Universidad Nacional de La Plata

Departamento de Economía
Facultad de Ciencias Económicas
Universidad Nacional de La Plata

Econometría

Tópicos de Econometría Aplicada
(Notas de Clase)

Walter Sosa Escudero

Trabajo Docente Nro. 2
Septiembre 1999

Tópicos de Econometría Aplicada

Notas de Clase

Walter Sosa Escudero
Universidad Nacional de La Plata

Septiembre de 1999

Correspondencia: Facultad de Ciencias Económicas, Departamento de Economía, 5o piso, Of. 519; La Plata, Argentina. e-mail: wsosa@feedback.net.ar

NOTA PRELIMINAR: Estas notas fueron escritas para ser distribuidas en cursos dictados en las Universidades Nacional de La Plata, de San Andres, Di Tella, y el Ministerio de Economía y Obras y Servicios Públicos de la Nación, Argentina. El objeto de las mismas es facilitar el desarrollo de los cursos y de ningún modo intentan cubrir los temas tratados con la profundidad de los textos y lecturas sugeridas en los programas de los cursos. A modo de ejemplo puede consultarse la página de uno de los cursos en donde estas notas fueron utilizadas:

`http://www.udesa.edu.ar/cursos/econometria/index.html`

Versión preliminar. Se agradecen comentarios.

Índice General

Capítulo 1

El modelo lineal general bajo los supuestos clásicos

En esta sección repasaremos los conceptos básicos del modelo lineal general con K variables bajo los supuestos clásicos. No es nuestro objetivo presentar un análisis detallado de este tema (para el cual existen excelentes referencias) sino algunos resultados básicos a modo de repaso. Aquellos interesados en realizar una revisión detenida de estos temas pueden consultar textos recientes como Johnston y DiNardo (1997) o Greene (1996). Esta sección supone conocimientos básicos de álgebra matricial, los cuales se presentan en forma resumida en los apéndices de los textos mencionados anteriormente. Schott (1997) y Harville (1997) son muy buenas referencias para aquellos que deseen un tratamiento mas extensivo.

1.1 Formulación del modelo

En esta especificación, la variable dependiente Y es una función lineal de K variables explicativas $(X_1, X_2, \ldots, X_K)$. u_i es un término aleatorio que representa el carácter no exacto de la relación entre Y y las variables explicativas. Para una muestra de n observaciones, el modelo puede escribirse como:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_K X_{Ki} + u_i \qquad i = 1, \ldots, n \tag{1.1}$$

en donde los $\beta_k, k = 1, \ldots, K$ son los coeficientes de la relación lineal. X_{ki} corresponde a la i-esima observación de la variable explicativa k. Trivialmente, la primer variable explicativa X_{1i} corresponde a una constante igual a 1 para todas las observaciones, por lo cual β_1 corresponde al *intercepto* de la relación lineal. Adicionalmente, haremos los siguientes supuestos, conocidos como *supuestos clásicos*:

1. $E(u_i) = 0$, para todo $i = 1, \ldots, n$. El término aleatorio tiene esperanza igual a cero para todas las observaciones. Este supuesto implica que en promedio la relación entre Y y las $X's$ es exactamente lineal, aunque las realizaciones particulares de los u_i's pueden ser distintas de cero.

2. $Var(u_i) = \sigma^2, i = 1, ..., n$. La varianza del término aleatorio es constante para todas las observaciones. Esto se conoce como supuesto de *homoscedasticidad*.

3. $Cov(u_i, u_j) = 0$, para todo $i \neq j$. Las covarianzas del término aleatorio entre dos observaciones distintas son iguales a cero. Si las observaciones se encuentran ordenadas a lo largo del tiempo esto implica que la correlación entre los términos aleatorios correspondientes a distintos periodos es nula. En este caso el supuesto se conoce como de no *autocorrelación* o no *correlación serial*.

4. Los vectores formados con las observaciones de las variables explicativas $(X_k, k = 1, \ldots, K)$ son no estocásticos y linealmente independientes. Esto último implica que ningún vector de observaciones de las variables explicativas puede ser obtenido como una combinación lineal de los restantes vectores. Por ejemplo, si en un modelo en donde la variable explicada es el consumo, incluyéramos el ingreso medido en pesos y el equivalente medido en marcos, obviamente el segundo puede ser obtenido como el producto del primero por un escalar. Por el contrario, si incluyéramos al ingreso y al ingreso al cuadrado como variables explicativas, esto no violaría el supuesto de independencia lineal ya que el ingreso al cuadrado no es una función *lineal* del ingreso. El supuesto de independencia lineal se conoce como de *no multicolinealidad*.

El modelo (1.1) puede se reexpresado en términos matriciales de la siguiente manera:

$$Y = X\beta + u \tag{1.2}$$

en donde Y es un vector columna de n observaciones con elemento característico Y_i. X es una matriz con n filas y k columnas, con elemento típico igual a X_{ki}, $k = 1, \ldots, K$ y $i = 1, \ldots, n$. Nótese que la primera columna de la matriz X es un vector con todas sus posiciones igual a uno. β es un vector de k parámetros desconocidos y u es un vector columna de n elementos.

Los supuestos clásicos pueden expresarse en términos matriciales como:

1. $E(u) = 0$. En este caso el operador esperanza $(E())$ afecta a un vector aleatorio (u) y tiene como elemento característico a la esperanza de cada posición $(E(u_i))$

2. $Var(u) = E(uu') = \sigma^2 I$, en donde I es la matriz identidad con dimensión n. Denotando con ω_{ij} al elemento i, j de la matriz $Var(u)$, los elementos

de la diagonal de la misma ($\omega_{ii}, i = 1, \ldots, n$) corresponden a la varianza da la i-esima observación, y los elementos $\omega_{ij}, i \neq j$ corresponden a las covarianzas entre las observaciones i y j, de lo cual surge que $Var(u)$ es una matriz simétrica.

3. X es una matriz no estocástica con rango K, lo cual denotaremos $\rho(X) = K$. Es importante notar que $\rho(X)$ implica $\rho(X'X) = K$. Este último resultado implica que la matriz inversa de $(X'X)$ existe. Es importante notar que este supuesto implica que el número de observaciones n tiene que ser necesariamente mayor o igual que el número de variables explicativas. De no ser este el caso, el rango fila de X (el máximo número de vectores fila de X linealmente independientes) sería necesariamente menor que K y por lo tanto $\rho(X) < K$.

De acuerdo a estos supuestos:

$$E(Y_i) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_K X_{Ki}$$

de modo que los coeficientes $\beta_k, k = 2, \ldots, K$ se interpretan como cambios marginales en el valor esperado de la variable dependiente Y que resultan de cambios marginales en las variables explicativas. Como mencionáramos anteriormente, β_1 corresponde a la ordenada al origen de la relación lineal. Si X_k es una variable binaria que toma valor 1 si un individuo pertenece a una cierta clase y 0 si no pertenece, es fácil observar que:

$$E(Y_i; X_{ki} = 1) - E(Y_i; X_{ki} = 0) = \beta_k$$

de modo que en el caso en que X_k sea una variable binaria, el coeficiente correspondiente a esa variable se interpreta como la diferencia en el valor esperado de la variable explicada entre individuos que pertenecen y que no pertenecen a la clase denotada por X_k. Por ejemplo, si Y_i midiera el ingreso de un individuo y X_k fuera un indicador binario que toma valor 1 si la persona es hombre y 0 si es mujer, β_k se interpreta como la diferencia en ingreso entre hombres y mujeres, manteniendo el resto de los factores explicativos constantes.

1.2 Estimación mínimo-cuadrática

El objetivo consiste en encontrar 'buenas' estimaciones para los parámetros desconocidos del modelo, β y σ^2, para lo cual debemos comenzar definiendo alguna noción de optimalidad bajo la cual el calificativo 'bueno' tenga sentido. Llamemos $\hat{\beta}$ al estimador de β. Definamos $\hat{Y} \equiv X\hat{\beta}$ como el estimador de $E(Y) = X\beta$. El vector de errores de estimación o *residuo* estará definido como $e \equiv Y - \hat{Y}$. La idea es encontrar estimadores de β que hagan que el vector e sea 'pequeño' en cierto sentido. Intuitivamente, sería deseable igualar todas las

coordenadas de este vector igual a cero, lo cual es virtualmente imposible dadas las características del problema. Para visualizar gráficamente este problema, consideremos el caso de una sola variable explicativa (X) ademas del intercepto ($K = 2$). En este caso, las observaciones de Y y X pueden ser graficadas como una nube de puntos en un espacio euclideo bidimensional. El problema de estimación puede ser visualizado como pasar una recta por las observaciones, y el error de estimación para cada observación i corresponde a la distancia vertical entre Y_i y el valor de la la recta evaluada en el punto X_i. Con solo dos observaciones ($n = 2$) es posible hacer que todos los e_i's sean iguales a cero ya que por ambos puntos pasa exactamente una linea recta. Obviamente, con $n > 2$ observaciones (que no caigan en una misma linea recta, lo cual es descartado por el supuesto $\rho(X) = K$) es imposible hacer que todos los e_i sean iguales a cero. Por esto es necesario introducir una noción de 'tamaño' de los e_i's en forma conjunta.

El estimador *mínimo cuadrático (MC)* de β es aquel que minimiza la suma de los residuos al cuadrado:

$$e'e = \sum_{i=1}^{n} e_i^2$$

Nótese que de acuerdo a este criterio los errores positivos importan lo mismo que los errores negativos ya que el signo de los mismos desaparece al elevar cada término al cuadrado.

En el caso del modelo lineal general:

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

Es fácil verificar que las condiciones de primer orden del problema de minimización son:

$$X'e = 0$$

o, alternativamente:

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

lo cual define un sistema lineal de K ecuaciones con K incógnitas (β). Bajo la condición de que $\rho(X) = K$ (lo cual implica que $\rho(X'X)$ también es K y que por lo tanto es invertible), el problema tiene una solución única igual a:

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

por lo que $\hat{\beta}$ es llamado el estimador mínimo cuadrático de β

1.2.1 Propiedades básicas

En esta subsección describiremos algunas propiedades básicas que surgen como consecuencia directa del proceso de minimización de la suma de cuadrados residuales.

1. $\text{Cov}(X_k, e) = 0, k = 1, \ldots, K$. Este resultado indica que la correlación muestral entre las variables explicativas y el vector de residuos e es nula.

2. $\sum_i^n e_i = 0$. Si la matriz de variables explicativas incluye una constante, de las condiciones de primer orden surge automáticamente que el proceso de minimización de la suma de cuadrados residuales impone como condición necesaria que la suma de los residuos sea igual a cero.

3. $\text{Cov}(\hat{Y}, e) = 0$ También surge como consecuencia del proceso de minimización que el vector de predicciones de Y $(\hat{Y})$ esta no correlacionado con el vector de residuos.

4. *Descomposición de la suma de cuadrados*: Es fácil demostrar que si $\hat{\beta}$ es el estimador mínimo cuadrático, se cumple la siguiente descomposición.

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{\hat{Y}}_i)^2 + \sum e_i^2$$

o,

$$\text{SCT} = \text{SCE} + \text{SCR}$$

en donde SCT, SCE y SCR se refieren a la suma de cuadrados *totales, explicados y residuales* respectivamente. Nótese que a diferencia de SCE y SCR, SCT no depende de las variables explicativas ni del estimador mínimo cuadrático. Esta expresión dice que la variabilidad total de la variable explicada Y puede descomponerse como la suma de la variabilidad explicada por la predicción basada en el estimador mínimo cuadrático (SCE), más la variabilidad atribuida a los residuos representada por la suma de los mismos al cuadrado (SCR).

De lo antedicho, resulta obvio proponer la siguiente expresión:

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

como una medida de *bondad del ajuste*. La misma indica la proporción de la variabilidad total que es explicada por el modelo lineal. Esta medida se conoce como *coeficiente de determinación*.

Es importante observar que $\hat{\beta}$ maximiza R^2. De la segunda expresión, como SCT no depende del estimador mínimo cuadrático ni de X, por construcción $\hat{\beta}$ minimiza SCR y, por lo tanto, maximiza R^2.

1.3 Propiedades estadísticas del estimador MC

Hasta el momento solo hemos utilizado el supuesto de que $\rho(X) = K$ para obtener una solución al problema de estimación y para derivar algunas propiedades elementales. A esta altura, el estimador obtenido es 'bueno' en el sentido de que minimiza una noción agregada de error: minimiza la suma de residuos al cuadrado. El paso siguiente consiste en explorar algunas propiedades estadísticas que se desprenden de los supuestos hasta ahora no utilizados e investigar si el estimador propuesto es bueno en algún otro sentido. Dentro del contexto clásico, procederemos mostrando algunas propiedades básicas del estimador obtenido para posteriormente demostrar que el estimador mínimo cuadrático es el mejor dentro de cierta clase de estimadores.

1. Comencemos observando que $\hat{\beta}$ se obtiene como una transformación lineal del vector de observaciones de la variable dependiente Y. En este caso, $\hat{\beta} = AY$ en donde $A = (X'X)^{-1}X'$ es la matriz que transforma a Y linealmente en $\hat{\beta}$. Nos referiremos a esta propiedad diciendo que $\hat{\beta}$ es un estimador *lineal*.

2. Bajo los supuestos clásicos, $\hat{\beta}$ es un estimador *insesgado* de β en el modelo (2), esto es, $E(\hat{\beta}) = \beta$

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y && (1.3)\\ &= (X'X)^{-1}X'(X\beta + u) && (1.4)\\ &= \beta + (X'X)^{-1}X'u && (1.5)\\ E(\hat{\beta}) &= \beta + (X'X)^{-1}X'E(u) && (1.6)\\ &= \beta && (1.7) \end{aligned}$$

   En la segunda linea reemplazamos Y por su definición en el modelo lineal y en la cuarta linea utilizamos el supuesto de que X es no estocástica. En última linea utilizamos el supuesto de que $E(u) = 0$.

3. $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$

$$\begin{aligned} V(\hat{\beta}) &= E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'] && (1.8)\\ &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] && (1.9)\\ &= E[((X'X)^{-1}X'u)((X'X)^{-1}X'u)'] && (1.10)\\ &= E[(X'X)^{-1}X'uu'X(X'X)^{-1}] && (1.11)\\ &= (X'X)^{-1}X'E(uu')X(X'X)^{-1} && (1.12)\\ &= (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} && (1.13)\\ &= \sigma^2(X'X)^{-1} && (1.14) \end{aligned}$$

La primer linea de la prueba corresponde a la definición de la matriz de varianzas $V(\hat{\beta})$. La segunda linea utiliza el resultado de que $\hat{\beta}$ es insesgado. En la tercer linea utilizamos el resultado (5) en la prueba anterior. Al pasar de (11) a (12) usamos el supuesto de que X es no estocástica y en la linea siguiente el supuesto de que $Var(u) = \sigma^2 I$. El resultado final se obtiene de simplificar en la anteúltima linea.

4. *Teorema de Gauss-Markov*: bajo los supuestos clásicos, el estimador mínimo cuadrático ($\hat{\beta}$) de β en el modelo lineal (2) es el mejor estimador lineal insesgado. Más específicamente, el Teorema dice que para todo vector c de K constantes:

$$Var(c'\beta^*) \geq Var(c'\hat{\beta})$$

en donde β^* es cualquier estimador lineal e insesgado de β. Esto es, cualquier combinación lineal de los coeficientes de un estimador lineal insesgado de β tiene varianza por lo menos tan grande como la correspondiente a la misma combinación lineal basada en el estimador mínimo-cuadrático.

Este teorema indica cuan 'bueno' puede ser el estimador MC si los supuestos clásicos se verifican. Dentro de cierta clase (la de los estimadores insesgados que son funciones lineales de Y) el estimador MC es el mejor en el sentido de que tiene la mínima varianza. Este resultado debe tomarse con precaución ya que no descartamos la posibilidad de que haya estimadores sesgados (y/o no lineales) que superen al estimador MC en varianza.

1.4 Estimación de σ^2

Todavía nos resta obtener un estimador para el parámetro σ^2. Propondremos:

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

el cual es un estimador insesgado de σ^2

1.5 Inferencia en el modelo lineal con K variables

A partir de los supuestos clásicos pudimos mostrar algunas propiedades básicas de los estimadores propuestos. El paso siguiente consiste en derivar propiedades de los mismos que nos permitan realizar inferencias acerca de los coeficientes del modelo.

Inicialmente, estaremos interesados en evaluar hipótesis *lineales* acerca del vector de parámetros β. Consideremos el siguiente ejemplo. Supongamos que estamos interesados en estimar los parámetros de una función de producción del tipo Cobb-Douglas:

$$Y = AK^{\beta_1} L^{\beta_2} e^u$$

en donde Y corresponde al producto, K al factor capital, L al trabajo, u es un término aleatorio y A es una constante. A, β_1 y β_2 son los parámetros del modelo. Tomando logaritmos naturales la función de producción puede ser expresada como:

$$y = \alpha + \beta_1 k + \beta_2 l + u$$

en donde las variables en minúsculas indican los logaritmos naturales de las variables en mayúsculas y $\alpha = \ln A$.

Algunas hipótesis interesantes a evaluar pueden ser las siguientes:

1. *Significatividad de las variables explicativas.* Por ejemplo, la hipótesis $H_0 : \beta_1 = 0$ corresponde a la hipótesis de que el factor capital no es una variable relevante para la determinación del producto.

2. *Igualdad de coeficientes.* Por razones económicas, podríamos estar interesados en evaluar $H_0 : \beta_1 = \beta_2$ (o $H_0 : \beta_1 - \beta_2 = 0$), la cual indicaria que las elasticidades del producto con respecto a los factores son iguales.

3. *Restricciones sobre los coeficientes.* Una cuestión de interes es determinar si los rendimientos son constantes a escala. Esto corresponde a evaluar $H_0 : \beta_1 + \beta_2 = 1$

4. *Relevancia del modelo lineal.* En términos generales, podríamos cuestionar si todas las variables explicativas son simultáneamente relevantes para la determinación del producto: $H_0 : \beta_1 = \beta_2 = 0$

Es fácil observar que todas las hipótesis anteriores implican una o más restricciones lineales sobre el vector de coeficientes β. En términos generales, un conjunto de q hipótesis lineales sobre el vector β puede ser expresado como:

$$H_0 : R\beta = r$$

en donde R es una matriz $q \times k$ y r es un vector $q \times 1$

Por ejemplo, las hipótesis descriptas anteriormente corresponden a:

1. $R = [0\ 1\ 0],\ r = 0,\ q = 1$

2. $R = [0\ 1\ -1],\ r = 0,\ q = 1$

3. $R = [0\ 1\ 1],\ r = 1,\ q = 1$
4. $R = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, r = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, q = 2$

Entonces, el objetivo es evaluar hipótesis del tipo $H_0 : R\beta - r = 0$ basándonos en un estimador de β, esto es, en: $R\hat{\beta} - r$. La idea consiste en computar $R\hat{\beta} - r$ basado en las observaciones de la muestra disponible y determinar si dicho valor es significativamente distinto de cero, para lo cual necesitamos conocer la distribución de dicho estadístico.

