Microeconometría I

Maestría en Econometría

Lecture 1

Variables Dependientes Limitadas y Cualitativas

1. Variables Discretas
   - Distintos Tipos de Variables Discretas
   - Clasificación de las Variables Categóricas
2. Modelo de Probabilidad Lineal
   - Características
3. Modelo Logit
   - Características
   - Estimación
   - Bondad del Ajuste
4. Modelo Probit
   - Características
   - Estimación
5. Relación entre Logit y Análisis Discriminante

Agenda

Distintos Tipos de Variables Discretas

- En esta primera parte del curso vamos a estudiar los métodos econométricos que existen para analizar variables dependientes que son cualitativas o asumen valores discretos.
- Ejemplos de estas variables son participación laboral, la demanda de bienes durables e inmuebles, elección entre diferentes marcas del mismo producto, grado de preferencia por determinado bien, número de patentes emitidas, número de empleados en diferentes firmas de un mismo sector, etc.
- La característica de todas estas variables es que asumen valores discretos.
- Estos valores pueden representar diferentes **categorías**, como es el caso de la elección entre diferentes marcas de un mismo producto, ó los valores no representan categorías pero son discretos, como el caso de las patentes emitidas.

Distintos Tipos de Variables Discretas

- Las variables discretas que adoptan valores que representan categorías se denominan variables categóricas.
- La variable categórica más simple es la variable binaria que adopta solamente dos categorías.
- Las variables categóricas con más de dos categorias pueden clasificarse en (a) ordenadas y (b) no ordenadas.

Agenda

Clasificación de las Variables Categóricas

- Un ejemplo de variable categórica no ordenada puede ser:

  $Y = 1$ si la persona viaja a su trabajo en *tren*
  $Y = 2$ si la persona viaja a su trabajo en *colectivo*
  $Y = 3$ si la persona viaja a su trabajo en *taxi*
  $Y = 4$ si la persona viaja a su trabajo en *auto*
  $Y = 5$ si la persona viaja a su trabajo en otro medio de transporte.

- En este caso, la variable categórica es no ordenada porque los valores que adopta la variable representan categorías sin ningún orden pre-establecido.

Clasificación de las Variables Categóricas

- Un ejemplo de variable categórica ordenada puede ser:

  $Y = 1$ si la calificación (rating) de la deuda del país es AAA
  $Y = 2$ si la calificación de la deuda del país es AA
  $Y = 3$ si la calificación de la deuda del país es A
  $Y = 4$ si la calificación de la deuda del país es menor a A

- En este caso, la variable categórica adopta valores que representan categorías con un orden pre-establecido. Nosotros sabemos que un país que tiene una categoría igual a 1 es porque tiene una mejor calificación de su deuda que un país que tiene una categoría igual a 2.

Agenda

Modelo de Probabilidad Lineal

- Nuestro objetivo es explicar el comportamiento de alguna de estas variables utilizando un modelo de regresión lineal.
- Y las primeras preguntas que vamos a tener que responder son:
  - Es posible estimar este modelo usando el método de mínimos cuadrados clásicos?.
  - Si no es posible, cuál es el mejor método de estimación alternativo?.
- Para analizar qué es lo que sucede cuando queremos estimar un modelo de variable dependiente categórica por mínimos cuadrados clásicos vamos a comenzar con la más simple de estas variables.
- El ejemplo más sencillo de variable categórica es el de una variable binaria, es decir, una variable que solo tiene dos categorías y que por convención adopta dos valores que se denotan con 0 y 1.

Modelo de Probabilidad Lineal

- Como ilustración considere la decisión de un trabajador entre trabajar y no trabajar.
- Entre las variables que influyen en esta decisión están: salario ofrecido, cantidad de hijos menores a 6 años en el hogar, si el cónyuge está desocupado, si se recibe algún tipo de ayuda social, etc.
- Las variables explicativas de la decisión de trabajar o no las vamos a agrupar en el vector x.

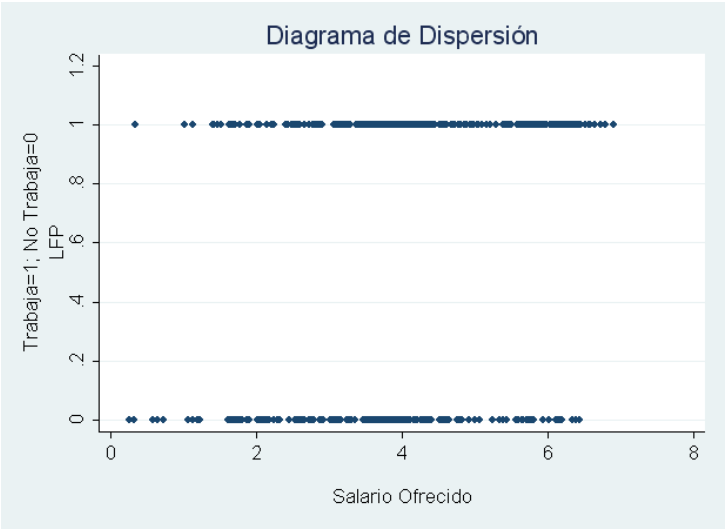
Modelo de Probabilidad Lineal

- Supongamos, para mantener el modelo simple, que x contiene una sola variable explicativa (el salario ofrecido) y definamos la siguiente ecuación de regresión:

$$y_i = \alpha + \beta\, x_i + u_i \tag{1}$$

  donde $y_i = 1$ si el trabajador i decide trabajar e $y_i = 0$ en otro caso.
- La teoría económica diría que a mayor salario ofrecido, el trabajador i debería tener más incentivos a trabajar.
- Como primera aproximación al tema veamos un diagrama de dispersión de las variables.

Modelo de Probabilidad Lineal



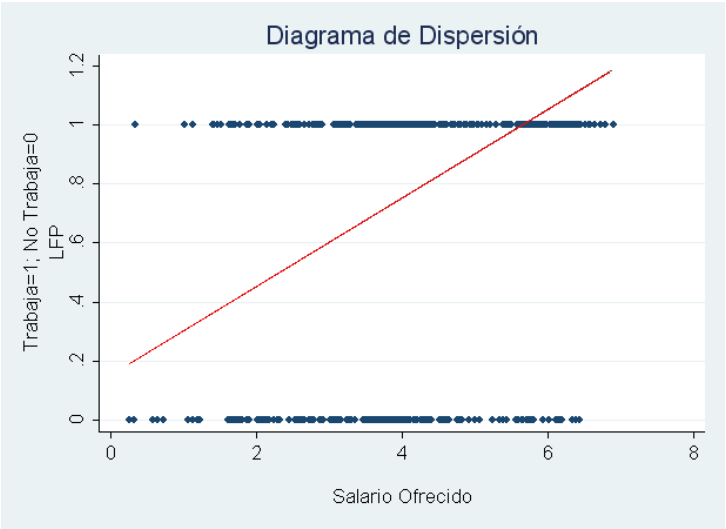
Modelo de Probabilidad Lineal

- La estimación por mínimos cuadrados clásicos del modelo es,

$$\begin{aligned} \hat{y}_i &= 0.1858 + 0.0968\, x_i \\ &\quad (0.063) \quad (0.015) \end{aligned}$$

  donde los números entre paréntesis son los desvíos estándar.
- Es decir, existe una relación directa entre el salario ofrecido y la decisión de trabajar. A medida que aumenta el salario ofrecido hay más incentivos a trabajar. Gráficamente:

Modelo de Probabilidad Lineal



Modelo de Probabilidad Lineal

- Como puede observarse en el gráfico anterior, un problema con esta estimación es que hay muchos valores ajustados para la variable dependiente que son MAYORES A UNO!
- Esta es una consecuencia del método de estimación elegido, ya que MCC no contiene ninguna restricción que nos diga que la estimación de la variable dependiente siempre tiene que ser cero o uno.
- El problema anterior puede plantearse de la siguiente manera. Cuál es la probabilidad de que una persona trabaje, dado el valor del salario ofrecido?
- Evidentemente, es poco probable que alguien pueda responder a esa pregunta. Sin embargo, podemos observar el valor de esta probabilidad ex-post.

Modelo de Probabilidad Lineal

- Esto es lo que observamos en la muestra. Es decir, un individuo, ex-post o trabaja (probabilidad de realización igual a uno) o no trabaja (probabilidad de realización igual a cero).
- Esto significa que lo que observamos es la realización de una variable latente (la probabilidad de trabajar).
- En realidad la pregunta relevante en este tipo de análisis es: cuáles son las características que afectan la probabilidad de trabajar?
- Es decir que a nosotros nos interesa el valor de la probabilidad ex-ante. Esto es:

$$Pr(y_i = 1 | Salario\ Ofrecido)$$

Modelo de Probabilidad Lineal

- Note que de acuerdo al modelo especificado anteriormente tenemos:

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

- Por lo tanto,

$$\begin{aligned} E(y_i|x_i) &= \alpha + \beta\, x_i + E(u_i|x_i) \\ &= \alpha + \beta\, x_i \end{aligned} \tag{2}$$

Modelo de Probabilidad Lineal

- Además, por definición de esperanza matemática, la esperanza matemática condicional de una variable es la suma de cada uno de los valores que adopta la variable multiplicados por su probabilidad de ocurrencia.
- En este caso:

$$\begin{aligned} E(y_i|x_i) &= 1 \times Pr(y_i = 1|x_i) + 0 \times Pr(y_i = 0|x_i) \\ &= Pr(y_i = 1|x_i) \end{aligned} \tag{3}$$

Modelo de Probabilidad Lineal

- Igualando las ecuaciones (2) y (3) se tiene,

$$Pr(y_i = 1|x_i) = \alpha + \beta\ x_i \tag{4}$$

- Como puede observarse en la ecuación (4) la probabilidad condicional de que el evento y_i ocurra (en este caso la decisión de trabajar) dado que conocemos el valor de la variable independiente, está expresada como una relación lineal en los parámetros del modelo.
- Debido a este hecho, los modelos de variable dependiente binaria reciben el nombre de Modelos de Probabilidad Lineal (MPL).

Modelo de Probabilidad Lineal

- Dado que y_i solo puede adoptar dos valores (cero ó uno) podemos obtener su distribución de probabilidad con la siguiente tabla:

y_i	$Pr(y_i \vert \cdot)$
0	$1 - \alpha - \beta\ x_i$
1	$\alpha + \beta\ x_i$

- Dada la distribución de la variable dependiente podemos utilizar el modelo para obtener la distribución de los errores. Esto es, sabemos que cuando $y_i = 0$ entonces $u_i = -\alpha - \beta\ x_i$ y cuando $y_i = 1$ entonces $u_i = 1 - \alpha - \beta\ x_i$. Por lo tanto,

Modelo de Probabilidad Lineal

u_i	$Pr(u_i \vert \cdot)$
$1 - \alpha - \beta\ x_i$	$\alpha + \beta\ x_i$
$-\alpha - \beta\ x_i$	$1 - \alpha - \beta\ x_i$

- Calculemos la esperanza matemática y la varianza de los errores del modelo.
-

$$\begin{aligned} E(u_i|\cdot) &= [1 - \alpha - \beta\ x_i] \times [\alpha + \beta\ x_i] \\ &+ [-\alpha - \beta\ x_i] \times [1 - \alpha - \beta\ x_i] \\ &= 0 \end{aligned}$$

Modelo de Probabilidad Lineal

- $$\begin{aligned} Var(u_i|\cdot) &= E(u_i^2|\cdot) - E(u_i|\cdot)^2 \\ &= E(u_i^2|\cdot) \\ &= [1 - \alpha - \beta\ x_i] \times [\alpha + \beta\ x_i] \end{aligned}$$

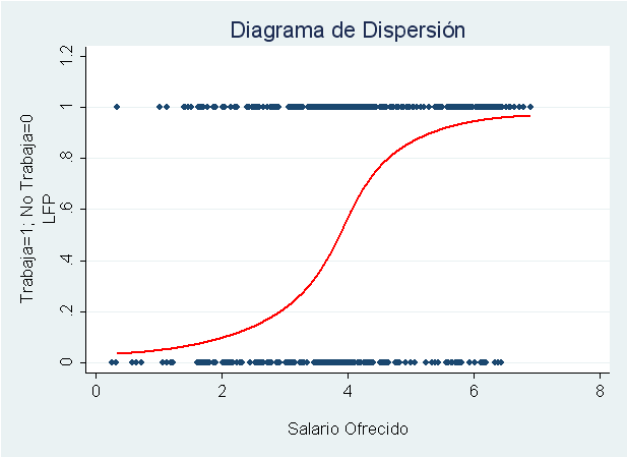
- De la ecuación anterior se desprende que, cuando la variable dependiente es binaria, los errores del modelo tienen heterocedasticidad.
- Esta característica puede generalizarse a cualquier modelo que tenga por variable dependiente una variable categórica.

Modelo de Probabilidad Lineal

- Ahora podemos resumir las características de los MPL:
  - Estos modelos reciben este nombre porque la variable dependiente puede interpretarse como una probabilidad.
  - Los valores estimados, por el método de MCC, de la variable dependiente pueden caer fuera del rango [0, 1].
  - Los errores del modelo son heterocedásticos por lo tanto MCC nos dará estimadores ineficientes.
- Otro de los problemas que sufren los MPL es el de la interpretación de los coeficientes estimados.
- En nuestro caso particular el coeficiente β mide cuánto afecta a la probabilidad de trabajar un cambio en el salario ofrecido.
- Económicamente uno pensaría que este efecto debiera ser chico para salarios pequeños (y grandes) y mayor para salarios intermedios. Es decir, el coeficiente β no debiera ser constante.

Modelo de Probabilidad Lineal

- La forma de resolver estos problemas es utilizando una función de probabilidad no lineal en los parámetros. Gráficamente,



Modelo de Probabilidad Lineal

- Las características de la curva de la figura anterior resuelve nuestros problemas ya que:
  - Empieza en cero y termina en uno. Esto es, solo adopta valores en el intervalo $[0, 1]$.
  - Tiene diferentes pendientes en distintos puntos. Para valores muy pequeños y muy grandes de la variable independiente la pendiente es chica y para valores intermedios la pendiente es más grande.
- Cualquier curva de probabilidad acumulada cumple con las características antes mencionadas. Por lo tanto, uno puede especificar el modelo utilizando la función de probabilidad acumulada de cualquier distribución.
- Las funciones más utilizadas son la distribución Normal y la distribución Logística.

Agenda

Modelo Logit

- El modelo Logit asume que la curva del gráfico anterior puede ser aproximada por la distribución logística.
- En términos matemáticos, una variable z se dice que tiene distribución logística cuando su función de distribución de probabilidad viene dada por:

$$F(z) = \frac{e^z}{1 + e^z}$$

- En economía, la forma tradicional de analizar este tipo de modelos es utilizando el concepto de variable latente.
- Supongamos que los individuos toman su decisión de trabajar o no sobre la base de la utilidad que les da el trabajo.

Modelo Logit

- Asumamos utilidades estocásticas como funciones lineales del salario ofrecido. Entonces:

$$U_{T,i} = \alpha_T + \beta_T \; x_i + u_{T,i}$$

es la utilidad para el individuo i de trabajar. Y

$$U_{D,i} = \alpha_D + \beta_D \; x_i + u_{D,i}$$

es la utilidad para el individuo i de no trabajar.

- Entonces, el individuo i trabajará si $U_{T,i} > U_{D,i}$.

Modelo Logit

- Por lo tanto, la probabilidad de trabajar para el individuo i puede expresarse como,

$$\begin{aligned} Pr[y_i = 1|\cdot] &= Pr[U_{T,i} > U_{D,i}|\cdot] \\ &= Pr[(\alpha_T - \alpha_D) + (\beta_T - \beta_D)\, x_i > u_{D,i} - u_{T,i}|\cdot] \\ &= Pr[u_i \leq \alpha + \beta\, x_i|\cdot] \end{aligned}$$

  donde $u_i = u_{D,i} - u_{T,i}$, $\alpha = \alpha_T - \alpha_D$ y $\beta = \beta_T - \beta_D$.
- Ahora tenemos expresado el modelo en términos de una probabilidad acumulada hasta $\alpha + \beta\, x_i$ para la variable aleatoria u_i.

Modelo Logit

- Entonces, si u_i tiene distribución logística tenemos:

$$F(\alpha + \beta x_i) = \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}}$$

- A diferencia de lo que ocurría con el MPL en este caso existe una relación no lineal en los parámetros del modelo y por lo tanto no puede usarse el método de mínimos cuadrados para la estimación.
- Una consecuencia del último punto es que los efectos de las variables explicativas sobre la variable dependiente no son lineales.

Modelo Logit: Interpretación de los Coeficientes

- En la función logística βx_i está en el exponente del número e en el numerador y en el denominador y por lo tanto no es inmediatamente claro cuál es el efecto sobre la probabilidad de trabajar de un cambio en el salario ofrecido.
- Este efecto no lineal puede verse calculando la derivada parcial de la probabilidad de trabajar con respecto al salario ofrecido,

$$\begin{aligned} \frac{\partial Pr[y_i = 1|\cdot]}{\partial x_i} &= F(\alpha + \beta x_i) \times [1 - F(\alpha + \beta x_i)] \times \beta \\ &= f(\alpha + \beta x_i) \times \beta \end{aligned}$$

Modelo Logit: Interpretación de los Coeficientes

- Esto muestra que el efecto, sobre la probabilidad de trabajar, de un cambio en el salario ofrecido depende no solo del valor de β sino también del valor tomado por la función de densidad de la logística.
- Recuerde que en el caso del MPL esta derivada era constante e igual a β. Esto nos permitía interpretar los coeficientes del modelo como el cambio marginal que se produce en la variable dependiente cuando cambia una de las variables independientes, manteniendo constante el resto de las variables.
- Ahora, la derivada refleja las diferentes pendientes de la curva logística. Esto significa que hay un valor para el cambio marginal de la probabilidad de trabajar para cada valor del salario ofrecido.

Modelo Logit: Interpretación de los Coeficientes

- Una forma de interpretar los coeficientes del modelo Logit es calculando los efectos marginales promedio sobre la muestra de n observaciones. Hay dos formas de hacer esto,
  - evaluando la función de densidad logística en la media de las variables,

$$\overline{\frac{\partial Pr[y_i = 1|\cdot]}{\partial x_i}} = f(\alpha + \beta \bar{x}) \times \beta$$

  - calculando el promedio de la función de densidad logística,

$$\overline{\frac{\partial Pr[y_i = 1|\cdot]}{\partial x_i}} = \frac{1}{n} \sum_{i=1}^{n} f(\alpha + \beta x_i) \times \beta$$

Modelo Logit: Interpretación de los Coeficientes

- Una segunda forma de interpretar los coeficientes del modelo Logit es a través del cociente de probabilidades ("odds ratio").
- Las chances ("tasa de probabilidad") de que un evento suceda se calculan como el cociente entre la probabilidad de que un evento suceda y la probabilidad de que no suceda. Por ejemplo, en el modelo Logit, las chances de trabajar ($y_i = 1$) son,

$$\frac{Pr(y_i = 1|x_i)}{Pr(y_i = 0|x_i)} = \frac{F(\alpha + \beta\, x_i)}{1 - F(\alpha + \beta\, x_i)} = e^{\alpha + \beta\, x_i}$$

- Fijando el salario ofrecido (x_i) en algún valor, se obtienen las chances de trabajar a ese salario ofrecido.

Modelo Logit: Interpretación de los Coeficientes

- Cuando x_i es una variable continua, como en nuestro caso, el cociente de probabilidades (CP) mide el cambio en las chances de trabajar de aumentar en una unidad el salario ofrecido:

$$CP = \frac{Pr(y_i = 1|x_i = x + 1)/Pr(y_i = 0|x_i = x + 1)}{Pr(y_i = 1|x_i = x)/Pr(y_i = 0|x_i = x)} = e^{\beta}$$

- El cociente de probabilidades nos dice que ante un aumento de un peso en el salario ofrecido esperamos ver alrededor de $(e^{\beta} - 1)\%$ de cambio en las chances de trabajar. Este valor del cambio porcentual no depende del valor que adopte el salario ofrecido.

Modelo Logit: Interpretación de los Coeficientes

- El cociente de probabilidades también se utiliza para calcular las chances de que el evento analizado suceda comparando dos grupos.
- Supongamos que $x_{i,1} = 1$ denota que el individuo i es de género femenino y $x_{i,1} = 0$ denota que i es de género masculino. Entonces, el cociente de probabilidades se define como,

$$CP = \frac{Pr(y_i = 1|x_{i,1} = 1)/Pr(y_i = 0|x_{i,1} = 1)}{Pr(y_i = 1|x_{i,1} = 0)/Pr(y_i = 0|x_{i,1} = 0)} = e^{\beta_1}$$

- El cociente de probabilidades nos dice cuantas más (menos) chances hay de que una mujer trabaje comparada con un hombre (con las mismas características).

Agenda

Modelo Logit: Estimación

- Para realizar la estimación del modelo debemos recurrir al método de máxima verosimilitud.
- Como sabemos la función de probabilidad de los errores y tenemos una muestra aleatoria (es decir, compuesta por variables aleatorias independientes) la función de verosimilitud es simplemente la multiplicación de las funciones de probabilidad para todas las observaciones que hay en la muestra.
- Entonces, en términos matemáticos la función de verosimilitud es,

$$L(\alpha, \beta; x_i) = \prod_{i=1}^{n} \left[\frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right]^{y_i} \times \left[1 - \frac{e^{\alpha+\beta x_i}}{1 + e^{\alpha+\beta x_i}} \right]^{1-y_i}$$

Modelo Logit: Estimación

- El logaritmo natural de la función de verosimilitud es,

$$l(\alpha, \beta; x_i) = \sum_{i=1}^{n} \left[y_i ln\{\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\} + (1-y_i) ln\{1 - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\} \right]$$

- Las condiciones de primer orden para la maximización de esta función son,

$$\begin{aligned} \frac{\partial l(\cdot)}{\partial \hat{\alpha}} &= \sum_{i=1}^{n} \left[y_i - \frac{e^{\hat{\alpha}+\hat{\beta} x_i}}{1+e^{\hat{\alpha}+\hat{\beta} x_i}} \right] = 0 \\ \frac{\partial l(\cdot)}{\partial \hat{\beta}} &= \sum_{i=1}^{n} \left[y_i - \frac{e^{\hat{\alpha}+\hat{\beta} x_i}}{1+e^{\hat{\alpha}+\hat{\beta} x_i}} \right] x_i = 0 \end{aligned}$$

Modelo Logit: Estimación

- Como se puede observar en las condiciones de primer grado las incógnitas de ambas ecuaciones ($\hat{\alpha}$, $\hat{\beta}$) entran en forma no lineal y por lo tanto no pueden resolverse por métodos lineales.
- Amemiya (1985) demostró que la función de verosimilitud del modelo Logit es globalmente cóncava por lo que las condiciones de segundo orden para un máximo se cumplen.
- Las condiciones de segundo orden vienen dadas por las siguientes expresiones,

$$\frac{\partial^2 l(\cdot)}{\partial \hat{\alpha}^2} = -\sum_{i=1}^{n} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \right) \times \left(1 - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \right)$$

Modelo Logit: Estimación

- $$\frac{\partial^2 l(\cdot)}{\partial \hat{\alpha} \partial \hat{\beta}} = -\sum_{i=1}^{n} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \right) \times \left(1 - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \right) x_i$$
$$\frac{\partial^2 l(\cdot)}{\partial \hat{\beta}^2} = -\sum_{i=1}^{n} \left(\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \right) \times \left(1 - \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}} \right) x_i^2$$

- Llamando P_i a $\frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$ para abreviar notación, estas condiciones de segundo orden pueden agruparse en la matriz Hesiana (la matriz de las segundas derivadas).

$$H(\hat{\alpha}, \hat{\beta}) = -\begin{bmatrix} \sum_{i=1}^{n} P_i(1-P_i) & \sum_{i=1}^{n} P_i(1-P_i)x_i \\ \sum_{i=1}^{n} P_i(1-P_i)x_i & \sum_{i=1}^{n} P_i(1-P_i)x_i^2 \end{bmatrix}$$

Modelo Logit: Estimación

- Los estimadores de máxima verosimilitud del modelo Logit son insesgados, consistentes y eficientes.
- La distribución asintótica es Normal,

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} \sim \text{Normal}\left[\begin{pmatrix} \alpha \\ \beta \end{pmatrix},\ [E\{-H(\hat{\alpha}, \hat{\beta})\}]^{-1}\right]$$

- La matriz de varianzas y covarianzas de los estimadores del modelo está dada por la inversa de la matriz Hesiana anterior con el signo opuesto.
- En la diagonal principal de esa matriz tenemos las varianzas de los coeficientes mientras que fuera de la diagonal principal tenemos las covarianzas entre los coeficientes.
- Volvamos a la modelización de la probabilidad de trabajar y ahora estimemos el modelo utilizando el Stata.

