primeras 4 condiciones de G-M.
- Sin embargo x_1^* no se observa, pero sí una medida x_1.
- Error de medida en la población:

  $$e_1 = x_1 - x_1^*$$
- Suponemos que $E[e_1] = 0$, aunque no afecta los resultados.
- También suponemos que u está incorrelado con x_1 y con x_2 : $E[y|x_1, x_1^*] = E[y|x_1^*]$.

Propiedades del EMCO con Errores de Medida en las Variables Independientes

Depende de qué supuestos hagamos sobre

$e_1 = x_1 - x_1^*$.

- e_1 está incorrelado con la medida observada x_1 :

  $Cov[x_1, e_1] = 0$.

- En este caso x_1^* tiene que estar correlado con e_1.

  $y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$

Propiedades del EMCO con Errores de Medida en las Variables Independientes

Depende de qué supuestos hagamos sobre

- donde el error está incorrelado con x_1 : EMCO es consistente, aunque

  $$Var[u - \beta_1 e_1] = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$$

  implica que es menos eficiente.

Propiedades del EMCO con Errores de Medida en las Variables Independientes

- El supuesto habitual es el Clásico Error en Variables, CEV,

  $Cov[x_1^*, e_1] = 0,$

  donde el error de medida está incorrelado con la variable explicativa inobservada.
- El supuesto viene de escribir

  $x_1 = x_1^* + e_1$

  y suponer que los dos componentes de x_1 están incorrelados.

Propiedades del EMCO con Errores de Medida en las Variables Independientes

- Entonces x_1 y e_1 deben estar correlados:

  $$Cov[x_1, e_1] = E[x_1 e_1] = E[x_1^* e_1] + E[e_1^2] = 0 + \sigma_{e_1}^2$$

- Esto va a causar problemas porque
  $Cov[x_1, u - \beta_1 e_1] = -\beta_1 Cov[x_1, e_1] = -\beta_1 \sigma_{e_1}^2$,

  por lo que los EMCO están sesgados y son inconsistentes

Propiedades del EMCO con Errores de Medida en las Variables Independientes (2)

- $plim(\hat{\beta}_1) = \beta_1 + \frac{Cov[x_1, u - \beta_1 e_1]}{Var[x_1]}$
- $plim(\hat{\beta}_1) = \beta_1 + \frac{-\beta_1 \sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} = \beta_1 (1 - \frac{\sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}})$
- $plim(\hat{\beta}_1) = \beta_1 (\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}})$

Problemas con los datos: Datos Faltantes

- A veces se colecciona una muestra aleatoria de personas, colegios, ciudades, y se descubre más tarde que hay información perdida sobre algunas variables clave para algunas unidades de la muestra.
- Si falta un dato para una observación de la variable dependiente o una de las independientes, entonces ese dato no se puede emplear para hacer una regresión múltiple y $n \downarrow$.
- ¿Hay otras consecuencias sobre el análisis estadístico de los datos?
- Depende sólo de porqué faltan los datos. Si los datos se perdieron aleatoriamente, entonces solamente n se reduce: baja la precisión de la estimación, pero no hay sesgos, por que RLM. 2 sigue siendo válida.

Muestras no Aleatorias

- Los datos faltantes son más problemáticos cuando producen una muestra no aleatoria de la población.
- Ejemplo: puede ocurrir que la probabilidad de que el dato de educación falte para aquellas personas que tienen menor nivel de educación.
- Ejemplo: puede que sea más fácil conseguir el dato de IQ para aquellos que tienen niveles más altos de IQ.
- En estos casos la muestra no es representativa de la población: el supuesto RLM.2 no es válido.