Debemos agregar el siguiente supuesto acerca de la distribución de los u_i:

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

el cual dice que los términos aleatorios u_i se distribuyen normalmente con media cero y varianza igual a σ^2 para todas las observaciones de la muestra. O, alternativamente, que:

$$u \sim N(0, \sigma^2 I)$$

lo cual indica que el vector u tiene distribución normal multivariada con media igual al vector cero y matriz de varianzas igual a $\sigma^2 I$. En el modelo lineal, Y resulta ser una función lineal de u ($Y = X\beta + u$) por lo que Y también tiene distribución normal multivariada:

$$Y \sim N(X\beta, \sigma^2 I)$$

En forma similar, $\hat{\beta}$ tiene también distribución normal multivariada ya que es una transformación lineal de Y ($\hat{\beta} = (X'X)^{-1}X'Y$), con esperanza $E(\hat{\beta}) = \beta$ y $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$, de modo que:

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

A fines prácticos, cualquier hipótesis lineal del tipo $R\beta - r = 0$ puede ser evaluada utilizando el siguiente estadístico:

$$F = \frac{(R\hat{\beta} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{e'e/(n - K)} \tag{1.15}$$

el cual bajo la hipótesis nula tiene distribución $F(q, n-K)$. Valores altos de este estadístico corresponden a valores altos de $R\hat{\beta} - r$ los cuales, de acuerdo a los valores críticos de la distribución $F(q, n - K)$ indicarían rechazo de la hipótesis nula $H_0 : R\beta - r = 0$.

Para las hipótesis discutidas anteriormente, el estadístico F corresponde a:

1. $H_0 : \beta_1 = 0$. $R\hat{\beta} - r = \hat{\beta}_1$ y $R(X'X)R' = c_{11}$ en donde c_{11} es el elemento $(1,1)$ de $Var(\hat{\beta})$. De modo que el estadístico F corresponde a:

$$F = \frac{\hat{\beta}_1^2}{\text{Var}(\hat{\beta}_1)^2}$$

y que:

$$t = \sqrt{F}$$

tiene, bajo la hipótesis nula, distribución 't' con $n - K$ grados de libertad.

2. En este caso es fácil verificar que el estadístico F corresponde a:

$$F = \frac{(\hat{\beta}_1 - \hat{\beta}_2)^2}{V(\hat{\beta}_1 - \hat{\beta}_2)}$$

y también que la raíz cuadrada del mismo tiene distribución 't' con $n - k$ grados de libertad bajo H_0.

3. Similarmente, el estadístico correspondiente será:

$$F = \frac{(\hat{\beta}_1 + \hat{\beta}_2 - 1)^2}{Var(\hat{\beta}_1 + \hat{\beta}_2 - 1)}$$

y también $\sqrt{F}$ tiene distribución $t(n - K)$.

4. Para este caso se puede mostrar que el estadístico correspondiente es:

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

en donde $R^2 = SCE/SCT = 1 - SCR/SCT$ es el coeficiente de determinación del modelo original.

Capítulo 2

Máxima Verosimilitud

En esta sección presentamos algunos resultados básicos sobre métodos de estimación e inferencia basados en el principio de máxima verosimilitud (MV). Un tratamiento más detallado puede encontrarse en Davidson y MacKinnon (1993). Acerca de las pruebas de resultados asintóticos, Newey y McFadden (1994) contiene abundantes detalles.

2.1 Conceptos básicos

Sea X una variable aleatoria con distribución $F(x;\theta_0)$ y función de densidad $f(x;\theta_0)$, en donde θ_0 es un vector de K parámetros desconocidos. Por simplicidad de presentación, primero consideraremos el caso de un solo parámetro ($K = 1$) y luego extenderemos los resultados al caso general. Una *muestra aleatoria* de n observaciones de X es un conjunto de n variables aleatorias independientes denotadas como $X_i;\ i = 1, \ldots, n$, en donde $X_i \sim f(x;\theta_0)$, es decir, las n variables aleatorias son independientes e idénticamente distribuidas (*i.i.d*). El objetivo consiste en estimar θ_0 a partir de la muestra aleatoria.

Sea $\mathbf{X} = (X_1, X_2, ..., X_n)$. En términos generales, un *estimador* de θ_0 (denotado como $\hat{\theta}_0$) es cualquier función de la muestra aleatoria:

$$\hat{\theta}_0 = T(\mathbf{X})$$

Un primer objetivo consistirá en explorar algún criterio que permita comparar estimadores de acuerdo a su 'calidad' y, que en consecuencia, permita decir que un estimador es preferible o no a otro. Es imposible definir la noción de que un estimador es mejor que otro en forma unívoca y uniforme. Una propuesta consiste en definir ciertas propiedades deseables que un estimador debería poseer y comparar estimadores de acuerdo a las mismas.

Propiedades de muestras pequeñas:

Las propiedades de muestras pequeñas se refieren a las propiedades del estimador entendido como una variable aleatoria, para un tamaño de muestra dado.

1. *Insesgadez:* $E(\hat{\theta}_0) = \theta_0$ Un estimador es insesgado si su valor esperado es igual al parámetro que se desea estimar.

2. *Eficiencia:* Esta propiedad se refiere a la comparación entre dos estimadores insesgados del mismo parámetro. Un estimador insesgado $\hat{\theta}_0$ es mas *eficiente* que otro estimador insesgado $\hat{\theta}^*$ si su varianza es menor, o sea, si $V(\hat{\theta}_0) \square V(\hat{\theta}^*)$.

3. *Distribución:* bajo ciertas condiciones generales[1] el estimador es una variable aleatoria ya que es una función de un vector aleatorio, y como tal puede ser caracterizada por su función de distribución, la cual se deriva de la distribución de las variables que componen la muestra aleatoria. Para muchas aplicaciones (test de hipótesis, etc.) será de utilidad conocer la distribución del estimador. En algunos casos es posible derivar analíticamente la distribución de $\hat{\theta}_0$ como una transformación de la distribución de $\mathbf{X}$ Por ejemplo, si X_i es una variable aleatoria normal con media μ y varianza σ^2, entonces la media muestral $\bar{X} = \sum X_i/n$ tiene distribución normal con media μ y varianza σ^2/n

Propiedades de muestras grandes

En muchas situaciones es complejo (sino imposible) conocer las propiedades de muestra pequeña, pero resulta ser que para tamaños de la muestra infinitamente grandes, es posible conocer ciertas propiedades de un estimador. Consideremos la siguiente secuencia de estimadores de θ_0: $\hat{\theta}_1 = T(X_1), \hat{\theta}_2 = T(X_1, X_2), \ldots, \hat{\theta}_n = T(X_1, X_2, \ldots, X_n)$ la cual se forma a través de ampliar el tamaño de la muestra. Las propiedades de muestras grandes de un estimador se refieren a las características de esta secuencia de estimadores que se verifican cuando la muestra tiende a ampliarse infinitamente. Intuitivamente, es deseable que cuando la muestra sea infinitamente grande, el estimador tienda a ser igual al parámetro que se desea estimar. Esta noción requiere especificar precisamente que significa que una variable aleatoria (el estimador) converga a un número real (el parámetro estimado).

1. *Consistencia:* un estimador de θ_0 es *consistente* si el límite probabilístico de $\hat{\theta}_n$ es igual a θ_0. Nótese que la secuencia de estimadores es una secuencia de variables aleatorias mientras que el límite de esta secuencia es un número real. Mas específicamente, $\hat{\theta}_n$ es un estimador consistente de θ_0 si para todo $\epsilon > 0$:

[1] Básicamente, que $\theta()$ sea una función medible. Ver Durrett (1996, p. 11)

$$\lim_{n\to\infty} P[|\hat{\theta}_n - \theta_0| > \epsilon] = 0$$

Esta noción equivale al concepto de convergencia en probabilidad de una variable aleatoria. Para entender como funciona este concepto, nótese que para ϵ y n dados, $|\hat{\theta}_n - \theta_0| > \epsilon$ define un evento cuya probabilidad de ocurrencia se puede determinar de acuerdo a la distribución de $\hat{\theta}_n$. Cuando n aumenta, la distribución de referencia (la de $\hat{\theta}_n$) cambia, de modo que el límite en la definición de consistencia se refiere al límite de una secuencia de probabilidades. Nótese que dicho límite deber ser cero para cualquier elección de ϵ, en particular para uno arbitrariamente pequeño, de modo que, intuitivamente, la noción de convergencia implica que la distribución del estimador tiende a colapsar en un punto, que es precisamente el valor que se desea estimar.

2. *Convergencia en distribución:* Sea $\hat{\theta}_n$ una secuencia de estimadores con distribución $G_n(\theta)$. $\hat{\theta}_n$ converge en distribución a una variable aleatoria $\hat{\theta}_\infty$ con distribución $G_\infty(\theta)$ si:

$$\lim_{n\to\infty} |G_n(\theta) - G_\infty(\theta)| = 0$$

para todos los puntos de continuidad de $G_\infty(\theta)$. En este caso, diremos que $G_\infty(\theta)$ es la *distribución límite* de $\hat{\theta}_n$. También haremos referencia a los momentos límite de la distribución. Por ejemplo, la media y la varianza de $G_\infty(\theta)$ serán la media y la varianza asintótica de $\hat{\theta}_n$.

Una cuestión básica es que si $\hat{\theta}_n$ es consistente para θ_0 entonces, trivialmente, $\hat{\theta}_n$ converge en distribución a una variable aleatoria degenerada cuya masa de probabilidad estará concentrada en el punto de convergencia. En muchos casos existirá una *transformación estabilizante* $h()$ tal que $h(\hat{\theta}_n)$ converga en distribución a una variable aleatoria con distribución no degenerada.

Los dos siguientes resultados ejemplifican esta situación:

Ley debil de grandes numeros (Kinchine): Sea $X_i, i = 1, \ldots, n$ una muestra aleatoria i.i.d. con $E(X_i) = \mu < \infty$. Sea $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ Entonces $\bar{X}_n$ es consistente para μ, o, equivalentemente, $\bar{X}_n$ converge en probabilidad a su esperanza matemática $E(X)$.

Teorema central del limite (Lindeberg-Levy): Sea $X_i, i = 1, \ldots, n$ una muestra aleatoria con $E(X_i) = \mu < \infty$ y $V(X_i) = \sigma^2 < \infty$. Entonces:

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{D} N(0, \sigma^2)$$

lo significa que la variable aleatoria $\sqrt{n}(\bar{X} - \mu)$ converge en distribución a otra variable aleatoria normalmente distribuida. Del primer resultado surge que si bien la media muestral converge en probabilidad a una constante (su esperanza), una transformación de la misma (su versión estandarizada) converge a una variable aleatoria no degenerada, cuya distribución es normal estándar.

Si los resultados anteriores son válidos para muestras infinitamente grandes, el mismo sugiere la siguiente aproximación para $\bar{X}_n$:

$$X_n \sim N(\mu, \sigma^2/n)$$

De modo que la media muestral se aproxima asintóticamente a una variable aleatoria normal con media μ y varianza σ^2/n.

De estos resultados surge una tercera propiedad deseable asintóticamente:

Normalidad asintótica y eficiencia asintótica: Sea $\hat{\theta}_n$ un estimador consistente con varianza asintótica $1/nV$. Sea θ_n^* cualquier otro estimador consistente con varianza asintótica $1/nV^*$. $\hat{\theta}_n$ es *asintóticamente eficiente* si $V^* \geq V$.

2.2 Función de verosimilitud

La *función de verosimilitud* de una variable aleatoria X con densidad $f(X;\theta)$ es:

$$L(\theta, X) = f(X, \theta)$$

Es importante notar que la función de verosimilitud considera como variables a X y al parámetro de interés. Análogamente, la función de verosimilitud para la muestra aleatoria $(X_1, X_2, \ldots, X_n)$ será:

$$\begin{aligned} L(\theta, \mathbf{X}) &= f(X_1, \ldots, X_n, \theta) \\ &= \Pi_{i=1}^n f(X_i, \theta) \\ &= \Pi_{i=1}^n L(\theta, X_i) \end{aligned}$$

dado que la muestra es aleatoria. El logaritmo de la función de verosimilitud es:

$$l(\theta, \mathbf{X}) = \sum_{i=1}^{n} l(\theta, X_i)$$

en donde $l(\theta, X_i)$ es el logaritmo de la función de verosimilitud de la variable aleatoria X_i.

El *score* de una variable aleatoria X es la derivada del logaritmo de la función de verosimilitud con respecto a θ:

$$s(\theta, X) = \frac{dl(\theta, X)}{d\theta}$$

El score de una muestra aleatoria, $s(\theta, \mathbf{X})$, es:

$$\begin{aligned} \frac{\partial l(\theta; \mathbf{X})}{\partial \theta} &= \frac{\partial [\sum_{i=1}^{n} l(\theta, X_i)]}{\partial \theta} \\ &= \sum_{i}^{n} \frac{\partial l(\theta, X_i)}{\partial \theta} \end{aligned}$$

La *información* de la variable aleatoria X, $I(\theta; X)$, se define como:

$$I(\theta) = E[s(\theta, X)^2]$$

en donde la esperanza es tomada con respecto a la distribución de X, es decir, considerando a X como una variable aleatoria y tomando a θ como un parámetro.

Algunos resultados básicos son los siguientes:

1. *Lema 1:* Sea X una variable aleatoria con densidad $f(x; \theta_0)$. Entonces $E(s(\theta_0, X)) = 0$, es decir, la esperanza del score igual a cero cuando es evaluada en el verdadero valor del parámetro.

   Prueba: El resultado a demostrar es:

   $$E(s(\theta_0, X)) = \int_{\Re} s(\theta_0; \mathbf{x}) f(x, \theta_0) dx = 0 \tag{2.1}$$

   Recordemos que:

   $$\int_{\Re} f(x, \theta_0) dx = 1$$

   por lo tanto, derivando con respecto a θ_0:

   $$\frac{d[\int_{\Re} f(x, \theta_0) dx)]}{d\theta_0} = 0$$

   si es posible intercambiar las operaciones de integración y diferenciación:

   $$\int_{\Re} \frac{df(x, \theta_0)}{d\theta_0} dx = 0$$

   De acuerdo a la definición del score, para una realización x de X:

   $$\begin{aligned} s(\theta_0, X) &= \frac{d \log f(x, \theta_0)}{d\theta_0} \\ &= \frac{df(x, \theta_0)/d\theta_0}{f(x, \theta_0)} \end{aligned}$$

por lo que

$$df(x,\theta_0)/d\theta = s(\theta_0;x)f(x,\theta_0)$$

reemplazando este resultado en (2.1) tenemos el resultado deseado.

Este resultado implica que para θ_0, la información $I(\theta_0)$ es igual la varianza del score:

$$V[s(\theta_0,X)] = E[s(\theta_0,X)^2] = I(\theta_0)$$

2. *Lema 2: (Information equality)* $I(\theta) = -E(H(\theta_0,X))$, en donde $H(\theta,X)$ es la primera derivada del score (o, de acuerdo a la definición, la segunda derivada de $l(\theta,X)$). Este resultado establece un nexo entre la varianza del score y la derivada segunda del logaritmo de la función de verosimilitud.

   Prueba: del resultado anterior:

$$\int_{\Re} s(\theta_0,x)f(x,\theta_0)dx = 0$$

   Derivando ambos miembros con respecto a θ_0, utilizando la regla del producto y omitiendo los argumentos de las funciones:

$$\begin{aligned} \int_{\Re}(sf' + s'f)dx &= 0 \\ \int_{\Re} sf'dx + \int_{\Re^n} s'fdx &= 0 \end{aligned}$$

   De la prueba del Lema 1, $f' = sf$, reemplazando f' arriba:

$$\begin{aligned} \int_{\Re} s^2 fd(x) + \int_{\Re} s'fdx &= 0 \\ E(s(\theta,X)^2) + \int_{\Re^n} H(\theta_0;x)dx &= 0 \\ I(\theta) + E(H(\theta;X)) &= 0 \end{aligned}$$

   lo que implica el resultado deseado.

2.3 Estimación máximo-verosímil

Dada una muestra aleatoria $X_i \sim f(x;\theta_0);\ i = 1,\ldots,n$, el *estimador máximo-verosímil (MV)* de θ_0, denotado como $\hat{\theta}$, es aquel para el cual:

$$L(\hat{\theta}, \mathbf{X}) \geq L(\theta, \mathbf{X}) \tag{2.2}$$

para todo θ. Intuitivamente, el estimador MV es el valor de los parámetros para los cuales la probabilidad de haber generado la muestra observada es máxima. Es fácil observar que dado que el logaritmo es una transformación monótona, el valor de $\hat{\theta}$ que satisface (2.2) tambien satisface:

$$l(\hat{\theta}, \mathbf{X}) \geq l(\theta, \mathbf{X}) \tag{2.3}$$

Dado que el converso de esta aseveración también es cierto, es equivalente utilizar (2.2) o (2.3) para resolver el problema de encontrar el estimador MV. Por simplicidad analítica trabajaremos con el logaritmo de la función de verosimilitud.

Si $l(\theta, \mathbf{X})$ es diferenciable y tiene un máximo local en un punto interior el estimador máximo verosímil $\hat{\theta}$ satisface:

$$s(\hat{\theta}, \mathbf{X}) = 0 \tag{2.4}$$

lo cual define una ecuación posiblemente no-lineal cuya solución caracteriza al estimador máximo verosímil. Si la función de verosimilitud es globalmente cóncava, encontrar el estimador MV equivale a encontrar la solución de esta ecuación.

2.3.1 Propiedades del estimador máximo-verosímil

Las condiciones bajo las cuales existe una solución al problema de máxima verosimilitud no garantizan automáticamente que dicha solución pueda expresarse como función de la muestra, es decir, que pueda despejarse analíticamente $\hat{\theta}$ de (2.4). En general, esto impide conocer las propiedades de muestra pequeña del estimador MV. Afortunadamente, bajo algunas condiciones generales (llamadas *condiciones de regularidad*), es posible verificar varias propiedades de muestra grande.

1. *Consistencia*: Si bien el desarrollo formal de una prueba general de consistencia excede los requisitos técnicos de estas notas, es útil dar alguna intuición acerca de como procede la misma. El desarrollo se basa en el tratamiento de Newey y McFadden (1994).

   El estimador máximo verosímil maximiza:

$$l(\theta) = n^{-1} \sum_{1=1}^{n} lnf(Z_i, \theta)$$

sujeto a que $\theta \in \Theta$. La única alteración con respecto al planteo original consiste en dividir la función objetivo por el número de observaciones, lo cual no altera la solución al problema. La idea central de la prueba consiste en mostrar que cuando n tiende a infinito, la función a maximizar, $l(\theta)$, converge a una función continua $l_0(\theta)$ que tiene un máximo único en θ_0, de modo que el maximizador de $l(\theta)$ (el estimador MV) converge al maximizador de la función límite, que es el verdadero parámetro θ_0.

Entonces, el primer paso consiste en obtener el límite de $l(\theta)$. Para un θ dado, $l(\theta)$ resulta ser un promedio de variables aleatorias independientes, de modo que será posible usar alguna versión adecuada de la ley de grandes números, lo cual implica que $l(\theta)$ converge en probabilidad a $l_0(\theta) = E[\ln f(Z, \theta)]$.

El segundo paso consiste en mostrar que esta función límite $l_0(\theta)$ tiene un máximo único en θ_0. Este resultado es consecuencia de la conocida *desigualdad de la información*: si $\theta \neq \theta_0$ implica que $f(Z|\theta) \neq f(Z|\theta_0)$, y $E[|lnf(Z|\theta)|] < \infty$, entonces $l_0(\theta)$ tiene un máximo único en θ_0. La prueba de este resultado es sencilla. Si $\theta \neq \theta_0$, podemos escribir:

$$l_0(\theta_0) - l_0(\theta) = E\left[-\ln \frac{f(Z;\theta)}{f(Z;\theta_0)}\right]$$

De acuerdo a la desigualdad (estricta) de Jensen

$$\begin{aligned} l_0(\theta_0) - l_0(\theta) &> -\ln E\left[\frac{f(Z;\theta)}{f(Z;\theta_0)}\right] \\ &> -\ln \int_{\Re} f(z;\theta)dz \\ &> 0 \end{aligned}$$

lo cual prueba el resultado deseado. El paso final consistiría en probar que la convergencia de $l(\theta)$ a $l(\theta_0)$ y el hecho de que esta última tenga un máximo único en θ_0 garantiza la convergencia de $\hat{\theta}$ a θ_0, lo cual requiere algunas nociones avanzadas de probabilidad. Formalmente, el resultado de consistencia puede expresarse de la siguiente forma:

Teorema (Consistencia del estimador MV) Si $Z_i, (i = 1, 2, ...)$ son i.i.d. con densidad $f(Z_i|\theta_0)$ y se verifican las siguientes condiciones:

(a) $\theta \neq \theta_0$ entonces $f(Z_i|\theta) \neq f(Z_i|\theta_0)$

(b) $\theta_0 \in \Theta$, con Θ compacto.