Agenda

Modelo Logit: Evaluación

- Una vez estimado el modelo, es importante (como sucedía en el caso del modelo de regresión lineal) chequear el modelo estimado utilizando medidas de bondad del ajuste y los contrastes de hipótesis habituales.
- Sin embargo, para los modelos de variable dependiente categórica no se puede calcular la medida de bondad del ajuste conocida como R^2.
- Una medida alternativa de bondad del ajuste se conoce como R^2 de McFadden y se define como:

$$R^2_{MF} = 1 - \frac{l(\hat{\alpha}, \hat{\beta})}{l(\hat{\alpha}_0)}$$

Modelo Logit: Evaluación

- Donde, $l(\hat{\alpha}, \hat{\beta})$ es el valor del logaritmo de la función de verosimilitud evaluada en los estimadores de MV y $l(\hat{\alpha}_0)$ es el valor del logaritmo de la función de verosimilitud de un modelo que tiene solo una constante.
- El R^2_{MF} tiene la misma interpretación que el R^2 común. Es decir, nos dice que porcentaje de la variabilidad de la variable dependiente está explicado por la regresión.
- El R^2_{MF} tiene la misma característica que el R^2 común: no sirve para comparar el ajuste de diferentes modelos con distinto número de variables explicativas.
- La medidad de bondad del ajuste que se utiliza para comparar ajustes es el R^2_{MF} ajustado.

Modelo Logit: Evaluación

- El R^2_{MF} ajustado se define como:

$$\bar{R}^2_{MF} = 1 - \frac{l(\hat{\alpha}, \hat{\beta}) - k}{l(\hat{\alpha}_0)}$$

  donde k es el número de parámetros a estimar.
- Al igual que el R^2 ajustado en la regresión lineal, el $\bar{R}^2_{MF}$ puede ser negativo.
- Otra forma de medir la bondad del ajuste es observar como clasifica a las observaciones el modelo en comparación con los datos realmente observados.
- En nuestro ejemplo, sería preguntarse cuán bien clasifica el modelo estimado a las personas que trabajan?

Modelo Logit: Evaluación

- Para responder a esta pregunta necesitamos saber cuándo nuestro modelo predice que $y_i = 1$ (es decir, cuándo $\hat{y}_i = 1$).
- Como nosotros tenemos una estimación de la probabilidad de trabajar una regla de clasificación sencilla podría ser: Siempre que la estimación de la probabilidad de trabajar sea mayor a un valor c (por ejemplo 0.5) entonces clasifique a esa persona como trabajando. En términos matemáticos:

$$Si\ F(\hat{\alpha} + \hat{\beta}\, x_i) = \frac{e^{\hat{\alpha}+\hat{\beta}\, x_i}}{1 - e^{\hat{\alpha}+\hat{\beta}\, x_i}} > c \Longrightarrow \hat{y}_i = 1$$

Modelo Logit: Evaluación

- En la práctica uno puede elegir cualquier valor para c. Una posibilidad es emplear la probabilidad empírica:

$$c = \frac{\#y_i = 1}{n}$$

- Una vez que clasificadas las observaciones se puede construir lo que se denomina la tabla de predicción-realización:

	Predicción	
Realización	$\hat{y}_i = 0$	$\hat{y}_i = 1$
$y_i = 0$	p_{00}	p_{10}
$y_i = 1$	p_{01}	p_{11}

Modelo Logit: Evaluación

- La fracción $p_{00} + p_{11}$ se denomina tasa de aciertos. Cuanto mayor es la tasa de aciertos mejor es el ajuste del modelo.
- La tasa de aciertos se define como la fracción de predicciones correctas en la muestra. Formalmente, si definimos a la variable aleatoria w_i como la indicadora de una predicción correcta (esto es $w_i = 1$ si $y_i = \hat{y}_i$ y $w_i = 0$ si $y_i \neq \hat{y}_i$) entonces la tasa de aciertos se define como $h = \frac{1}{n} \sum_{i=1}^{n} w_i$
- En la población, la proporción de unos es p. Supongamos que nosotros hacemos una predicción completamente aleatoria. Esto es, predecimos un 1 con probabilidad p y predecimos un 0 con probabilidad $(1 - p)$. En este escenario, la probabilidad de hacer una predicción correcta es $q = p^2 + (1 - p)^2$.

Modelo Logit: Evaluación

- Usando las propiedades de la distribución binomial para el número de predicciones (hechas en forma aleatoria) correctas, la tasa de aciertos "aleatoria" (h_a) tiene $E(h_a) = q$ y $Var(h_a) = q(1-q)/n$. Entonces la habilidad predictiva del modelo estimado puede evaluarse comparándola con esta tasa de aciertos "aleatoria".
- La idea es hacer un test de hipótesis cuya hipótesis nula sea que las predicciones del modelo estimado no son mejores que las predicciones hechas en forma aleatoria y cuya hipótesis alternativa sea que las predicciones del modelo son mejores que las hechas aleatoriamente.

Modelo Logit: Evaluación

- Bajo la hipótesis nula, la tasa de aciertos, h se distribuye normalmente con media q y varianza $q(1-q)/n$. Por lo tanto el estadístico de contraste es:

$$z = \frac{h-q}{\sqrt{q(1-q)/n}} = \frac{nh-nq}{\sqrt{nq(1-q)}}$$

  En la práctica $q = p^2 + (1-p)^2$ se estima con $\hat{p}^2 + (1-\hat{p})^2$, con $\hat{p}$ siendo la proporción de unos en la muestra.
- La regla de decisión es rechazar la hipótesis nula siempre que el valor de probabilidad de z sea menor al nivel de significación del test ó siempre que z sea mayor al valor crítico de la normal estándar.

Modelo Logit: Evaluación

- Otros índices de interés son los denominados sensitividad y especificidad
- La sensitividad es la probabilidad de predecir un "éxito" entre los "éxitos": $Pr(\hat{y}_i = 1 | y_i = 1)$
- La especificidad es la probabilidad de predecir un "fracaso" entre los "fracasos": $Pr(\hat{y}_i = 0 | y_i = 0)$
- La probabilidad de predecir un "éxito falso" ó "falso positivo" es uno menos la especificidad:
  $Pr(\hat{y}_i = 1 | y_i = 0) = 1 - Pr(\hat{y}_i = 0 | y_i = 0)$
- Los falsos positivos corresponden a lo que llamamos error de tipo I.
- Es claro que una mejor bondad de ajuste se obtiene con una alta sensitividad y especificidad.

Modelo Logit: Evaluación

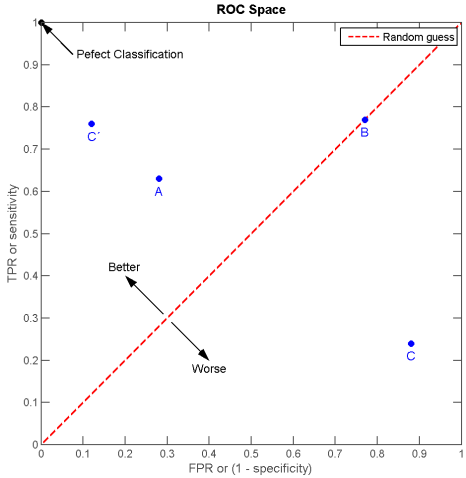
- En una tabla de contingencia (similar a la de predicción-realización) se pueden ver estas definiciones:

Realización	Predicción	
	$\hat{y}_i = 0$	$\hat{y}_i = 1$
$y_i = 0$	$Pr(\hat{y}_i = 0 \mid y_i = 0)$ Verdadero Negativo	$Pr(\hat{y}_i = 1 \mid y_i = 0)$ Falso Positivo
$y_i = 1$	$Pr(\hat{y}_i = 0 \mid y_i = 1)$ Falso Negativo	$Pr(\hat{y}_i = 1 \mid y_i = 1)$ Verdadero Positivo

- Una forma de resumir la bondad del ajuste con estas dos medidas es graficando la curva ROC (Relative (Reciever) Operating Characteristic).
- La curva ROC es una representación gráfica de la sensibilidad frente a (1 − especificidad) para un sistema clasificador binario según se varía el umbral de clasificación.

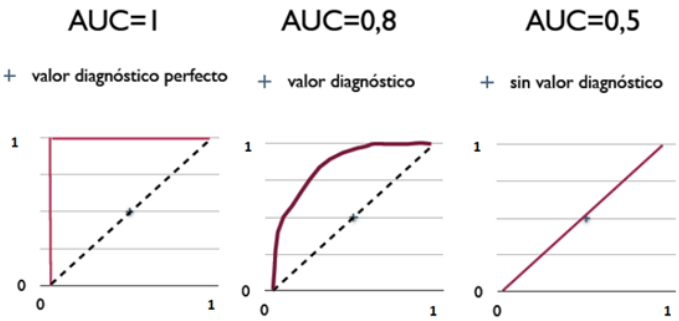
Modelo Logit: Evaluación

- Gráficamente:



Modelo Logit: Evaluación

- Una forma de resumir la curva ROC es calcular el área bajo la curva (AUC)
- Diferentes curvas ROC y sus AUC



Agenda

Modelo Probit

- Ahora asumimos que los errores del modelo en lugar de describirse con la función de distribución logística, pueden describirse con la función de distribución Normal estándar.
- Cuando representamos los errores con la distribución Normal, el modelo resultante recibe el nombre de Modelo Probit.
- Una variable z se dice que tiene distribución normal cuando su función de distribución de probabilidad acumulada tiene la siguiente forma:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt$$

Modelo Probit

- En términos de nuestro modelo,

$$\Phi(\hat{\alpha} + \hat{\beta}\, x_i) = \int_{-\infty}^{\hat{\alpha}+\hat{\beta}\, x_i} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\, dt$$

- Al igual de lo que ocurría con el modelo Logit en este caso existe una relación no lineal en los parámetros del modelo y por lo tanto, el método de mínimos cuadrados no puede usarse para estimar el modelo Probit.
- Esta relación no lineal puede ser vista calculando la derivada parcial de la probabilidad de trabajar con respecto al salatio ofrecido. Por ejemplo,

$$\frac{\partial Pr[y_i = 1|\cdot]}{\partial x_i} = \phi(\hat{\alpha} + \hat{\beta}\, x_i) \times \beta$$

Modelo Probit

- La ecuación anterior muestra que el efecto de un cambio en el salario ofrecido sobre la probabilidad de trabajar depende no solo del valor de β sino también del valor tomado por la función de densidad de la Normal estándar (ϕ).
- La derivada parcial refleja las diferentes pendientes de la curva de distribución acumulada de la normal estándar. Esto significa que hay un valor para el cambio marginal de la probabilidad de trabajar para cada valor del salario ofrecido.
- La forma de interpretar los coeficientes del modelo Probit es la misma que en el modelo Logit. Hay que calcular los efectos marginales promedio sobre la muestra de n observaciones con cualquiera de las dos formas descriptas anteriormente.

Agenda

Modelo Probit: Estimación

- Para realizar la estimación del modelo debemos recurrir al método de máxima verosimilitud.
- Como sabemos la función de probabilidad de los errores y sabemos que tenemos una muestra aleatoria (es decir, compuesta por variables aleatorias independientes) la función de verosimilitud es simplemente la multiplicación de las funciones de probabilidad para todas las observaciones que hay en la muestra.
- En términos matemáticos la función de verosimilitud del modelo Probit es,

$$L(\hat{\alpha}, \hat{\beta}; x_i) = \prod_{i=1}^{n} \left[\Phi(\hat{\alpha} + \hat{\beta}\, x_i)\right]^{y_i} \times \left[1 - \Phi(\hat{\alpha} + \hat{\beta}\, x_i)\right]^{1-y_i}$$

Modelo Probit: Estimación

- El logaritmo natural de la función de verosimilitud es,

$$l(\hat{\alpha}, \hat{\beta}; x_i) = \sum_{i=1}^{n} \left[y_i ln\{\Phi(\hat{\alpha} + \hat{\beta}\, x_i)\} + (1 - y_i) ln\{1 - \Phi(\hat{\alpha} + \hat{\beta}\, x_i)\} \right]$$

- Las condiciones de primer orden para la maximización de esta función son,

$$\begin{aligned} \frac{\partial l(\cdot)}{\partial \hat{\alpha}} &= \sum_{i=1}^{n} \left[\frac{y_i - \Phi(\hat{\alpha} + \hat{\beta}\, x_i)}{\Phi(\hat{\alpha} + \hat{\beta}\, x_i)[1 - \Phi(\hat{\alpha} + \hat{\beta}\, x_i)]} \phi(\hat{\alpha} + \hat{\beta}\, x_i) \right] = 0 \\ \frac{\partial l(\cdot)}{\partial \hat{\beta}} &= \sum_{i=1}^{n} \left[\frac{y_i - \Phi(\hat{\alpha} + \hat{\beta}\, x_i)}{\Phi(\hat{\alpha} + \hat{\beta}\, x_i)[1 - \Phi(\hat{\alpha} + \hat{\beta}\, x_i)]} \phi(\hat{\alpha} + \hat{\beta}\, x_i) \right] x_i = 0 \end{aligned}$$

Modelo Probit: Estimación

- Como se puede observar en las condiciones de primer orden, las incógnitas de ambas ecuaciones ($\hat{\alpha}$, $\hat{\beta}$) entran en forma no lineal y por lo tanto no pueden resolverse por métodos lineales.
- Amemiya (1985) demostró que la función de verosimilitud del modelo Probit es globalmente cóncava por lo que las condiciones de segundo orden para un máximo se cumplen.
- Los estimadores de máxima verosimilitud del modelo Probit son insesgados, consistentes y eficientes.
- La matriz de varianzas y covarianzas de los estimadores del modelo está dada por la inversa de la matriz de las condiciones de segundo orden con el signo opuesto.
- En la diagonal principal de esta matriz tenemos las varianzas de los coeficientes mientras que fuera de la diagonal principal tenemos las covarianzas entre los coeficientes.

Modelo Probit: Estimación

- Una vez que los parámetros del modelo han sido estimados, pueden realizarse los contrastes de hipótesis habituales.
- También pueden utilizarse todas las medidas de bondad del ajuste mencionadas para el caso del modelo Logit.
- Como existe una relación entre la distribución Normal y la distribución Logística, también existe una relación entre los coeficientes estimados del Probit y del Logit.
- Amemiya (1981) sugiere la siguiente relación entre las estimaciones del modelo Probit y Logit:

$$\hat{\beta}_{Probit} = 0.625\hat{\beta}_{Logit}$$

Análisis Discriminante

- Al igual que el método para variables categóricas, el análisis discriminante tiene el objetivo de asignar nuevos objetos (observaciones) a grupos previamente definidos.
- Para fijar ideas, supongamos que estamos interesados en asignar un nuevo objeto a una de dos clases. Vamos a llamar a estas clases π_1 y π_2.
- Los objetos son clasificados o separados sobre la base de observar, por ejemplo, p variables $X_1, X_2, \ldots, X_p$.
- Los valores de las X's difieren en alguna medida entre las dos clases y por lo tanto uno podría pensar en diferenciar a las dos clases en función de la probabilidad de pertenecer a cada población.

Análisis Discriminante

- Esto es, si $f_1(X)$ y $f_2(X)$ representan a las probabilidades de pertenecer a las clases π_1 y π_2, respectivamente, uno puede hablar de clasificar a los datos como provenientes de dos poblaciones diferentes, utilizando estas funciones de probabilidad.
- Específicamente, el conjunto de datos se divide en dos regiones R_1 y R_2, tal que si una nueva observación cae en R_1 se asigna a la población π_1 y si cae en R_2 se asigna a π_2.
- Sin embargo, las reglas de clasificación no están exentas de error. Esto puede deberse a que no hay una clara distinción entre las características medidas de ambas poblaciones y por lo tanto los grupos pueden superponerse.
- Es posible, entonces, que uno pueda clasificar incorrectamente un objeto de la población π_2 en π_1 y viceversa.

Análisis Discriminante

- Un buen procedimiento de clasificación debería resultar en pocas equivocaciones. En otras palabras, la probabilidad de clasificar incorrectamente un objeto debe ser pequeña.
- El análisis discriminante utiliza la misma regla de clasificación que la mencionada arriba para el modelo logit. Si X_0 es una nueva observación y si $f_1(X_0)/f_2(X_0) > 1$ debemos asignar X_0 a π_1. Por otro lado, si $f_1(X_0)/f_2(X_0) < 1$ debemos asignar X_0 a π_2.
- Asumamos que conocemos que $f_1(X)$ y $f_2(X)$ son funciones de densidad normales multivariantes, la primera con vector de medias μ_1 y matriz de varianzas y covarianzas Σ_1, y la segunda con vector de medias μ_2 y matriz de varianzas y covarianzas Σ_2.

Análisis Discriminante

- Además, supongamos que $\Sigma_1 = \Sigma_2 = \Sigma$. Entonces, las funciones de probabilidad conjunta de $X' = [X_1, X_2, ..., X_p]$ para las poblaciones $\pi_1 y \pi_2$, vienen dadas por:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{[-\frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i)]}, \quad i = 1, 2.$$

- Utilizando la regla de clasificación anterior, asignamos una nueva observación a la población π_1 si

$$e^{[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)]+[\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)]} > 1, \text{ ó}$$

$$[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)] + [\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)] > 0$$

Análisis Discriminante

- Reordenando los términos de la última expresión,

$$[-\frac{1}{2}(x-\mu_1)'\Sigma^{-1}(x-\mu_1)] + [\frac{1}{2}(x-\mu_2)'\Sigma^{-1}(x-\mu_2)] =$$

$$(\mu_1-\mu_2)'\Sigma^{-1}x - \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1+\mu_2) > 0 \qquad (5)$$

- En la práctica, los parámetros poblacionales μ_1, μ_2 y Σ son desconocidos y deben ser reemplazados por sus estimaciones muestrales.
- Supongamos que tenemos n_1 observaciones muestrales de las variables $X' = [X_1, X_2, \ldots, X_p]$ provenientes de π_1 y n_2 observaciones de las mismas variables provenientes de π_2.
- Entonces los respectivos datos de ambas poblaciones son,

Análisis Discriminante

- $$X_1 = \begin{bmatrix} x_{11}' \\ x_{12}' \\ \vdots \\ x_{1n_1}' \end{bmatrix} \quad X_2 = \begin{bmatrix} x_{21}' \\ x_{22}' \\ \vdots \\ x_{2n_2}' \end{bmatrix}$$

  con estos datos muestrales, se pueden calcular los vectores de medias y las matrices de varianzas y covarianzas.

- $$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1i}, \quad S_1 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)(x_{1i} - \bar{x}_1)'$$

  $$\bar{x}_2 = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2i}, \quad S_2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n_2}(x_{2i} - \bar{x}_2)(x_{2i} - \bar{x}_2)'$$

Análisis Discriminante

- Las matrices de varianzas y covarianzas muestrales S_1 y S_2 deben combinarse para obtener una estimación de la varianza conjunta. En particular, el promedio ponderado es un estimador insesgado de Σ si las muestras son aleatorias.

$$S = \frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)} S_1 + \frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)} S_2$$

- Sustituyendo en (1), μ_1 por $\bar{x}_1$, μ_2 por $\bar{x}_2$ y Σ por S obtenemos la regla de clasificación muestral del análisis discriminante para dos poblaciones:
Asigne X_0 a π_1 si

$$(\bar{x}_1 - \bar{x}_2)' S^{-1} X_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 + \bar{x}_2) > 0$$

Análisis Discriminante

- Note que la regla de clasificación implica comparar dos números,

$$\hat{y} = (\bar{x}_1 - \bar{x}_2)' S^{-1} X_0 = \hat{\alpha}' X_0,$$

- con

$$\begin{aligned} \hat{m} &= \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 + \bar{x}_2) \\ &= \frac{1}{2}(\bar{\hat{y}}_1 + \bar{\hat{y}}_2) \end{aligned}$$

- donde

$$\bar{\hat{y}}_1 = (\bar{x}_1 - \bar{x}_2)' S^{-1} \bar{x}_1 = \hat{\alpha}' \bar{x}_1,$$

$$\bar{\hat{y}}_2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} \bar{x}_2 = \hat{\alpha}' \bar{x}_2.$$

Análisis Discriminante

- Entonces, la regla de asignación del análisis discriminante para dos poblaciones normales consiste en crear dos grupos utilizando los valores de $\hat{y}$, calculado a través de una combinación lineal apropiada de las observaciones de las muestras de las poblaciones π_1 y π_2, y luego asignar una nueva observación, X_0, a π_1 ó π_2, dependiendo de si $\hat{y} = \hat{\alpha}' X_0$ cae a la derecha o a la izquierda del punto medio entre las medias de $\hat{y}$ en los dos grupos, $\hat{m} = \frac{1}{2}(\bar{\hat{y}}_1 + \bar{\hat{y}}_2)$.
- Los coeficientes del vector $\hat{\alpha} = S^{-1}(\bar{x}_1 - \bar{x}_2)$ se denominan coeficientes discriminantes.
- Estos coeficientes no son únicos, ya que cualquier múltiplo de los mismos también sirve para discriminar. Esto es, para cualquier $c \neq 0$, el vector $c\hat{\alpha}$ discrimina entre dos poblaciones de la misma manera que lo hace $\hat{\alpha}$.

Análisis Discriminante

- El vector $\hat{\alpha}$ usualmente se "normaliza" para facilitar su interpretación. Las dos formas más comunes de normalización son las siguientes:
  - Defina
  $$\hat{\alpha}^* = c\hat{\alpha} = \frac{\hat{\alpha}}{\sqrt{\hat{\alpha}'\hat{\alpha}}}$$
  de forma tal que $\hat{\alpha}^*$ tenga largo unitario.
  - Defina
  $$\hat{\alpha}^* = c\hat{\alpha} = \frac{\hat{\alpha}}{\hat{\alpha}_1}$$
  de forma tal que el primer elemento del nuevo vector $\hat{\alpha}^*$ sea 1.
- Las magnitudes de $\hat{\alpha}_1^*, \hat{\alpha}_2^*, \ldots, \hat{\alpha}_p^*$ en la primera normalización caen en el intervalo $[-1, 1]$. En la segunda normalización, $\hat{\alpha}_1^* = 1$ y $\hat{\alpha}_2^*, \ldots, \hat{\alpha}_p^*$ están expresados como múltiplos de $\hat{\alpha}_1^*$.

Análisis Discriminante

- Restringir a que los $\hat{\alpha}_i^*$ caigan en el intervalo $[-1, 1]$ facilita la comparación entre los coeficientes.
- Similarmente, expresar a los coeficientes como múltiplos de $\hat{\alpha}_1^*$ permite evaluar la importancia relativa de las variables $X_2, X_3, \ldots, X_p$ como discriminantes.
- Qué se puede hacer cuando no se conoce la distribución de probabilidad de las variables?
- Una alternativa consiste en construir una combinación lineal de las variables (una suma ponderada) de forma tal que esta combinación discrimine de la mejor manera a los grupos.
- Podemos luego comparar como difieren los grupos con respecto a esta combinación lineal y también observar los pesos relativos de cada variable para determinar su importancia relativa en la discriminación.

Análisis Discriminante

- La función discriminante de Fisher es el método por el cual se determina la combinación lineal.
- La idea de Fisher es transformar las observaciones multivariantes de X en observaciones univariantes de $\hat{y}$, tal que las $\hat{y}$'s derivadas de las poblaciones π_1 y π_2 estuvieran tan separadas como fuera posible.
- Fisher sugirió tomar combinaciones lineales de X para crear $\hat{y}$. Esto es,

$$\begin{aligned} \hat{y}_{1i} &= \hat{\alpha}_1 X_{1i} + \hat{\alpha}_2 X_{2i} + \cdots + \hat{\alpha}_p X_{pi} \quad i = 1, 2, \ldots, n_1 \\ \hat{y}_{2i} &= \hat{\alpha}_1 X_{1i} + \hat{\alpha}_2 X_{2i} + \cdots + \hat{\alpha}_p X_{pi} \quad i = 1, 2, \ldots, n_2 \end{aligned}$$

- La separación de estos dos conjuntos de valores de $\hat{y}$ se establece en función de la diferencia entre $\bar{\hat{y}}_1$ e $\bar{\hat{y}}_2$ expresada en unidades de desvíos estándar.