(c) $\ln f(Z_i|\theta)$ continua para todo $\theta_0 \in \Theta$ con probabilidad uno.

(d) $E[\sup_{\theta_0 \in \Theta} |\ln f(Z|\theta)|] < \infty$

entonces $\hat{\theta}$ converge en probabilidad a θ_0.

Prueba: ver Newey y McFadden (1994).

Las condiciones 1 y 4 implican la desigualdad de la información mostrada anteriormente. 2, 3, y 4 permiten utilizar una ley uniforme de grandes números que garantiza que $\hat{l}(\theta)$ converga uniformemente en probabilidad a una función continua $l_0(\theta)$. Luego, 2 y los resultados anteriores implican que $\hat{\theta}$ converge en probabilidad a θ_0[2].

2. *Normalidad asintótica*: Bajo condiciones generales, el estimador máximo verosímil tiene distribución asintótica normal. Más específicamente:

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0,1)$$

También daremos alguna intuición acerca de la prueba de este resultado. El estimador máximo verosímil satisface:

$$s(\hat{\theta}) = 0$$

Consideremos una expansión de Taylor de $s(\hat{\theta})$ en θ_0:

$$s(\hat{\theta}) \approx s(\theta_0) + (\hat{\theta} - \theta_0)s'(\theta_0) = 0$$

En este punto es importante notar que esta aproximación anterior tiende a ser exacta cuando $\hat{\theta}$ esta infinitamente cerca de θ_0, lo cual, de acuerdo al resultado anterior, ocurre cuando n tiende a infinito. Despejando en la expresión anterior:

$$\begin{aligned}(\hat{\theta} - \theta_0) &\approx \frac{-s(\theta_0)}{s'(\theta_0)} \\ n^{1/2}(\hat{\theta} - \theta_0) &\approx \frac{-n^{-1/2}s(\theta_0)}{n^{-1}s'(\theta_0)}\end{aligned}$$

Consideremos primero el numerador. Su esperanza es:

$$E[-n^{-1/2}s(\theta_0)] = -n^{-1/2}\sum_{i=1}^{n} E[s(\theta_0, X_i)] = 0$$

por el Lema 1. Su varianza es:

[2]El supuesto de compacidad puede resultar un tanto restrictivo en la práctica. El mismo puede ser reemplazado por supuestos acerca de la concavidad de la función de verosimilitud. Ver Newey y McFadden (1994)

$$V[-n^{-1/2}s(\theta_0)] = \frac{\sum_{i=1}^{n} V[s(\theta_0, X_i)]}{n} = I(\theta_0)$$

también por el mismo lema. Consideremos ahora el denominador:

$$\frac{1}{n}s'(\theta_0) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \log f(x_i;\theta_0)}{\partial\theta^2}$$

De acuerdo a la ley de grandes números, la expresión anterior converge en probabilidad a:

$$E\left[\frac{\partial^2 \log f(x_i;\theta_0)}{\partial\theta^2}\right] = -I(\theta_0)$$

,de acuerdo al Lema 2. Juntando estos dos resultados, para n grande:

$$n^{1/2}(\hat{\theta} - \theta_0) \approx \frac{n^{-1/2}s(\theta_0)}{I(\theta_0)}$$

Reescribamos este resultado como:

$$\sqrt{nI(\theta_0)}(\hat{\theta} - \theta_0) \approx \sqrt{n}\frac{\frac{\sum s(x_i;\theta_0)}{n}}{\sqrt{I(\theta_0)}}$$

El numerador es un promedio de variables aleatorias ($s(x_i;\theta_0)$) con media igual a cero y varianza igual a $I(\theta_0)$, de acuerdo al lema anterior. Por lo tanto, de acuerdo al Teorema Central del Límite, la expresión tiene distribución asintótica normal estándar.

Entonces, de acuerdo a la terminología introducida anteriormente, el estimador máximo verosímil $\hat{\theta}$ tiene distribución asintótica normal con media θ_0 y varianza asintótica igual a $1/nI(\theta_0)$

3. *Eficiencia asintótica*: $\hat{\theta}$ es asintóticamente eficiente, o sea, su varianza es menor que la de cualquier otro estimador consistente. La prueba de este resultado excede los propósitos de estas notas. Ver Davidson y MacKinnon (1993, Capitulo 8).

4. *Invarianza*: Si $\hat{\theta}$ es el estimador MV de θ_0, entonces el estimador MV de $h(\theta_0)$ es $h(\hat{\theta})$. La prueba de este resultado, si bien sencilla, también será omitida.

Capítulo 3

Modelos de Elección Binaria

En esta sección estudiaremos modelos estadísticos para casos en los que la variable de interés 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.

3.1 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 (1 si fue admitida, cero si no lo fue), calificación obtenida en los examenes GRE y TOEFL, 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:

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

Esta es una breve lista de preguntas que el método a analizar intenta responder:

1. Dado un determinado nivel de GRE, TOEFL y de las notas en la carrera de grado, ¿cuál es la probabilidad de ser admitido al programa?
2. ¿En cuánto mejoran las chances de admsión si mejora el GRE?
3. ¿Es realmente importante el TOEFL en el proceso de admisión?
4. ¿Cuán confiables son las respuestas a las preguntas anteriores?
5. Aún 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?

3.2 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 (1 si ocurre y 0 si no). Se dispone de una muestra aleatoria de n observaciones $Y_i, i = 1, \ldots, n$. Llamemos Ω_i al conjunto de información relevante asociado con el 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 Ω_i:

$$P_i = Pr(Y_i = 1|\Omega_i)$$

Es importante notar que dado que Y_i toma solo los valores cero y uno, ésta probabilidad condicional es también la esperanza de Y_i condicional en Ω_i:

$$E(Y_i|\Omega_i) = 1P_i + 0(1 - P_i) = P_i$$

Supongamos que Ω_i esta constituido por un vector fila de k variables explicativas X_i. Un primer intento de modelación podría consistir en postular una relación lineal entre Y_i y X_i, por ejemplo:

$$Y_i = X_i\beta + u_i \qquad \text{con} \qquad E[u_i|X_i] = 0$$

entonces:

$$E[Y_i|X_i] = P_i = X_i\beta$$

En este caso el vector de parámetros β podría ser consistentemente estimado utilizando el mínimos cuadrados ordinarios. El proceso de estimación consistiría

simplemente en regresar el vector de ceros y unos de las realizaciones de Y, en las variables explicativas.

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 lineal no es homoscedástico ya que la varianza condicional ($\text{Var}(u_i|X_i)$) es igual a $X_i\beta(1 - X_i\beta)$, la cual varía según las observaciones.[1]

3.3 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 esten restringidos al intervalo [0,1]. Una forma 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:

$$F(-\infty) = 0,\ F(\infty) = 1,\ f(x) = dF(x)/dx > 0$$

O sea, F(.) es una función diferenciable monótona creciente con dominio real y rango (0,1). Nuestro modelo no-lineal sería el siguiente:

$$y_i = F(X_i\beta) + u_i \tag{3.1}$$

con u_i definida como $u_i \equiv E[y_i|X_i] - F(X_i\beta)$. 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* $(X_i\beta)$ que luego es transformado por la función $F(.)$ de manera tal que los valores de la misma sean consistentes con los de una probabilidad.[2]

2. ¿Cómo elegir la función F(.)? Nótese que la función de distribución de cualquier variable aleatoria continua tiene las propiedades de F(.). En esta dirección es que buscaremos una forma funcional para F(.).

[1] Asi y todo, el modelo lineal de probabillidad ha sido recientemente revitalizado en Heckman y Snyder (1997).

[2] 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)

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*:

$$P_i = F(X_i\beta) = \Lambda(X_i\beta) = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}}$$

y corresponde al modelo *logit*. En la siguiente sección discutiremos una interpretación más intuitiva de estos modelos.[3]

3.4 La interpretación de variables latentes

Una forma alternativa de representar nuestro modelo de elección binaria es la siguiente:

$$y_i^* = X_i\beta - \epsilon_i,\ \epsilon_i \sim F(u) \tag{3.2}$$

$$y_i = \begin{cases} 1 & \text{si } y_i^* > 0 \\ 0 & \text{si } y_i^* \ \square\ 0 \end{cases} \tag{3.3}$$

Una interpretación de esta especificación es la siguiente. y_i^* es una variable aleatoria *latente* no observable por el econometrista, quien sólo observa la variable y_i, la cual adopta valores cero o uno de acuerdo a que y_i^* sea positiva o no.

En términos de nuestro ejemplo original de admisión a un programa, una historia consistente con esta interpretación podría ser la siguiente: el proceso de admisión se basa en la construcción de un índice lineal basado en el TOEFL, el GRE y las notas de la carrera ($X_i\beta$). También se sabe que existe un elemento discrecional que maneja el comité de admisión, que sube o baja este índice de acuerdo a factores que no son observables por el econometrista (ϵ_i). Estos dos factores se suman y un índice resulta (y_i^*, que tampoco es observado por el econonometrista!) en el cuál se basa el proceso de admisión: el alumno es admitido solo si el índice y_i^* es positivo.

Es sencillo verificar que esta interpretación es consistente con nuestra formulación original. Nuestro objeto de interés es P_i, la probabilidad de que el evento ocurra condicional en el conjunto de información caracterizado por $X_i\beta$:

[3] Es importante notar que todavía no hemos dicho nada acerca de que variable en el modelo tiene distribución normal o logística, solamente hemos tomado la forma funcional de las mismas por pura conveniencia analítica. (Pregunta: ¿cuál es la distribución de u_i condicional en X_i en el modelo probit? ¿y en el modelo logit?)

$$
\begin{aligned}
P_i &= Pr[y = 1] \\
&\quad Pr[y^* > 0] \\
&\quad Pr[X\beta - \epsilon > 0] \\
&\quad Pr[\epsilon < X\beta] \\
&\quad F(X\beta)
\end{aligned}
$$

que es nuestra formulación original

Comentarios:

1. De acuerdo a esta interpretación, especificar la función F() en (1) es equivalente a especificar la distribución del término aleatorio en (2).

2. ¿Qué distribución tiene u condicional en X? Este punto se presta a confusión y es válido aclararlo cuanto antes (lo pregunté antes!!!). De acuerdo a la definición de u y teniendo en cuenta que y puede adoptar solo valores cero o uno, una vez fijado el valor de X u puede adoptar solo dos valores, $1 - F(X_i\beta)$y $-F(X_i\beta)$ con probabilidades $F(X_i\beta)$ y $1 - F(X_i\beta)$ respectivamente, de modo que u sigue la distribución de Bernoulli. Por ejemplo, en el modelo logit la variable que tiene distribución logística es ϵ y no u, la cual tiene distribución de Bernoulli.

3.5 Como 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 la interpretación de los mismos. 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 en nuestra función de probabilidad condicional:

$$\frac{\partial P_i}{\partial X_k} = \beta_k \, f(X_i\beta) \tag{3.4}$$

De acuerdo a (4), el efecto marginal tiene dos componentes multiplicativos, el primero indica como un cambio en una variable explicativa afecta al *índice lineal* $(X\beta)$ y el segundo muestra como la variación en el índice se manifiesta en cambios en la probabilidad a través de cambios en la función $F()$.

Si $f()$ es una función de densidad simétrica y unimodal, alcanzará su máximo en la media. Supongamos por un instante que las variables explicativas X_i se

hallan expresadas como desviaciones con respecto a sus medias. En este caso $X\beta = 0$ corresponde al 'individuo promedio'. Dado que $f()$ tiene un máximo global en cero, (4) indica que un cambio marginal en una variable explicativa tiene efecto máximo para el individuo promedio cuando las variables se hallan expresadas como desviaciones respecto de la media.

En términos de nuestro ejemplo original, y habiendo encontrado que el GRE es una variable significativa en el modelo, de acuerdo a lo anterior estaríamos diciendo que mejorar 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.

La interpretación en términos del modelo de variables latentes es un poco mas compleja, y previamente debemos resolver un problema de identificación. Supongamos que el verdadero proceso generador de datos esta dado por (2)-(3) y que la varianza de ε es igual a σ^2. Si dividimos ambos miembros de (2) por σ, el modelo resultante es:

$$y_i^*/\sigma = X_i(\beta/\sigma) - \epsilon_i/\sigma,\ \epsilon_i \sim F(u) \tag{3.5}$$

$$y_i = \begin{cases} 1 & \text{si } y_i^* > 0 \\ 0 & \text{si } y_i^* \ \square\ 0 \end{cases} \tag{3.6}$$

Nótese que las observaciones de (3) no se alteran con esta transformación, por lo que el modelo dado por (4)-(5) es escencialmente indistinguible de (2)-(4). El punto central de esta discusión es que no es posible estimar en forma separada a β y σ sino en forma conjunta (β/σ). En definitiva, en la interpretación de variables latentes no es posible distinguir el efecto de β del de un cambio de escala, por lo que los veraderos parámetros pueden ser identificados solo en forma estandarizada.

3.6 Estimación e inferencia

Por sencillez analítica describiremos el método de *máxima verosimilitud (MV)* cuyas propiedades se hayan extensamente estudiadas. Mas adelante mencionaremos otras alternativas de estimación.[4]

De acuerdo al modelo especificado en (2)-(3), y_i sigue la distribución binomial con $y_i = 1$ con probabilidad $F(X_i\beta)$ e $y_i = 0$ con probabilidad $1 - F(X_i\beta)$. MV requiere perfecto conocimiento de estas probabilidades, por lo que es necesario especificar completamente la función F(), excepto por el vector de parámetros β. El logaritmo de la función de verosimilitud es:

[4] Una revisión rápida de MV es presentada en el apéndice y para un estudio mas detallado referirse a textos como Lindgren (1993) o para un tratamiento mas avanzado Lehmann (1983)

$$l(y,\beta) = \sum_{i=1}^{n} (y_i \log(F(X\beta)) + (1-y_i)\log(1-\log(F(X\beta))) \qquad (3.7)$$

Las condiciones de primer orden para la existencia de un máximo son:

$$\sum_{i=1}^{n} \frac{(y_i - F_i) f_i X_{ki}}{F_i(1-F_i)} = 0, \qquad k = 1,\ldots,K \qquad (3.8)$$

Se puede verificar fácilmente que la función de verosimilitud es estrictamente cóncava[5] para el caso de los modelos logit y probit, por lo cual la solución al problema de maximización de (7), si existen, definen un máximo único. También es posible verificar que para estos dos casos se satisfacen las condiciones de regularidad (Lehmann (1983) p.409) que garantizan que el estimador MV $\hat{\beta}$ sea consistente y asintóticamente normal, con matriz de covarianzas (asintótica) igual a la inversa de la matriz de información.[6]

El problema de maximización de (7) no siempre tiene una solución finita. Un caso frecuente en la práctica ocurre cuando existe un *clasificador perfecto*. Por simplicidad, supongamos que el modelo tiene una sola variable explicativa X además de una constante, o sea, $F(X\beta) = F(\alpha + \beta X)$. Supongamos que existe un valor x^* tal que $Y_i = 1$ cuando $X_i > x^*, i = 1,\ldots,n$ y $Y_i = 0$ en caso contrario. En este caso, el signo de $X_i - x^*$ predice perfectamente los valores de Y_i, por lo cual la variable X es llamada un clasificador perfecto de Y. Es sencillo observar que en este caso (5) no tiene solución. Primeramente, es fácil notar que (5) no puede tomar valores mayores que cero. Consideremos la secuencia de estimadores $\hat{\beta}_k = \gamma_k$ y $\hat{\alpha}_k = -x^*\hat{\beta}_k$, $k = 1,2,\ldots$, en donde γ_k es cualquier secuencia con $\lim_{k\to\infty} \gamma_k = \infty$. En este caso:

$$F(\hat{\alpha} + \hat{\beta}X_i) = F(\gamma_k(X_i - x^*))$$

Cuando $k \to \infty$, $F(\gamma_k(X_i - x^*))$ tiende a 1 si $X_i > x^*$ y a 0 en caso contrario, por lo que el logaritmo de la función de verosimilitud tiende a 0, el cual es una cota superior, de modo que esta función alcanza un máximo en el límite de esta secuencia de estimadores, ya que (7) es globalmente cóncava. Esto implica que no existen valores finitos de α y β que maximicen (7).[7]

3.7 Logits o Probits?

Una simple inspección visual permite observar que no existen mayores diferencias entre la distribución logística y la normal, excepto en las colas de la

[5] Ver Amemiya (1985, pp. xx)

[6] Una discusión detallada de los aspectos computacionales de la obtención del estimador MV y su matriz de covarianzas puede encontrarse en textos como Amemiya (1985) o Greene (1993)

[7] Ver Davidson y MacKinnon (1993, p. 521) para una generalización de este resultado

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 distibución 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 compararlos con los del modelo probit. Amemiya(1981), basándose en un método de prueba y error, propone como factor 1/1.6

3.8 Tests de especificación

Test de significatividad de coeficientes individuales.

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 MV del coeficiente de la k-esima variable explicativa y $\hat{se}(\hat{\beta}_k)$ es el la raíz cuadrada del estimador de la varianza de $\hat{\beta}_k$. Nótese que solo se conoce la distribución asintótica de este estadístico, la cual esta dada por la distribución normal estándar (y no la distribución 't' como en el caso del modelo lineal bajo el supuesto de normalidad).

Test de significatividad de un grupo de coeficientes.

Un test sencillo se puede basar en el principio de razón de verosimilitud. El test se basa en el siguiente estadístico:

$$LR = -2[\ln \hat{L}_r - \ln \hat{L}_{nr}] \sim \chi^2(r) \tag{3.9}$$

En donde $\hat{L}$ y $\hat{L}_r$ son, respectivamente, el valor de la verosimilitud en el el modelo restringido y sin restringir. Este estadístico tiene distribución asintótica χ^2 con r grados de libertad, en donde r es el número de restricciones.

3.9 Bondad del ajuste

En forma análoga al caso del modelo lineal estimado por mínimos cuadrados, vale preguntarse si existe una medida de 'bondad del ajuste' similar al estadístico R^2.

Una medida análoga frecuentemente utilizada en la práctica consiste en comparar el valor adoptado por la función de verosimilitud bajo el modelo estimado con el valor obtenido en un modelo en donde la única variable explicativa es una constante. Muchos programas reportan el estadístico:

$$LRI = 1 - \frac{\ln L}{lnL_0}$$

, a veces llamado 'pseudo R^2', en donde L es el valor máximo de la función de verosimilitud bajo el modelo original y L_0 es el valor correspondiente al modelo con solo una constante. Es fácil comprobar que este estadístico toma valores entre cero y uno. Valores igual a 1 occurren cuando $\ln L = 0$ el cual corresponde al caso de clasificación perfecta discutido anteriormente. Por el contrario, valores muy cercanos a cero provienen de casos en donde la función de verosimilitud adopta valores máximos similares bajo el modelo original y bajo el modelo con solo una constante, indicando que la ganancia (en términos de verosimilitud) por incorporar variables explicativas es baja. Valores cercanos a uno indican que la diferencia entre las verosimilitudes es significativa, de modo que el modelo que incorpora variables explicativas es superior al modelo con solo una constante. Es importante señalar que no es posible dar a este estadístico una interpretación en término de proporciones similar a la asociada con el estadístico R^2 en el modelo lineal estimado por mínimos cuadrados . En este caso el R^2 pude interpretarse como la proporción de la variabilidad total en la variable dependiente que es capturada por incorporar al modelo variables explicativas. Esta interpretación se basa en la descomposición de la suma de cuadrados de la variable a explicar en términos de la suma de cuadrados asociada con la regresión y con los residuos, la cual es una consecuencia directa del método de mínimos cuadrados, el cual proyecta ortogonalmente el vector de variables dependientes en el espacio generado por los vectores formados con las observaciones de las variables independientes. El estimador MV de los modelos discutidos anteriormente no se basa en dicha proyección, por lo que valores entre cero y uno del 'pseudo R^2' no tienen una interpretación natural.

Otra forma muy popular de evaluar la capacidad explicativa de los modelos binarios se basa en computar predicciones $\hat{Y}$ de la variable Y de la siguiente forma: $\hat{Y} = 1$ si $F(X_i\hat{\beta}) > F^*$ y 0 en caso contrario. Usualmente se toma $F^* = 0.5$, lo que equivale a predecir que el evento ocurre ($\hat{Y}_i = 1$) si la probabilidad predicha es mayor que 0.5. La medida de bondad del ajuste consiste en reportar la proporción de predicciones correctas sobre el total de observaciones. Esta medida es un tanto arbitraria y, en consecuencia, debe ser interpretada con sumo cuidado. Si bien la predicción de las probabilidades tiene cierto sentido estadístico, la predicción de Y a través de la probabilidad es arbitraria. Es fácil crear un modelo que predice correctamente por lo menos el 50% de las observaciones. Por ejemplo, si la proporción observada de ocurrencia del evento es mayor que 50%, digamos 70%, un modelo que predice $\hat{Y}_i = 1$ para todas

las observaciones acertará exactamente el 70% de los casos, aunque el modelo sea decididamente inútil para cualquier otro propósito que no sea predecir. También es importante remarcar que el estimador máximo verosímil no maximiza la cantidad de predicciones correctas sino, precisamente, la verosimilitud de las observaciones con respecto a la densidad postulada. Existen numerosas alternativas propuestas[8], pero ninguna de ellas parece dar una respuesta concluyente. El problema consiste en que el criterio de estimación (maximizar la verosimilitud) no necesariamente implica maximizar el ajuste[9].

En definitiva, si se trata de elegir un 'buen modelo', es importante definir de antemano que es lo que se persigue como objetivo. Si la forma funcional del modelo está correctamente especificada y si se cumplen las condiciones de regularidad, el método de máxima-verosimilitud produce estimadores (y no *estimaciones*) que son consistentes y asintóticamente eficientes, de lo cual no es posible hacer consideraciones acerca de las propiedades de dichos estimadores en muestras pequeñas.

3.10 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:

1. Existe una amplia literatura relacionada con diversos tests de especificación para modelos binarios. De manera similar al modelo lineal general, todos los supuestos bajo los cuales el estimador MV es óptimo deberían ser sometidos a pruebas estadísticas. Existen tests de autocorrelación, heteroscedasticidad, normalidad, variables omitidas, etc. Godfrey (1989) y Greene (1993) presentan algunos resultados básicos.

2. El problema de heteroscedasticidad tiene consecuencias un tanto mas graves que en el caso del modelo lineal general. Es posible demostrar (Yatchew y Griliches, 1984) que bajo la presencia de errores heteroscedásticos, el estimador MV es inconsistente.

3. El supuesto de correcta especificación de la distribución del término de error (en la interpretación de variables latentes) es también crucial. Las propiedades de optimalidad del método de máxima verosimilitud (consistencia y eficiencia asintótica) no se mantienen cuando la distribución del término de error ha sido incorrectamente especificada. Por ello es muy importante someter a prueba estadística este tipo de supuestos.

[8] Amemiya (1981) es una muy buena referencia sobre el tema

[9] A diferencia del caso del modelo lineal bajo normalidad, en donde el estimador máximo-verosímil es igual al estimador mínimo-cuadrático, el cual maximiza el R^2

4. Una alternativa es dejar de lado el método de máxima verosimilitud y obtener estimadores basados en supuestos menos restrictivos que los que requiere el método MV. En particular, los métodos *semiparamétricos* han recibido considerable atención. Dichos métodos se basan en algunas características de la función de distribución y no requieren perfecto conocimiento de la misma ni tampoco que el término de error sea homoscedástico. El método de *'maximum score'* (Manski (1986)) es un ejemplo de este tipo de técnicas. Estos estimadores presentan una serie de propiedades deseables (consistencia, robustez) aunque existe cierta perdida de eficiencia con respecto al caso ideal en el cual la distribución del término de error es conocida (en cuyo caso el método MV será eficiente). Históricamente los econometristas aplicados intentaron evitar el uso de este tipo de técnicas debido a la complejidad computacional de las mismas. Actualmente existen diversos programas que permiten obtener estimaciones semiparamétricas con la misma facilidad (desde el punto de vista del usuario) que las que se obtienen a partir de métodos completamente paramétricos.

3.11 Aplicaciones

En esta sección presentamos dos aplicaciones. La primera es nuestro modelo de admisión a un programa de doctorado. La segunda corresponde a un trabajo reciente de Donald y Sappington (1995), quienes estiman un modelo de adopción de políticas regulatorias.

3.11.1 Proceso de admisión

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 GRE, TOEFL y notas de la carrera (PROM), se puede expresar de la siguiente manera:

$$P_i = F(X_i\beta)$$

con:

$$X_i\beta = \beta_0 + \beta_1 \mathrm{GRE}_i + \beta_2 \mathrm{TOEFL}_i + \beta_3 \mathrm{PROM}_i$$

en donde hemos incluido una constante como varaiable 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.

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 intersante es que de acuerdo al

Tabla 1: Resultados para el modelo de admisión

	Logit			Probit			
	Coef	Err.Std	t	Coef	Err.Std	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

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 presenta los signos esperados: un aumento en el GRE o en PROM aumentan la probabilidad de admisión. El modelo probit presenta resultados similares. Cuando los coeficientes del modelo logit son reescalados presentan valores similares a los del modelo probit. La hipótesis de que solo el GRE es relevante en el proceso de admisión es evaluada con un test de razón de verosimilitud. El modelo es reestimado eliminando las variables PROM y TOEFL y el valor de la función de verosimilitud de este modelo restringido es comparada con el correspondiente valor del modelo original (sin restringir) de acuerdo al estadístico descripto en (6). El valor obtenido es $L1 = 38.6659$, el cual excede el valor crítico de la distribución $\chi^2(2)$ para un nivel de significatividad igual a 0.05. Concluimos que la hipótesis nula de que solo el GRE es relevante puede ser rechazada, de acuerdo a los resultados de nuestro test. En síntesis, de acuerdo con nuestro modelo empírico, el Toefl no es una variable relevante en el proceso de admisión. Un modelo que incorpora solo el gre como regresor no está correctamente especificado dado que la variable prom también resulta ser significativa.

En la siguente tabla calculamos las derivadas reemplazando los resultados obtenidos en (4). Estas derivadas son evaluadas en las medias de las variables independientes.

Tabla 2: Derivadas calculadas en las medias

	Probit	Logit	Medias
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 aumentara 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 significa-

tivamente distinto de cero, de acuerdo a los resultados obtenidos anteriormente.

3.11.2 Adopción de políticas regulatorias

Donald y Sappington (1995) estudian el proceso de adopción de políticas regulatorias en empresas de telecomunicaciones. Dichos autores presentan un esquema de análisis en el cual dos tipos de políticas regulatorias pueden ser adoptadas: regulación por *tasas de retorno* y regulación *basada en incentivos*. En el primer caso, la política regulatoria consiste en fijar un tasa máxima de retorno sobre las inversiones llevadas a cabo por una empresa de telecomunicaciones. Existen varias formas de implementar un esquema de regulación por incentivos. Un ejemplo podría ser una regulación por precios máximos. La principal diferencia entre estos esquemas es que en el caso de regulación por incentivos la firma se apropia de los beneficios asociados con actividades destinadas a reducir costos.

La pregunta que Donald y Sappington intentan analizar consiste en determinar porque diferentes regiones (estados, provincias, países) adoptan diferentes políticas regulatorias. Mas específicamente, que factores determinan la adopción de un determinado régimen regulatorio.

El modelo teórico estudiado por dichos autores sugiere que una política regulatoria basada en incentivos es mas probable que sea adoptada cuando: 1) Los beneficios asociados con la política de tasa de retorno sean muy altos o muy bajos, 2) La firma perciba una clara señal de que va a poder apropiarse de los beneficios generado por la política de incentivos, 3) Los beneficios asociados con la adopción de una política de incentivos sean significativamente altos, 4) Los costos de transacción de cambiar la política regulatoria sean relativamente bajos.

A partir de estas consideraciones, Donald y Sappington (1995) elaboran un modelo para la probabilidad de adoptar una política regulatoria basada en un esquema de incentivos. La estimación del modelo econométrico se basa en una base de datos de 47 estados en los EEUU observadas en el año 1991. Estos autores utilizan seis variables explicativas: AROR (tasa de retorno permitida antes de adoptar la política de incentivos), BYPASS (un indicador de competitividad basado en el uso de servicios alternativos no provistos por la firma en cuestión), URBGROW (tasa de crecimiento poblacional urbano), DEMOCRAT (proporción de gobiernos demócratas en los últimos años), LRATES (indicador de la tasa promedio cargada a los usurarios de los servicios de telecomunicaciones) y ELECT (indicador que refleja si los oficiales de las comisiones reguladoras son elegidos por la población o designados por el gobierno).

AROR es una proxy de cuan restrictiva fue la política de tasa de retorno en el pasado. BYPASS y URBGROW intentan medir la rentabilidad de la firma en cuestión, DEMOCRAT mide costos de transacción. Valores de DEMOCRAT cercanos a uno o cero indican persistencia de un determinado partido en el poder. Valores intermedios indican cambios relativamente frecuentes en la afiliación política del gobierno. Donald y Sappington interpretan que los costos de

cambios de régimen deberían ser inferiores en estados en donde se producen frecuentes cambios de partido. Algo similar ocurre con LRATES. Dichos autores conjeturan que una política de incentivos es mas probable que sea adoptada en estados en donde las tasas de servicio son elevadas. La última variable incluida es ELECT.

Tabla 3: Resultados del modelo de adopción de políticas regulatorias Tabla III en Donald and Sappington (1995, pp. 252)

	Coeficiente	Err. Std	p
INTERCEPTO	220.324	126.723	.082
AROR	-476.291	264.488	.072
$AROR^2$	247.072	136.370	.070
BYPASS	-1.17365	0.55023	.033
URBGROW	26.1704	13.0542	.045
DEMOCRAT	13.0746	5.76986	.023
$DEMOCRAT^2$	-9.59229	5.09325	.060
LRATES	4.28620	2.48637	.085
ELECT	-0.017586	1.18510	.988

Log de la función de verosimilitud = -17.19; R^2=.49; Porcentaje de predicciones correctas = 85

Las variables AROR y DEMOCRAT son incorporadas en forma cuadrática. Los resultados obtenidos tienden a confirmar las predicciones del modelo teórico. Los coeficientes asociados con las variables $AROR$ y $AROR^2$ son en forma conjunta significativamente distintos de cero, de acuerdo al test de razón de verosimilitud, lo que sugiere una no-linealidad en la relación entre las tasas de retorno permitidas y la probabilidad de adoptar un esquema de incentivos. La probabilidad de adoptar una política de incentivos es mayor cuando las tasas de retorno previamente permitidas son o muy altas o muy bajas. Los coeficientes asociados con BYPASS y URBGROW son también significativamente distintos de cero y presentan los signos esperados: la probabilidad de adoptar una política de incentivos es mayor en zonas de mas alto crecimento urbano y en donde la competitividad es mas baja. Los coeficientes asociados con la variable DEMOCRAT sugieren que una política de incentivos tiene mayor probabilidad de ser adoptada en estados en donde se producen frecuentes cambios de afiliación política del partido gobernante. El coeficiente asociado con LRATES sugiere que es mas probable que se adopte un esquema de incentivos para aquellas firmas que cargan tasas de servicios relativamente mas altas. Por último, el coeficiente negativo de ELECT no resulta ser significativamente distinto de cero. Para evaluar la especificación del modelo, los autores presentan un test de heteroscedasticidad, el cual no provee evidencia suficiente para rechazar la hipótesis nula de residuos homoscedásticos.

3.12 Bibliografía

Existe una extensa bibliografía sobre el tema. Greene (1993) presenta una revisión completa y actualizada de modelos binarios. Davidson y MacKinnon (1993) o Amemiya (1985) presentan discusiones un tanto mas técnicas y algunos detalles sobre métodos númericos de estimación e inferencia. Maddala (1983) es un texto entero dedicado al tema de variables dependientes limitadas. Lee (1996) presente una revisión de temas recientes, enfatizando métodos de simulación y semiparamétricos. El survey de McFadden (1984) contiene bibliografía detallada sobre el tópico. Anderson, et al. (1992) presentan un análisis completo del uso de modelos de elección en la teoría de mercados con productos diferenciados. Pudney(1989) presenta un análisis detallado del modelo de elección discreta. McCullagh and Nelder (1989) tratan el tema desde la perspectiva de los modelos lineales generalizados.

Capítulo 4

Modelos para Datos en Paneles

En esta sección analizaremos modelos econométricos utilizados cuando se dispone de *datos en paneles*: observaciones tomadas para varios individuos (o empresas, o paises, etc.) en varios períodos. Analizaremos el *modelo de componente de errores*, el cual es una extensión simple del modelo lineal general. Desde este punto de vista, los métodos de estimación e inferencia utilizados no difieren significativamente de los habituales (mínimos cuadrados y sus generalizaciones). Tampoco cambia la interpretación de los coeficientes básicos del modelo. La principal dificultad asociada a las técnicas de datos en panel radica en la interpretación de los distintas versiones del modelo de componente de errores.

4.1 Motivación

En una primera impresión uno estaría inclinado a creer que la disponibilidad de datos en paneles solo implica un incremento en el tamaño de la muestra. Pero en realidad, este aumento en la muestra proviene de agregar individuos en varios períodos. Alternativamente, cuando se dispone de este tipo de información, se podría pensar en estimar distintos modelos de series de tiempo, uno para cada país o persona, o distintos modelos de corte transversal, uno por período. Es válido preguntarse en que situaciones esta agregación de datos es posible sin modificar los métodos para series de tiempo o corte transversal utilizados cuando no se dispone de datos en paneles. Mas específicamente, es muy posible que diferentes individuos presenten diferentes características no observables que agreguen una complicación adicional al problema a analizar. Por otro lado, es válido intuir que la disponibilidad de datos en paneles permite analizar en forma parsimoniosa ciertos aspectos que no pueden ser explorados con modelos simples de series de tiempo o corte transversal.

A modo de motivación, consideremos el siguiente caso. Supongamos que estamos interesados en construir un modelo simple para la tasa de criminalidad (R) utilizando como posibles variables explicativas el gasto en seguridad (G), la tasa de desempleo de la economía (U) y un indicador de eficiencia judicial (E). Esta incompleta lista de variables intenta captar los costos e incentivos que encuentran los individuos de una sociedad para dedicarse a la actividad delictiva. En términos generales y como es habitual, planteamos la existencia de una relación lineal del siguiente tipo

$$R = \beta_0 + \beta_1 G + \beta_2 U + \beta_3 E + u$$

en donde u es un término de error. Supongamos que, eventualmente, dispondríamos de datos de series de tiempo y corte transversal para las provincias de un país. También supongamos que si bien la tasa de criminalidad R y el gasto en seguridad G varían por provincia y en el tiempo, la tasa de desempleo U solo lo hace en el tiempo pero no por provincias y el indicador de eficiencia legislativa solo varía por provincias. En definitiva, la tasa de desempleo es una característica estrictamente temporal del problema y la eficiencia legislativa es una característica provincial.

Si solo dispusiéramos de datos de series de tiempo para una provincia dada, la versión estimable de nuestro modelo basada en una muestra de T períodos ($t = 1, \ldots, T$) para una provincia sería:

$$R_t = \beta_0 + \beta_1 G_t + \beta_2 U_t + \beta_3 E_t + \delta_t \qquad t = 1, \ldots, T$$

la cual puede reescribirse como:

$$R_t = \beta_0^* + \beta_1 G_t + \beta_2 U_t + \delta_t$$

con $\beta_0^* = \beta_0 + \beta_3 E_t$ ya que E_t no varia en el tiempo.

Análogamente, si dispusiéramos de datos de corte transversal para un período dado, podríamos estimar la siguiente versión del modelo:

$$R_i = \beta_0^{**} + \beta_1 G_i + \beta_3 E_i + \mu_i \qquad t = 1, \ldots, T$$

con $\beta_0^{**} = \beta_0 + \beta_2 U$ ya que U es constanste para todas las provincias.

Comparemos la interpretación de los términos de error δ_t y μ_i, y de los interceptos (β_0^*, β^{**}) en los modelos anteriormente descriptos. El intercepto del modelo de series temporales (β_0) capta el efecto de factores relevantes en la determinación de la tasa de criminalidad que no varían en el tiempo, y el término aleatorio δ_t mide el efecto de factores relevantes que varían en el tiempo pero que no son observables por el econometrista. De esta manera, con los datos disponibles en el caso de serie de tiempo no sería posible identificar el efecto de la eficiencia legislativa en la tasa de criminalidad de la provincia estudiada ya que el mismo es indistinguible de cualquier otro factor relevante que no varía en

el tiempo, los cuales son absorbidos por el intercepto. En el caso del modelo de corte transversal el intercepto representa factores relevantes que determinan la tasa de criminalidad pero que no varían por provincias, y el término aleatorio μ_i representa factores relevantes que varían por provincia y que son no observables por el econometrista. En síntesis, los modelos de series de tiempo no pueden utilizar información acerca de variables que varían solamente según individuos, y los modelos de corte transversal no pueden utilizar información que varíe solamente en el tiempo.

Afortunadamente, la disponibilidad de datos en paneles permitiría identificar estos efectos y el objeto de esta nota consiste en estudiar modelos para estas situaciones. Si estuviéramos dispuestos a suponer que el efecto del gasto en seguridad sobre la tasa de criminalidad es homogéneo en el tiempo para todas las provincias, la disponibilidad de datos en paneles nos permitiría estimar un único modelo de la siguiente forma:

$$g_{it} = \beta_0 + \beta_1 y_{it} + \beta_2 p_{it} + \beta_3 s_{it} + u_{it}$$

De la discusión anterior surge que, potencialmente, el término de error en el caso de datos en paneles debería tener una estructura particular que refleje shocks que varían según individuos pero no en el tiempo y/o shocks temporales que no varíen según individuos. Esto da origen al modelo de componente de errores estudiado en la siguiente sección.

4.2 El modelo de componentes de errores

De lo discutido anteriormente, el modelo de datos en paneles podría expresarse de la siguiente manera:

$$y_{it} = X_{it}\beta + u_{it}$$

$$u_{it} = \mu_i + \delta_t + e_{it} \qquad i = 1, \ldots, N;\ t = 1, \ldots, T$$

en donde X_{it} es un vector fila con K variables explicativas siendo la primera de ellas una constante igual a 1. β es un vector de K parámetros que son nuestro objetivo de estimación.