Análisis Discriminante

- En términos matemáticos,

$$\text{Separación} = \frac{|\bar{\hat{y}}_1 - \bar{\hat{y}}_2|}{s_{\hat{y}}}$$

- Donde,

$$s^2_{\hat{y}} = \frac{\sum_{i=1}^{n_1}(\hat{y}_{1i} - \bar{\hat{y}}_1)^2 + \sum_{i=1}^{n_2}(\hat{y}_{2i} - \bar{\hat{y}}_2)^2}{n_1 + n_2 - 2}$$

- El objetivo es seleccionar la combinación lineal de las X's para alcanzar la máxima separación entre las medias muestrales $\bar{\hat{y}}_1$ e $\bar{\hat{y}}_2$. Para ello, maximizar la ecuación anterior es lo mismo que maximizar,

$$\frac{\text{Distancia entre las medias de } \hat{y} \text{ al cuadrado}}{\text{Varianza de } \hat{y}} = \frac{(\bar{\hat{y}}_1 - \bar{\hat{y}}_2)^2}{s^2_{\hat{y}}}$$

Análisis Discriminante

En la expresión anterior, note que,

$$
\begin{aligned}
s^2_{\hat{y}} &= \frac{\sum_{j=1}^{n_1}(\hat{y}_{1j}-\bar{\hat{y}}_1)^2+\sum_{j=1}^{n_2}(\hat{y}_{2j}-\bar{\hat{y}}_2)^2}{n_1+n_2-2} \\
&= \frac{\sum_{j=1}^{n_1}(\hat{\alpha}'x_{1j}-\hat{\alpha}'\bar{x}_1)(x'_{1j}\hat{\alpha}-\bar{x}'_1\hat{\alpha})}{n_1+n_2-2} \\
&+ \frac{\sum_{j=1}^{n_2}(\hat{\alpha}'x_{2j}-\hat{\alpha}'\bar{x}_2)(x'_{2j}\hat{\alpha}-\bar{x}'_2\hat{\alpha})}{n_1+n_2-2} \\
&= \frac{\hat{\alpha}'[\sum_{j=1}^{n_1}(x_{1j}-\bar{x}_1)(x_{1j}-\bar{x}_1)']\hat{\alpha}+\hat{\alpha}'[\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)(x_{2j}-\bar{x}_2)']\hat{\alpha}}{n_1+n_2-2} \\
&= \frac{\hat{\alpha}'\{[\sum_{j=1}^{n_1}(x_{1j}-\bar{x}_1)(x_{1j}-\bar{x}_1)']+[\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)(x_{2j}-\bar{x}_2)']\}\hat{\alpha}}{n_1+n_2-2} \\
&= \frac{\hat{\alpha}'[(n_1-1)S_1+(n_2-1)S_2]\hat{\alpha}}{n_1+n_2-2} \\
&= \hat{\alpha}'S\hat{\alpha}
\end{aligned}
$$

Análisis Discriminante

- y que,

$$(\bar{\hat{y}}_1 - \bar{\hat{y}}_2)^2 = (\hat{\alpha}'\bar{x}_1 - \hat{\alpha}'\bar{x}_2)^2 = [\hat{\alpha}'(\bar{x}_1 - \bar{x}_2)]^2 = [\hat{\alpha}'d]^2,$$

  donde $d = (\bar{x}_1 - \bar{x}_2)$.
- Por lo tanto el problema se reduce a encontrar los coeficientes $\hat{\alpha}$ que maximicen,

$$\frac{(\text{Distancia entre las medias muestrales de } \hat{y} \text{ al cuadrado})}{(\text{Varianza muestral de } \hat{y})} = \frac{[\hat{\alpha}'d]^2}{\hat{\alpha}'S\hat{\alpha}}$$

- Realizando la maximización sobre todos los posibles $\hat{\alpha}$ se puede mostrar que el máximo se alcanza con los coeficientes $\hat{\alpha} = (\bar{x}_1 - \bar{x}_2)'S^{-1}$ que son los mismos coeficientes que determinamos anteriormente para dos poblaciones normales.

Análisis Discriminante

- Note que el procedimiento de Fisher no asume que las poblaciones son normales, sin embargo si asume implicitamente que las matrices de varianzas y covarianzas de las dos poblaciones son iguales.
- Con estos supuestos Fisher llega a la misma regla de clasificación que asumiendo poblaciones normales. Esto es, Asigne X_0 a π_1 si

$$(\bar{x}_1 - \bar{x}_2)' S^{-1} X_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 + \bar{x}_2) > 0$$

Logit versus Discriminante

- Para ver la relación entre los coeficientes discriminantes y los del modelo Logit escribamos $Pr[y_i = 1|x]$ usando la definición de la probabilidad condicional. Esto es,

$$\begin{aligned} Pr[y_i = 1|x] &= \frac{Pr[y_i = 1 \wedge x]}{Pr[x]} \\ &= \frac{Pr[x|y_i = 1]Pr[y_i = 1]}{Pr[x]} \end{aligned}$$

de la misma manera,

$$Pr[y_i = 0|x] = \frac{Pr[x|y_i = 0]Pr[y_i = 0]}{Pr[x]}$$

- Ahora calculemos la tasa de probabilidad,

$$\frac{Pr[y_i = 1|x]}{Pr[y_i = 0|x]} = \frac{Pr[x|y_i = 1]}{Pr[x|y_i = 0]} \times \frac{Pr[y_i = 1]}{Pr[y_i = 0]}$$

Logit versus Discriminante

- Asumiendo que las probabilidades iniciales son iguales (i.e. $Pr[y_i = 1] = Pr[y_i = 0]$) y tomando logaritmos naturales en ambos miembros, la ecuación anterior queda,

$$log\left\{\frac{Pr[y_i = 1|x]}{Pr[y_i = 0|x]}\right\} = log\left\{\frac{Pr[x|y_i = 1]}{Pr[x|y_i = 0]}\right\}$$

- Reemplazando el lado izquierdo por el logaritmo de la tasa de probabilidad del modelo Logit, tenemos

$$\beta_0 + \beta_1 x = log\left\{\frac{Pr[x|y_i = 1]}{Pr[x|y_i = 0]}\right\}$$

- Note que el lado derecho de la ecuación anterior es el cociente de las funciones de probabilidad de las variables explicativas para los dos grupos bajo análisis. Esto es lo que llamamos $f_1(x)$ y $f_2(x)$ en el análisis discriminante.

Logit versus Discriminante

- Si las distribuciones de probabilidad de las variables explicativas son Normales Multivariantes con la misma matriz de varianzas y covarianzas en ambos grupos, el lado derecho de la ecuación anterior puede reemplazarse por el lado izquierdo de la ecuación (5) para obtener,

$$\beta_0 + \beta_1 x = -\frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2) + (\mu_1 - \mu_2)'\Sigma^{-1}x$$

- Igualando los coeficientes de ambos lados de la ecuación se obtiene,

$$\beta_0 = -\frac{1}{2}(\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 + \mu_2)$$

y

$$\beta_1 = (\mu_1 - \mu_2)'\Sigma^{-1}$$

Logit versus Discriminante

- Entonces las estimaciones de los coeficientes del modelo Logit vienen dadas por,

$$\hat{\beta}_0 = -\frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 + \bar{x}_2) = -\frac{1}{2}\hat{\alpha}(\bar{x}_1 + \bar{x}_2)$$

y

$$\hat{\beta}_1 = (\bar{x}_1 - \bar{x}_2)' S^{-1}$$

Microeconometría I

Maestría en Econometría

Lecture 2

Modelos Logit y Probit

1. Temas Avanzados de Logit y Probit
   - Medidas de Diagnóstico
     - Método "informal" para contrastar por error en la especificación de la forma funcional
     - Método formal para contrastar por error en la especificación del argumento de la forma funcional
   - Heterocedasticidad
   - Endogeneidad
     - Variable endógena continua
     - Variable endógena dicotómica

Agenda

1. Temas Avanzados de Logit y Probit
   - Medidas de Diagnóstico
     - Método "informal" para contrastar por error en la especificación de la forma funcional
     - Método formal para contrastar por error en la especificación del argumento de la forma funcional
   - Heterocedasticidad
   - Endogeneidad
     - Variable endógena continua
     - Variable endógena dicotómica

Forma Funcional

- Recuerde que abandonamos el MPL esencialmente porque la forma funcional de la probabilidad no era correcta.
- En su lugar especificamos la probabilidad de ocurrencia del evento analizado con la siguiente función:

$$Pr(y_i = 1|x) = G(\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)$$

donde $G(\cdot)$ es la función de distribución logística (modelo Logit) o la función de distribución normal estándar (modelo Probit).

- Primero, no hay ninguna garantía de que las funciones de distribución de la normal o la logística sean las formas funcionales adecuadas.
- Segundo, es posible que la función asumida para el modelo $\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K$ pueda ser incorrecta.

Forma Funcional

- Es decir, hay dos fuentes de error en la especificación del modelo:
  - la función $G(\cdot)$, usualmente normal o logística, puede ser incorrecta.
  - el argumento de la función $G(\cdot)$ puede tener una forma funcional incorrecta.
- Necesitamos algunos métodos que nos ayuden a detectar este tipo de errores en la forma funcional.
- Vamos a desarrollar dos de esos métodos:
  - un método "informal" semi-paramétrico.
  - un método formal y paramétrico.

Forma Funcional

- Supongamos que acabamos de estimar el siguiente modelo Probit:

$$Pr(y_i = 1|x) = \Phi(\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K)$$

  (si es un modelo Logit, solo reemplace $\Phi(\cdot)$ por $F(\cdot)$).
- La hipótesis nula es que la forma funcional es correcta.
- Esto es, la función $\Phi(\cdot)$ es correcta y su argumento $\beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K$ también.

Forma Funcional

- Dada la estimación de los parámetros $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_K$ calculemos el argumento de la función:

$$\hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_K x_K = x\hat{\beta}$$

- Bajo la hipótesis nula tenemos:

$$\begin{aligned} \widehat{Pr(y = 1|x)} = \widehat{E(y|x)} &= \Phi(x\hat{\beta}) \\ y &= \Phi(x\hat{\beta}) + \nu \end{aligned}$$

Forma Funcional

- Se sigue que si regresamos y contra $x\hat{\beta}$ permitiendo una forma funcional $\Lambda(\cdot)$ general (i.e. no necesariamente normal):

$$y = \Lambda(x\hat{\beta}) + \nu$$

la función estimada $\widehat{\Lambda(x\hat{\beta})}$ debería parecerse a la función de distribución de la normal estándar.

- Para permitir una forma funcional flexible para $\Lambda(\cdot)$ se utiliza un procedimiento semi-paramétrico.
- Un método fácil de implementar en Stata es lowless smoothing.
- Una explicación muy informal del procedimiento es como sigue.

Lowless Smoothing

- Para cada observación en los datos calcular el valor esperado de y_i dado x_i desde una regresión que
  1. use solo observaciones de x que estén "cerca" de x_i; y
  2. use ponderadores determinados por la cercanía de cada x_j con x_i (más cerca de x_i, mayor ponderador).
- Una vez que este procedimiento ha sido implementado para todos los datos de la muestra, podemos graficar las n estimaciones de la esperanza matemática de y dado x en el eje vertical y x en el eje horizontal.
- Este gráfico es nuestra estimación de $\Lambda(\cdot)$
- Comparando con el gráfico de $\widehat{\Phi(\cdot)}$ podemos decidir informalmente si la distribución normal estándar es una buena especificación.

Forma Funcional

- Obviamente, existen procedimientos formales para contrastar la forma funcional.
- Estos procedimientos son muy difíciles de implementar empíricamente.
- La referencia clásica es:

  Horowitz, J. L. 1993. "Semiparametric Estimation of a Work-trop Mode Choice Model," *Journal of Econometrics* 58, pp. 49-70.

Forma Funcional

- El punto de partida es el mismo que en el contraste anterior: $x\hat{\beta}$.
- Bajo la hipótesis nula:

$$Pr(y_i = 1|x) = \Phi(x\beta)$$

- Considere la especificación alternativa

$$Pr(y_i = 1|x) = \Phi\left[(x\beta) + \gamma_1(x\beta)^2 + \gamma_2(x\beta)^3\right]$$

- Bajo la hipótesis nula de que el modelo Probit está bien especificado: $\gamma_1 = \gamma_2 = 0$.

Forma Funcional

- Si esta hipótesis nula se rechaza, entonces hay evidencia estadística para creer que el modelo Probit inicial está mal especificado.
- La hipótesis nula se puede contrastar estimando un modelo Probit con y como variable dependiente y con $(x\beta)$, $(x\beta)^2$ y $(x\beta)^3$ como variables explicativas.
- En esta estimación hay que imponer un coeficiente unitario sobre la variable $(x\beta)$ y asegurarse de que el modelo no tenga constante.

Forma Funcional

- Si se rechaza el modelo Probit (Logit), qué se puede hacer?
- Dos opciones:
  - Cambiar la forma funcional de $\Lambda(\cdot)$.
  - Cambiar la forma funcional del argumento de la función. Es muy probable que agregando términos de ordenes superiores en las variables explicativas resuelva el problema.
- Ir a ejemplo con el Stata.

Agenda

1. Temas Avanzados de Logit y Probit
   - Medidas de Diagnóstico
     - Método "informal" para contrastar por error en la especificación de la forma funcional
     - Método formal para contrastar por error en la especificación del argumento de la forma funcional
   - Heterocedasticidad
   - Endogeneidad
     - Variable endógena continua
     - Variable endógena dicotómica

Heterocedasticidad

- Los estimadores de los modelos Logit y Probit no son consistentes bajo heterocedasticidad.
- En presencia de heterocedasticidad la matriz de varianzas y covarianzas de los coeficientes estimados no es apropiada.
- Este es un problema serio ya que la mayoría de las estimaciones de estos modelos se hace con datos de corte transversal donde el problema de la heterocedasticidad es más frecuente.
- Para ilustrar el problema considere el siguiente modelo de variable latente con una sola variable explicativa:

$$y_i^* = \psi_0 + \psi_1 x_{i1} + u_i$$

- Supongamos que el error u_i es heterocedástico.

Heterocedasticidad

- Consideremos una heterocedasticidad multiplicativa:

$$u_i \sim \text{Normal}(0, x_{i1}^2)$$

- Recordemos que no se observa la variable latente y_i^*, lo que observamos es,

$$y_i = 1 \quad \text{si} \quad y_i^* > 0$$
$$y_i = 0 \quad \text{si} \quad y_i^* \leq 0$$

- Entonces,

$$y_i = 1 \text{ si } \psi_0 + \psi_1 x_{i1} + u_i > 0.$$

Heterocedasticidad

- Siguiendo con nuestro análisis,

$$\begin{aligned} Pr(y_i = 1|x_i) &= Pr(y_i^* > 0|x_i) \\ &= Pr(\psi_0 + \psi_1 x_{i1} + u_i > 0) \\ &= Pr(\psi_0 + \psi_1 x_{i1} + \sqrt{x_{i1}^2} e_i > 0) \end{aligned}$$

  donde $e_i \sim \text{Normal}(0, 1)$.

- Por lo tanto,

$$\begin{aligned} Pr(y_i = 1|x_i) &= Pr\left(e_i < -\frac{1}{x_{i1}}(\psi_0 + \psi_1 x_{i1})\right) \\ &= 1 - \Phi\left(-\frac{1}{x_{i1}}(\psi_0 + \psi_1 x_{i1})\right) \\ &= \Phi\left(\frac{1}{x_{i1}}(\psi_0 + \psi_1 x_{i1})\right) \end{aligned}$$

Heterocedasticidad

- Es decir,

$$Pr(y_i = 1|x_i) = \Phi\left(\psi_0 \frac{1}{x_{i1}} + \psi_1\right)$$

- Podemos ver que la presencia de heterocedasticidad ha alterado radicalmente la forma funcional del modelo.
- Dado el modelo subyacente para la variable latente

$$y_i^* = \psi_0 + \psi_1 x_{i1} + u_i$$

- Uno estaría tentado a especificar el modelo Probit como,

$$Pr(y_i = 1|x_i) = \Phi\left(\psi_0 + \psi_1 x_{i1}\right),$$

pero esta no es la especificación correcta en presencia de heterocedasticidad.

Heterocedasticidad

- Pensemos en el efecto marginal de x_{i1}, la especificación correcta es

$$Pr(y_i = 1|x_i) = \Phi\left(\psi_0 \frac{1}{x_{i1}} + \psi_1\right)$$

- El efecto marginal correcto es,

$$\frac{\partial\, Pr(y_i = 1|x_i)}{\partial\, x_{i1}} = \phi\left(\psi_0 \frac{1}{x_{i1}} + \psi_1\right) \times \left(-\psi_0 \left(\frac{1}{x_{i1}}\right)^2\right).$$

- El signo del efecto marginal es el opuesto al signo de ψ_0 (i.e. la constante el el modelo de variable latente) y no depende del signo de ψ_1 (el coeficiente de pendiente en el modelo de variable latente).

Heterocedasticidad

- Del desarrollo anterior se sigue que si ψ_0 y ψ_1 son positivos, el efecto marginal de x_{i1} sobre la probabilidad de ocurrencia del evento analizado tiene el signo opuesto al efecto marginal de x_{i1} en el modelo de variable latente.
- Por supuesto que este último resultado depende crucialmente de la existencia de heterocedasticidad multiplicativa de la forma planteada y por lo tanto el punto anterior no puede verse como un resultado general.
- El punto principal es que si el error del modelo de variable latente es heterocedástico, entonces se altera la forma funcional del Probit.
- Exactamente cómo se altera depende de la forma de la heterocedasticidad.

Heterocedasticidad

- Supongamos que especificamos el Probit incorrectamente como

$$Pr(y_i = 1|x_i) = \Phi\left(\eta_0 + \eta_1 x_{i1}\right)$$

- Será la estimación del coeficiente η_1 una buena estimación de ψ_1 en el modelo de variable latente

$$y_i^* = \psi_0 + \psi_1 x_{i1} + u_i?$$

- La respuesta es no. Y este es un ejemplo de como la presencia de heterocedasticidad lleva a estimadores no consistentes de los parámetros del modelo de variable latente.
- Cómo habría que proceder si creemos que la heterocedasticidad es un problema en nuestro modelo?

Heterocedasticidad

- Una posibilidad es utilizar el comando hetprob del Stata que estima un modelo Probit generalizado:

$$\begin{aligned} y^* &= \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K + e \\ y^* &= x\beta + e \end{aligned}$$

donde,

$$\sigma_e^2 = \left[e^{z\gamma}\right]^2,$$

con z un vector de variables (sin constante) que se piense afectan la varianza de e y γ sus correspondientes coeficientes.

Heterocedasticidad

- Entonces,

$$\begin{aligned} Pr(y = 1|x,\, z) &= Pr(y^* > 0|x,\, z) \\ &= Pr(x\,\beta + e > 0|x,\, z) \\ &= Pr(x\,\beta + e^{z\,\gamma} u > 0|x,\, z) \end{aligned}$$

  donde u sigue una normal estándar (una, valga la redundancia, normalización).

- Por lo tanto,

$$\begin{aligned} Pr(y = 1|x,\, z) &= Pr\left(u > \frac{-x\,\beta}{e^{z\,\gamma}}\right) \\ &= 1 - \Phi\left(\frac{-x\,\beta}{e^{z\,\gamma}}\right) \\ &= \Phi\left(\frac{x\,\beta}{e^{z\,\gamma}}\right) \end{aligned}$$

Heterocedasticidad

- Por supuesto que si un regresor, x_k está incluido en x y en z el efecto marginal es más complejo,

$$\frac{\partial\, Pr(y = 1|x,\, z)}{\partial\, x_k} = \phi\left(\frac{x\,\beta}{e^{z\,\gamma}}\right) \times \left(\frac{\beta_k - (x\beta)\gamma_k}{e^{z\,\gamma}}\right).$$

- Esto muestra que el efecto marginal no tiene necesariamente el signo de β_k.

Heterocedasticidad

- La función de verosimilitud es,

$$L(\hat{\beta}; x, z) = \prod_{i=1}^{n} \left[\Phi \left(\frac{x\beta}{e^{z\gamma}} \right) \right]^{y_i} \times \left[1 - \Phi \left(\frac{x\beta}{e^{z\gamma}} \right) \right]^{1-y_i}$$

- Y su logaritmo natural es,

$$l(\hat{\beta}; x, z) = \sum_{i=1}^{n} \left[y_i ln\{ \Phi \left(\frac{x\beta}{e^{z\gamma}} \right) \} + (1 - y_i) ln\{ 1 - \Phi \left(\frac{x\beta}{e^{z\gamma}} \right) \} \right]$$

Heterocedasticidad

- Las condiciones de primer orden para la maximización son,

$$\frac{\partial l(\cdot)}{\partial \hat{\beta}} = \sum_{i=1}^{n} \left[\frac{y_i - \Phi\left(\frac{x\beta}{e^{z\gamma}}\right)}{\Phi\left(\frac{x\beta}{e^{z\gamma}}\right)\left[1 - \Phi\left(\frac{x\beta}{e^{z\gamma}}\right)\right]} \phi\left(\frac{x\beta}{e^{z\gamma}}\right) \right] e^{-z\gamma} x = 0$$

$$\frac{\partial l(\cdot)}{\partial \hat{\gamma}} = \sum_{i=1}^{n} \left[\frac{y_i - \Phi\left(\frac{x\beta}{e^{z\gamma}}\right)}{\Phi\left(\frac{x\beta}{e^{z\gamma}}\right)\left[1 - \Phi\left(\frac{x\beta}{e^{z\gamma}}\right)\right]} \phi\left(\frac{x\beta}{e^{z\gamma}}\right) \right] e^{-z\gamma} z(-x\beta) = 0$$

- Estas son ecuaciones no lineales en las incógnitas.
- Se puede contrastar por heterocedasticidad con un test LR.

Heterocedasticidad

- H_0 : Homocedasticidad vs. H_1 : Heterocedasticidad

- Estadístico de contraste:

$$LR = -2 \times [\log L(\text{Probit}) - \log L(\text{HetProb})] \sim \chi^2_q.$$

Donde q es la dimensión de γ.

- Este contraste aparece es la salida de la estimación del Stata utilizando hetprob.
- Una interpretación alternativa de este contraste sugeriría que la forma funcional del modelo Probit es incorrecta.
- Una posibilidad aquí es incorporar como variables explicativas en el modelo Probit inicial las variables z elevadas al cuadrado.

Agenda

1. Temas Avanzados de Logit y Probit
   - Medidas de Diagnóstico
     - Método "informal" para contrastar por error en la especificación de la forma funcional
     - Método formal para contrastar por error en la especificación del argumento de la forma funcional
   - Heterocedasticidad
   - Endogeneidad
     - Variable endógena continua
     - Variable endógena dicotómica

Endogeneidad

- Todos los problemas provocados por la correlación entre alguna de las variables explicativas y el error de la ecuación que se estudian en los modelos lineales se trasladan a los modelos Probit y Logit.
- Sin embargo, la estimación y la interpretación de los modelos Probit y Logit con variables instrumentales no es completamente directa.
- En principio si se quiere estimar un modelo Probit o Logit con variables endógenas hay que imponer algunos supuestos bastante fuertes.
- Estos supuestos hacen que el único modelo que se puede estimar es el modelo Probit.
- Además, la estimación depende de si la variable endógena en el modelo es continua o dicotómica.

Endogeneidad

- Comencemos ilustrando el caso de una variable potencialmente endógena continua.
- Escribamos el modelo en forma de variable latente:

$$
\begin{aligned}
y_1^* &= z_1\delta_1 + \alpha_1 y_2 + u_i && (1)\\
y_2 &= z_1\delta_{21} + z_2\delta_{22} + v_2 = z\delta_2 + v_2 && (2)\\
y_1 &= 1[y_1^* > 0] && (3)
\end{aligned}
$$

  donde (u_1, v_2) tienen distribución bivariada normal con media cero y son independientes de z.
- Las ecuaciones (1) y (3) constituyen el modelo estructural.
- La ecuación (2) es la forma reducida de y_2, que es endógena si u_1 y v_2 están correlacionados.