A la luz de la discusión de la sección anterior, el término de error u_{it} presenta tres componentes. El primero (μ_i) respresenta factores no observables que difieren por individuos (o provincias) pero no en el tiempo. Por ejemplo, podría ser un indicador de la capacidad empresarial de una firma, la habilidad natural de un individuo, las regulaciones propias de cada país o, como en nuestro ejemplo inicial, un indicador de eficiencia legislativa provincial, los cuales varían por individuos pero permanecen inalterados durante el período analizado. El segundo componente (δ_t) representa shocks que varían en el tiempo pero no por

individuos. En el caso del modelo de combustible podría tratarse de shocks no observables que afectan a todos los países simultáneamente, por ejemplo, un índice global de estabilidad política, la tasa de desempleo de la economía, etc.

El tercer componente (e_{it}) representa la visión mas tradicional del término de error, representando shocks puramente aleatorios que afectan a un individuo en un determinado período específicamente.

Las distintas versiones del modelo de componente de errores surgen de diferentes formas de especificar el término de error u_{it}. Por razones pedagógicas, en lo que sigue supondremos que solo hay efectos individuales, o sea, $\delta_t = 0$. El tratamiento de el caso general en donde ambos efectos se hallan presentes es una simple extensión del caso de efectos individuales. Ver Baltagi (1995, Cap. 3) para una exposición detallada de este caso general.

Comencemos por la especificación mas sencilla. Cuando $\mu_i = 0$ y:

$$E(e_{it}|X_{it}) = 0$$

$$E(e_{it}e_{hs}) = \begin{cases} \sigma^2 & \text{si } i = h \text{ y } t = s \\ 0 & \text{si } i \neq h \text{ o } t \neq s \end{cases}$$

Bajo esta especificación, el término de error u_{it} satisface todos los supuestos del modelo lineal general bajo los cuales, según el teorema de Gauss-Markov, el estimador de *mínimos cuadrados ordinarios (MCC)* es el mejor estimador lineal e insesgado. El modelo a estimar sería el siguiente:

$$y_{it} = X_{it}\beta + e_{it} \qquad i = 1, \ldots, N;\ t = 1, \ldots, T$$

Bajo el supuesto adicional de normalidad de los e_{it} el estimador MC es también el estimador máximo-verosímil de los coeficientes lineales, y como consecuencia, el estimador resultante es asintóticamente eficiente. En definitiva, cuando no existe heterogeneidad no observable en el problema (asociada con individuos o con factores temporales), el método de mínimos cuadrados clásicos puede ser una alternativa válida.

El segundo tipo de especificación consiste en suponer que μ_i es una constante diferente para cada individuo, de modo que el modelo lineal es el mismo para todos los individuos excepto por el intercepto. En este caso los parámetros de interés (el vector β) pueden ser consistente y eficientemente estimados agregando $N - 1$ variables binarias, una por cada individuo menos una. El modelo sería:

$$y_{it} = X_{it}\beta + d_{1t}\mu_1 + \cdots + d_{(N-1)t}\mu_{N-1} + e_{it} \tag{4.1}$$

en donde para cada individuo j, la variable binaria d_{it} adopta el valor uno si $i = j$ y cero si $i \neq j$ [1].

[1] Si incorporaramos N variables binarias en vez de $N-1$, la suma de todas ellas sería igual a 1 para cada individuo en cada momento, por lo cual la primera variable explicativa de la matriz X sería perfectamente colinear con las variables binarias

En esta especificación la forma de resolver el problema de heterogeneidad no observable es a través de la agregación de $N-1$ variables adicionales las cuales capturan el efecto sobre el intercepto en cada individuo. El estimador MCC de β en (1) es conocido como el estimador de *efectos fijos*. Nuestro modelo tiene ahora $K+(N-1)$ parámetros.

Una tercer alternativa consiste en tratar a μ_i como una variable aleatoria no observable que varía sólo a traves de los individuos pero no en el tiempo:

$$y_{it} = X_{it}\beta + u_{it}$$

$$u_{it} = \mu_i + e_{it}$$

con:

$$E(\mu_i|X_{it}) = 0,\ E(e_{it}|X_{it}) = 0$$

$$E(u_{it}|X_{it}) = 0,\ V(\mu_i) = \sigma_\mu^2,\ V(e_{it}) = \sigma_e^2$$

Con esta informacion podemos construir la matriz de covarianzas del vector que contiene a los términos de error, cuyo elemento i,j será:

$$E(u_{it}u_{hj}) = \begin{cases} 0 & \text{si } i \neq h \\ \sigma_\mu^2 & \text{si } i = h \text{ y } t \neq j \\ \sigma_\mu^2 + \sigma_e^2 & \text{si } i = h \text{ y } t = j \end{cases}$$

O sea que bajo la especificación de *efectos aleatorios*, la matriz de covarianzas no es diagonal: existe una correlación entre los shocks para un mismo individuo originada por la presencia del efecto aleatorio específico para cada individuo. El estimador MCC sigue siendo insesgado pero no eficiente debido a la presencia de autocorrelación inducida por el efecto aleatorio, el cual es constante para cada individuo. De acuerdo a la teoría clásica, el estimador de mínimos cuadrados generalizados (MCG) producirá un estimador de varianza mínima dentro de los estimadores lineales insesgados.

4.3 Estimación e inferencia en el modelo de componentes de errores

En términos de estimación resta discutir como obtener estimadores para los modelos de efectos fijos y aleatorios.

4.3.1 Estimación

a) Estimador de efectos fijos

El modelo de efectos fijos (1) puede expresarse en forma matricial de la siguiente forma:

$$y = X\beta + Z\mu + e$$

en donde y es un vector columna con NT observaciones ordenadas primero por individuos y luego en el tiempo (por ejemplo, las observaciones del primer individuo ocupan las primeras T posiciones del vector, las del segundo las observaciones $T+1$ a $T+N+1$, etc.). X es una matrix (NT x K) con las variables explicativas ordenadas de la misma manera, Z es una matriz NT x $(N-1)$ en donde cada columna es la variable binaria definida anteriormente. Finalmente, e es un vector columna NT.

Es importante remarcar que el vector de parámetros de interés es β. Si e satisface los supuestos del modelo lineal clásico, el método de mínimos cuadrados producirá los mejores estimadores lineales insesgados de β. En definitiva, se trata de aplicar MCC al modelo original incorporando $N-1$ variables binarias.

Un problema práctico es que MCC implica invertir una matriz con rango $K+(N-1)$, lo que puede crear ciertas dificultades computacionales si el número de individuos en la muestra es muy grande (que es la situación típica en modelos microeconómicos).

Es sencillo mostrar que el estimador mínimo cuadrático de β en (1) es idéntico al estimador MCC de β en el siguiente modelo:

$$(y_{it} - \bar{y_i}) = (X_{it} - \bar{X_i})\beta + e_{it} - \bar{e} \tag{4.2}$$

en donde $\bar{h_i} = \sum_{t=1}^{T} h_{it}/T$, $h = y, X_k, e$. El modelo de efectos fijos se puede expresar como :

$$Y_{it} = X_{it}\beta + \mu_i + e_{it} \tag{4.3}$$

para todo $i, \ldots, N, t, \ldots, T$. Sumando a ambos miembros a través de las observaciones de corte transversal y dividiendo por N obtenemos:

$$\bar{y_i} = \bar{X_i}\beta + \mu_i + \bar{e_i} \tag{4.4}$$

para todo i, t. Esto puede prestarse a confusión. La operación anterior es realizada para cada observación del modelo y le asigna a cada observación original el promedio por individuo. Como resultado de tomar promedios por corte transversal, para cada observación correspondiente al mismo individuo el promedio *dentro* de cada corte es constante y, trivialmente, el efecto fijo también lo es.

Restando (4) y (3) obtenemos el resultado deseado. En síntesis, los parámetros del modelo original son idénticos a los parámetros del modelo transformado y, por lo tanto, el estimador de efectos fijos se puede computar como el estimador de mínimos cuadrados ordinarios en un modelo en el cual las variables han sido transformadas tomándolas como desviaciones con respecto a la media de cada invididuo.

En forma similar, la matriz de covarianzas del estimador $\hat{\beta}$ obtenido en (2) puede ser consistente y eficientemente estimada utilizando el estimador mínimo cuadrático del modelo con las variables transformadas:

$$\hat{V}(\hat{\beta}) = s^2 (X^{*'} X^*)^{-1}$$

en donde s^2 es un estimador de σ_e^2 y X^* es la matriz X habiendo sustraído las medias por individuo. s^2 puede obtenerse como:

$$s^2 = s_F^2 \frac{NT - K}{N(T-1) - K}$$

en donde s_F^2 es el estimador de MCC de la varianza del error en el modelo transformado (2). La correción por grados de libertad es necesaria porque el modelo transformado estima solo K parámetros mientras que el modelo original tiene $K + (N-1)$ parámetros.

Comentarios

1. Nótese la forma en la que opera el estimador de efectos fijos. Al reexpresar todas las observaciones como desviaciones de la media de cada individuo el efecto fijo por individuo desaparece dado que es constante en el tiempo para cada individuo.

2. Como consecuencia de este proceso, cualquier variable *observable* que no varía en el tiempo es también anulada por la transformación, de modo que los parámetros asociados con este tipo de variables no pueden ser identificados (desde el punto de vista de efectos fijos, el efecto de una variable que no varía en el tiempo es indistinguible del efecto fijo). En referencia al modelo de seguridad, el estimador de efectos fijos por provincias no permitiría identificar el efecto del índice de eficiencia legislativa ya que el mismo es indistinguible del efecto fijo.

3. El estimador de efectos fijos implica una enorme pérdida de grados de libertad ya que se estiman $K + (N-1)$ parámetros.

4. Es importante remarcar que en la versión original (1), sólo el vector de parámetros β puede ser consistentemente estimado. El vector μ no puede ser consistentemente estimado si $N \to \infty$ dado que el número de parámetros a estimar aumenta con el tamaño de la muestra (ver Matyas (1996) para más detalles).

b) Estimador de efectos aleatorios

Llamemos - a la matriz de covarianzas del término de error u, el cual está ordenado primero por individuos y después por períodos. El estimador de mínimos cuadrados generalizados es:

$$\hat{\beta}_{MCG} = (X'\text{-}\ ^{-1}X)^{-1}X'\text{-}\ ^{-1}y$$

Expresado de esta manera, el proceso de estimación implica invertir la matriz - , que tiene dimensiones NT x NT lo cual puede crear serios problemas computacionales. Se puede demostrar (ver apéndice) que el estimador MCG ($\hat{\beta}_{MCG}$) es igual al estimador MCC de β en la siguiente regresión:

$$(y_{it} - \alpha\bar{y}_i) = (X_{it} - \alpha\bar{X}_i)\beta + \text{residuo}$$

con:

$$\alpha = 1 - \frac{\sigma_e^2}{(T\sigma_\mu^2 + \sigma_e^2)^{1/2}}$$

Comentarios:

1. Para la implementación de este método necesitamos estimaciones de σ_μ^2 y σ_e^2. Existen varios métodos para obtener dichos estimadores. Consideremos los estimadores MCC de β en los siguientes modelos de regresión:

$$y_{it} - \bar{y}_i = (X_{it} - \bar{X}_i)\beta + \text{residuo}$$

$$\bar{y}_i = \bar{X}_i\beta + \text{residuo}$$

El primero es el estimador de efectos fijos anteriormente analizado. El segundo se conoce en la literatura como estimador *between* y se basa en las medias para cada individuo. El primero utiliza NT observaciones y el segundo sólo N observaciones. Llamemos e_W y e_B a los vectores de residuos de los dos modelos (el primero tiene NT componentes y el segundo N). Denominemos s_W^2 y s_B^2 a los estimadores MCC de la varianza del error obtenidos en cada modelo:

$$s_W^2 = \frac{e_W' e_W}{N(T-1) - (K-1)}$$

$$s_B^2 = \frac{e_B' e_B}{N - K}$$

Se puede demostrar facilmente (Matyas (1996) pp.61) que s_W^2 estima consistentemente a σ_e^2 y que s_B^2 es un estimador consistente de $\sigma_\mu^2 + 1/T\sigma^2$, por lo que los estimadores de las varianzas ($\hat{\sigma}_e^2$ y $\hat{\sigma}_\mu^2$) pueden obtenerse resolviendo el siguiente sistema de ecuaciones:

$$\begin{aligned} \hat{\sigma}_e^2 &= s_W^2 \\ \hat{\sigma}_\mu^2 + 1/T\hat{\sigma}_e^2 &= s_B^2 \end{aligned}$$

Estos estimadores son insesgados y eficientes dentro de la familia de estimadores cuadráticos.

2. Es interesante ver como los modelos MCC y de efectos fijos aparecen como situaciones especiales del caso de efectos aleatorios. Cuando $\alpha = 0$ el estimador MGC es idéntico al de MCC. Esto ocurre cuando $\sigma_\mu^2 = 0$, o sea, cuando el término de error no contiene un componente individual. Cuando $\alpha = 1$ obtenemos el estimador de efectos fijos.

4.3.2 Tests de especificación

a) Test de efectos fijos

Un test sencillo para evaluar la hipótesis $H_0 : \mu_1 = \cdots = \mu_{N-1} = 0$ puede basarse en un test de significatividad conjunta de las variables binarias correspondientes en el modelo (1). Bajo el supuesto de normalidad de los residuos, el familiar estadístico:

$$F = \frac{(SCRR - SCRN)/(N + T - 2)}{SCRN/(N-1)(T-1) - K}$$

se distribuye como $F(N + T - 2, (N-1)(T-1) - K)$ bajo la hipótesis nula. $SCRR$ y $SCRN$ representan las sumas de los cuadrados de los residuos en los modelos restringido y sin restringir, respectivamente. En síntesis, un test simple de la hipótesis de ausencia de efectos fijos puede basarse en correr una regresión sin variables binarias y otra incluyéndolas, y comparando sus sumas de cuadrados de los residuos utilizando un test F. Valores demasiado grandes de este estadístico permitirían rechazar la hipótesis nula de que no hay efectos fijos individuales.

b) Test de efectos aleatorios

Aquí estamos interesados en evaluar la hipótesis $H_0 : \sigma_\mu^2 = 0$ en el modelo de efectos aleatorios. Breusch y Pagan (1980) proponen un test de Multiplicador de Lagrange basado en los residuos del estimador MCC. Dichos autores muestran que el estadístico:

$$LM = \frac{NT}{2(T-1)} \left\{ \frac{\sum_{i=1}^{N} (\sum_{t=1}^{T} e_{it})^2}{\sum_{i=1}^{N} \sum_{t=1}^{T} e_{it}^2} - 1 \right\}^2$$

tiene distribución asintótica $\chi^2(1)$ bajo la hipótesis nula, en donde e son los residuos del estimador MCC. Este estadistico puede expresarse en términos matriciales como:

$$LM = \frac{NT}{2(T-1)} \left\{ \frac{e'(J_N \quad I_T)e}{e'e} - 1 \right\}^2$$

en donde J_N es una matriz N xN con todos sus elementos iguales a 1, I_T es la matriz identidad con dimensión T, e es el vector de residuos del método MCC y es el producto de Kronecker. Valores significativamente distintos de cero de este estadístico sugerirían rechazar la hipótesis nula de ausencia de efectos aleatorios.

4.4 Efectos Fijos o Aleatorios?

Discernir entre los modelos de efectos fijos y aleatorios es, posiblemente, el problema mas complicado en la implementación de un modelo de datos en paneles. Comencemos remarcando que dichos modelos le asignan al término de error un rol completamente distinto y en consecuencia producen estimaciones basadas en procedimientos radicalmente diferentes. En el primer caso el efecto fijo es incorporado a la esperanza condicional de la variable explicada y en consecuencia es estimado conjuntamente con los otros coeficientes correspondientes al resto de las variables explicativas. Como consecuencia, el efecto fijo es indistinguible de cualquier otra variable que no varía por individuos. Adicionalmente, el modelo de efectos fijos incorpora a los efectos individuales como variables explicativas lo que implica una considerable pérdida de grados de libertad. Por el contrario, la especificación de efectos aleatorios trata al efecto fijo como una variable aleatoria omitida, la cual pasa a formar parte del término aleatorio lo cual solo altera la estructura de la matriz de covarianzas.

Mundlak (1978) sostiene que en realidad la distinción entre efectos fijos y aleatorios se refiere exclusivamente al tratamiento dado a los mismos. Todos los efectos individuales son aleatorios y la distinción entre fijos y aleatorios en realidad se refiere a si el análisis es *condicional* en los mismos (efectos fijos) o no condicionales (efectos aleatorios). Hsiao (1986, pp.41-48) presenta una interesante discusión de estos aspectos.

Desde un punto de vista práctico las estimaciones resultantes de ambos modelos pueden diferir significativamente, sobre todo teniendo en cuenta que el estimador de efectos fijos implica incorporar $N-1$ variables adicionales. La decisión entre efectos fijos y aleatorios puede basarse estrictamente en cuestiones de conveniencia práctica y en esa dirección discutiremos algunos resultados que pueden ayudar a decidir entre un modelo u otro. Por empezar y como remarcaramos anteriormente, el efecto fijo es indistinguible de cualquier variable que no varía por individuo, por lo que si el objetivo consiste en identificar el efecto de dichas

variables, el modelo de efectos aleatorios permitirá estimar en forma única los coeficientes asociados con este tipo de variables.

Segundo, a lo largo de nuestra discusión del modelo de efectos aleatorios hemos supuesto que los componentes del término de error no están correlacionados con las variables explicativas. De no ser este el caso, las propiedades de los estimadores analizados se altera significativamente. Si el término de error es tratado como un componente aleatorio cuya correlación con las variables explicativas no es nula, MCC y MCG producirán estimadores *inconsistentes*. Es importante observar que aún cuando el efecto aleatorio esté correlacionado con las variables explicativas, el estimador de efectos fijos preserva la propiedad de consistencia. Esto se debe a que la tranformación operada para obtener dicho estimador elimina el efecto específico por individuos.

Teniendo en cuenta que el estimador de efectos fijos es siempre consistente y que el estimador de efectos aleatorios solo lo es cuando las variables explicativas no están correlacionadas con el término aleatorio, una prueba de exogeneidad de las variables explicativas con respecto al efecto aleatorio puede basarse en el *test de Hausman* el cual se basa en una comparación de los estimadores de efectos fijos y aleatorios. Bajo la hipótesis nula de exogeneidad de las variables explicativas con respecto al efecto aleatorio, las estimaciones de β obtenidas en el modelo de efectos fijos deberían ser similares a las del método de efectos aleatorios ya que ambos producen estimadores consistentes. Por el contrario, bajo la hipótesis alternativa de que las variables explicativas están correlacionadas con el efecto aleatorio, solo el estimador de efectos fijos preserva la propiedad de consistencia por lo cual sería esperable, en muestras grandes, encontar diferencias significativas en las estimaciones obtenidas por los distintos métodos.

El estadístico correspondiente es:

$$H = (\hat{\beta}_A - \hat{\beta}_F)'(\text{-}_F - \text{-}_A)^{-1}(\hat{\beta}_A - \hat{\beta}_F)$$

en donde $\hat{\beta}_A$ y $\hat{\beta}_F$ corresponden al estimador de β del modelo de efectos aleatorios y fijos respectivamente, y -_A y -_F son las matrices de varianzas estimadas de los estimadores. Este estadístico tiene distribución asintótica χ^2 con K grados de libertad bajo la hipótesis nula de que el estimador de efectos aleatorios es consistente. Valores significativamente altos de este test sugieren que el estimador de efecto aleatorios es inconsistente, lo cual conduciría a rechazar la hipótesis nula de exogeneidad de los regresores. De ser este el caso, se podría proceder con el estimador de efectos fijos el cual, como mencionáramos anteriormente, es siempre consistente. Una alternativa mas eficiente consiste en utilizar un análogo del método de mínimos cuadrados en dos etapas, tal como sugieren Hausman y Taylor (1981).