Endogeneidad

- Note que si la ecuación estructural se especifica como un Logit habria que pensar que tipo de distribución bivariada podrian tener (u_1, v_2).
- Esta distribución al no ser normal bivariada seguramente sea mucho más compleja de analizar y por eso en la literatura se analiza solo el modelo Probit.
- Un segundo punto a tener en cuenta es que si v_2 tiene distribución normal, entonces y_2 también la tiene y no puede comportarse como una variable discreta (i.e. la variable potencialmente endógena es continua).
- Finalmente, necesitamos asumir que la varianza de u_1 es igual a uno para poder identificar δ_1 y α_1 en (1).

Endogeneidad

- Para ver la necesidad del último supuesto, supongamos que $u_1 \sim N(0, \sigma_1^2)$.
- Usando las ecuaciones (1) y (3) tenemos,

$$\begin{aligned} Pr(y_1 = 1|y_2, z_1) &= Pr(y_1^* > 0|y_2, z_1) \\ &= Pr(u_1 > -z_1\delta_1 - \alpha_1 y_2|y_2, z_1) \\ &= Pr(u_1 \leq z_1\delta_1 + \alpha_1 y_2|y_2, z_1) \\ &= \Phi(z_1\frac{\delta_1}{\sigma_1} + \frac{\alpha_1}{\sigma_1}y_2) \end{aligned}$$

- Esto muestra que la varianza de u_1 y los parámetros de pendiente del modelo no pueden identificarse por separado.

Endogeneidad

- Las variables en z_1 y z_2 se asumen exógenas.
- Note que las variables en z_2 son los instrumentos (restricciones de exclusión).
- Esto implica que y_2 es endógena en (1) si y solo si la correlación entre u_1 y v_2 es diferente de cero.
- En este contexto, la estimación del modelo puede hacerce con el procedimiento de dos etapas de Rivers y Vuong (1988).
- Normalidad bivariante de (u_1, v_2) con $Var(u_1) = 1$ implica que,

$$u_1 = \frac{Cov(u_1,\ v_2)}{Var(v_2)} v_2 + e_1 = \theta_1 v_2 + e_1$$

donde e_1 es independiente de z y v_2 (y por lo tanto de y_2).

Endogeneidad

- Note que $var(u_1) = 1$ implica

$$\begin{aligned} Var(e_1) &= 1 - \frac{[Cov(u_1,\ v_2)]^2}{Var(v_2)} \\ &= 1 - \frac{[Cov(u_1,\ v_2)]}{\sqrt{Var(v_2)}\sqrt{1}} \frac{[Cov(u_1,\ v_2)]}{\sqrt{Var(v_2)}\sqrt{1}} \\ &= 1 - \rho_1^2 \end{aligned}$$

  donde $\rho_1 = corr(u_1,\ v_2)$.

- Re-escribiendo el modelo de variable latente tenemos

$$y_1^* = z_1\delta_1 + \alpha_1 y_2 + \theta_1 v_2 + e_1$$

  y piense que v_2 es una variable sobre la que podemos condicionar en el modelo Probit.

Endogeneidad

- Entonces

$$Pr(y_1 = 1|z,\ y_2,\ v_2) = \Phi\left(\frac{z_1\delta_1 + \alpha_1 y_2 + \theta_1 v_2}{\sqrt{1-\rho_1^2}}\right).$$

- Esto es, un Probit sobre z_1, y_2 y v_2 estima consistentemente $\delta_1/\sqrt{1-\rho_1^2}$, $\alpha_1/\sqrt{1-\rho_1^2}$ y $\theta_1/\sqrt{1-\rho_1^2}$.

- Como $\rho_1^2 < 1$ cada uno de los coeficientes estimados es mayor que el coeficiente que se estimaría si y_2 fuera exógeno.

Endogeneidad

- En la práctica no se observa v_2.
- Rivers y Vuong sugieren el siguiente procedimiento de dos etapas.
  1. Estime por MCC una regresión de y_2 sobre z y obtenga los residuos $\hat{v}_2$.
  2. Estime un Probit de y_1 sobre z_1, y_2 y $\hat{v}_2$.
- Una característica de este procedimiento es que si y_2 es exógeno entonces el coeficiente sobre $\hat{v}_2$ es cero.
- Esto sugiere contrastar por exogeneidad comparando $H_0 : \theta_1 = 0$ con $H_1 : \theta_1 \neq 0$ usando el test t usual en la estimación del Probit en el segundo paso.
- Sin embargo los errores estándar y estadísticos t usuales del Probit no son estrictamente válidos debido a la presencia de $\hat{v}_2$.

Endogeneidad

- Una alternativa a este procedimiento de dos etapas es estimar el sistema (1)-(3) simultaneamente.
- Esto requiere escribir la función de verosimilitud de observar y_1 e y_2 y maximizarla.
- A diferencia del procedimiento de Rivers y Vuong, el método de estimación simultánea provee de estimaciones directas de δ_1 y α_1.
- Una ventaja del enfoque simultáneo es que permite calcular directamente los errores estándar de los coeficientes estimados.
- Este procedimiento de estimación simultánea se denomina Probit de variables instrumentales o IV Provit y es la estimación que se obtiene usando el comando ivprobit en Stata.

Endogeneidad

- Consideremos el caso en el que el Probit contiene una variable explicativa binaria endógena.

$$y_1 = 1[z_1\delta_1 + \alpha_1 y_2 + u_1 > 0] \quad (4)$$
$$y_2 = 1[z\delta_2 + v_2 > 0] \quad (5)$$

  donde $(u_1,\ v_2)$ es independiente de z y se distribuye normal bivariada con media cero. Cada componente tiene varianza unitaria y $\rho_1 = corr(u_1,\ v_2)$.
- Si $\rho_1 \neq 0$ entonces u_1 e y_2 están correlacionados y una estimación Probit de la ecuación (4) nos dará estimadores inconsistentes de δ_1 y α_1.

Endogeneidad

- Al igual que en el caso de una variable endógena continua, la normalización apropiada para poder identificar a los parámetros de pendiente en la ecuación (4) es que $Var(u_1) = 1$.
- Para obtener la distribuci'on conjunta de (y_1, y_2) condicionada en z recuerde que,

$$f(y_1, y_2|z) = f(y_1|y_2, z)f(y_2|z)$$

- Note que,

$$Pr(y_1 = 1|v_2, z) = \Phi[(z_1\delta_1 + \alpha_1 y_2 + \rho_1 v_2)/(1 - \rho_1^2)^{1/2}].$$

- Como $y_2 = 1$ si y solo si $v_2 > -z\delta_2$ para poder calcular los momentos de la distribución truncada necesitamos asumir que v_2 tiene distribución normal estándar y es independiente de z.

Endogeneidad

- Entonces, la función de densidad de v_2 dado que $v_2 > -z\delta_2$ es

$$\phi(v_2)/Pr(v_2 > -z\delta_2) = \phi(v_2)/\Phi(z\delta_2)$$

- Por lo tanto,

$$\begin{aligned} Pr(y_1 = 1|y_2 = 1, z) &= E[Pr(y_1 = 1|v_2, z)|y_2 = 1, z] \\ &= E\{\Phi[(z_1\delta_1 + \alpha_1 y_2 + \rho_1 v_2)/(1 - \rho_1^2)^{1/2}]|y_2 = 1, z\} \end{aligned}$$

$$= \frac{1}{\Phi(z\delta_2)} \int_{-z\delta_2}^{\infty} \Phi[(z_1\delta_1 + \alpha_1 y_2 + \rho_1 \nu_2)/(1 - \rho_1^2)^{1/2}]\phi(\nu_2) d\nu_2$$

- ν_2 en el integral es el argumento de integración.

Endogeneidad

- Obviamente $Pr(y_1 = 0|y_2 = 1,\, z) = 1 - Pr(y_1 = 1|y_2 = 1,\, z)$.
- Similarmente $Pr(y_1 = 1|y_2 = 0,\, z)$ es,

$$\frac{1}{1 - \Phi(z\delta_2)} \int_{-z\delta_2}^{\infty} \Phi[(z_1\delta_1 + \alpha_1 y_2 + \rho_1\nu_2)/(1 - \rho_1^2)^{1/2}]\phi(\nu_2)d\nu_2$$

- Combinando los cuatro posibles resultados de $(y_1,\, y_2)$ junto con el modelo Probit para y_2 y tomando logaritmos se obtiene la función de verosimilitud del modelo.
- En la práctica se puede estimar el modelo de las ecuaciones (4)-(5) utilizando un modelo Probit bivariado.
- En Stata el comando es biprobit.

Microeconometría I

Maestría en Econometría

Lecture 3

Variables Categóricas

1. Variables Categóricas No Ordenadas
   - Clasificación de las Variables Explicativas
   - Modelo Logit Multinomial
   - Modelo Logit Condicional
   - Modelo Logit General
   - Modelo Probit Multinomial, Condicional y General
2. Variables Categóricas Ordenadas
   - Modelo Logit Ordenado
   - Modelo Probit Ordenado
3. Variables Categóricas Secuenciales
   - Modelo Logit Secuencial
   - Modelo Probit Secuencial

Agenda

Variables Categóricas No Ordenadas

- Hasta ahora hemos considerado modelos cuya variable dependiente es binaria. Sin embargo, en la práctica, son mucho más comúnes aquellos modelos que tratan de explicar una variable categórica con más de dos opciones.
- Por ejemplo, es usual tratar de explicar el tipo de trabajo que tiene un individuo sobre la base de características personales y del trabajo en si.
- Para extender el análisis que hemos hecho consideremos el caso de una variable dependiente categórica no ordenada.
- Es decir, ahora asumimos que un individuo i puede elegir entre J categorías, con $J > 2$.
- La elección del individuo i se denota por y_i, que ahora puede adoptar valores discretos: $1, 2, \ldots, J$.

Variables Categóricas No Ordenadas

- Como en los modelos de variable dependiente binaria, vamos a relacionar la elección de las categorías con diferentes variables explicativas.
- En general, tenemos tres tipos de variables explicativas:
  1. variables que son diferentes entre individuos pero iguales entre categorías (edad, ingreso, género, etc.). Vamos a llamar a estas variables x_i.
  2. variables diferentes entre individuos y además diferentes entre categorías (por ejemplo, en la elección de un tipo de trabajo una de estas variables podría ser el salario de cada tipo de trabajo j que enfrenta el individuo i). Llamaremos a estas variables $w_{i,j}$.
  3. variables iguales entre individuos pero diferentes entre categorías (por ejemplo, diferentes características de cada trabajo j). Llamaremos a estas variables z_j.

Modelo Multinomial

- Cuando la variable dependiente se explica utilizando solo el primer tipo de variables explicativas, x_i (edad, género, etc.) el modelo recibe el nombre de MULTINOMIAL.
- Un posible modelo en términos de utilidades estocásticas es,

$$U_i^j = x_i' \beta_j + \epsilon_{i,j}$$

Donde, U_i^j representa la utilidad que tiene para el individuo i seleccionar la categoría j. Las x_i representan diferentes características del individuo i; y los β_j ponderan las diferentes caraterísticas del individuo para dar la utilidad total de la categoría j. Los $\epsilon_{i,j}$ son las preferencias específicas del individuo i no modeladas (i.e. los errores del modelo).

Modelo Multinomial

- Se asume que el individuo i selecciona aquella categoría que le brinda mayor utilidad:

$$\begin{aligned} p_{i,j} &= Pr[y_i = j|\cdot] = Pr[U_i^j > U_i^k|\cdot] \ \forall k \neq j \\ &= Pr[x_i'\beta_j + \epsilon_{i,j} > x_i'\beta_k + \epsilon_{i,k}|\cdot] \ \forall k \neq j \\ &= Pr[\epsilon_{i,k} - \epsilon_{i,j} \leq x_i'(\beta_j - \beta_k)|\cdot] \ \forall k \neq j \\ &= Pr[\epsilon_i \leq x_i'(\beta_j - \beta_k)|\cdot] \ \forall k \neq j \end{aligned}$$

- Dependiendo de la función de distribución asumida (logística o normal) el modelo recibe el nombre de Logit Multinomial o Probit Multinomial.

Agenda

Modelo Logit Multinomial

- Asumiendo que $\epsilon_{i,j}$ tiene distribución doble exponencial, se puede demostrar (Amemiya, 1995 pg. 297):

$$p_{i,j} = Pr[y_i = j|\cdot] = \frac{e^{x_i'(\beta_j - \beta_k)}}{\sum_{l=1}^{J} e^{x_i'(\beta_l - \beta_k)}} \quad j = 1, 2, \ldots, J$$

  y $\sum_{j=1}^{J} p_{i,j} = 1$.
- La última condición asegura que cada individuo selecciona alguna de las categorías de la variable dependiente. Esto implica que una de las categorías actúa como base.

Modelo Logit Multinomial

- Es decir, en el denominador de la expresión anterior tenemos,

$$\sum_{l=1}^{J} e^{x_i'(\beta_l-\beta_k)} = e^{x_i'(\beta_1-\beta_k)} + e^{x_i'(\beta_2-\beta_k)} + \cdots + e^{x_i'(\beta_k-\beta_k)} + \cdots + e^{x_i'(\beta_J-\beta_k)} = 1 + \sum_{l=1}^{J} e^{x_i'(\beta_l-\beta_k)} \ \forall l \neq k.$$

- Si pérdida de generalidad, asumamos que la categoría base es la categoría J.

Modelo Logit Multinomial

- Entonces, el modelo Logit multinomial queda,

$$p_{i,j} = Pr[y_i = j|\cdot] = \frac{e^{x_i'\alpha_j}}{1 + \sum_{l=1}^{J-1} e^{x_i'\alpha_l}} \quad j = 1, 2, \ldots, J-1$$

y

$$p_{i,J} = Pr[y_i = J|\cdot] = \frac{1}{1 + \sum_{l=1}^{J-1} e^{x_i'\alpha_l}}$$

donde $\alpha_j = \beta_j - \beta_J$ tiene la interpretación de ser el efecto del coeficiente de la categoría j sobre la categoría base.

- No es obvio que se pueda hacer una interpretación directa de estos parámetros. El efecto de x_i sobre la probabilidad de elegir la categoría j es claramente una función no lineal de los α_j.

Modelo Logit Multinomial: Interpretación

- El efecto de un cambio en x_i sobre la probabilidad de que se cumpla el evento analizado surge de la derivada parcial de $Pr[y_i = j|\cdot]$ con respecto a x_i:

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial x_i} = Pr[y_i = j|\cdot]\{\alpha_j - \sum_{l=1}^{J-1} \alpha_l Pr[y_i = l|\cdot]\}$$

- El signo de la derivada depende ahora no solo del signo del coeficiente que acompaña a x_i, sino también del signo del término entre llaves.
- Este último punto contrasta con el modelo Logit para variables binarias donde las probabilidades eran monótonas crecientes o decrecientes.

Modelo Logit Multinomial: Interpretación

- Note que para $J = 2$ la derivada parcial anterior se reduce a:

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial x_i} = Pr[y_i = j|\cdot]\{1 - Pr[y_i = j|\cdot]\}\alpha_j$$

- Además de la derivada parcial, la semi-elasticidad de x_i, también se utiliza para interpretar el modelo:

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial x_i} x_i = Pr[y_i = j|\cdot]\{\alpha_j - \sum_{l=1}^{J-1} \alpha_l Pr[y_i = l|\cdot]\} x_i$$

- La semi-elasticidad mide el cambio en puntos porcentuales en la probabilidad de que la categoría j sea elegida ante un cambio porcentual en x_i.

Agenda

Modelo Logit Condicional

- En el modelo Logit multinomial, la elección de los individuos está relacionada con variables explicativas específicas de esos individuos.
- En otros casos, uno puede tener variables explicativas que adoptan diferentes valores para cada categoría. La versión del modelo Logit que se ajusta a este tipo de variables recibe el nombre de modelo Logit Condicional (McFadden 1973).
- En el modelo Logit condicional, la probabilidad de que el individuo i seleccione la categoría j es:

$$p_{i,j} = Pr[y_i = j|\cdot] = \frac{e^{\gamma_0 + w'_{i,j}\beta}}{\sum_{l=1}^{J} e^{\gamma_0 + w'_{i,l}\beta}} \quad j = 1, 2, \ldots, J$$

y $\sum_{j=1}^{J} p_{i,j} = 1$.

Modelo Logit Condicional: Interpretación

- Igual que antes, fijamos a J como categoría base. Al hacerlo, la constante del modelo se hace igual a cero, $\gamma_0 = 0$; y las probabilidades del modelo condicional quedan:

$$p_{i,j} = Pr[y_i = j|\cdot] = \frac{e^{(w_{i,j} - w_{i,J})'\beta}}{1 + \sum_{l=1}^{J-1} e^{(w_{i,l} - w_{i,J})'\beta}} \quad j = 1, 2, \ldots, J-1$$

y

$$p_{i,J} = Pr[y_i = J|\cdot] = \frac{1}{1 + \sum_{l=1}^{J-1} e^{(w_{i,l} - w_{i,J})'\beta}}$$

- Como sucedía en el modelo multinomial, las probabilidades son funciones no lineales de los parámetros del modelo y por lo tanto la interpretación de los mismos no es directa.

Modelo Logit Condicional: Interpretación

- El efecto de un cambio en $w_{i,j}$ sobre la probabilidad de selección surge de la derivada parcial de $Pr[y_i = j|\cdot]$ con respecto a $w_{i,j}$:

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial w_{i,j}} = Pr[y_i = j|\cdot]\{1 - Pr[y_i = j|\cdot]\}\beta$$

- La derivada parcial de la probabilidad de que el individuo i elija la categoría j con respecto a $w_{i,l}$ para $l \neq j$. Esto es,

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial w_{i,l}} = -Pr[y_i = j|\cdot]Pr[y_i = l|\cdot]\beta$$

- El signo de estas derivadas está determinado por el signo de β y entonces la probabilidad varía monotónicamente.

Modelo Logit Condicional: Interpretación

- En el ejemplo de la elección de diferentes tipos de trabajo donde $w_{i,j}$ es el salario del trabajo j que enfrenta la persona i, las dos derivadas anteriores muestran que para $\beta > 0$, un incremento en el salario del trabajo j incrementa la probabilidad de elección de j y disminuye la probabilidad de elección de otros trabajos ($l \neq j$).
- De las dos derivadas anteriores surgen las semi-elasticidades directas y cruzadas. El cambio porcentual en la probabilidad de que la categoría j sea elegida ante un cambio porcentual en $w_{i,j}$ es:

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial w_{i,j}} w_{i,j} = Pr[y_i = j|\cdot]\{1 - Pr[y_i = j|\cdot]\} w_{i,j}\beta$$

Modelo Logit Condicional: Interpretación

- El cambio porcentual en la probabilidad de elegir la categoría j debido a un cambio porcentual en $w_{i,l}$ es:

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial w_{i,l}} w_{i,l} = -Pr[y_i = j|\cdot]Pr[y_i = l|\cdot]w_{i,l}\beta$$

Agenda

Modelo Logit General

- Una especificación más general se da cuando se combinan los dos modelos anteriores (el multinomial y el condicional) y se agrega el último tipo de variable explicativa. Esto es, variables z_j que adoptan valores diferentes entre categorías pero iguales entre individuos.
- Por lo tanto, el modelo especificado queda:

$$p_{i,j} = Pr[y_i = j|\cdot] = \frac{e^{x_i'\alpha_j+\gamma_0+w_{i,j}'\beta+z_j'\delta}}{\sum_{l=1}^{J} e^{x_i'\alpha_l+\gamma_0+w_{i,l}'\beta+z_l'\delta}} \quad j = 1, 2, \ldots, J$$

y $\sum_{j=1}^{J} p_{i,j} = 1$.

Modelo Logit General

- Igual que en los dos modelos anteriores, para identificar los parámetros del modelo hay que tomar una categoría como base. Fijando a J como esta categoría las probabilidades del modelo general quedan:

$$p_{i,j} = Pr[y_i = j|\cdot] = \frac{e^{x_i'\alpha_j + (w_{i,j} - w_{i,J})'\beta + (z_j - z_J)'\delta}}{1 + \sum_{l=1}^{J-1} e^{x_i'\alpha_l + (w_{i,l} - w_{i,J})'\beta + (z_l - z_J)'\delta}}$$

para $j = 1, 2, \ldots, J-1$, y

$$p_{i,J} = Pr[y_i = J|\cdot] = \frac{1}{1 + \sum_{l=1}^{J-1} e^{x_i'\alpha_l + (w_{i,l} - w_{i,J})'\beta + (z_l - z_J)'\delta}}$$

Modelo Logit General: Interpretación

- Como sucedía en los modelos anteriores, las probabilidades son funciones no lineales de los parámetros del modelo y por lo tanto la interpretación de los mismos no es directa.
- El efecto de un cambio en z_j sobre la probabilidad de selección surge de la derivada parcial de $Pr[y_i = j|\cdot]$ con respecto a z_j:

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial z_j} = Pr[y_i = j|\cdot]\{1 - Pr[y_i = j|\cdot]\}\delta$$

- El signo de esta derivada está totalmente determinado por el signo de δ y por lo tanto la probabilidad varía monotónicamente.

Modelo Logit General: Interpretación

- El cambio porcentual en la probabilidad de que la categoría j sea elegida ante un cambio porcentual en z_j determina la cuasi-elasticidad:

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial z_j} z_j = Pr[y_i = j|\cdot]\{1 - Pr[y_i = j|\cdot]\} z_j \delta$$

Modelo Logit General: Interpretación

- Para cualquiera de los tres modelos se puede definir la tasa de probabilidad entre dos alternativas j y k.
- En el modelo multinomial, el logaritmo natural de la tasa de probabilidad entre j y k viene dado por,

$$log\left(\frac{Pr[y_i = j|\cdot]}{Pr[y_i = k|\cdot]}\right) = x_i'(\alpha_j - \alpha_k)$$

- En el modelo condicional, el logaritmo natural de la tasa de probabilidad entre j y k viene dado por,

$$log\left(\frac{Pr[y_i = j|\cdot]}{Pr[y_i = k|\cdot]}\right) = (w_{i,j} - w_{i,k})'\beta$$

Modelo Logit General: Interpretación

- En el modelo general, el logaritmo natural de la tasa de probabilidad entre j y k viene dado por,

$$log\left(\frac{Pr[y_i = j|\cdot]}{Pr[y_i = k|\cdot]}\right) = x_i'(\alpha_j - \alpha_k) + (w_{i,j} - w_{i,k})'\beta + (z_j - z_k)'\delta$$

- Note que la tasa de probabilidad entre la elección de las alternativas j y k no está afectada por el resto de las alternativas.
- Esta propiedad de los tres modelos se conoce como la independencia de alternativas irrelevantes.
- Esto es, comparando las alternativas j y k, las otras opciones son irrelevantes.

Modelo Logit General: Estimación

- La estimación de los parámetros de cualquiera de los modelos descriptos se realiza mediante el método de maximización de la función de verosimilitud.
- La función de verosimilitud para cualquiera de los tres modelos anteriores es la misma, excepto porque la forma funcional de la probabilidad es diferente.
- La función de verosimilitud viene dada por:

$$L(\theta) = \prod_{i=1}^{n} \prod_{j=1}^{J} Pr[y_i = j|\cdot]^{I[y_i=j]}$$

donde $I[\cdot]$ es la función indicador que asume el valor 1 cuando el argumento de la función es verdadero y asume el valor 0 cuando el argumento de la función es falso.

Modelo Logit General: Estimación

- El logaritmo de la función de verosimilitud es:

$$l(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} I[y_i = j] log(Pr[y_i = j|\cdot])$$

- El estimador de máxima verosimilitud del modelo es el valor de $\hat{\theta}$ que maximiza la función anterior.
- Las condiciones de primer orden para un máximo vienen dadas por,

$$\begin{aligned} \frac{\partial l(\hat{\theta})}{\partial \hat{\theta}} &= \sum_{i=1}^{n} \sum_{j=1}^{J} I[y_i = j] \frac{\partial log(Pr[y_i = j|\cdot])}{\partial \hat{\theta}} \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} \frac{I[y_i = j]}{Pr[y_i = j|\cdot]} \frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\theta}} = 0 \end{aligned} \quad (1)$$

Modelo Logit General: Estimación

- Para el modelo general $\hat{\theta}' = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots \hat{\alpha}_{J-1}, \hat{\beta}', \hat{\delta}')$.
- Derivando con respecto a $\hat{\alpha}_j$ tenemos,

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\alpha}_j} = Pr[y_i = j|\cdot](1 - Pr[y_i = j|\cdot])x_i \quad j = 1\,2, \ldots, J-1$$

y

$$\frac{\partial Pr[y_i = l|\cdot]}{\partial \hat{\alpha}_j} = -Pr[y_i = j|\cdot]Pr[y_i = l|\cdot]x_i \quad j = 1\,2, \ldots, J-1 \neq l$$

Modelo Logit General: Estimación

- Reemplazando en (1),

$$
\begin{aligned}
\frac{\partial l(\hat{\theta})}{\partial \hat{\alpha}_j} &= \sum_{i=1}^{n} \sum_{j=1}^{J} \frac{I[y_i = j]}{Pr[y_i = j|\cdot]} \frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\alpha}_j} \\
&= \sum_{i=1}^{n} \left\{ \frac{I[y_i = j]}{Pr[y_i = j|\cdot]} \frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\alpha}_j} \right. \\
&+ \left. \sum_{l \neq j} \frac{I[y_i = l]}{Pr[y_i = l|\cdot]} \frac{\partial Pr[y_i = l|\cdot]}{\partial \hat{\alpha}_j} \right\} \\
&= \sum_{i=1}^{n} \left(I[y_i = j] - Pr[y_i = j|\cdot] \right) x_i = 0 \qquad (2)
\end{aligned}
$$

Modelo Logit General: Estimación

- Derivando con respecto a $\hat{\beta}$ tenemos,

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\beta}} = Pr[y_i = j|\cdot](w_{i,j} - \sum_{l=1}^{J-1} Pr[y_i = l|\cdot]w_{i,l})$$

- Reemplazando en (1),

$$\begin{aligned} \frac{\partial l(\hat{\theta})}{\partial \hat{\beta}} &= \sum_{i=1}^{n}\sum_{j=1}^{J} \frac{I[y_i = j]}{Pr[y_i = j|\cdot]} \frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\beta}} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J} I[y_i = j] \left(w_{i,j} - \sum_{l=1}^{J-1} Pr[y_i = l|\cdot]w_{i,l} \right) = 0 \end{aligned} \tag{3}$$

Modelo Logit General: Estimación

- Finalmente, derivando con respecto a $\hat{\delta}$ tenemos,

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\delta}} = Pr[y_i = j|\cdot](z_j - \sum_{l=1}^{J-1} Pr[y_i = l|\cdot]z_l)$$

- Reemplazando en (1),

$$\begin{aligned}\frac{\partial l(\hat{\theta})}{\partial \hat{\delta}} &= \sum_{i=1}^{n}\sum_{j=1}^{J} \frac{I[y_i = j]}{Pr[y_i = j|\cdot]} \frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\delta}} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J} I[y_i = j]\left(z_j - \sum_{l=1}^{J-1} Pr[y_i = l|\cdot]z_l\right) = 0\end{aligned} \tag{4}$$

Modelo Logit General: Estimación

- Las ecuaciones (2), (3) y (4) representan las condiciones de primer orden para la maximización de la función de verosimilitud.

$$\frac{\partial l(\hat{\theta})}{\partial \hat{\alpha}_j} = \sum_{i=1}^{n} \left(I[y_i = j] - Pr[y_i = j|\cdot] \right) x_i = 0 \quad j = 1, 2, \ldots, J-1$$

$$\frac{\partial l(\hat{\theta})}{\partial \hat{\beta}} = \sum_{i=1}^{n} \sum_{j=1}^{J} I[y_i = j] \left(w_{i,j} - \sum_{l=1}^{J-1} Pr[y_i = l|\cdot] w_{i,l} \right) = 0$$

$$\frac{\partial l(\hat{\theta})}{\partial \hat{\delta}} = \sum_{i=1}^{n} \sum_{j=1}^{J} I[y_i = j] \left(z_j - \sum_{l=1}^{J-1} Pr[y_i = l|\cdot] z_l \right) = 0$$

Modelo Logit General: Estimación

- Es inmediatamente obvio que no se pueden resolver las condiciones de primer orden despejando las incógnitas por tratarse de ecuaciones no lineales en las mismas.
- Para obtener los estimadores de máxima verosimilitud del modelo hay que recurrir a algoritmos no lineales.
- El procedimiento de maximización requiere que se cumplan las condiciones de segundo orden representadas por la siguiente matriz Hesiana:

Modelo Logit General: Estimación

- $$H(\hat{\theta}) = \begin{bmatrix} \frac{\partial^2 l(\cdot)}{\partial \alpha_1^2} & \cdots & \frac{\partial^2 l(\cdot)}{\partial \alpha_1 \partial \alpha_{J-1}} & \frac{\partial^2 l(\cdot)}{\partial \alpha_1 \partial \beta} & \frac{\partial^2 l(\cdot)}{\partial \alpha_1 \partial \delta} \\ \vdots & \ddots & \vdots & \vdots & \vdots \\ \frac{\partial^2 l(\cdot)}{\partial \alpha_{J-1} \partial \alpha_1} & \cdots & \frac{\partial^2 l(\cdot)}{\partial \alpha_{J-1}^2} & \frac{\partial^2 l(\cdot)}{\partial \alpha_{J-1} \partial \beta} & \frac{\partial^2 l(\cdot)}{\partial \alpha_{J-1} \partial \delta} \\ \frac{\partial^2 l(\cdot)}{\partial \beta \partial \alpha_1} & \cdots & \frac{\partial^2 l(\cdot)}{\partial \beta \partial \alpha_{J-1}} & \frac{\partial^2 l(\cdot)}{\partial \beta^2} & \frac{\partial^2 l(\cdot)}{\partial \beta \partial \delta} \\ \frac{\partial^2 l(\cdot)}{\partial \delta \partial \alpha_1} & \cdots & \frac{\partial^2 l(\cdot)}{\partial \delta \partial \alpha_{J-1}} & \frac{\partial^2 l(\cdot)}{\partial \delta \partial \beta} & \frac{\partial^2 l(\cdot)}{\partial \delta^2} \end{bmatrix}$$

- Puede demostrarse que la función de verosimilitud es globalmente cóncava y por lo tanto las condiciones de segundo orden se satisfacen.
- El estimador de máxima verosimilitud del vector θ es insesgado, consistente y eficiente.

Modelo Logit General: Evaluación

- El estimador de máxima verosimilitud tiene distribución normal,

$$\hat{\theta} \xrightarrow{d} N(\theta, [E\{-H(\hat{\theta})\}]^{-1})$$

- Como medidas de bondad del ajuste se pueden usar los $Pseudo - R^2$, como el R^2_{MF} de McFadden:

$$R^2_{MF} = 1 - \frac{l(\hat{\theta})}{l(\hat{\theta}_0)}$$

- Donde, $l(\hat{\theta})$ es el valor del logaritmo de la función de verosimilitud evaluada en los estimadores de MV y $l(\hat{\theta}_0)$ es el valor del logaritmo de la función de verosimilitud de un modelo que tiene solo una constante.

Modelo Logit General: Evaluación

- Otra forma de medir la bondad del ajuste es observar como clasifica a las observaciones el modelo en comparación con los datos realmente observados.
- Para generar esta clasificación debemos estimar las probabilidades de seleccionar cada categoría:

$$\hat{p}_{i,j} = \widehat{Pr[y_i = j|\cdot]} = \frac{e^{x_i'\hat{\alpha}_j + (w_{i,j} - w_{i,J})'\hat{\beta} + (z_j - z_J)'\hat{\delta}}}{1 + \sum_{l=1}^{J-1} e^{x_i'\hat{\alpha}_j + (w_{i,j} - w_{i,J})'\hat{\beta} + (z_j - z_J)'\hat{\delta}}}$$

para $j = 1, 2, \ldots, J-1$, y

$$\hat{p}_{i,J} = \widehat{Pr[y_i = J|\cdot]} = \frac{1}{1 + \sum_{l=1}^{J-1} e^{x_i'\hat{\alpha}_j + (w_{i,j} - w_{i,J})'\hat{\beta} + (z_j - z_J)'\hat{\delta}}}$$

- El siguiente paso consiste en trasladar estas estimaciones en elecciones discretas.

Modelo Logit General: Evaluación

- En la práctica la regla es asignarle a $\hat{y}_i$ el valor j correspondiente a la categoría con la mayor probabilidad. Esto es,

$$\hat{y}_i = j \quad si \ \hat{p}_{i,j} = max\{\hat{p}_{i,1}, \ \hat{p}_{i,2}, \ \ldots \hat{p}_{i,J}\}$$

- Después se puede construir la tabla de predicción-realización.

	Predicción				
Realización	$\hat{y}_i = 1$	$\cdots$	$\hat{y}_i = j$	$\cdots$	$\hat{y}_i = J$
$y_i = 1$	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1J}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$y_i = j$	p_{j1}	$\cdots$	p_{jj}	$\cdots$	p_{jJ}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$y_i = J$	p_{J1}	$\cdots$	p_{Jj}	$\cdots$	p_{JJ}

Modelo Logit General: Evaluación

- La proporción $h = p_{11} + \cdots + p_{jj} + \cdots + p_{JJ}$ se intrepreta como la tasa de aciertos.
- La tasa de aciertos puede ser comparada con una predicción totalmente aleatoria, donde para cada individuo i, la alternativa j se predice con probabilidad igual a la proporción de observaciones en la categoría j en la muestra: n_j/n. La tasa esperada de aciertos con estas predicciones aleatorias es $\hat{q} = \sum_{j=1}^{J} (n_j/n)^2$.
- El modelo general tiene mejores predicciones que las hechas aleatoriamente siempre que

$$z = \frac{h - \hat{q}}{\sqrt{\hat{q}(1-\hat{q})/n}} = \frac{nh - n\hat{q}}{\sqrt{n\hat{q}(1-\hat{q})}}$$

es lo suficientemente grande (mayor a 1.645 al 5% de significación).

Agenda

1. Variables Categóricas No Ordenadas
   - Clasificación de las Variables Explicativas
   - Modelo Logit Multinomial
   - Modelo Logit Condicional
   - Modelo Logit General
   - Modelo Probit Multinomial, Condicional y General
2. Variables Categóricas Ordenadas
   - Modelo Logit Ordenado
   - Modelo Probit Ordenado
3. Variables Categóricas Secuenciales
   - Modelo Logit Secuencial
   - Modelo Probit Secuencial

Modelo Probit Multinomial, Condicional y General

- Finalmente, si en los tres modelos desarrollados, el Multinomial, el Condicional y el General en lugar de $F(\cdot)$ y $f(\cdot)$ se utilizan $\Phi(\cdot)$ y $\phi(\cdot)$ tenemos los modelos Probit Multinomial, Condicional y General, respectivamente.

Agenda

1. Variables Categóricas No Ordenadas
   - Clasificación de las Variables Explicativas
   - Modelo Logit Multinomial
   - Modelo Logit Condicional
   - Modelo Logit General
   - Modelo Probit Multinomial, Condicional y General
2. Variables Categóricas Ordenadas
   - Modelo Logit Ordenado
   - Modelo Probit Ordenado
3. Variables Categóricas Secuenciales
   - Modelo Logit Secuencial
   - Modelo Probit Secuencial

Modelo Logit Ordenado

- Ahora nuestro interés se basa en una variable dependiente discreta ordenada.
- Este tipo de variables surge frecuentemente en marketing cuando una empresa, antes de lanzar un producto al mercado, realiza un estudio para que distintas personas evalúen las características del mismo.
- A las personas se les pide, por ejemplo, que indiquen si el producto les gusta mucho, poco o nada. En este caso tenemos tres categorías ordenadas.
- Otro caso surge cuando los individuos mismos son asignados a categorías ordenadas de acuerdo a su actitud acerca de cierto fenómeno y el objetivo del investigador es ver que variables explican la clasificación de los individuos en esas categorías.

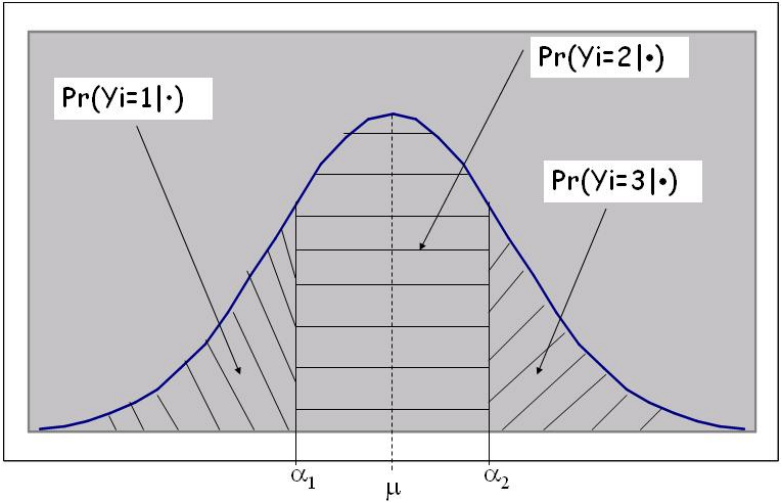
Modelo Logit Ordenado

- Por ejemplo, personas que son clientes del mismo banco o compañía financiera y que son asignadas a tres categorías de acuerdo a su perfil de riesgo.
- Aquellos que solo tienen plazos fijos corresponden a una categoría de bajo riesgo; aquellas personas que compran bonos del gobierno pertenecen a una categoría de mayor riesgo; y aquellos que invierten en la bolsa son las de más alto riesgo.
- Tal como hicimos con los otros modelos asumamos que la variable y_i representa las diferentes categorías a las que, por ejemplo, pueden asignarse los individuos.
- En nuestro caso de los clientes del banco y_i adoptará tres valores: 1 si es un cliente que asume poco riesgo, 2 si es un cliente que asume algo de riesgo y 3 si es un cliente que asume mucho riesgo en sus inversiones.

Modelo Logit Ordenado

- El objetivo de nuestro análisis es relacionar esta variable categórica ordenada con algunas variables explicativas.
- Tal como hicimos en los casos anteriores el ajuste se realizará asumiendo un modelo de probabilidad para la elección de la categoría j.
- A diferencia de lo que ocurría con los modelos multinomiales y condicionales ahora debemos preservar el orden de las categorías de la variable dependiente al calcular las probabilidades.
- Esto se puede lograr, por ejemplo, dividiendo a la distribución de probabilidad que se decida usar en partes ordenadas como sigue:

Modelo Logit Ordenado



Modelo Logit Ordenado

- La figura anterior asigna probabilidades a tres categorías. La generalización al caso de J categorías se realiza de la siguiente manera:
  - Definamos los valores $\alpha_0 < \alpha_1 < \alpha_2 < \cdots < \alpha_{J-1} < \alpha_J$.
  - Usualmente los límites inferior y superior se determinan como $\alpha_0 = -\infty$ y $\alpha_J = +\infty$.
  - Las categorías se asignan de la siguiente manera: $\alpha_{j-1} < y_i < \alpha_j$, implica $y_i = j$, $j = 1, 2, \ldots, J$.
- Esto es,

$$\begin{aligned} Pr[y_i = j|\cdot] &= Pr[\alpha_{j-1} < y_i < \alpha_j] = Pr[\alpha_{j-1} < x_i'\beta + u_i < \alpha_j] \\ &= Pr[\alpha_{j-1} - x_i'\beta < u_i < \alpha_j - x_i'\beta] \\ &= F(\alpha_j - x_i'\beta) - F(\alpha_{j-1} - x_i'\beta) \end{aligned}$$

para $j = 1, 2, 3, \ldots, J$.

Modelo Logit Ordenado

- Note que para que los parámetros de este modelo estén identificados las x_i no deben tener un término constante.
- Cuando la función de distribución utilizada es la logística, el modelo recibe el nombre de Logit Ordenado.
- En este caso,

$$F(\alpha_j - x_i'\beta) = \frac{e^{\alpha_j - x_i'\beta}}{\sum_{j=1}^{J} e^{\alpha_j - x_i'\beta}}$$

- No es obvio que se pueda hacer una interpretación directa de los parámetros del modelo ordenado. El efecto de x_i sobre la probabilidad de elegir la categoría j es claramente una función no lineal de los α_j y los β.

Modelo Logit Ordenado

- La derivada parcial de la probabilidad de selección con respecto a alguna variable explicativa es:

$$\begin{aligned}\frac{\partial Pr[y_i = j|\cdot]}{\partial x_i} &= \frac{\partial F(\alpha_j - x_i'\beta)}{\partial x_i} - \frac{\partial F(\alpha_{j-1} - x_i'\beta)}{\partial x_i} \\ &= [f(\alpha_{j-1} - x_i'\beta) - f(\alpha_j - x_i'\beta)]\beta\end{aligned}$$

donde $f(\cdot)$ denota la función de densidad de la distribución logística.

- Puede verse que esta última expresión no sólo depende del signo de β, sino también del valor de $f(\alpha_{j-1} - x_i'\beta) - f(\alpha_j - x_i'\beta)$.

Modelo Logit Ordenado

- Como en los modelos anteriores la estimación se realiza mediante la maximización de la función de verosimilitud.

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n}\prod_{j=1}^{J} Pr[y_i = j|\cdot]^{I[y_i=j]} \\
&= \prod_{i=1}^{n}\prod_{j=1}^{J} [F(\alpha_j - x_i'\beta) - F(\alpha_{j-1} - x_i'\beta)]^{I[y_i=j]}
\end{aligned}
$$

- Aquí, θ representa a $(\alpha_1, \alpha_2, \ldots, \alpha_{J-1})$ y β. La función indicador $I(y_i = j)$ adopta el valor 1 cuando $y_i = j$ y 0 para los demás casos.

Modelo Logit Ordenado

- El logaritmo de la FV es,

$$l(\theta) = \sum_{i=1}^{n} \sum_{j=1}^{J} I[y_i = j] log\{F(\alpha_j - x_i'\beta) - F(\alpha_{j-1} - x_i'\beta)\}$$

- El estimador de máxima verosimilitud del modelo es el valor de $\hat{\theta}$ que maximiza la función anterior.

Modelo Logit Ordenado

- Las condiciones de primer orden para un máximo vienen dadas por,

$$\begin{aligned}\frac{\partial l(\hat{\theta})}{\partial \hat{\theta}} &= \sum_{i=1}^{n}\sum_{j=1}^{J} I[y_i = j]\frac{\partial log(Pr[y_i = j|\cdot])}{\partial \hat{\theta}} \\ &= \sum_{i=1}^{n}\sum_{j=1}^{J} \frac{I[y_i = j]}{Pr[y_i = j|\cdot]}\frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\theta}} = 0\end{aligned}$$

- Donde,

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\beta}} = [f(\alpha_{j-1} - x_i'\beta) - f(\alpha_j - x_i'\beta)]x_i$$

Modelo Logit Ordenado

- Y,

$$\frac{\partial Pr[y_i = j|\cdot]}{\partial \hat{\alpha}_s} = \begin{cases} f(\alpha_s - x_i'\beta) & \text{si } s = j \\ -f(\alpha_s - x_i'\beta) & \text{si } s = j-1 \\ 0 & \text{en otros casos.} \end{cases}$$

- Reemplazando las expresiones de las derivadas de la probabilidad en la derivada del logaritmo de la función de verosimilitud se obtienen las condiciones de primer orden para el Logit ordenado.
- La función de verosimilitud es globalmente cóncava por lo tanto las condiciones de segundo orden se cumplen.
- Como pasaba en los modelos anteriores, aquí se pueden utilizar como medidas de bondad del ajuste los denominados $Pseudo - R^2$.

Modelo Logit Ordenado

- Uno de esos estadísticos es el R^2_{MF} de McFadden:

$$R^2_{MF} = 1 - \frac{l(\hat{\theta})}{l(\hat{\theta}_0)}$$

- Donde, $l(\hat{\theta})$ es el valor del logaritmo de la función de verosimilitud evaluada en los estimadores de MV y $l(\hat{\theta}_0)$ es el valor del logaritmo de la función de verosimilitud de un modelo que tiene solo una constante.
- Otra forma de medir la bondad del ajuste es observar como clasifica a las observaciones el modelo en comparación con los datos realmente observados.

Modelo Logit Ordenado

- Para generar esta clasificación debemos estimar las probabilidades de seleccionar cada categoría:

$$\hat{p}_{i,j} = \widehat{Pr[y_i = j|\cdot]} = F(\hat{\alpha}_j - x_i'\hat{\beta}) - F(\hat{\alpha}_{j-1} - x_i'\hat{\beta})$$

- El siguiente paso consiste en trasladar estas estimaciones en elecciones discretas.
- En la práctica la regla es asignarle a $\hat{y}_i$ el valor j correspondiente a la categoría con la mayor probabilidad. Esto es,

$$\hat{y}_i = j \quad si \ \hat{p}_{i,j} = max\{\hat{p}_{i,1}, \ \hat{p}_{i,2}, \ \ldots \hat{p}_{i,J}\}$$

Modelo Logit Ordenado

- Una vez estimados estos valores se puede construir la tabla de predicción-realización como hacíamos en el caso de una variable binaria.

	Predicción				
Realización	$\hat{y}_i = 1$	$\cdots$	$\hat{y}_i = j$	$\cdots$	$\hat{y}_i = J$
$y_i = 1$	p_{11}	$\cdots$	p_{1j}	$\cdots$	p_{1J}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$y_i = j$	p_{j1}	$\cdots$	p_{jj}	$\cdots$	p_{jJ}
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
$y_i = J$	p_{J1}	$\cdots$	p_{Jj}	$\cdots$	p_{JJ}

- La proporción $h = p_{11} + \cdots + p_{jj} + \cdots + p_{JJ}$ se intrepreta como la tasa de aciertos.

Modelo Logit Ordenado

- La tasa de aciertos puede ser comparada con una predicción totalmente aleatoria, donde para cada individuo i, la alternativa j se predice con probabilidad igual a la proporción de observaciones en la categoría j en la muestra: n_j/n. La tasa esperada de aciertos con estas predicciones aleatorias es $\hat{q} = \sum_{j=1}^{J} (n_j/n)^2$.
- El modelo general tiene mejores predicciones que las hechas aleatoriamente siempre que

$$z = \frac{h - \hat{q}}{\sqrt{\hat{q}(1-\hat{q})/n}} = \frac{nh - n\hat{q}}{\sqrt{n\hat{q}(1-\hat{q})}}$$

es lo suficientemente grande (mayor a 1.645 al 5% de significación).

Agenda

1. Variables Categóricas No Ordenadas
   - Clasificación de las Variables Explicativas
   - Modelo Logit Multinomial
   - Modelo Logit Condicional
   - Modelo Logit General
   - Modelo Probit Multinomial, Condicional y General
2. Variables Categóricas Ordenadas
   - Modelo Logit Ordenado
   - Modelo Probit Ordenado
3. Variables Categóricas Secuenciales
   - Modelo Logit Secuencial
   - Modelo Probit Secuencial

Modelo Probit Ordenado

- Finalmente, si en lugar de $F(\cdot)$ y $f(\cdot)$ se utilizan $\Phi(\cdot)$ y $\phi(\cdot)$ tenemos el modelo Probit Ordenado.