Es importante remarcar que esta versión del test de Hausman no es un test de efectos fijos versus aleatorios sino un test de la hipótesis de exogeneidad de las variables explicativas con respecto al efecto aleatorio no observable. En todo caso, puede ser interpretado como un test de validez del estimador de efectos

aleatorios.

4.5 Extensiones

El análisis de modelos para datos en paneles es uno de los tópicos con mayor difusión en econometría. En estas notas hemos cubierto los principios básicos del modelo de componente de errores. Como mencionáramos en la introducción de esta sección, la extensión básica consiste en incorporar efectos temporales, lo cual se hace en una manera análoga al caso de efectos individuales. Textos como Baltagi (1995) o Matyas y Sevestre (1996) tratan el tema en detalle. Consideremos brevementa algunas extensiones básicas.

1. En forma similar al modelo lineal clásico, es importante someter a pruebas estadísticas los supuestos implícitos en el modelo estimado. En adición a los tests de efectos aleatorios y fijos y de exogeneidad mencionados anteriormente, es importante evaluar las hipótesis de heteroscedasticidad y autocorrelación de los residuos. Baltagi (1995, capítulo 5) contiene abundantes detalles sobre la implementación e interpretación de dichos tests en el modelo de componente de errores. Bera, et al. (1996) presentan una familia de tests de multiplicador de Lagrange para este modelo.

2. En estas notas hemos hecho el supuesto implícito de que los paneles son *balanceados*, eso es, que el número de observaciones temporales es el mismo para todos los individuos. El caso de paneles *no balanceados*, en donde el número de observaciones temporales varía según los individuos, pude ser fácilmente tratado como una simple extensión del caso original, aunque con cierta complicación en lo que se refiere a la estimación de las varianzas de los componentes de errores. Ver Baltagi (1995, cap. 9) para más detalles. Programas como Limdep v.7 proveen estimadores para este caso.

3. Otros temas clásicos tienen su contraparte para datos en paneles, a saber: modelos dinámicos (Arrellano y Bond, 1991), modelos para variables dependientes binarias (Chamberlain, 1980), modelos de duración (Petersen, 1988), ecuaciones simultáneas (Baltagi, 1981).

4.6 Aplicaciones

Retomemos el ejemplo de Baltagi y Griffin (1983) mencionado en la Introducción. La variable dependiente es un índice de consumo de nafta por auto. Como variables independientes utilizan el ingreso per cápita (LINCOMEP), un indicador del precio relativo de la nafta (LRPMG) y el stock de autos per capita (LCARPCAP). Todas las variables se encuentran medidas en logaritmos. En la Tabla 1 presentamos resultados de los tres modelos estudiados: mínimos cuadrados clásicos utilizando el panel original (MCC), el estimador de efectos fijos y

el de efectos aleatorios. Se presentan los coeficientes estimados y sus errores estándar.

Tabla 1: Resultados del modelo de demanda de nafta
Baltagi y Griffin (1983)

	MCC		Ef. Fijo		Ef. Aleat.	
	Coef.	Err. Std	Coef.	Err. Std	Coef.	Err. Std
LINCOMEP	0,8896	0,0358	0,6626	0,0734	0,5550	0,0572
LRPMG	-0,8918	0,3032	-0,3217	0,0441	-0,4204	0,0387
LCARPCAP	-0,7634	0,1861	-0,6405	0,0297	-0,6068	0,0247

F (ef. fijos) = 83.960, LM (ef. aleatorios) = 1465.55

Los distintos métodos de estimación presentan resultados diferentes, aunque todos los signos coinciden. Las elasticidades ingreso son postivas y las elasticidades precio negativas. El coeficiente negativo asociado con el stock de autos per-cápita sugiere que el uso de cada automóvil se reduce con el aumento del stock, con la consiguiente reducción de la demanda por nafta.

Los resultados obtenidos para el modelo de efectos fijos y aleatorios son similares y ambos difieren de los resultados del método MCC. El test de significatividad conjunta de todas las variables binarias en el modelo de efectos fijos resulta ser significativamente distinto de cero, de acuerdo al estadístico F. Por otra parte, el test de Breusch-Pagan sugiere que la hipótesis nula de ausencia de efectos aleatorios debe ser rechazada, por lo cual las estimaciones de los errores estándar del método MCC son sesgadas. Estos resultados sugieren la presencia de efectos individuales (por países) que afectan a la demanda de combustible, los cuales no son considerados por el método de MCC. El coeficiente α es 0.8923 lo cual es consistente con el hecho de que el estimador de efectos aleatorios se asemeje al de efectos fijos mas que al de MCC.

4.7 Bibliografía

Como consecuencia de la considerable atención que ha recibido el análisis de datos en paneles en los últimos años, existe una abundante literatura sobre este tópico. Sevestre y Matyas (1996) es la referencia mas actualizada y comprensiva. Es un extenso manual que recopila los principales resultados teóricos y algunas aplicaciones en diversas áreas de econometría aplicada. Baltagi (1995) es otro texto reciente, con abundantes detalles y ejemplos. La monografía de Hsiao (1986) fue durante muchos años la referencia clásica en el tema. El survey de Chamberlain (1984) también resulta muy útil. Para más detalles acerca del estado actual de las investigaciones en el tema, consultar la edición especial del Journal of Econometrics (1995).

Capítulo 5

Datos Censurados y Truncados

5.1 Motivación

Consideremos el problema estudiado por Mroz (1987). Uno de los puntos principales de su trabajo es la estimación de una función de oferta de trabajo para el sector femenino en EEUU. Según su especificación básica, la variable 'cantidad de horas trabajadas' es modelada como una función lineal de su ingreso, su experiencia laboral y el ingreso del marido. La estimación se basa en una base de datos de 753 mujeres para el año 1975, de las cuales solo 428 declararon haber trabajado durante el período de la muestra. El problema consiste en que la variable 'salario' sólo se observa positivamente para aquellas mujeres que trabajaron, el resto presenta cero como salario. Esto es consistente con algunos resultados de teoría económica. Un resultado habitual de los modelos de búsqueda de trabajo es que, en equilibrio, los agentes entran al mercado laboral sólo si el salario de mercado excede un determinado salario de reserva. Desde este punto de vista, en la base de datos analizada por Mroz sólo se observan horas de trabajo positivas para aquellas personas con salario de reserva inferior al de mercado, las cuales decidieron trabajar.

Intuitivamente, habría dos alternativas posibles. Primeramente, se podría pensar en estimar los parámetros del modelo utilizando sólo información para aquellas mujeres que han trabajado durante el período de la muestra, es decir, descartando la información correspondiente a las mujeres para las cuales se observa salario igual a cero. Otra posibilidad es tomar toda la muestra considerando que las mujeres que no trabajan tienen salario cero. El objetivo mas importante de esta sección es mostrar porque estas estrategias pueden conducir a resultados erróneos, y como puede utilizarse toda la información disponible para obtener estimadores consistentes.

En el caso de Mroz, los datos son *censurados*, en el sentido de que si bien disponemos de observaciones para todos los individuos de la muestra, el valor verdadero de la variable independiente es observado sólo para ciertos individuos (los que trabajan, en este caso) mientras que para el resto observamos un valor 'censurado' (cero, en nuestro ejemplo). Si los datos fueran *truncados* la situación sería todavía mas restrictiva: sólo observaríamos información (tanto de la variable dependiente como de las independientes) para ciertos individuos de la muestra original. En el caso de Mroz, la muestra de datos censurados incluye 753 individuos mientras que la muestra truncada no incluiría ningún tipo de información sobre las mujeres que no trabajaron.

5.2 Datos truncados vs. censurados

Comencemos haciendo explícita la diferencia entre datos truncados y censurados. Una muestra se considera *truncada* cuando ciertas observaciones son sistemáticamente *excluidas* de la muestra. En el caso de datos *censurados* ninguna observación es excluida, pero cierta información en la misma es sistemáticamente *alterada*.

Mas formalmente, un modelo sencillo de datos truncados podría ser el siguiente. Supongamos que existe una muestra de N_1 observaciones de la variable latente y_i^* (no observable) y que x_i es un vector fila de K variables explicativas que se encuentran relacionadas a través del siguiente modelo lineal con término aleatorio u_i:

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

En el modelo de datos truncados observamos pares (y_i, x_i) sólo para aquellas observaciones que satisfacen cierto criterio, por ejemplo, observamos pares $(y_i = y_i^*, x_i)$ sólo para los casos en los cuales $y_i^* > 0$. En este caso los datos son el resultado de una submuestra de la variable original: sólo disponemos de $N_2 < N_1$ observaciones para individuos para los cuales $y_i^* > 0$.

Una especificación similar para el modelo de datos censurados podría ser la siguiente:

$$y_i^* = x_i\beta + u_i \tag{5.1}$$

$$y_i = \begin{cases} y_i^* & \text{si } y^* > 0 \\ 0 & \text{si } y^* \square\ 0 \end{cases} \tag{5.2}$$

Nótese la diferencia con el caso anterior. En este caso disponemos de observaciones para todos los individuos en la muestra original ($N_2 = N_1$), pero ciertas observaciones han sido alteradas de acuerdo al mecanismo descripto en (1)-(2): la variable dependiente observada toma el valor cero cuando $y^* \square\ 0$. Un ejemplo clásico de datos truncados podría ser una base de datos tributaria en donde se dispone de información solo para individuos que superan un mínimo

ingreso no imponible. El mismo ejemplo con datos censurados podría consistir en una base de datos en donde los individuos con ingresos menores que cierto umbral reportan ingreso igual a cero.

5.3 Datos truncados

5.3.1 Distribuciones truncadas

Antes de proceder al análisis de modelos con datos truncados es conveniente estudiar algunos resultados relacionados con distribuciones truncadas.

Sea X una variable aleatoria continua con función de densidad $f(x)$. Supongamos que dicha variable ha sido truncada en el punto a, mas específicamente, nuestras observaciones son X tales que $X > a$. La función de densidad de la variable truncada es la función de densidad condicional:

$$f(x|x > a) = \frac{f(x)}{Pr[x > a]}$$

Por ejemplo, supongamos que X tiene distribución normal con media μ y varianza σ^2. La función de densidad de la variable aleatoria X truncada en a será:

$$\begin{aligned} f(x|x > a) &= \frac{f(x)}{1 - \Phi(a)} && (5.3) \\ &= \frac{(2\pi\sigma^2)^{-1/2} \exp(-(x-\mu)^2/(2\sigma^2))}{1 - \Phi(\alpha)} && (5.4) \\ &= \frac{(1/\sigma)\phi((x-\mu)/\sigma)}{1 - \Phi(\alpha)} && (5.5) \end{aligned}$$

en donde $\phi()$ y $\Phi()$ son, respectivamente, las funciones de densidad y de distribución de una variable aleatoria normal estándar. Además se puede verificar (Johnson et al (1994)) que:

$$E(x|x > a) = \mu + \sigma\lambda(\alpha) \tag{5.6}$$

en donde:

$$\begin{aligned} \alpha &= (a - \mu)/\sigma \\ \lambda(\alpha) &= \frac{\phi(\alpha)}{\Phi(\alpha)} \end{aligned}$$

si x es truncada de modo que $x > a$. Si x es truncada de modo que $x < a$ el signo de λ se invierte. λ es una variable conocida como la inversa de la razón de Mills. Para el caso de la distribución normal, el resultado (6) muestra dos

características importantes del problema de datos censurados. Primero, si el mecanismo de truncamiento es tal que $x > a$, la esperanza (condicional) de la variable truncada se corre a la derecha. La magnitud de este desplazamiento depende de la combinación de dos factores: de cuan concentrada esté la masa de probabilidad alrededor de la media (σ) y de cuan lejos esté el punto de truncamiento con respecto a la media. Esta última magnitud está medida por la inversa de la razón de Mills. A medida que el punto de truncamiento a se desplaza a la izquierda de μ el efecto sobre la media tiende a cero ya que en el caso normal es posible mostrar que λ tiende a cero cuando a tiende a menos infinito.

5.3.2 El modelo lineal truncado

Supongamos que nuestro modelo de interés es el siguiente:

$$y_i = x_i\beta + u_i \qquad u_i \sim N(0, \sigma^2), \qquad E(u_i|x_i) = 0$$

Si los datos no fueron truncados:

$$y_i \sim N(x_i\beta, \sigma^2) \qquad E(y_i|x_i) = x_i\beta$$

Si los datos han sido truncados de manera que solo disponemos observaciones para individuos para los cuales $y_i \geq a$, entonces, de acuerdo a (6), para la muestra que disponemos:

$$\begin{aligned} E(y_i|x_i; y_i \geq a) &= x_i\beta + E(u_i|x_i, y_i \geq a) \\ &= x_i\beta + E(u_i|x_i, u_i \geq a - x_i\beta) \\ &= x_i\beta + \sigma\lambda(\alpha_i) \end{aligned}$$

con $\alpha_i = (a - x_i\beta)/\sigma$. A partir de este resultado es fácil observar porque el estimador mínimo cuadrático de β basado sólo en observaciones truncadas es sesgado. Dicho estimador solo considera el término $x_i\beta$ omitiendo el término $\sigma\lambda_(\alpha_i)$. El sesgo del estimador MCC está asociado a la omisión de dicha variable. Mas específicamente, en términos matriciales:

$$\begin{aligned} \hat{\beta} &= \beta + (X'X)^{-1}X'u \\ E(\hat{\beta}|X, Y > a) &= \beta + (X'X)^{-1}X'E(u|X, Y > a) \\ &= \beta + (X'X)^{-1}X'\sigma\lambda(\alpha) \end{aligned}$$

en donde X, Y y u corresponden a las representaciones matriciales de y_i, x_i y u_i, y $\lambda(\alpha)$ es un vector de n posiciones con las inversas de las razones de Mills para cada observación.

Bajo el supuesto de normalidad de los errores, es posible estimar los parámetros del modelo utilizando el método de máxima verosimilitud. De acuerdo a (5):

$$f(y_i|x_i; y_i \geq a) = \frac{1/\sigma\phi((y_i - x_i\beta)/\sigma)}{1 - \Phi((a - x_i\beta)/\sigma}$$

Entonces, el logaritmo de la función de verosimilitud será:

$$l(\beta, \sigma^2) = -\frac{n}{2}(\ln(2\pi) + \ln\sigma^2) - \frac{1}{2\sigma^2}\sum_i (y - x_i\beta)^2 - \sum_i \ln(1 - \Phi((a - x_i)/\sigma))$$

la cual puede ser maximizada con respecto a los parámetros de interés utilizando técnicas numéricas. Como es habitual, la matriz de varianzas del estimador MV puede obtenerse a través de un estimador de la inversa de la matriz de información.

5.4 Datos Censurados

El modelo de datos censurados que estudiaremos es el siguiente:

$$y_i^* = x_i\beta + u_i \qquad u_i \sim N(0, \sigma^2) \tag{5.7}$$

$$y_i = \begin{cases} y_i^* & \text{si } y^* > 0 \\ 0 & \text{si } y^* \square\ 0 \end{cases} \tag{5.8}$$

Este es el modelo introducido en (1)-(2) con el supuesto adicional de normalidad, y es conocido en la literatura como modelo *tobit*[1]. Este es un modelo lineal para la variable latente, de la cual observamos su verdadero valor sólo si es mayor que cero. El mecanismo de censura asigna el valor cero cuando y_i^* no es positiva.

Utilizando un argumento similar al caso de datos truncados, es posible mostrar que el estimador MCC de β utilizando datos censurados es sesgado. Bajo el supuesto de normalidad es posible obtener estimadores máximo verosímiles. Nótese que la función de distribución de y_i es *mixta* en el sentido de que es discreta en cero y continua en el resto de los valores. De acuerdo a (7)-(8):

$$\begin{aligned} p(y_i = 0) &= P(y_i^* \square\ 0) \\ &= P(x_i\beta + u_i \square\ 0) \\ &= \Phi(-(1/\sigma)x_i\beta) \end{aligned}$$

Cuando y_i es positiva su densidad es:

[1] El modelo de datos censurados fue introducido por Tobin (1958) para estimar una función de demanda deseada de bienes durables

$$f(y_i) = \sigma^{-1}\phi((y_i - x_i\beta)/\sigma)$$

Con esta información podemos construir la función de verosimilitud, y el logartimo de la miama será:

$$l = \sum_{y_i=0} \ln(\Phi(-(1/\sigma)x_i\beta)) + \sum_{y_i>0} ln(\sigma^{-1}\phi((y_i - x_i\beta)/\sigma))$$

la cual puede ser maximizada con respecto a los parámetros de interés utilizando métodos numéricos. Esta función de verosimilitud es un tanto particular en el sentido de que la primera parte suma probabilidades y la segunda densidades. Amemiya (1973) mostro que el estimador máximo-verosímil resultante tiene las propiedades habituales. Como es usual, el estimador de la matriz de covarianzas del estimador máximo verosímil puede obtenerse a través de un estimador de la inversa de la matriz de información.

5.5 Ejemplo numérico

A modo de ilustración, consideremos el siguiente ejemplo. Se genera una muestra de 100 observaciones del siguiente modelo lineal:

$$y_i = -5 + x_i + e_i \tag{5.9}$$

En donde las observaciones x_i fueron obtenidas de una muestra aleatoria de la distribución uniforme en el intervalo $[0, 10]$; e_i se obtiene de la distribución normal estándar. Con esta información se construyen las observaciones para y_i. Primeramente, se estiman los parámetros del modelo lineal (9) utilizando las 100 observaciones originales. Luego utilizamos una submuestra truncada, consistente en pares (x_i, y_i) para los cuales $y_i > 0$, esto es, descartamos toda la información para la cual y es cero o negativa. Esta submuestra tiene tamaño xx. El tercer método utiliza información censurada: las observaciones para las cuales y_i □ 0 son reemplazadas por cero. El tamaño de la muestra es el mismo que el original. Luego estimamos el modelo utilizando el procedimiento de máxima verosimilitud para el modelo tobit descripto anteriormente, usando la muestra censurada. Finalmente aplicamos el estimador máximo verosímil utilizando sólo datos truncados.

La Tabla 1 sintetiza los resultados de los distintos procedimientos de estimación. Como es de esperar, el estimador MCC aplicado sobre la muestra original produce estimaciones muy cercanas a los verdaderos valores de los parámetros. Los resultados obtenidos utilizando datos truncados o censurados presentan un notorio sesgo comparados con los del método MCC sobre la muestra original. Nótese como la estimación máximo-verosímil del modelo tobit usando datos censurados tiende a producir resultados mas cercanos a los

verdaderos valores de los parámetros. Lo mismo sucede con el método máximo verosímil aplicado a datos truncados. Este ejemplo ilustra cuan erróneos pueden ser los resultados obtenidos si el problema de censura es ignorado.

Tabla 1: Efecto de datos censurados y truncados

	intercepto	x
MCC	-4.8638	0.9754
	0.2094	0.0370
Datos Censurados	-1.1770	0.4808
	0.1861	0.0329
Datos Truncados	-3.0754	0.7637
	0.6731	0.0904
Tobit MV	-5.1528	1.0178
	0.5561	0.0780
Truncado MV	-4.1531	0.8938
	0.9437	0.1210

5.6 El método de 2 etapas de Heckman

Una variante comúnmente utilizada para estimar consistentemente los parámetros del modelo de datos censurados es el *método en dos etapas de Heckman*. Recordemos que para nuestro modelo de datos censurados (7)-(8):

$$E(y_i|y_i > 0; x_i) = x_i\beta + \sigma\frac{\phi(-x_i\beta/\sigma)}{1 - \Phi(-x_i\beta/\sigma)}$$

El sesgo asociado con el método de MCC aplicado sobre la muestra truncada proviene de ignorar el segundo término. Heckman propone el siguiente método en dos etapas para la obtención de estimadores consistentes de β. En una primera etapa obtendremos estimadores consistentes de β/σ con el objeto de construir una estimación de λ_i. En la segunda etapa reestimaremos el modelo truncado pero incluyendo esta variable adicional. Es posible demostrar que el estimador resultante es consistente para β.

Con la información del modelo censurado podemos construir la siguiente variable:

$$\tilde{y}_i = \begin{cases} 1 & \text{si } y_i^* > 0 \\ 0 & \text{en caso contrario} \end{cases} \tag{5.10}$$

(10) junto con (7) constituyen el modelo *probit* estudiado anteriormente, excepto que en el modelo probit $\sigma^2 = 1$. Dividiendo ambos términos de (7) por σ, el modelo transformado es:

$$\frac{y_i^*}{\sigma} = \frac{x_i\beta}{\sigma} + \frac{u_i}{\sigma}$$

con $u_i/\sigma \sim N(0,1)$. Este modelo transformado es idéntico al modelo probit y, con la información disponible, el vector de parámetros β/σ puede ser consistentemente estimado utilizando los métodos antes descriptos. Con esta infomación, podemos estimar λ_i. En la segunda etapa se reestima el modelo original incorporando λ_i como regresor adicional, utilizando la muestra truncada.

El procedimiento en dos etapas de Heckman puede ser resumido de la siguiente manera:

1. 1a Etapa: Estimar $\gamma = \beta/\sigma$ utilizando el modelo probit (7), (10) usando todas las observaciones de la muestra. La variable binaria $\tilde{y}_i$ es la variable dependiente. Construir la variable $\lambda_i = \phi()/\Phi()$.

2. 2a Etapa: Estimar el modelo original regresando la variable y_i en el vector de variables explicativas x_i adicionando la variable λ_i obtenida en el paso anterior. En esta etapa utilizamos la muestra truncada, es decir, sólo las observaciones no censuradas.

Intuitivamente, la primera etapa del método de Heckman estima un modelo de la probabilidad de que los datos sean no censurados, con el propósito de obtener estimaciones de la variable λ_i. En la segunda etapa el modelo truncado es estimado, pero incorporando esta variable adicional, lo cual corrige el problema de sesgo encontrado en el método de MCC. Un problema con este estimador es que el modelo lineal de la segunda etapa es heteroscedástico, por lo que el estimador MC de la matriz de varianzas y covarianzas es sesgado (Heckman, 1979). Existen varias alternativas para obtener estimadores consistentes de la matriz de covarianzas, la mas usual es usar el estimador consistente de White (Davidson y MacKinnon, 9. 544) disponible en la mayoría de los paquetes econométricos.

5.7 Extensiones

1. El supuesto de normalidad del término aleatorio es de crucial importancia. Las propiedades de optimalidad de los estimadores máximo-verosímiles desaparecen cuando la distribución de los errores es incorrectamente especificada. Ver Bera (1982) para un test de normalidad en el modelo de datos truncados y censurados.

2. Similarmente al caso de variables binarias, los métodos semiparamétricos ofrecen una alternativa robusta a los estimadores máximo-verosímiles. Los métodos de regresión por cuantiles (Koenker y Basset, 1978) pueden ser adaptados al caso de datos censurados (Powell, 1984). Dichos estimadores se basan en las propiedades de equivariancia de la mediana muestral, extendiendo la idea al caso de cuantiles condicionales. Ver Powell (1994) o Lee (1996) para una revisión reciente de estos temas.

3. En forma similar al caso de los modelos binarios, el estimador máximo-verosímil del modelo Tobit es inconsistente cuando el término de error no es homoscedástico. Ver Petersen y Waldman (1881) para la estimación de un modelo de datos censurados con heteroscedasticidad. Pagan y Vella (1989) presentan una familia de tests para el modelo tobit basada en el principio de momentos condicionales.

5.8 Aplicaciones

La aplicación presentada en esta sección se basa en el estudio de Mroz (1987). Como señaláramos en la introducción, el propósito del trabajo es estudiar funciones de oferta de trabajo del sector femenino. En nuestro ejemplo estimaremos dicha función de oferta utilizando como variable dependiente la cantidad de horas trabajadas y como variables explicativas utilizaremos la cantidad de hijos menores de 6 y 18 años (KL6 y K618, respectivamente, la edad (WA), la cantidad de años de educación recibida (WE), los ingresos del marido (PRIN) y el salario (LOGWAGE). Esta última variable es, para aquellas mujeres que trabajaron, igual a su salario mientras que para aquellas que no lo hicieron se trata de una estimación (ver Berndt, pp. 675 para más detalles acerca de la estimación de los salarios no observados). De las 753 observaciones disponibles sólo 453 mujeres declararon haber trabajado durante el período de análisis. La Tabla 2 presenta estimaciones de la función de oferta utilizando los métodos estudiados en esta sección.

Comenzamos estimando la función de oferta utlizando solo información para aquellas mujeres que han trabajado, esto es, usando solo información truncada, utilizando el método de mínimos cuadrados ordinarios. Luego estimamos el mismo modelo utilizando la formulación tobit utilizando el método de máxima verosimilitud. Por último se utiliza el método en dos etapas de Heckman. Los coeficientes estimados por el método de MCC y por el método de Heckman son notoriamente similares, aunque los errores estándar obtenidos del método de Heckman son algo inferiores. Los signos de KL6 y de K618 sugieren que la cantidad de horas trabajadas decrece con el número de hijos menores de 6 y 18 años. Similarmente, el nivel de educación y la edad afectan negativamente a la oferta de horas trabajadas, aunque ninguna de estas variables es significativamente distinta de cero. Lo mismo ocurre con el salario, medido por LOGWAGE.

Los resultados obtenidos con el modelo Tobit son significativemente diferentes a los anteriores. La mayoría de los signos se corresponde con los obtenidos a través del método MCC y del método de Heckman. Todas los coeficientes resultan ser significativamente distintos de cero. El coeficiente de la variable salario es ahora significativamente distinto de cero y positivo.

Tabla 1: Estimaciones del ejemplo basado en Mroz (1987)

	MCC			Tobit			Heckman		
	Coef.	Err.Std	t	Coef.	Err.Std	t	Coef.	Err.Std	t
Constante	2114.70	340.13	6.22	1172.00	477.98	2.45	2104.80	505.46	4.16
KL6	-342.50	100.01	-3.43	-1045.20	125.02	-8.36	-351.90	371.08	-0.95
K618	-115.02	30.83	-3.73	-100.35	42.29	-2.37	-115.28	32.08	-3.59
WA	-7.73	5.53	-1.40	-36.51	76.40	-4.78	-8.07	14.04	-0.58
WE	-14.45	17.97	-0.80	104.92	25.74	4.08	-13.22	50.11	-0.26
PRIN	0.00	0.00	-1.16	-0.02	0.00	-4.54	0.00	0.09	-0.49
LOGWAGE	-17.41	54.22	-0.32	199.39	88.75	2.25	-14.58	120.52	-0.12

5.9 Bibliografía

Los capítulos correspondientes de Greene (1993) y Davidson y MacKinnon (1993) y Long (1997) pueden ser una buena introducción al tema. Amemiya (1985) y Maddala (1986) presenta una visión mas detallada. Para enfoques recientes puede consultarse Lee (1996) o Powell (1995), especialmente para el enfoque de regresiones por cuantiles censuradas.

Capítulo 6

Modelos de duración

6.1 Motivación

Si bien los modelos de duración tienen una larga trayectoria en biología y estadística médica, su uso en economía aplicada es relativamente reciente. El ejemplo clásico en donde se usa este tipo de técnica es el análisis de duración del desempleo. Consideremos la siguiente situación. Supongamos que un individuo fue despedido de su trabajo, tras lo cual dedica sus esfuerzos a buscar un nuevo empleo. Supongamos que a las dos semanas de haber sido despedido, el individuo sigue desempleado. Una pregunta de interés para el individuo es saber cual es la probabilidad de encontrar trabajo en la semana siguiente teniendo en cuenta que todavía (pasadas dos semanas de búsqueda) no encontró trabajo. Llamemos a esta probabilidad p_1. Obviamente, la misma pregunta es relevante si el individuo sigue desempleado al cabo de, digamos, dos meses, y denotemos a esta probabilidad como p_2. Es interesante comparar estas dos probabilidades. Cuál de ellas es mayor? Existen varias razones para argumentar que cualquiera de ellas puede serlo. Por ejemplo, se podría decir que dos semanas es un período de búsqueda relativamente corto y que, por el contrario, en dos meses el individuo debería haber explorado lo suficiente el mercado como para darse cuenta de sus posibilidades de empleo. Desde esta perspectiva, p_1 es mayor que p_2. También puede arguirse que a medida que pasa el tiempo el individuo baja sus pretensiones, y por lo tanto estaría mas dispuesto a considerar a los dos meses ofertas que a las dos semanas de haber estado despedido rechazaba. Desde esta perspectiva podría arguirse que p_2 es mayor que p_1. Obviamente, el análisis de estas probabilidades condicionales podría repetirse para cualquier período, lo que nos permitiría caracterizar la evolución de la misma en el tiempo. El análisis de duración permite caracterizar importantes aspectos de la dinámica del proceso de búsqueda de trabajo, tales como: Cuál es la duración promedio del desempleo? De qué factores depende esta duración? Cómo evoluciona

la probabilidad condicional de encontrar empleo en un determinado período teniendo en cuenta que el individuo sigue desempleado?

Para analizar tales preguntas se podría pensar en construir una base de datos basada en el siguiente experimento natural. Se le pediría a un grupo de individuos que se reporten a una oficina en el momento de ser despedidos y se les preguntarían, por ejemplo, algunas características personales (edad, sexo, salario anterior, etc.) que pueden influenciar el proceso de búsqueda de empleo. Luego se les pediría que se reporten con cierta frecuencia (por ejemplo, semanal) indicando si han encontrado trabajo o si siguen el proceso de búsqueda. Idealmente, la base de datos resultante sería una tabla indicando el tiempo que le tomó a cada indivudo encontrar empleo y sus características personales, las cuales pueden ser utilizadas como variables explicativas de la duración del desempleo. Pero es lógico pensar que en el momento en que el investigador decide estimar un modelo de duración del desempleo se encuentre con que algunos individuos abandonaron el 'experimento' (dejan de reportarse) y que otros siguen desempleados al momento de realizar el estudio. Ambos casos corresponden a observaciones *censuradas por la derecha*. Por ejemplo, supongamos que el investigador decide realizar el estudio al año de haber comenzado el experimento natural, y que un grupo de individuos sigue desempleado. Para estas personas no se observa directamente la duración del desempleo sino que dichas personas estuvieron buscando trabajo un año *como mínimo*. Por otro lado, si una persona abandonó, dejó de reportarse a los 6 meses de haber estado buscando trabajo, también sabemos que esta persona estuvo desempleada 6 meses como mínimo. Ambos casos corresponden a observaciones censuradas, para las cuales la verdadera duración del desempleo no se observa. Esta situación ejemplifica una característica inherente del análisis de duración: salvo raras ocasiones, la misma mecánica del proceso de generación de datos de duración hace que se deba considerar la posibilidad de que existan datos censurados.

En esta nota presentamos algunas nociones básicas del análisis de duración[1]. Por razones pedagógicas adoptaremos un enfoque estrictamente paramétrico, pero es importante aclarar que en este tipo de estudios son notorias las ventajas de utilizar técnicas no-paramétricas y semiparamétricas. Así y todo preferimos introducir el tema usando métodos paramétricos, los cuales gozan de mayor popularidad en la práctica de la econometría.

6.2 Conceptos básicos

En términos generales, la variable de interés en estos modelos es el tiempo que tarda un sistema en pasar de un estado a otro. Generalmente dicha transición se halla asociada a la ocurrencia de un suceso (encontrar trabajo, quiebra de

[1] En la literatura de biometría este tipo de análisis habitualmente se refiere al tiempo de supervivencia de un individuo luego de un tratamiento médico, de ahí que a este tópico se lo llame *análisis de supervivencia*.

una firma, resolución de un conflicto laboral, desaparición de un producto del mercado, etc.) que indica la finalización del evento cuya duración se intenta estudiar. Esta variable aleatoria, llamada *duración*, toma valores positivos y supondremos que es absolutamente continua, y la denominaremos con T.

En general, la distribución de probabilidad de una variable aleatoria continua puede ser caracterizada por alguna de las siguientes funciones:

1. Función de distribución: $F(t) = P(T \square t)$
2. Función de supervivencia: $S(t) = P(T \geq t) = 1 - F(t)$
3. Función de densidad: $f(t) = dF(t)/dt = -dS(t)/dt$

   En el caso en donde la variable T representa la duración de un evento, $F(t)$ indica la probabilidad de que el evento dure como máximo hasta t y $S(t)$ indica la probabilidad de que el evento dure por lo menos hasta t.
4. Función de riesgo: como señaláramos en la introducción a esta sección, en muchas ocasiones estaremos interesados en conocer la probabiilidad de que el evento concluya en un intervalo Δ a partir de t conociendo que no ha concluido hasta ese momento. Dicha probabilidad será:

   $$h(t, \Delta) = Pr[t \square T \square t + \Delta | T \geq t]$$

   La función de riesgo se define como:

   $$h(t) = \lim_{\Delta \to 0} \frac{h(t, \Delta)}{\Delta}$$

   e indica la probabilidad de que el evento concluya en el instante siguiente al momento t.

   De acuerdo a la definición de probabilidad condicional:

$$\begin{aligned} h(t) &= \lim_{\Delta \to 0} \frac{Pr[t \square T \square t + \Delta | T \geq t]}{\Delta} && (6.1) \\ &= \lim_{\Delta \to 0} \frac{Pr[t \square T \square t + \Delta]}{Pr[T \geq t]\Delta} && (6.2) \\ &= \lim_{\Delta \to 0} \frac{F(t + \Delta) - F(t)}{\Delta} \frac{1}{S(t)} && (6.3) \\ &= \frac{f(t)}{S(t)} && (6.4) \end{aligned}$$

   Entonces, la función de riesgo puede ser definida como el cociente de la densidad sobre la función de supervivencia.

5. Relación función de riesgo - survival: Tal como sugiriéramos anteriormente, las funciones de densidad $f(t)$ y de supervivencia $S(t)$ permiten caracterizar completamente la distribución de probabilidad de una variable aleatoria continua ya que es inmediato reconstruir la función de distribución $F(t)$ a partir de cualquiera de estas dos funciones. En esta subsección mostramos que también es posible caracterizar la distribución de una variable aleatoria continua a través de la función de riesgo, para lo cual debemos poder derivar $F(t)$ a partir de ésta.

   Por definición:

$$\begin{aligned} h(t) &= f(t)/S(t) && (6.5)\\ &= \frac{-dS(t)/dt}{S(t)} && (6.6)\\ &= -\frac{d\ln S(t)}{dt} && (6.7) \end{aligned}$$

   integrando ambos miembros en el intervalo $[0,t]$

$$\begin{aligned} \int_0^t h(s)ds &= -\int_0^t -\frac{d\ln S(s)}{ds}ds && (6.8)\\ &= -[\ln S(t) - \ln S(0)] && (6.9)\\ &= -\ln S(t) && (6.10) \end{aligned}$$

   dado que $S(0) = 1$. Entonces:

$$S(t) = exp(-\int_0^t h(s)ds)$$

   Este resultado muestra como es posible reconstruir la función de supervivencia $S(t)$ (e, inmediatamente, $F(t)$) a partir de la función de riesgo, e implica que nada se pierde si construimos el modelo de duración basado exclusivamente en esta última. En definitiva, la función de riesgo también puede caracterizar la distribución de una variable aleatoria continua.

6.3 El modelo sin variables explicativas

El objetivo consiste en estimar un modelo simple de las características de la duración de un evento. Como paso inicial, intentaremos especificar un simple modelo paramétrico sin utilizar variables explicativas. Como adelantáramos en la Introducción, la presencia de datos censurados es un hecho habitual en el análisis de duración.

Consideremos el caso de datos censurados (a la derecha). La estimación se basa en una muestra de n observaciones de la variable T, denotadas como $t_i, i = 1, \ldots, n$. Cuando t_i no es censurada eso significa que el evento efectivamente término en t_i. Cuando es censurada solo sabemos que el evento duro *por lo menos* hasta t_i. En forma similar al modelo Tobit estudiado anteriormente, cada observación contribuye a la función de verosimilitud con su 'probabilidad' de ocurrencia: las observaciones no censuradas contribuyen con $f(t_i|\theta)$ y las censuradas con $S(t_i|\theta)$. θ es un vector de parámetros desconocidos, los cuales son el objeto de la estimación. Aquí también es importante notar que la distribución censurada de la muestra tiene una distribución mixta.

Sea $\delta_i, i = 1, \ldots n$ un indicador binario con valor igual a 1 si la duración t_i no fue censurada y 0 si lo fue. De acuerdo a esta información, el logaritmo de la función de verosimilitud es:

$$l(\theta) = \sum_{i|y_i=1} \ln f(t_i;\theta) + \sum_{i|y_i=0} \ln S(t_i;\theta) \tag{6.11}$$

$$= \sum_{i=1}^{n} [y_i \ln f(t_i;\theta) + (1-y_i) \ln S(t_i;\theta)] \tag{6.12}$$

la cual puede ser reescrita en términos de la función de riesgo, de la siguiente manera. De la definición de $h(t)$ obtenemos: $f(t) = h(t)S(t)$. Reemplazando en (1);

$$l(\theta) = \sum_{i=1}^{n} [\ln(h()S()) + (1-y_i)S()] \tag{6.13}$$

$$= \sum_{i=1}^{n} y_i \ln h() + \sum_{i=1}^{n} [y_i \ln S() + (1-y_i) \ln S()] \tag{6.14}$$

$$= \sum_{i=1}^{n} y_i h() + \sum_{i=1}^{n} \ln S() \tag{6.15}$$

en donde los argumentos de las funciones $h()$ y $S()$ han sido suprimidos para facilitar la notación. De esto se sigue que una vez especificada la función de riesgo $h(t)$, los parámetros de interés pueden ser estimados por el método de máxima verosimilitud maximizando (2).

6.4 Algunos ejemplos

Supongamos que T tiene distribución *exponencial* con:

$$F(t) = 1 - e^{-\lambda t}$$

de la cual se obtiene que:

$$\begin{aligned} S(t) &= e^{-\lambda t} && (6.16)\\ f(t) &= \lambda e^{-\lambda t} && (6.17)\\ h(t) &= \lambda && (6.18) \end{aligned}$$

Esta distribución tiene función de riesgo constante y es conocida como *memoryless*: la probabilidad instantánea de que el evento concluya, condicional en el pasado de la misma no varía en el tiempo. El pasado no contribuye a aumentar o disminuir esta probabilidad condicional. También se dice que en el caso exponencial no hay *dependencia de la duración*, es decir la función de riesgo es independiente del tiempo.

Otra alternativa es la distribución de *Weibull*, la cual presenta:

$$\begin{aligned} F(t) &= 1 - e^{-\lambda t^{\alpha}} && (6.19)\\ S(t) &= e^{-\lambda t^{\alpha}} && (6.20)\\ f(t) &= \lambda \alpha t^{\alpha-1} e^{-\lambda t^{\alpha}} && (6.21)\\ h(t) &= \lambda \alpha t^{\alpha-1} && (6.22) \end{aligned}$$

Se trata de una generalización del caso exponencial, el cual corresponde a $\alpha = 1$. La función de riesgo correpondiente a la distribución de Weibull es creciente en el tiempo si $\alpha > 1$ y decreciente si $\alpha < 1$. En el primer caso diremos que la dependencia de la duración es positiva y en el segundo, negativa. En el contexto de duración de desempleo, en el caso de dependencia positiva, la probabilidad condicional de encontrar trabajo es creciente en el tiempo.