Agenda

1. Variables Categóricas No Ordenadas
   - Clasificación de las Variables Explicativas
   - Modelo Logit Multinomial
   - Modelo Logit Condicional
   - Modelo Logit General
   - Modelo Probit Multinomial, Condicional y General
2. Variables Categóricas Ordenadas
   - Modelo Logit Ordenado
   - Modelo Probit Ordenado
3. Variables Categóricas Secuenciales
   - Modelo Logit Secuencial
   - Modelo Probit Secuencial

Modelo Logit Secuencial

- Este es un caso fácil de analizar porque la maximización de la función de verosimilitud en este caso particular puede realizarse maximizando la función de verosimilitud de variables binarias que ya estudiamos.
- A modo de ilustración, considere que los individuos pueden ser clasificados en cuatro categorías educativas como sigue:
  - $y_i = 1$ si la persona i tiene hasta educación primaria completa.
  - $y_i = 2$ si la persona i tiene más que educación primaria completa y hasta secundaria completa.
  - $y_i = 3$ si la persona i tiene más que educación secundaria completa y hasta universitaria completa.
  - $y_i = 4$ si la persona i tiene educación post-universitaria.
- Entonces las probabilidades correspondientes a las distintas categorías son:

Modelo Logit Secuencial

- $$\begin{aligned} Pr[y_i = 1|\cdot] &= F(x_i'\beta_1) \\ Pr[y_i = 2|\cdot] &= [1 - F(x_i'\beta_1)] \times F(x_i'\beta_2) \\ Pr[y_i = 3|\cdot] &= [1 - F(x_i'\beta_1)] \times [1 - F(x_i'\beta_2)] \times F(x_i'\beta_3) \\ Pr[y_i = 4|\cdot] &= [1 - F(x_i'\beta_1)] \times [1 - F(x_i'\beta_1)] \times [1 - F(x_i'\beta_3)] \end{aligned}$$

- Entonces, β_1 puede estimarse maximizando un modelo de variable binaria definida adoptando el valor 1 para aquellos que tienen hasta primaria completa y 0 para el resto.
- β_2 puede estimarse maximizando un modelo de variable binaria para una sub-muestra integrada por todas las personas que terminaron la escuela primaria.

Modelo Logit Secuencial

- Para estas personas se define una variable binaria que adopta el valor 1 para aquellos que tienen educación hasta secundaria completa y 0 para el resto.
- β_3 puede estimarse maximizando un modelo de variable binaria para una sub-muestra integrada por todas las personas que terminaron el colegio secundario.
- Para estas personas se define una variable binaria que adopta el valor 1 para aquellos que tienen educación hasta universitaria completa y 0 para el resto.

Agenda

Modelo Probit Secuencial

- Si las probabilidades se definen con $\Phi(\cdot)$ en lugar de $F(\cdot)$ tenemos el modelo Probit Secuencial.

Microeconometría I

Maestría en Econometría

Lecture 4

Agenda

1. M-Estimation
   - Introduction
   - Identification, Uniform Convergence, and Consistency
   - Asymptotic Normality
2. Two-Step M-Estimation
   - Consistency
   - Asymptotic Normality of Two-Step M-Estimators
   - Estimating the Asymptotic Variance
   - Adjustments for when we cannot ignore the first-stage estimation

Agenda

M-Estimation

- M-estimation methods include maximum likelihood, nonlinear least squares, least absolute deviations, quasi-maximum likelihood, and many other procedures used by econometricians.
- In a nonlinear regression model, we have a random variable, y, and we would like to model $E(y|x)$ as a function of the explanatory variables x, a K-vector.
- We already know how to estimate models of $E(y|x)$ when the model is linear in its parameters: OLS produces consistent, asymptotically normal estimators.
- What happens if the regression function is nonlinear in its parameters?

M-Estimation

- Generally, let $m(x;\theta)$ be a parametric model for $E(y|x)$, where m is a known function of x and θ, and θ is a $P \times 1$ parameter vector.
- This is a parametric model because $m(x;\theta)$ is assumed to be known up to a finite number of parameters.
- The dimension of the parameters, P, can be less than or greater than K. The parameter space, Θ, is a subset of $\mathbb{R}^P$
- This is the set of values of y that we are willing to consider in the regression function. Unlike in linear models, for nonlinear models the asymptotic analysis requires explicit assumptions on the parameter space

M-Estimation

- An example of a nonlinear regression function is the logistic function, $m(x;\theta) = \exp(x\theta)/[1+\exp(x\theta)]$. The logistic function is nonlinear in θ.
- We say that we have a correctly specified model for the conditional mean, $E(y|x)$, if, for some $\theta_{\mathrm{o}} \in \Theta$,

$$\mathrm{E}(y \mid \mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}}) \tag{1}$$

- We introduce the subscript "o" on theta to distinguish the parameter vector appearing in $E(y|x)$ from other candidates for that vector.
- Often, the value θ_{o} is called the true value of theta.

M-Estimation

- Equation (1) is the most general way of thinking about what nonlinear least squares is intended to do: estimate models of conditional expectations.
- As a statistical matter, equation (1) is equivalent to a model with an additive, unobservable error with a zero conditional mean:

$$y = m(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}}) + u, \quad \mathrm{E}(u \mid \mathbf{x}) = 0, \tag{2}$$

- Given equation (1), we obtain equation (2) by defining the error to be $u \equiv y - m(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}})$.
- We formalize the first nonlinear least squares (NLS) assumption as follows:
  Assumption NLS.1: For some $\boldsymbol{\theta}_{\mathrm{o}} \in \boldsymbol{\Theta}, \mathrm{E}(y \mid \mathbf{x}) = m(\mathbf{x}, \boldsymbol{\theta}_{\mathrm{o}})$.

M-Estimation

- If we let $\mathbf{w} \equiv (\mathbf{x}, y)$, then $\boldsymbol{\theta}_{\text{o}}$ indexes a feature of the population distribution of $\mathbf{w}$, namely, the conditional mean of y given x.
- More generally, let $\mathbf{w}$ be an M-vector of random variables with some distribution in the population.
- We let $\mathscr{W}$ denote the subset of $\mathbb{R}^M$ representing the possible values of $\mathbf{w}$.
- Let $\boldsymbol{\theta}_{\text{o}}$ denote a parameter vector describing some feature of the distribution of $\mathbf{w}$ (i.e. a conditional mean).
- We assume that $\boldsymbol{\theta}_{\text{o}}$ belongs to a known parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^P$.
- We assume that our data come as a random sample of size N from the population; we label this random sample $\{\mathbf{w}_i : i = 1, 2, \ldots\}$, where each $\mathbf{w}_i$ is an M-vector.

M-Estimation

- What allows us to estimate $\boldsymbol{\theta}_{\mathrm{o}}$ when it indexes $E(y|x)$? It is the fact that $\boldsymbol{\theta}_{\mathrm{o}}$ is the value of $\boldsymbol{\theta}$ that minimizes the expected squared error between y and $m(x;\boldsymbol{\theta})$.
- That is, $\boldsymbol{\theta}_{\mathrm{o}}$ solves the population problem

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathrm{E}\left\{[y - m(\mathbf{x},\boldsymbol{\theta})]^2\right\}, \tag{3}$$

  where the expectation is over the joint distribution of $(\mathbf{x}, y)$.
- Because $\boldsymbol{\theta}_{\mathrm{o}}$ solves the population problem in expression (3), the analogy principle suggests estimating $\boldsymbol{\theta}_{\mathrm{o}}$ by solving the sample analogue.

M-Estimation

- In other words, we replace the population moment $\mathrm{E}\left\{[(y - m(\mathbf{x}, \boldsymbol{\theta})]^2\right\}$ with the sample average.
- The NLS estimator of $\boldsymbol{\theta}_{\mathrm{o}}$, $\hat{\boldsymbol{\theta}}$, solves

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} \left[y_i - m\left(\mathbf{x}_i, \boldsymbol{\theta}\right)\right]^2 \tag{4}$$

  For now, we assume that a solution to this problem exists.
- The NLS objective function in expression (3) is a special case of a more general class of estimators. Let $q(\mathbf{w}, \boldsymbol{\theta})$ be a function of the random vector $\mathbf{w}$ and the parameter vector $\boldsymbol{\theta}$.

M-Estimation

- An M-estimator of $\boldsymbol{\theta}_{\mathrm{o}}$ solves the problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} N^{-1} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right), \tag{5}$$

  assuming that a solution, call it $\hat{\boldsymbol{\theta}}$, exists. The estimator clearly depends on the sample $\{\mathbf{w}_i : i = 1, 2, \ldots\}$, but we suppress that fact in the notation.
- The parameter vector $\boldsymbol{\theta}_{\mathrm{o}}$ is assumed to uniquely solve the population problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})], \tag{6}$$

Agenda

1. M-Estimation
   - Introduction
   - Identification, Uniform Convergence, and Consistency
   - Asymptotic Normality
2. Two-Step M-Estimation
   - Consistency
   - Asymptotic Normality of Two-Step M-Estimators
   - Estimating the Asymptotic Variance
   - Adjustments for when we cannot ignore the first-stage estimation

M-Estimation

- How do we translate the fact that $\boldsymbol{\theta}_{\mathrm{o}}$ solves the population problem (6) into consistency of the M-estimator $\hat{\boldsymbol{\theta}}$ that solves problem (5)?
- Heuristically, the argument is as follows. Since for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\{q(\mathbf{w}_i, \boldsymbol{\theta}) : i = 1, 2, \ldots\}$ is just an i.i.d. sequence, the law of large numbers implies that

$$N^{-1} \sum_{i=1}^{N} q(\mathbf{w}_i, \boldsymbol{\theta}) \xrightarrow{p} \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})], \tag{7}$$

  under very weak finite moment assumptions.
- Since $\hat{\boldsymbol{\theta}}$ minimizes the function on the left hand side of (7) and $\boldsymbol{\theta}_{\mathrm{o}}$ minimizes the function on the right, it seems plausible that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_{\mathrm{o}}$.

M-Estimation

- There are essentially two issues to address.
- The first is identifiability of $\boldsymbol{\theta}_{\text{o}}$, which is purely a population issue.
- The second is the sense in which the convergence in equation (7) happens across different values of $\boldsymbol{\theta}$ in $\boldsymbol{\Theta}$.
- For nonlinear regression, we showed how $\boldsymbol{\theta}_{\text{o}}$ solves the population problem (3). However, we did not argue that $\boldsymbol{\theta}_{\text{o}}$ is always the unique solution to problem (3).
- Whether or not this is the case depends on the distribution of $\mathbf{x}$ and the nature of the regression function:

  Assumption NLS.2: $\mathrm{E}\left\{\left[m\left(\mathbf{x}, \boldsymbol{\theta}_{\text{o}}\right) - m(\mathbf{x}, \boldsymbol{\theta})\right]^2\right\} > 0, \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \boldsymbol{\theta} \neq \boldsymbol{\theta}_{\text{o}}$

- Assumption NLS.2 plays the same role as the assumption of no multicollinearity in OLS.

M-Estimation

- For the general M-estimation case, we assume that $q(\mathbf{w}, \boldsymbol{\theta})$ has been chosen so that $\boldsymbol{\theta}_{\mathrm{o}}$ is a solution to problem (6).
- Identification requires that $\boldsymbol{\theta}_{\mathrm{o}}$ be the unique solution:

$$\mathrm{E}\left[q\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right] < \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})], \qquad \text{all } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}_{\mathrm{o}}, \tag{8}$$

- The second component for consistency of the M-estimator is convergence of the sample average $N^{-1} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right)$ to its expected value.
- It is not enough to simply invoke the usual weak law of large numbers at each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

M-Estimation

- Instead, uniform convergence in probability is sufficient. Mathematically,

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| N^{-1} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}\right) - \mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta})] \right| \xrightarrow{p} 0, \tag{9}$$

- Uniform convergence clearly implies pointwise convergence, but the converse is not true: it is possible for equation (7) to hold but equation (9) to fail.
- To state a formal result concerning uniform convergence, we need to be more careful in stating assumptions about the function $q(\cdot, \cdot)$ and the parameter space $\boldsymbol{\Theta}$.
- Technically, we should assume that $q(\cdot, \boldsymbol{\theta})$ is a Borel measurable function on $\mathscr{W}$ for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

M-Estimation

- The next assumption concerning q is practically more important.
- We assume that, for each$\mathbf{w} \in \mathscr{W}$, $q(\mathbf{w}, \cdot)$ is a continuous function over the parameter space $\boldsymbol{\Theta}$.
- We can now state a theorem concerning uniform convergence appropriate for the random sampling environment. This result, known as the uniform weak law of large numbers (UWLLN), dates back to LeCam (1953).
  Theorem 1 (Uniform Weak Law of Large Numbers): Let $\mathbf{w}$ be a random vector taking values in $\mathscr{W} \subset \mathbb{R}^M$, let $\boldsymbol{\Theta}$ be a subset of $\mathbb{R}^P$ and let $q : \mathscr{W} \times \boldsymbol{\Theta} \to \mathbb{R}$ be a real valued function. Assume that (a) $\boldsymbol{\Theta}$ is compact; (b) for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $q(\cdot, \boldsymbol{\theta})$ is Borel measurable on $\mathscr{W}$; (c) for each $\mathbf{w} \in \mathscr{W}$, $q(\mathbf{w}, \cdot)$ is continuous on $\boldsymbol{\Theta}$; and (d) $|q(\mathbf{w}, \boldsymbol{\theta})| \leq b(\mathbf{w})$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where b is a nonnegative function on $\mathscr{W}$ such that $\mathrm{E}[b(\mathbf{w})] < \infty$. Then equation (9) holds.

M-Estimation

- Theorem 2 (Consistency of M-Estimators): Under the assumptions of Theorem 1, assume that the identification assumption (8) holds. Then a random vector, $\hat{\boldsymbol{\theta}}$, solves problem (5), and $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_{\mathrm{o}}$.
- Lemma 1: Suppose that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_{\mathrm{o}}$, and assume that $r(\mathbf{w}, \boldsymbol{\theta})$ satisfies the same assumptions on $q(\mathbf{w}, \boldsymbol{\theta})$ in Theorem 2. Then

$$N^{-1} \sum_{i=1}^{N} r\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right) \xrightarrow{p} \mathrm{E}\left[r\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right], \tag{10}$$

That is $N^{-1} \sum_{i=1}^{N} r\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right)$ is a consistent estimator of $\mathrm{E}\left[r\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right]$.

Agenda

M-Estimation

- The simplest asymptotic normality proof proceeds as follows.
- Assume that $\boldsymbol{\theta}_{\mathrm{o}}$ is in the interior of $\boldsymbol{\Theta}$, which means that $\boldsymbol{\Theta}$ must have nonempty interior (this assumption is true in most applications). Then, since $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_{\mathrm{o}}, \hat{\boldsymbol{\theta}}$ is in the interior of $\boldsymbol{\Theta}$ with probability approaching one.
- If $q(\mathbf{w}, \cdot)$ is continuously differentiable on the interior of $\boldsymbol{\Theta}$, then (with probability approaching one) $\hat{\boldsymbol{\theta}}$ solves the first-order condition

$$\sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right) = \mathbf{0}, \tag{11}$$

  where $\mathbf{s}(\mathbf{w}, \boldsymbol{\theta})$ is the $P \times 1$ vector of partial derivatives of $q(\mathbf{w}, \boldsymbol{\theta}) : \mathbf{s}(\mathbf{w}, \boldsymbol{\theta})' = \nabla_{\theta} q(\mathbf{w}, \boldsymbol{\theta}) \equiv [\partial q(\mathbf{w}, \boldsymbol{\theta})/\partial \theta_1, \partial q(\mathbf{w}, \boldsymbol{\theta})/\partial \theta_2, \ldots, \partial q(\mathbf{w}, \boldsymbol{\theta})/\partial \theta_P]$. (That is, $\mathbf{s}(\mathbf{w}, \boldsymbol{\theta})$ is the transpose of the gradient of $q(\mathbf{w}, \boldsymbol{\theta})$).
- We call $\mathbf{s}(\mathbf{w}, \boldsymbol{\theta})$ **the score of the objective function** $q(\mathbf{w}, \boldsymbol{\theta})$.

M-Estimation

- If $q(\mathbf{w}, \boldsymbol{\theta})$ is twice continuously differentiable, then each row of the left-hand side of equation (11) can be expanded about $\boldsymbol{\theta}_{\mathrm{o}}$ in a mean-value expansion:

$$\sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \hat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_{\mathrm{o}}\right) + \left(\sum_{i=1}^{N} \ddot{\mathbf{H}}_i\right)\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right), \tag{12}$$

- The notation $\ddot{\mathbf{H}}_i$ denotes the $P \times P$ Hessian of the objective function, $q\left(\mathbf{w}_i, \boldsymbol{\theta}\right)$, with respect to $\boldsymbol{\theta}$, but with each row of $\mathbf{H}\left(\mathbf{w}_i, \boldsymbol{\theta}\right) \equiv \partial^2 q\left(\mathbf{w}_i, \boldsymbol{\theta}\right) / \partial \boldsymbol{\theta} \partial \hat{\boldsymbol{\theta}}' \equiv \nabla_{\theta}^2 q\left(\mathbf{w}_i, \boldsymbol{\theta}\right)$ evaluated at a different mean value.

M-Estimation

- Combining equations (11) and (12) and multiplying through by $1/\sqrt{N}$ gives

$$\mathbf{0} = N^{-1/2} \sum_{i=1}^{N} \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_{\mathrm{o}}\right) + \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i \right) \sqrt{N} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}} \right), \tag{13}$$

- Using Lemma 1 we get $N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i \xrightarrow{p} \mathrm{E}\left[\mathbf{H}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right]$.
- If $\mathbf{A}_{\mathrm{o}} \equiv \mathrm{E}\left[\mathbf{H}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right]$ is nonsingular, then $N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i$ is nonsingular w.p.a. 1 and $\left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i \right)^{-1} \xrightarrow{p} \mathbf{A}_{\mathrm{o}}^{-1}$.
- Therefore, we can write

$$\sqrt{N} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}} \right) = \left(N^{-1} \sum_{i=1}^{N} \ddot{\mathbf{H}}_i \right)^{-1} \left[-N^{-1/2} \sum_{i=1}^{N} \mathbf{s}_i \left(\boldsymbol{\theta}_{\mathrm{o}} \right) \right], \tag{14}$$

  where $\mathbf{s}_i \left(\boldsymbol{\theta}_{\mathrm{o}} \right) \equiv \mathbf{s}\left(\mathbf{w}_i, \boldsymbol{\theta}_{\mathrm{o}}\right)$.

M-Estimation

- Since $o_p(1) \cdot O_p(1) = o_p(1)$ we have,

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right) = \mathbf{A}_o^{-1}\left[-N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i\left(\boldsymbol{\theta}_o\right)\right] + o_p(1), \tag{15}$$

- This is an important equation. It shows that $\sqrt{N}\left(\boldsymbol{\theta} - \boldsymbol{\theta}_o\right)$ inherits its limiting distribution from the average of the scores, evaluated at $\boldsymbol{\theta}_o$. The matrix $\mathbf{A}_0^{-1}$ simply acts as linear transformation.
- Absorbing this linear transformation into $\mathbf{s}_i\left(\boldsymbol{\theta}_o\right)$, we can write

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o\right) = N^{-1/2}\sum_{i=1}^{N}\mathbf{e}_i\left(\boldsymbol{\theta}_o\right) + o_p(1), \tag{16}$$

where $\mathbf{e}_i\left(\boldsymbol{\theta}_o\right) \equiv -\mathbf{A}_o^{-1}\mathbf{s}_i\left(\boldsymbol{\theta}_o\right)$; this is sometimes called the **influence function representation** of $\boldsymbol{\theta}$, where $\mathbf{e}(\mathbf{w}, \boldsymbol{\theta})$ is the influence function.

M-Estimation

- THEOREM 3 (Asymptotic Normality of M-Estimators): In addition to the assumptions in Theorem 2, assume (a) $\boldsymbol{\theta}_{\mathrm{o}}$ is in the interior of $\boldsymbol{\Theta}$; (b)$\mathbf{s}(\mathbf{w}, \cdot)$ is continuously differentiable on the interior of $\boldsymbol{\Theta}$ for all $\mathbf{w} \in \mathscr{W}$; (c) Each element of $\mathbf{H}(\mathbf{w}, \boldsymbol{\theta})$ is bounded in absolute value by a function $b(\mathbf{w})$, where $\mathrm{E}[b(\mathbf{w})] < \infty$; (d) $\mathbf{A}_0 \equiv \mathrm{E}\left[\mathbf{H}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathbf{o}}\right)\right]$ is positive definite; (e) $\mathrm{E}\left[\mathbf{s}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathbf{o}}\right)\right] = \mathbf{0}$; and (f) each element of $\mathbf{s}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathbf{o}}\right)$ has finite second moment. Then

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right) \xrightarrow{d} \text{Normal}\left(0, \mathbf{A}_{\mathrm{o}}^{-1}\mathbf{B}_0\mathbf{A}_{\mathrm{o}}^{-1}\right), \tag{17}$$

  where $\mathbf{A}_{\mathbf{o}} \equiv \mathrm{E}\left[\mathbf{H}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right]$ and $\mathbf{B}_{\mathrm{o}} \equiv \mathrm{E}\left[\mathbf{s}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\mathbf{s}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)'\right] = \mathrm{Var}\left[\mathbf{s}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}\right)\right]$
- Thus,

$$\mathrm{Avar}(\hat{\boldsymbol{\theta}}) = \mathbf{A}_{\mathrm{o}}^{-1}\mathbf{B}_{\mathrm{o}}\mathbf{A}_{\mathrm{o}}^{-1}/N, \tag{18}$$

Two-Step M-Estimators

- Sometimes applications of M-estimators involve a first-stage estimation (an example is OLS with generated regressors).
- Let $\hat{\gamma}$ be a preliminary estimator, usually based on the random sample $\{\mathbf{w}_i : i = 1, 2, \ldots, N\}$.
- A two-step M-estimator θ of θ_0 solves the problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{i=1}^{N} q\left(\mathbf{w}_i, \boldsymbol{\theta}; \hat{\gamma}\right), \tag{19}$$

where q is now defined on $\mathscr{W} \times \boldsymbol{\Theta} \times \boldsymbol{\Gamma}$, and $\boldsymbol{\Gamma}$ is a subset of $\mathbb{R}^J$.

Agenda

1. M-Estimation
   - Introduction
   - Identification, Uniform Convergence, and Consistency
   - Asymptotic Normality
2. Two-Step M-Estimation
   - Consistency
   - Asymptotic Normality of Two-Step M-Estimators
   - Estimating the Asymptotic Variance
   - Adjustments for when we cannot ignore the first-stage estimation

Two-Step M-Estimators

- For the general two-step M-estimator, when will $\hat{\boldsymbol{\theta}}$ be consistent for $\boldsymbol{\theta}_{\mathrm{o}}$?
- In practice, the important condition is the identification assumption.
- To state the identification condition, we need to know about the asymptotic behavior of $\hat{\gamma}$.
- A general assumption is that $\hat{\gamma} \xrightarrow{p} \gamma^*$, where γ^* is some element in Γ.
- The identification condition for the two-step M-estimator is

$$\mathrm{E}\left[q\left(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}; \boldsymbol{\gamma}^*\right)\right] < \mathrm{E}\left[q\left(\mathbf{w}, \boldsymbol{\theta}; \boldsymbol{\gamma}^*\right)\right] \text{ all } \boldsymbol{\theta} \in \boldsymbol{\Theta}, \quad \boldsymbol{\theta} \neq \boldsymbol{\theta}_{\mathrm{o}}, \tag{20}$$

Two-Step M-Estimators

- The consistency argument is essentially the same as that underlying Theorem 2. If $q(\mathbf{w}_i, \boldsymbol{\theta}; \gamma)$ satisfies the UWLLN over $\boldsymbol{\Theta} \times \boldsymbol{\Gamma}$ then expression (19) can be shown to converge to $\mathrm{E}[q(\mathbf{w}, \boldsymbol{\theta}; \gamma^*)]$ uniformly over $\boldsymbol{\Theta}$. Along with identification, this result can be shown to imply consistency of $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}_{\mathrm{o}}$.

Agenda

1. M-Estimation
   - Introduction
   - Identification, Uniform Convergence, and Consistency
   - Asymptotic Normality
2. Two-Step M-Estimation
   - Consistency
   - Asymptotic Normality of Two-Step M-Estimators
   - Estimating the Asymptotic Variance
   - Adjustments for when we cannot ignore the first-stage estimation

Two-Step M-Estimators

- With the two-step M-estimator, there are two cases worth distinguishing.
- The first occurs when the asymptotic variance of $\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right)$ does not depend on the asymptotic variance of $\sqrt{N}\left(\hat{\gamma} - \gamma^{*}\right)$.
- The second occurs when the asymptotic variance of $\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right)$ should be adjusted to account for the first-stage estimation of γ^{*}.
- We first derive conditions under which we can ignore the first-stage estimation error.

Two-Step M-Estimators

- first derive conditions under which we can ignore the first-stage estimation error.
- Using arguments similar to those used to derive the asymptotic normality of M-estimators, it can be shown that, under standard regularity conditions,

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right) = \mathbf{A}_{\mathrm{o}}^{-1}\left(-N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\hat{\boldsymbol{\gamma}}\right)\right) + \mathrm{o}_p(1), \tag{21}$$

  where now $\mathbf{A}_{\mathbf{o}} = \mathrm{E}\left[\mathbf{H}\left(\mathbf{w}, \boldsymbol{\theta}_{\mathbf{o}}; \boldsymbol{\gamma}^*\right)\right]$.
- In obtaining the score and the Hessian, we take derivatives only with respect to $\boldsymbol{\theta}$; $\boldsymbol{\gamma}^*$ simply appears as an extra argument.

Two-Step M-Estimators

- Now if,

$$N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\hat{\boldsymbol{\gamma}}\right)=N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\boldsymbol{\gamma}^{*}\right)+\mathrm{o}_p(1), \tag{22}$$

- Then $\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\mathrm{o}}\right)$ behaves the same asymptotically whether we used $\hat{\gamma}$ or its plim in defining the M-estimator.
- When does equation (22) hold?

Two-Step M-Estimators

- Assuming that $\sqrt{N}(\hat{\gamma} - \gamma^*) = \mathrm{O}_p(1)$ (which is standard).
- A mean value expansion similar to (12) gives

$$N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i(\boldsymbol{\theta}_{\mathrm{o}};\hat{\gamma}) = N^{-1/2}\sum_{i=1}^{N}\mathbf{s}_i(\boldsymbol{\theta}_0;\gamma^*) + \mathbf{F}_{\mathrm{o}}\sqrt{N}(\hat{\gamma} - \gamma^*) + \mathrm{o}_p(1), \quad (23)$$

  where $\mathbf{F}_0$ is the $P \times J$ matrix $\mathbf{F}_{\mathrm{o}} \equiv \mathrm{E}[\nabla_\gamma \mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}; \boldsymbol{\gamma}^*)]$.
- Therefore if

$$\mathrm{E}[\nabla_\gamma \mathbf{s}(\mathbf{w}, \boldsymbol{\theta}_{\mathrm{o}}; \boldsymbol{\gamma}^*)] = \mathbf{0}, \quad (24)$$

  then equation (22) holds.
- And the asymptotic variance of the two-step M-estimator is the same as if $\boldsymbol{\gamma}^*$ were plugged in.

Two-Step M-Estimators

- There are many problems for which assumption (24) does not hold.
- These problems include some the methods for correcting for endogeneity in Probit and Tobit models.
- In such cases we need to make an adjustment to the asymptotic variance of $\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\mathrm{o}}\right)$.
- Assume

$$\sqrt{N}\left(\hat{\gamma} - \gamma^{*}\right) = N^{-1/2} \sum_{i=1}^{N} \mathbf{r}_{i}\left(\gamma^{*}\right) + \mathrm{o}_{p}(1), \tag{25}$$

where $\mathbf{r}_{i}\left(\gamma^{*}\right)$ is a $J \times 1$ vector with $\mathrm{E}\left[\mathbf{r}_{i}\left(\gamma^{*}\right)\right] = \mathbf{0}$.

Two-Step M-Estimators

- Now using equation (23) we can write

$$\sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\mathrm{o}}\right)=\mathbf{A}_{\mathrm{o}}^{-1}N^{-1/2}\sum_{i=1}^{N}\left[-\mathbf{g}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\boldsymbol{\gamma}^*\right)\right]+\mathrm{o}_p(1), \tag{26}$$

  where $\mathbf{g}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\boldsymbol{\gamma}^*\right)\equiv\mathbf{s}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\boldsymbol{\gamma}^*\right)+\mathbf{F}_{\mathrm{o}}\mathbf{r}_i\left(\boldsymbol{\gamma}^*\right)$.
- Since $\mathbf{g}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\boldsymbol{\gamma}^*\right)$ has zero mean, the standardized partial sum in equation (26) can be assumed to satisfy the central limit theorem.
- Define the $P\times P$ matrix

$$\mathbf{D}_{\mathrm{o}}\equiv\mathrm{E}\left[\mathbf{g}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\boldsymbol{\gamma}^*\right)\mathbf{g}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\boldsymbol{\gamma}^*\right)'\right]=\mathrm{Var}\left[\mathbf{g}_i\left(\boldsymbol{\theta}_{\mathrm{o}};\boldsymbol{\gamma}^*\right)\right], \tag{27}$$

- Then

$$\mathrm{Avar}\ \sqrt{N}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_{\mathrm{o}}\right)=\mathbf{A}_{\mathrm{o}}^{-1}\mathbf{D}_{\mathrm{o}}\mathbf{A}_{\mathrm{o}}^{-1}, \tag{28}$$

Agenda

1. M-Estimation
   - Introduction
   - Identification, Uniform Convergence, and Consistency
   - Asymptotic Normality
2. Two-Step M-Estimation
   - Consistency
   - Asymptotic Normality of Two-Step M-Estimators
   - Estimating the Asymptotic Variance
   - Adjustments for when we cannot ignore the first-stage estimation

Two-Step M-Estimators

- We first consider estimating the asymptotic variance of $\hat{\boldsymbol{\theta}}$ in the case where there are no nuisance parameters.
- Under regularity conditions that ensure uniform converge of the Hessian, the estimator

$$N^{-1}\sum_{i=1}^{N}\mathbf{H}\left(\mathbf{w}_i,\hat{\boldsymbol{\theta}}\right)\equiv N^{-1}\sum_{i=1}^{N}\hat{\mathbf{H}}_i, \tag{29}$$

  is consistent for $\mathbf{A}_o$, by Lemma 1.
- By Lemma 1, under standard regularity conditions we have

$$N^{-1}\sum_{i=1}^{N}\mathbf{s}\left(\mathbf{w}_i,\hat{\boldsymbol{\theta}}\right)\mathbf{s}\left(\mathbf{w}_i,\hat{\boldsymbol{\theta}}\right)'\equiv N^{-1}\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'\xrightarrow{p}\mathbf{B}_o. \tag{30}$$

Two-Step M-Estimators

- Combining equations (29) and (30) we can consistently estimate $\text{Avar}\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{o}}\right)$ by

$$\text{Avar}\sqrt{N}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{\text{o}}\right) = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}, \tag{31}$$

- The asymptotic standard errors are obtained from the matrix

$$\hat{\mathbf{V}} \equiv \widehat{\text{Avar}(\hat{\boldsymbol{\theta}})} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}/N. \tag{32}$$

- Which can be expressed as

$$\left(\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}\left(\sum_{i=1}^{N}\hat{\mathbf{s}}_i\hat{\mathbf{s}}_i'\right)\left(\sum_{i=1}^{N}\hat{\mathbf{H}}_i\right)^{-1}. \tag{33}$$

Agenda

1. M-Estimation
   - Introduction
   - Identification, Uniform Convergence, and Consistency
   - Asymptotic Normality
2. Two-Step M-Estimation
   - Consistency
   - Asymptotic Normality of Two-Step M-Estimators
   - Estimating the Asymptotic Variance
   - Adjustments for when we cannot ignore the first-stage estimation

Two-Step M-Estimators

- When assumption (24) is violated, the asymptotic variance estimator of $\hat{\boldsymbol{\theta}}$ must account for the asymptotic variance of $\hat{\boldsymbol{\gamma}}$.
- We need to estimate equation (28).
- We already know how to consistently estimate $\mathbf{A}_{\mathrm{o}}$ using equation (29).
- Estimation of $\mathbf{D}_{\mathrm{o}}$ is also straightforward.
- First we need to estimate $\mathbf{F}_{\mathrm{o}}$,

$$\hat{\mathbf{F}} = N^{-1} \sum_{i=1}^{N} \nabla_{\gamma} \mathbf{s}_i(\hat{\boldsymbol{\theta}}; \hat{\boldsymbol{\gamma}}), \tag{34}$$

Two-Step M-Estimators

- Next, replace $\mathbf{r}_i(\gamma^*)$ with $\hat{\mathbf{r}}_i \equiv \mathbf{r}_i(\hat{\gamma})$.
- Then

$$\hat{\mathbf{D}} \equiv N^{-1} \sum_{i=1}^{N} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' \xrightarrow{p} \mathbf{D}_o, \tag{35}$$

  where $\hat{\mathbf{g}}_i = \hat{\mathbf{s}}_i + \hat{\mathbf{F}}\hat{\mathbf{r}}_i$.
- The asymptotic variance of the two-step M-estimator is,

$$\left(\sum_{i=1}^{N} \hat{\mathbf{H}}_i \right)^{-1} \left(\sum_{i=1}^{N} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' \right) \left(\sum_{i=1}^{N} \hat{\mathbf{H}}_i \right)^{-1}. \tag{36}$$

Microeconometría I

Maestría en Econometría

Lecture 5

Tobit and Selection Models

Agenda

Introduction

- We consider two closely related topics: regression when the dependent variable of interest is incompletely observed and regression when the dependent variable is completely observed but is observed in a selected sample that is not representative of the population.
- This includes limited dependent variable models, latent variable models, generalized Tobit models, and selection models.

Introduction

- All these models share the common feature that even in the simplest case of population conditional mean linear in regressors, OLS regression leads to inconsistent parameter estimates because the sample is not representative of the population.
- Leading causes of incompletely observed data are truncation and censoring
- For truncated data some observations on both the dependent variable and regressors are lost. For example, income may be the dependent variable and only low-income people are included in the sample.

Introduction

- For censored data information on the dependent variable is lost, but not data on the regressors. For example, people of all income levels may be included in the sample, but for confidentiality reasons the income of high-income people may be top-coded and reported only as exceeding, say, \$100,000 per year.
- A leading example of truncation and censoring is the Tobit model, named after Tobin (1958), who considered linear regression under normality.

Introduction

- Let y^* denote a variable that is incompletely observed.
- For truncation from below, y^* is only observed if y^* exceeds a threshold.
- For simplicity, let that threshold be zero.
- Then we observe $y = y^*$ if $y^* > 0$.
- Since negative values do not appear in the sample, the truncated mean exceeds the mean of y^*.

Introduction

- Let y^* denote a variable that is incompletely observed.
- For censoring from below at zero, y^* is not completely observed when $y^* = 0$, but it is known that $y^* < 0$ and for simplicity y is then set to 0.
- Since negative values are scaled up to zero, the censored mean also exceeds the mean of y^*.
- Clearly, sample means in truncated or censored samples cannot be used without adjustment to estimate the original population mean.

Introduction

- With luck, truncation and censoring might lead only to a shift up or down in the intercept, leaving slope coefficients unchanged; however, this is not the case.
- For example, if $E[y^*|x] = x\beta$ in the original model then truncation or censoring leads to $E[y|x]$ being nonlinear in x and β so that OLS gives inconsistent estimates of β and hence inconsistent estimates of marginal effects.

Example

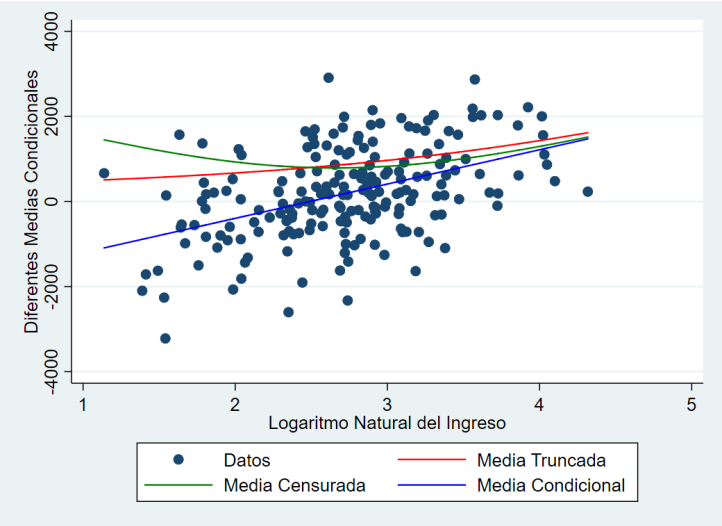
- As an illustration we consider the following labor supply example with simulated data.
- The relationship between desired annual hours worked, y^*, and hourly wage, w, is specified to be of linear-log form with data-generation process:

$$y^* = -2500 + 1000 \ln w + \epsilon,$$

$$\epsilon \sim \text{Normal}(0, 1000^2)$$

$$\ln w \sim \text{Normal}(2.75, 0.60^2)$$

Example


Diferentes Medias Condicionales
4000
2000
0
-2000
-4000
1
2
3
4
5
Logaritmo Natural del Ingreso
Datos
Media Truncada
Media Censurada
Media Condicional

Example

- It is clear that censored and truncated conditional means are nonlinear in x even if the underlying population mean is linear.
- OLS estimation using truncated or censored data will lead to inconsistent estimation of the slope parameter, since by visual inspection of last figure a linear approximation to the nonlinear truncated and censored means will have flatter slope than that for the original untruncated mean.

Agenda

Censoring and Truncation Mechanisms

- Let y denote the observed value of the dependent variable.
- The departure from usual analysis is that y is the incompletely observed value of a latent dependent variable y^*, where the observation rule is

$$y = g(y^*),$$

for some specified function $g(\cdot)$.

Censoring

- With censoring we always observe the regressors x, completely observe y^* for a subset of the possible values of y^*, and incompletely observe y for the remaining possible values of y^*.
- If censoring is from below (or from the left), we observe

$$y = \begin{cases} y^* & \text{If } y^* > L \\ L & \text{If } y^* \leq L \end{cases}$$

- Example: all consumers may be sampled with some having positive durable goods expenditures ($y^* > 0$) and others having zero expenditures ($y^* = 0$).

Censoring

- If censoring is from above (or from the right) we observe

$$y = \begin{cases} y^* \text{ If } y^* < U \\ U \text{ If } y^* \geq U \end{cases}$$

- Example: annual income data may be top-coded at U = \$100,000.

Truncation

- Truncation entails additional information loss as all data on observations at the bound are lost.
- With truncation from below we observe only

$$y = y^* \text{ if } y^* > L.$$

- Example: only consumers who purchased durable goods may be sampled ($L = 0$).

Truncation

- With truncation from above we observe only

$$y = y^* \text{ if } y^* < U.$$

- Example, only low-income individuals may be sampled.

Agenda

Censored and Truncated MLE

- If the conditional distribution of y^* given regressors x is specified, then the parameters of this distribution can be consistently and efficiently estimated by ML estimation based on the conditional distribution of the censored or truncated y.
- Let $f^*(y^*|x)$ and $F^*(y^*|x)$ denote the conditional probability density function (or probability mass function) and cumulative distribution function of the latent variable y^*.

Censored and Truncated MLE

- One can always obtain $f(y|x)$ and $F(y|x)$, the corresponding conditional pdf and cdf of the observed dependent variable y, since $y = g(y^*)$ is a transformation of y^*.
- Consider ML estimation given censoring from below.
- For $y > L$ the density of y is the same as that for y^*, so $f(y|x) = f^*(y|x)$.
- For $y = L$, the lower bound, the density is discrete with mass equal to the probability of observing $y^* \leq L$, or $F^*(L|x)$.

Censored MLE

- Thus for censoring from below

$$f(y|x) = \begin{cases} f^*(y|x) & \text{If } y > L \\ F^*(L|x) & \text{If } y = L \end{cases}$$

- Similar to analysis for binary outcome models, it is notationally convenient to introduce an indicator variable

$$d = \begin{cases} 1 & \text{If } y > L \\ 0 & \text{If } y = L \end{cases} \tag{1}$$

Censored MLE

- Then the conditional density given censoring from below can be written as

$$f(y|x) = f^*(y|x)^d F^*(L|x)^{1-d}$$

- For a sample of N independent observations, the censored MLE maximizes

$$\ln L_N(\theta) = \sum_{i=1}^{N} \{d_i \ln f^*(y_i|x_i, \theta) + (1 - d_i) \ln F^*(L_i|x_i, \theta)\} \tag{2}$$

Censored MLE

- For generality the censoring lower bound L_i is permitted to vary across individuals, though usually $L_i = L$.
- The censored MLE is consistent and asymptotically normal, provided the original density of the uncensored variable $f^*(y^*|x, \theta)$ is correctly specified.
- When censoring is instead from above, the log-likelihood is similar to (2), except now $d = 1$ if $y < U$ and $d = 0$ otherwise, and $F^*(L|x, \theta)$ is replaced by $1 - F^*(U|x, \theta)$.

Truncated MLE

- For truncation from below at L, and suppressing dependence on x, the conditional density of the observed y is

$$\begin{aligned} f(y) &= f^*(y|y > L) \\ &= f^*(y)/Pr[y|y > L] \\ &= f^*(y)/[1 - F^*(L)]. \end{aligned}$$

- The truncated MLE therefore maximizes

$$\ln L_N(\theta) = \sum_{i=1}^{N} \{\ln f^*(y_i|x_i, \theta) - \ln [1 - F^*(L_i|x_i, \theta)]\} \tag{3}$$

Truncated MLE

- If instead truncation is from above, the log-likelihood is (3), except that $[1 - F^*(L|x, \theta)]$ is replaced by $F^*(U|x, \theta)$.
- Ignoring censoring or truncation leads to inconsistency.
- If truncation is ignored the MLE maximizes

$$\sum_i \ln f^*(y_i|x_i, \theta),$$

which is the wrong likelihood function as it drops the second term in (3).

Poisson Truncated MLE Example

- Assume that y^* is Poisson distributed, so that $f^*(y) = e^{-\mu}\mu^y/y!$ and $\ln f^*(y) = -\mu + y \ln \mu - \ln y!$, with mean $\mu = \exp(x'\beta)$.
- Suppose the number of cigarettes smoked is modeled, but data are only available for people who smokes.
- Then the data are truncated from below at zero and we only observe $y = y^*$ if $y^* > 0$.
- Then $F^*(0) = Pr[y^* \leq 0] = Pr[y^* = 0] = e^{-\mu}$.

Poisson Truncated MLE Example

- From (3) the truncated MLE for β maximizes

$$\begin{aligned} \ln L_N(\beta) &= \sum_{i=1}^{N} \{ -\exp x_i'\beta + y x_i'\beta - \ln y! \\ &- \ln [1 - \exp(-\exp(x_i'\beta))] \} \end{aligned}$$

Poisson Truncated MLE Example

- Suppose instead that data are censored from above at 10 because of top-coding, so that we observe $y = y^*$if $y^* < 10$ and that $y = 10$ if $y^* \geq 10$. Then $Pr[y^* \geq 10] = 1 - Pr[y^* < 10] = 1 - \sum_{k=0}^{9} f^*(k)$. Then the censored MLE for β maximizes

$$\begin{aligned} \ln L_N(\beta) &= \sum_{i=1}^{N} \Big\{ d_i[-\exp x_i'\beta + y x_i'\beta - \ln y!] \\ &+ (1 - d_i) \ln [\sum_{k=0}^{9} \exp\left(-\exp\left(x_i'\beta\right)\right)(\exp x_i'\beta)^k / k!] \Big\} \end{aligned}$$

Agenda

Tobit Model

- Truncation and censoring arise most often in econometrics in the linear regression model with normally distributed error, when only positive outcomes are completely observed.
- This model is called the Tobit model after Tobin (1958), who applied it to individual expenditures on consumer durable goods.

Tobit Model

- The censored normal regression model, or Tobit model, is one with censoring from below at zero where the latent variable is linear in regressors with additive error that is normally distributed and homoskedastic.

$$y^* = x'\beta + \epsilon, \tag{4}$$

where

$$\epsilon \sim \text{Normal}[0, \sigma^2]$$

- This implies that the latent variable $y^* \sim N[x'\beta, \sigma^2]$.

Tobit Model

- The observed y is defined by

$$y = \begin{cases} y^* \text{ If } y^* > L \\ L \text{ If } y^* \leq L \end{cases}$$

  with $L = 0$.
- The conditional density given above is

$$f(y|x) = f^*(y|x)^d F^*(L|x)^{1-d}$$

Tobit Model

- with $f^*(y|x) \sim N[x'\beta, \sigma^2]$ and

$$\begin{aligned} F^*(0|x) &= Pr[y^* \leq 0] \\ &= Pr[x'\beta + \epsilon \leq 0] \\ &= \Phi(-x'\beta/\sigma) \\ &= 1 - \Phi(x'\beta/\sigma) \end{aligned}$$

- The censored density can be expressed as

$$f(y) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y - x'\beta)^2\right\}\right]^d \left[1 - \Phi(\frac{x'\beta}{\sigma})\right]^{1-d} \quad (5)$$

Tobit Model

- The Tobit MLE $\hat{\theta} = (\hat{\beta}', \hat{\sigma}^2)'$ maximizes the censored log-likelihood function (2).
- Given (5) this becomes

$$\begin{aligned} \ln L_N(\beta, \sigma^2) &= \sum_{i=1}^{N} \left\{ d_i \left(-\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y_i - x_i'\beta)^2 \right) \right. \\ &- \left. (1 - d_i) \ln \left(\Phi(\frac{x'\beta}{\sigma}) \right) \right\}, \end{aligned} \tag{6}$$

Tobit Model

- The first-order conditions are

$$\frac{\partial \ln L_N}{\partial \beta} = \sum_{i=1}^{N} \frac{1}{\sigma^2} \left(d_i (y_i - x_i'\beta) - (1 - d_i) \frac{\sigma \phi_i}{1 - \Phi_i} \right) x_i = 0 \tag{7}$$

$$\frac{\partial \ln L_N}{\partial \sigma^2} = \sum_{i=1}^{N} \left\{ d_i \left(\frac{1}{2\sigma^2} + \frac{(y_i - x_i'\beta)^2}{2\sigma^4} \right) + (1 - d_i) \frac{\phi_i x_i'\beta}{1 - \Phi_i} \frac{1}{2\sigma^3} \right\} = 0 \tag{8}$$

where $\phi_i = \phi(x_i'\beta/\sigma)$ and $\Phi_i = \Phi(x_i'\beta/\sigma)$.

Tobit Model

- A very major weakness of the Tobit MLE is its heavy reliance on distributional assumptions.
- If the error ϵ is either heteroskedastic or nonnormal the MLE is inconsistent.
- Consistent estimation with heteroskedastic normal errors is possible by specifying a model for heteroskedasticity, say $\sigma_i^2 = exp(z_i'\gamma)$.
- Consistency then requires normal errors and correct specification of the functional form of the heteroskedasticity.

Tobit Model

- If data are truncated, rather than censored, from below at zero then the Tobit MLE $\hat{\theta} = (\hat{\beta}', \hat{\sigma}^2)'$ maximizes the truncated normal log-likelihood function

$$\begin{aligned} \ln L_N(\beta, \sigma^2) &= \sum_{i=1}^{N} \left\{ -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y_i - x_i'\beta)^2 \right. \\ &- \left. \ln \left(1 - \Phi(\frac{x'\beta}{\sigma}) \right) \right\}, \end{aligned} \tag{9}$$

Censored and Truncated Means in Linear Regression

- Censoring and truncation in the linear regression model (4) lead to observed dependent variable y that:
  1. has distribution with conditional mean other than $x'\beta$,
  2. conditional variance other than σ^2 even if ϵ is homoskedastic,
  3. and distribution that is nonnormal even if ϵ is normally distributed.
- Truncated mean. The effects of truncation are intuitively predictable. Left-truncation excludes small values, so the mean should increase, whereas with right-truncation the mean should decrease.
- Since truncation reduces the range of variation, the variance should decrease.

Censored and Truncated Means in Linear Regression

- For left-truncation at zero we only observe y if $y^* > 0$. If we suppress dependence of expectations on x for notational simplicity, the left-truncated mean becomes

$$\begin{aligned} E[y] &= E[y^*|y^* > 0] \\ &= E[x'\beta + \epsilon | x'\beta + \epsilon > 0] \\ &= E[x'\beta | x'\beta + \epsilon > 0] + E[\epsilon | x'\beta + \epsilon > 0] \\ &= x'\beta + E[\epsilon | \epsilon > -x'\beta] \end{aligned} \tag{10}$$

- As expected the truncated mean exceeds $x'\beta$, since $E[\epsilon|\epsilon > c]$ for any constant c will exceed $E[\epsilon]$.