6.5 El modelo con variables explicativas

La adición de variables epxlicativas en el modelo de duración no es trivial, y depende de la interpretación deseada. En estas notas presentaremos algunas estrategias paramétricas comúnmente utilizadas, y su relación en el contexto de la distribución de Weibull, la cual, como dijéramos anteriormente, incluye a la exponencial como caso particular. Sea x un vector de p variables explicativas

a) El modelo de riesgo proporcional: en esta especificación las variables explicativas afectan directamente a la función de riesgo en forma proporcional. La incorporación de variables explicativas se efectúa a través de la función de riesgo, la cual resulta especificada como:

$$h(t|x) = h_0(t) \exp(\beta' x) \qquad (6.23)$$

en donde β es un vector de p coeficientes lineales. $h_0(t)$ es la *función de riesgo base*. La función de riesgo base depende del tiempo solo a través de la función de riesgo base, la cual controla el comportamiento temporal de la función de riesgo. El segundo factor controla el efecto de las variables explicativas en forma multiplicativa. Esta distinción es importante. Por ejemplo, si la dependencia de la duración es positiva o negativa depende exclusivamente de la función de riesgo base.

Nótese que la función de riesgo base es la misma para cualquier individuo en la muestra y, por lo tanto, para dos individuos con variables explicativas igual a x y x^* el cociente de las funciones de riesgo es:

$$\frac{h(t|x^*)}{h(t|x)} = exp(\beta'(x^* - x))$$

,el cual es constante en el tiempo. La interpretación de los coeficientes está dada por la siguiente derivada:

$$\frac{\partial \ln h(t|x)}{\partial x_k} = \beta_k$$

,$k = 1, \ldots, p$. Entonces β_k da el cambio proporcional en la función de riesgo que resulta de un cambio marginal en la k-esima variable explicativa. En otros términos, cuando las variables explicativas están medidas en niveles, los coeficientes estimados de esta especificación tienen la interpretación de semielasticidades de la función de riesgo con respecto a las variables explicativas. Cuando x_k es una variable binaria que toma valor igual a 1 si el individuo pertenece a cierto grupo y 0 si no pertenece, la interpretación adecuada es la siguiente. Supongamos que x y x^* difieren solo en la k-esima variable explicativa, la cual es una variable binaria. Entonces, a partir del resultado anterior:

$$\frac{h(t|x^*)}{h(t|x)} = e^{\beta_k}$$

Este resutlado proporciona el riesgo de que el evento ocurra si el individuo pertenece a un cierto grupo (indicado por x_k) en relación al riesgo de que el evento ocurra si el individuo no pertence a dicho grupo.

b) El modelo de riesgo acelerado: Otra posibillidad consiste en incorporar las variables explicativas a través de la función de supervivencia. La especificación es:

$$S(t|x) = S_0[exp(\theta' x)t] \tag{6.24}$$

en donde θ es un vector de p coeficidentes. Nótese que en este caso, el efecto de las variables explicativas es alterar la escala temporal del evento. La función de supervivencia es, por definición, una función monótona decreciente de t.

Entonces, $exp(\theta' x)$ es un factor de aceleración que indica como las variables explicativas afectan a la escala temporal. Para observar este efecto en forma mas clara, consideremos el siguiente caso en que las variables explicativas se miden como desvíos con respecto a sus medias. Supongamos que la única variable explicativa es la edad, y que la probabilidad de que un individuo con edad promedio permanezca desempleado un mínimo de 4 semanas 0.5 ($S(4, x = 0) = S_0[4] = 0.5$). Esto es, la probabilidad no condicional de estar empleado al menos 4 semanas es 0.5 para el individuo promedio. También supongamos que para el mismo individuo la probabilidad de supervivencia a 5 semanas es $S(5, x = 0) = S_0[5] = 0.3$.

Supongamos que $\theta = 0.2231$, entonces, para un individuo un año mayor que el promedio, la probabilidad de supervivencia será $S(t, x = 1) = S_0[exp(0.2231)t]$, entonces, la probabilidad de supervivencia a 4 semanas es:

$$S(4, x = 1) = S_0[exp(0.2231)4] = S_0[5] = 0.3$$

Esto es, la probabilidad de supervivencia a 4 semanas para un individuo un año mayor que el promedio es la misma que la probabilidad de supervivencia a *cinco* semanas para el individuo promedio. Aquí se aprecia que el efecto de la edad es *acelerar* la escala temporal del evento.

b) La representación log-lineal: para varias distribuciones, el logaritmo de la duración T puede ser expresada en forma de regresión:

$$\ln T = \gamma' x + \sigma\omega$$

en donde γ es un vector de p coeficientes, ω es un 'término de error' y σ es un parámetro de escala que controla la varianza del término de error $\sigma\omega$. En este caso, la interpretación de los coeficietnes es análoga al modelo de regresión clásico: γ_k es el efecto proporcional sobre la duración de cambiar x_k marginalmente, esto es, los coeficientes γ son semielasticidades de la duración del evento con respecto a las variables explicativas.

Para entender la interrelación entre las distintas representaciones, consideremos el caso del modelo de Weibull. Como mostraramos anteriormente, el caso sin variables explicativas tiene las siguientes funciones de supervivencia y de riesgo:

$$S(t) = exp(-\lambda t^{\alpha}); \qquad h(t) = \lambda\alpha x^{\alpha-1}$$

La representación log-lineal puede ser fácilmente obtenida a través de la siguiente reparametrización. Definamos $\lambda = exp(-\mu/\sigma)$ y $\alpha = 1/\sigma$. Entonces, se puede mostrar que $y = \ln T$ puede ser expresado como:

$$y = \mu + \sigma\omega$$

en donde ω es un término de error que tiene la distribución de *valor extremo*. En esta especificación, las variables explicativas pueden ser naturalmente incoporadas como:

$$y = \mu_0 + \gamma' x + \sigma\omega \tag{6.25}$$

Reemplazando $\mu = \mu_0 + \gamma' x$ en la defincón de λ proporcionada anteriormente y utilizando la expresión de la función de riesgo para el caso Weibull, podemos obtener la siguiente reparametrización de la función de riesgo en término de las variables explicativas:

$$h(t|x) = \alpha\lambda_0 t^{\alpha-1}\exp(\beta' x)$$

con: $\alpha = 1/\sigma, \lambda_0 = \exp(-\mu/\sigma)$ and $\beta = -\gamma/\sigma$. Esto corresponde a la representación de riesgo proporcional descripta anteriormente. La representación de riesgo acelerado puede ser fácilmente obtenida utilizando la definición de la función de superviviencia:

$$h(t|x) = exp(\theta' x)h_0[exp(\theta' x)]$$

con $\theta = \beta/\alpha$ o $\theta = -\gamma$.

Lamentablemente, la distribución de Weibull (incluyendo a la exponencial como caso particular) es la única distribución que tiene una representación de riesgo proporcional y de riesgo acelerado.

La estimación de los parámetros puede basarse en cualquiera de las representaciones descriptas ya que los parámetros de las restantes pueden ser fácilmente recuperados utilizando las relaciones discutidas anteriormente. Adicionalmente, estas diferentes reparametrizaciones proveen diferentes formas de intepretar los coeficientes obtenidos.

Resulta interesante observar el efecto de las variables explicativas sobre ciertos momentos de la distribución de Weibull. En el caso sin variables explicativas, se puede mostrar que la esperanza de la distribución de Weibull es:

$$E(T) = \frac{\Gamma(1 + 1/\alpha)}{\lambda^{1/\alpha}}$$

en donde $\Gamma()$ es la funcion Gama, $\Gamma(x) \equiv \int_0^\infty t^{x-1}\exp(-t)dt$. Utilizando la reparametrización log-lineal:

$$E(T) = \frac{\Gamma(1 + 1/\alpha)}{exp(-(\mu_0 + \gamma' x))}$$

entonces,

$$\frac{\partial \ln E(T)}{\partial x_k} = \gamma_k = -\beta_k\sigma$$

entonces, los coeficientes de la especificación log-linea miden la semielasticidad de la duración esperada con respecto a las variables explicativas, lo cual es también medido por menos los coeficientes de la especificación de riesgo proporcional (reescalados por σ).

Un resultado similar puede ser obtenido para la mediana. Para el caso sin variables explicativas, es posible mostrar que:

$$M(T) = (-\ln(0.5)/\lambda)^{1/\alpha}$$

Tomando logaritmos:

$$\ln M(T) = K - 1/\alpha \ln \lambda$$

con $K = 1/\alpha \ln(-\ln(0.5))$. Usando la reparametrización log-lineal:

$$\ln M(T) = K + \mu_0 + \gamma' x$$

Entonces:

$$\frac{\partial \ln M(T)}{\partial x_k} = \gamma_k = -\beta_k \sigma$$

En forma similar, cuando x_k es una variable binaria y x y x^* difieren solo en la k-esima variable explicativa:

$$M(T|x^*) = e^{\gamma_k} M(T|x)$$

6.6 Bibliografía

Existen varios libros de texto que cubren las técnicas de análisis de supervivencia, la mayoría de ellos con referencias a problemas de biometría. Cox y Oakes (1984) es una introducción clara y concisa. Klein y Moeschberger (1997) es un excelente texto introductorio con abundantes ejemplos empíricos. Kalbfleisch y Prentice (1980) es la referencia clásica y es muy claro. Lamentablemente estos textos son escritos explícitamente para investigadores en biología y medicina. El único texto escrito para econometristas es Lancaster (1990), que presenta un análisis detallado de los modelos de duración con énfasis en aplicaciones en economía laboral. Amemiya (1985) y Greene (1997) tratan el tópico, aunque no con demasiados detalles. Kiefer (1988) es un muy útil *survey* de modelos básicos y sus aplicaciones. Algunas aplicaciones recientes son Greesntein y Wade (1998) y Meyer (1990)

Bibliografía

[1] Amemiya, T., 1981, Qualitative Response Models: A Survey, Journal of Economic Literature. Vol. XIX, pp.1483-1536.

[2] Amemiya, T., 1985, Advanced Econometrics, Harvard University Press, CA Massachusetts.

[3] Anderson, S. de Palma, A. and Thisse, J.F, 1992, Discrete Choice Theory of Product Differentiation, MIT Press, Cambridge.

[4] Arellano, M. and Bond, S., 1991, Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, Review of Economic Studies, 58, pp.277-297.

[5] Baltagi, B., 1995, Econometric Analysis of Panel Data, John Wiley and Sons, West Sussex, England

[6] Baltagi, B., 1998, Panel Data Methods, Ch. 9 in Ullah, A. y Giles, D. (eds.), 1998, Handbook of Applied Economic Statistics, Marcel Dekker, New York

[7] Baltagi, B., and Griffin, J. ,1983, Gasoline demand in the OECD. An application of pooling and testing procedures, European Economic Review, 22, pp. 117-137. North Holland.

[8] Bera, A., Sosa Escudero, W. and Yoon, M., 1997, Tests for the Error-Component Model under Local Misspecification, mimeo, University of Illinois at Urbana-Champaign.

[9] Berndt, E., 1991, The Practice of Econometrics. Classic and Contemporary, Addison-Wesley Publishing Co. Massachusetts.

[10] Bertranou, F., 1998, Application of Count Data Regression Models to Health Care Services Utilization Data from Argentina, mimeo, University of Pittsburgh.

[11] Bickel, P. and Doksum, K., 1981, Mathematical Statistics, New York: Holden Day.

[12] Breusch, T. and Pagan, A., 1980, The Langrange Multiplier Test and its Applications to Model Specification in Econometrics, Review of Economic Studies, 47, 239-53.

[13] Cameron, C and Trivedi, P., 1997, Regression Analysis of Count Data, Cambridge: Cambridge University Press.

[14] Chamberlain, G., 1980, Analysis of Covariance with Qualitative Data, Review of Economic Studies, XLVII, pp. 225-238.

[15] Chamberlain, G., 1984, Panel Data, in Griliches, Z. and Intrilligator, M. (eds.), Handbook of Econometrics, Vol. II, Edited by Griliches & Intriligator. Elsevier, Amsterdam.

[16] Dabos, M. , 1996, Crisis Bancaria y Medicion del Riesgo de Default: Metodos y el caso de los Bancos Cooperativos en Argentina, mimeo, Universidad de San Andres.

[17] Dabos, M. y Sosa Escudero, W., 1998, Estimating and predicting bank failure in Argentina, mimeo, Universidad de San Andres.

[18] Davidson, R. and MacKinnon, J., 1993, Estimation and Inference in Econometrics, Oxford University Press, Oxford.

[19] Deaton, A., 1997, The Analysis of Household Surveys, Baltimore: Johns Hopkins University Press for the World Bank

[20] Donald, S. and Sappington, D., 1995, Explaining the choice among regulatory plans in the U.S. telecommunications industry, Journal of Economics & and Management Strategy, Vol. 4, No. 2, Summer 1995, 237-265.

[21] Durrett,R., 1996, Probability: Theory and Examples, 2nd Edition, Duxbury Press, Belmont, California.

[22] Efron, B., and Tibshirani, R., 1993, An Introduction to the Bootstrap, Chapman & Hall, London.

[23] Engle, R. and McFadden, D. (eds), 1994, Handbook of Econometrics, vol. 4, Elsevier Science, Amsterdam.

[24] Galiani, S., Lamarche, C., Porto, A. y Sosa Escudero, W., 1999, Persistance and the Determinants of Unemployment. Argentina 1984-1997, mimeo, Universidad Nacional de La Plata

[25] Gasparini, L., 1998, Measuring unfairness, mimeo, Universidad Nacional de La Plata.

[26] Godfrey, L., 1988, Misspecification Tests in Econometrics, Cambridge University Press, Cambridge.

[27] Greene, W. ,1997, Econometric Analysis, 3rd Ed, Macmillan, New York.

[28] Greenstein, S. and Wade, J., 1996 *Dynamic modeling of the product life cycle in the commercial mainframe computer market, 1968-1982*, mimeo, University of Illinois at Urbana-Champaign.

[29] Griliches, Z. and Intrilligator, M. (eds), 1986, Handbook of Econometrics, vol. 1, 2 and 3. Elsevier, Amsterdam.

[30] Gujarati, D., 1995, Basic Econometrics, 3rd edition, McGraw-Hill, New York.

[31] Hardle, W., 1990, Applied Nonparametric Regression, Cambridge University Press, New York

[32] Hardle, W., 1994, Applied Nonparametric Methods, en Engle y McFadden (1994) Handbook of Econometrics, vol. 4, Elsevier Science, Amsterdam.

[33] Harville, D., 1997, Matrix Algebra From a Statistician's Perspective, Springer, New York.

[34] Hausman, J. and Taylor, W., 1981, Panel Data and Unobsevable Individual Effects, Econometrica, Vol 49, No. 6.

[35] Heckman, J. y Snyder, J., 1997, Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of, RAND Journal of Economics. Vol. 28, pp. s142-s189.

[36] Johnston, J. y J. DiNardo, 1997, Econometric Methods, 4th edition, McGraw-Hill, New York

[37] Jones, M., Sanguinetti, P. and Tommasi, M., 1998, Politics, Institutions and Fiscal Performance in the Argentine Provinces, mimeo, Universidad de San Andres.

[38] Kalbfleisch, J.D., & R.L. Prentice, 1980 *The Statistical Analysis of Failure Time Data*. New York: Wiley.

[39] Kiefer, N., 1988, 'Economic Duration Data and Hazard Functions', *Journal of Economic Literature*, 26, pp. 646-679.

[40] Klein, J. and Moeschberger, M., 1997 *Survival Analysis. Techniques for Censored and Truncated Data*. New York: Springer-Verlag.

[41] Koenker, R., 1977, Was Bread Giffen?, Review of Economics and Statistics.

[42] Koenker, R. and Bassett, G., 1978, Regression Quantiles, Econometrica, Vol.46, No. 1, pp.33-50.

[43] Lancaster, T. ,1990, The Econometric Analysis of Transition Data, Cambridge University Press.

[44] Lee, M., 1996, Method of Moments and Semiparametric Econometrics for Limited Dependent Variable Models, Springer-Verlag, New York.

[45] Lehmann, E., 1983, Theory of Point Estimation, Champan & Hall, Cambridge.

[46] Lindgren, B., 1993, Statistical Theory, Chapman & Hall, Cambridge.

[47] Maddala, G.S., 1983, Limited Dependent and Qualitative Variables in Econometrics, Cambridge University Press.

[48] MacKie-Mason, J. and Varian, H. ,1997, Economic FAQ's About the Internet, en Internet Economics

[49] Marchionni, M., 1998, La Cobertura en Salud en Argentina. Un Analisis Empirico en base al Modelo Logistico Multinomial, Tesis de Maestria en Economia, Universidad Nacional de La Plata.

[50] Matyas, L and Sevestre, P. (eds), 1996, The Econometrics of Panel Data, Kluwer Academic Publishers, Dordrecth, The Netherlands.

[51] McCullagh, P. y Nelder, J., 1989, Generalized Linear Models, 2nd edition, Chapman & Hall, Cambridge.

[52] McFadden, D., 1984, Econometric Analysis of Qualitative Response Models, in Griliches, Z. and Intrilligator, M. (eds.), Handbook of Econometrics, Vol. II, Edited by Griliches & Intriligator. Elsevier, Amsterdam.

[53] Moore, M., 1996, Death and tobacco taxes, RAND Journal of Economics, vol. 27, No.2 , Summer 1996, 1997, pp. 415-428.

[54] Mroz, T., 1987, The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions, Econometrica, Vol. 55, No.4 (July 1987), pp. 765-799.

[55] Mundlak, Y., 1978, On the Pooling of Time Series and Cross section Data, Econometrica, Vol. 46, No.1.

[56] Newey, W. and McFadden, D., 1994, Large Sample Estimation and Hypothesis Testins, in Engle, r. and McFadden, D. (eds.) Handbook of Econometrics, Vol.4, North Holland, Amsterdam.

[57] Pagan, A. and Vella,, 1989, Diagnostic Tests for models based on individual data: a survey, Journal of Applied Econometrics, Vol 4, S29-S59.

[58] Pessino, C., 1995, Returns to Education in Greater Buenos Aires 1986-1993: from Hyperinflation to Stabilization, Documento de Trabajo CEMA No. 104.

[59] Petersen, T., 1986 Fitting parametric survival models with time-dependent covariates. *Applied Statistics* 35: 3,pp. 281-288.

[60] Petersen, T., 1986 Estimating fully parametric hazard rate models with time-dependent covariates. *Sociological Methods & Research*, 14: 3, pp. 219-246

[61] Powell, J., 1984, Least Absolute Deviations Estimation for the Censored Regression Model, Journal of Econometrics, 25, pp. 303-325.

[62] Powell, J., 1994, Estimation of Semiparametric Models, in Engle, R. and McFadden, D. (eds.) Handbook of Econometrics, Vol.4, North Holland, Amsterdam.

[63] Pudney, S, 1989, Modelling Individual Choice. The Econometrics of Corners, Kinks and Holes, Basil Blackwell, Cambridge, Massachusetts.

[64] Schott, J., Matrix Analysis for Statistics, Wiley & Sons, New York.

[65] Scott Long, J., 1997, Regression Models for Categorical and Limited Dependent Variables, SAGE Publications.

[66] Sosa Escudero, W. y Marchionni, M., 1999, Household Structure, Gender, and the Educational Attainment of Children in Argentina, mimeo, World Bank

[67] Train, K., 1986, Qualitative Choice Analysis. Theory, Econometrics and and Applicaton to Automobile Demand, The MIT Press, Cambridge.

[68] Welsh, A., 1996, Aspects of Statistical Inference, Wiley and Sons.

[69] Yatchew, A., and Griliches, Z., 1984, Specification Error in Probit Models, Review of Economics and Statistics, pp. 134-139.