Censored and Truncated Means in Linear Regression

- For data left-censored at zero suppose we observe $y = 0$, rather than merely that $y^* \leq 0$.
- The censored mean is obtained by first conditioning the observable y on the binary indicator d defined in (1) with $L = 0$ and then unconditioning.
- Suppressing dependence on x for notational simplicity again, we have the left-censored mean

$$\begin{aligned} E[y] &= E_d[E_{y|d}[y|d]] \\ &= Pr[d = 0] \times E[y|d = 0] + Pr[d = 1] \times E[y|d = 1] \\ &= 0 \times Pr[y^* \leq 0] + Pr[y^* > 0] \times E[y^*|y^* > 0] \\ &= Pr[y^* > 0] \times E[y^*|y^* > 0] \end{aligned}$$

Censored and Truncated Means in Linear Regression

- Remember $Pr[y^* > 0] = 1 - Pr[y^* \leq 0] = Pr[\epsilon > -x'\beta]$ and $E[y^*|y^* > 0]$ is given by (10).
- Summarizing, for the linear regression model with censoring or truncation from below at zero, the conditional means are given by

$$\text{latent variable:} \quad E[y^*|x] = x'\beta \tag{11}$$

$$\text{left truncated at 0:} \quad E[y|x, y > 0] = x'\beta + E[\epsilon|\epsilon > -x'\beta] \tag{12}$$

$$\text{left censored at 0:} \quad E[y|x] = Pr[\epsilon > -x'\beta]\{x'\beta + E[\epsilon|\epsilon > -x'\beta]\} \tag{13}$$

Censored and Truncated Means in Linear Regression

- It is clear that even though the original conditional mean is linear, censoring or truncation leads to conditional means that are nonlinear so that OLS estimates will be inconsistent.
- One possible approach to take is a parametric one of assuming a distribution for ϵ. This leads to expressions for $E[\epsilon|\epsilon > -x'\beta]$ and $Pr[\epsilon > -x'\beta]$ and hence the truncated or censored conditional mean.

Censored and Truncated Means in the Tobit Model

- For the Tobit model the regression error ϵ is normal
- **Truncated Moments of the Standard Normal:** Suppose $z \sim N[0,1]$. Then the left-truncated moments of z are
  - $E[z|z > c] = \phi(c)/[1 - \Phi(c)]$, and $E[z|z > -c] = \phi(c)/\Phi(c)$,
  - $E[z^2|z > c] = 1 + c\phi(c)/[1 - \Phi(c)]$, and
  - $V[z|z > c] = 1 + c\phi(c)/[1 - \Phi(c)] - \phi(c)^2/[1 - \Phi(c)]^2$

Censored and Truncated Means in the Tobit Model

- Appling this result to (10), the error term has truncated mean

$$\begin{aligned} E[\epsilon|\epsilon > -x'\beta] &= \sigma E[\frac{\epsilon}{\sigma}|\frac{\epsilon}{\sigma} > \frac{-x'\beta}{\sigma}] \\ &= \sigma\phi(\frac{x'\beta}{\sigma})/[\Phi(\frac{x'\beta}{\sigma})] \\ &= \sigma\lambda(\frac{x'\beta}{\sigma}) \end{aligned} \tag{14}$$

where $\lambda(z) = \phi(z)/\Phi(z)$ is the inverse Mills ratio.

Censored and Truncated Means in the Tobit Model

- Then the conditional means in (11)-(13) specialize to

$$\text{latent variable:} \quad E[y^*|x] = x'\beta \tag{15}$$

$$\text{left truncated at 0:} \quad E[y|x, y > 0] = x'\beta + \sigma\lambda(\frac{x'\beta}{\sigma}) \tag{16}$$

$$\text{left censored at 0:} \quad E[y|x] = \Phi(\frac{x'\beta}{\sigma})\{x'\beta + \sigma\phi(\frac{x'\beta}{\sigma})\} \tag{17}$$

Censored and Truncated Means in the Tobit Model

- The variance is similarly obtained. Defining $w = x'\beta/\sigma$, we have

$$\begin{aligned} &\text{latent variable:} && V[y^*|x] = \sigma^2 && (18)\\ &\text{left truncated at 0:} && V[y|x, y > 0] = \sigma^2[1 - w\lambda(w) - \lambda(w)^2] && \\ &\text{left censored at 0:} && V[y|x] = \sigma^2\Phi(w)\{w^2 + w\lambda(w) + && \\ & && \quad +1 - \Phi(w)[w + \lambda(w)]\}^2 && \\ & && && (19) \end{aligned}$$

- Clearly truncation and censoring induce heteroskedasticity, and for truncation $V[y|x] < \sigma^2$ so that truncation reduces variability, as expected.

Marginal Effects in the Tobit Model

- The marginal effect is the effect on the conditional mean of the dependent variable of changes in the regressors.
- This effect varies according to whether interest lies in the latent variable mean $x'\beta$ or the truncated or censored means given in (15)-(17).

$$\text{latent variable:} \quad \partial E[y^*|x]/\partial x = \beta \tag{20}$$

$$\text{left truncated at 0:} \quad \partial E[y|x, y > 0]/\partial x = [1 - w\lambda(w) - \lambda(w)^2]\beta$$

$$\text{left censored at 0:} \quad \partial E[y|x]/\partial x = \Phi(w)\beta \tag{21}$$

Tobit en Stata

[R] tobit — Tobit regression

Syntax

tobit *depvar* [*indepvars*] [*if*] [*in*] [*weight*] , **ll**[(#)] **ul**[(#)] [*options*]

options	description
Model	
noconstant	suppress constant term
* **ll**[(#)]	left-censoring limit
* **ul**[(#)]	right-censoring limit
offset(*varname*)	include *varname* in model with coefficient constrained to 1
SE/Robust	
vce(*vcetype*)	*vcetype* may be **oim**, **robust**, **cluster** *clustvar*, **bootstrap**, or **jackknife**
Reporting	
level(#)	set confidence level; default is **level(95)**
display_options	control spacing and display of omitted variables and base and empty cells
Maximization	
maximize_options	control the maximization process; seldom used
+ **coeflegend**	display coefficients' legend instead of coefficient table

* You must specify at least one of **ll**[(#)] or **ul**[(#)].
+ **coeflegend** does not appear in the dialog box.
indepvars may contain factor variables; see fvvarlist.
depvar and *indepvars* may contain time-series operators; see tsvarlist.
bootstrap, **by**, **jackknife**, **nestreg**, **rolling**, **statsby**, **stepwise**, and **svy** are allowed; see prefix.
Weights are not allowed with the **bootstrap** prefix.
aweights are not allowed with the **jackknife** prefix.
vce() and weights are not allowed with the **svy** prefix.
aweights, **fweight**s, **pweight**s, and **iweight**s are allowed; see weight.
See **[R] tobit postestimation** for features available after estimation.

Sample Selection Models

- Selection may be due to self-selection, with the outcome of interest determined in part by individual choice of whether or not to participate in the activity of interest.
- It can also result from sample selection, with those who participate in the activity of interest deliberately oversampled - an extreme case being sampling only participants.
- In either case, similar issues arise and selection models are usually called sample selection models.

A Bivariate Sample Selection Model (Type II Tobit)

- Let y_1^* denote the outcome of interest. In the standard truncated Tobit model this outcome is observed if $y_1^* > 0$.
- A more general model introduces a different latent variable, y_2^*, and the outcome y_1^* is observed if $y_2^* > 0$.
- For example, y_2^* determines whether or not to work and y_1^* determines how much to work, and $y_1^* \neq y_2^*$ since there are fixed costs to work such as commuting costs that are more important in determining participation than hours of work once working.

A Bivariate Sample Selection Model (Type II Tobit)

- The bivariate sample selection model comprises a participation equation that

$$y_2 = \begin{cases} 1 \text{ if } y_2^* > 0, \\ 0 \text{ if } y_2^* \leq 0 \end{cases} \tag{22}$$

- and a resultant outcome equation that

$$y_1 = \begin{cases} y_1^* \text{ if } y_2^* > 0, \\ - \text{ if } y_2^* \leq 0 \end{cases} \tag{23}$$

- This model specifies that y_1 is observed when $y_2^* > 0$, whereas y_1 need not take on any meaningful value when $y_2^* \leq 0$.

A Bivariate Sample Selection Model (Type II Tobit)

- In general the model can be written for a random draw from the population as

$$y_1 = x_1\beta_1 + u_1 \quad (24)$$
$$y_2 = 1[x\delta_2 + v_2 > 0] \quad (25)$$

- **Assumption:** (a) (x, y_2) is always observed in the population, but y_1 is observed only when $y_2 = 1$; (b) $(u_1; v_2)$ is independent of x with zero mean; (c) $v_2 \sim Normal(0; 1)$; and (d) $E(u_1|v_2) = \gamma_1 v_2$.
- Assumption (d) requires linearity in the population regression of u_1 on v_2. It always holds if $(u_1; v_2)$ is bivariate normal.

A Bivariate Sample Selection Model (Type II Tobit)

- To derive an estimating equation, let $(y_1; y_2; x; u_1; v_2)$ denote a random draw from the population.
- Since y_1 is observed only when $y_2 = 1$, what we can hope to estimate is $E(y_1|x; y2 = 1)$ [along with $P(y_2 = 1|x)$].
- From (24)

$$E(y_1|x; v_2) = x_1\beta_1 + E(u_1|x; v_2) = x_1\beta_1 + E(u_1|v_2) = x_1\beta_1 + \gamma_1 v_2 \qquad (26)$$

- If $\gamma_1 = 0$ (which implies that u_1 and v_2 are uncorrelated) then $E(y_1|x; v_2) = E(y_1|x) = E(y_1|x_1) = x_1\beta_1$.
- In other words, if $\gamma_1 = 0$, then **there is no sample selection problem**, and β_1 can be consistently estimated by OLS using the selected sample.

A Bivariate Sample Selection Model (Type II Tobit)

- If $\gamma_1 \neq 0$, then using iterated expectations in (26)

$$E(y_1|x; y_2) = x_1\beta_1 + \gamma_1 E(v_2|x; y_2) = x_1\beta_1 + \gamma_1 h(x; v_2) \tag{27}$$

  where $h(x; v_2) = E(v_2|x; y_2)$.
- Since the selected sample has $y_2 = 1$, we need only find $h(x; 1)$. But $h(x; 1) = E(v_2|v_2 > -x\delta_2) = \lambda(x\delta_2)$, where $\lambda(\cdot) \equiv \phi(\cdot)/\Phi(\cdot)$ is the inverse Mills ratio.
- We can write

$$E(y_1|x; y_2 = 1) = x_1\beta_1 + \gamma_1\lambda(x\delta_2) \tag{28}$$

A Bivariate Sample Selection Model (Type II Tobit)

- Equation (28) makes it clear that an OLS regression of y_1 on x_1 using the selected sample omits the term $\lambda(x\delta_2)$ and generally leads to inconsistent estimation of β_1.
- Equation (28) also suggests a way to consistently estimate β_1.
- Following Heckman (1979), we can consistently estimate β_1 and γ_1 using the selected sample by regressing y_1 on x_1, and $\lambda(x\delta_2)$.
- The problem is that δ_2 is unknown, so we cannot compute the additional regressor $\lambda(x\delta_2)$.
- However, a consistent estimator of δ_2 is available from the first-stage probit estimation of the selection equation.

A Bivariate Sample Selection Model (Type II Tobit)

- Heckman's Procedure (Heckit estimator): (a) Obtain the probit estimate $\hat{\delta}_2$ from the model

$$P(y_{i2} = 1|x_i) = \Phi(x_i\delta_2) \tag{29}$$

using **all N observations.** Then obtain the estimated inverse Mills ratios $\hat{\lambda}_{i2} \equiv \lambda(x_i\hat{\delta}_2)$ (at least for $i = 1, \ldots, N_1$).
(b) Obtain $\hat{\beta}_1$ and $\hat{\gamma}_1$ from the OLS regression on the selected sample, y_{i1} on x_{i1}; and $\hat{\lambda}_{i2}$; for $i = 1, \ldots, N_1$.

- These estimators are consistent and $\sqrt{N}$ asymptotically normal.

A Bivariate Sample Selection Model (Type II Tobit)

- When $\gamma_1 \neq 0$, obtaining a consistent estimate for the asymptotic variance of $\hat{\beta}_1$ is complicated for two reasons.

1. If $\gamma_1 \neq 0$, then $Var(y1|x; y_2 = 1)$ is not constant. As we know, heteroskedasticity itself is easy to correct for using the robust standard errors. However,
2. we should also account for the fact that $\hat{\hat{\delta}}_2$ is an estimator of δ_2.

- The adjustment to the variance of $(\hat{\beta}_1; \hat{\gamma}_1)$ because of the two-step estimation is cumbersome, it is not enough to simply make the standard errors heteroskedasticity robust.

A Bivariate Sample Selection Model (Type II Tobit)

- Replacing parts (c) and (d) in the Assumption above with the stronger assumption that $(u_1; v_2)$ is bivariate normal with mean zero, $Var(u_1) = \sigma_1^2$, $Cov(u_1; v_2) = \sigma_{12}$, and $Var(v_2) = 1$, then partial maximum likelihood estimation can be used.

$$\begin{bmatrix} v_2 \\ u_1 \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma_1^2 \end{pmatrix}\right] \tag{30}$$

- Given participation and outcome equations, for $y_2^* > 0$ we observe y_1, with probability equal to the probability that $y_2^* > 0$ times the conditional probability of y_1^* given that $y_2^* > 0$.
- For positive y_1 the density of observables is $f^*(y_1^*|y_2^* > 0) \times Pr[y_2^* > 0]$.
- For $y_2^* \leq 0$ all that is observed is that this event has occurred, and the density is the probability of this event occurring.

A Bivariate Sample Selection Model (Type II Tobit)

- The bivariate sample selection model therefore has likelihood function

$$L = \prod_{i=1}^{n} [Pr(y_{2i}^* \leq 0)]^{1-y_{2i}} \{f(y_{1i}|y_{2i}^* > 0) \times Pr[y_{2i}^* > 0]\}^{y_{2i}} \tag{31}$$

  where the first term is the discrete contribution when $y_{2i}^* \leq 0$, since then $y_{2i} = 0$, and the second term is the continuous contribution when $y_2^* i > 0$.
- The classic early application of this model was to labor supply, where y_2 is the unobserved desire or propensity to work, whereas y_1 is actual hours worked.

▸ Go to Example

Sample Selection: Type III Tobit Model

- Consider the case where the selection equation is of the censored Tobit form.
- The population model is

$$y_1 = x_1\beta_1 + u_1 \quad (32)$$
$$y_2 = \max(0; x\delta_2 + v_2) \quad (33)$$

  where (x, y_2) is always observed in the population but y_1 is observed only when $y_2 > 0$.
- A standard example occurs when y_1 is the log of the hourly wage offer and y_2 is weekly hours of labor supply.

Sample Selection: Type III Tobit Model

- **Assumption:** (a) (x, y_2) is always observed in the population, but y_1 is observed only when $y_2 > 0$; (b) $(u_1; v_2)$ is independent of x; (c) $v_2 \sim Normal(0; \tau_2^2)$; and (d) $E(u_1|v_2) = \gamma_1 v_2$.

Sample Selection: Type III Tobit Model

- The starting point is equation (26), just as in the probit selection case.
- Define the selection indicator as $s_2 = 1$ if $y_2 > 0$, and $s_2 = 0$ otherwise.
- Since s_2 is a function of x and v_2, it follows immediately that

$$E(y_1|x; v_2; s_2 = 1) = x_1\beta_1 + \gamma_1 v_2 \tag{34}$$

- This equation means that, if we could observe v_2, then an OLS regression of y_1 on x_1, and v_2 using the selected subsample would consistently estimate $(\beta_1; \gamma_1)$.
- v_2 cannot be observed when $y_2 = 0$ (because when $y_2 = 0$, we only know that $v_2 \leq x\delta_2$, for $y_2 > 0$, $v_2 = y_2 - x\delta_2$.
- If we knew δ_2, we would know v_2 whenever $y_2 > 0$.
- It seems reasonable that, because δ_2 can be consistently estimated by Tobit on the whole sample, we can replace v_2 with consistent estimates.

Sample Selection: Type III Tobit Model

- **Estimation Procedure:** (a) Estimate equation (33) by standard Tobit using all N observations. For $y_{i2} > 0$ (say $i = 1, 2, \ldots N_1$), define

$$\hat{v}_{i2} = y_{i2} - xi\hat{\delta}_2 \tag{35}$$

  (b) Using observations for which $y_{i2} > 0$, estimate $(\beta_1; \gamma_1)$ by the OLS regression: y_{i1} on x_{i1}, and $\hat{v}_{i2}$ $i = 1, 2, \ldots N_1$
- This regression produces consistent and $\sqrt{N}$ asymptotically normal estimators of $(\beta_1; \gamma_1)$.

Sesgo de Selección por Truncamiento Incidental

- En economía un caso emblemático de aplicación de máxima verosimilitud es el modelo de oferta de trabajo de las mujeres (Gronau, 1974; Heckman, 1976). Este modelo consiste de dos ecuaciones, una ecuación de salarios que representa la diferencia entre el salario de mercado de una persona y su salario de reserva en función de características tales como la edad, educación, número de hijos etc.
- La segunda ecuación es una ecuación de horas deseadas de trabajo que depende del salario, de la presencia de hijos pequeños en el hogar, del estado civil, etc.

Sesgo de Selección por Truncamiento Incidental

- El problema del truncamiento es que en la segunda ecuación observamos las horas reales solo si la persona está trabajando. Esto es, solo si el salario de mercado excede al salario de reserva. En este caso se dice que la variable horas en la segunda ecuación está incidentalmente truncada.
- Definiciones: Suponga que y y z tienen una distribución bivariada con correlación ρ. Nosotros estamos interesados en la distribución de y dado que z excede un determinado valor. Esto es, la función de densidad conjunta de y y z es:

$$f(y, z|z > a) = \frac{f(y, z)}{Pr(z > a)}$$

Sesgo de Selección por Truncamiento Incidental

- Teorema 20.4 (Greene, 1997, Cap. 20, página 975): Si y y z tienen una distribución normal bivariada con medias μ_y y μ_z, desvíos estándar σ_y y σ_z y correlación ρ, entonces:

$$E(y|z > a) = \mu_y + \rho\sigma_y\lambda(\alpha_z)$$

$$Var(y|z > a) = \sigma_y^2[1 - \rho^2\delta(\alpha_z)],$$

donde, $\alpha_z = (a - \mu_z)/\sigma_z$, $\lambda(\alpha_z) = \phi(\alpha_z)/[1 - \Phi(\alpha_z)]$ y $\delta(\alpha_z) = \lambda(\alpha_z)[\lambda(\alpha_z) - \alpha_z]$.

Sesgo de Selección por Truncamiento Incidental

- Para poner el ejemplo de la oferta de trabajo de las mujeres en un marco general de análisis, digamos que la ecuación que determina la selección muestral es:

$$z_i^* = \gamma' w_i + u_i,$$

donde z_i^* es la diferencia entre el salario de mercado y el salario de reserva de la persona i.

- La ecuación de interés es,

$$y_i = \beta' x_i + \epsilon_i,$$

donde y_i es la oferta de trabajo (en horas) de la persona i.

- La regla es que y_i es observada solo cuando z_i^* es mayor a cero.

Sesgo de Selección por Truncamiento Incidental

- Asumamos que u_i y ϵ_i tienen distribución normal bivariada con medias cero y correlación ρ. Aplicando el teorema 20.4 tenemos,

$$\begin{aligned} E(y_i|y_i\text{es observada}) &= E(y_i|z_i^* > 0) \\ &= E(y_i|u_i > -\gamma' w_i) \\ &= \beta' x_i + E(\epsilon_i|u_i > -\gamma' w_i) \\ &= \beta' x_i + \rho\sigma_\epsilon \lambda_i(\alpha_u) \\ &= \beta' x_i + \beta_\lambda \lambda_i(\alpha_u) \end{aligned}$$

  donde $\alpha_u = -\gamma' w_i/\sigma_u$ y $\lambda_i(\alpha_u) = \phi(\gamma' w_i/\sigma_u)/\Phi(\gamma' w_i/\sigma_u)$.
- Entonces, la ecuación de interés puede escribirse como,

$$y_i|z_i^* > 0 = \beta' x_i + \beta_\lambda \lambda_i(\alpha_u) + v_i$$

  donde v_i es un término de error con media cero.

Sesgo de Selección por Truncamiento Incidental

- Como queda claro de este desarrollo, estimar por MCC la ecuación de interés usando solo los datos observados, produce estimadores inconsistentes de β por el argumento estándar de variables omitidas (i.e. estamos omitiendo $\lambda_i(\cdot)$).
- Cómo podemos obtener estimaciones consistentes de la ecuación de horas de trabajo utilizando solo los datos observados.
- En principio tenemos un problema similar al de la variable habilidad omitida en la ecuación del salario. En este caso, la variable $\lambda_i(\cdot)$ no es observada.
- Note que aún cuando observáramos $\lambda_i(\cdot)$, MCC no nos daría estimadores eficientes porque los errores de la ecuación de interés, v_i, son heterocedásticos (i.e. $Var(v_i) = \sigma_\epsilon^2(1 - \rho^2\delta_i)$ de acuerdo al teorema 20.4 de Greene).

Sesgo de Selección por Truncamiento Incidental

- Una posible solución es estimar la ecuación de selección para obtener los $\hat{\gamma}$ y construir la variable omitida como $\hat{\lambda}_i = \phi(\hat{\gamma}' w_i)/\Phi(\hat{\gamma}' w_i)$. Luego en un segundo paso estimar la ecuación de interés por MCC en una regresión de y sobre x y $\hat{\lambda}$.
- El único problema de esta solución es que la variable dependiente de la ecuación de selección, z_i^*, no es observada. Lo que podemos observar es $z_i = 1$ si $z_i^* > 0$, es decir si la persona está trabajando; o $z_i = 0$ si $z_i^* < 0$ si la persona no está trabajando.
- Es decir que el modelo que podemos estimar es:

$$z_i = \gamma' w_i + u_i, \tag{36}$$

donde z_i es una variable binaria.

Sesgo de Selección por Truncamiento Incidental

- Heckman (1979) sugirió utilizar el siguiente procedimiento en dos etapas.
  1. Estimar la ecuación de selección usando un Probit para obtener estimaciones de γ. Luego para cada observación de la muestra se construye,

$$\hat{\lambda}_i = \frac{\phi(\hat{\gamma}' w_i)}{\Phi(\hat{\gamma}' w_i)}$$

     También vamos a necesitar $\hat{\delta}_i = \hat{\lambda}_i(\hat{\lambda}_i + \hat{\gamma}' w_i)$.
  2. Estime β y β_λ por MCC en una regresión de y sobre x y $\hat{\lambda}$.

Sesgo de Selección por Truncamiento Incidental

- Para poder hacer inferencia estadística en la regresión del segundo paso hay que tener en cuenta dos problemas: heterocedasticidad en v_i y el hecho de que una de las variables explicativas de la regresión está construída a partir de una estimación anterior.
- Heckman (1979) deriva la verdadera matriz de varianzas y covarianzas de los estimadores del segundo paso.
- Recuerde que $\widehat{Var(v_i)} = \hat{\sigma}_\epsilon^2(1 - \hat{\rho}^2\hat{\delta}_i)$ usando el teorema (20.4) de Greene. Donde $\hat{\sigma}_\epsilon^2 = \frac{e'e}{n} + \bar{\delta}\hat{\beta}_\lambda^2$, y $\bar{\delta} = \frac{1}{n}\sum_i \hat{\delta}_i$.
- Entonces la estimación correcta de la matriz de varianzas y covarianzas del segundo paso es,

$$Var[\hat{\beta}, \hat{\beta}_\lambda] = \hat{\sigma}_\epsilon^2(\sum_i x_i^{*'} x_i^*)^{-1}\left[\sum_i (1 - \hat{\rho}^2\hat{\delta}_i)x_i^* x_i^{*'} + Q\right](\sum_i x_i^{*'} x_i^*)^{-1}$$

Sesgo de Selección por Truncamiento Incidental

- Donde,

$$Q = \hat{\rho}^2(X^{*\prime}\Delta W)[Avar(\hat{\gamma})](W'\Delta X^*)$$

y Δ es una matriz diagonal con $\hat{\delta}_i$ en la diagonal principal.

▸ Return to Sample Selection

Microeconometría I

Maestría en Econometría

Lecture 6

Duration Models

Agenda

Introduction

- Econometric models of durations are models of the length of time spent in a given state before transition to another state.
- A **state** is a classification of an individual entity at a point in time, **transition** is movement from one state to another, and a **spell length or duration** is the time spent in a given state.
- A typical regression example is determining the effect of higher unemployment benefit levels on the average length of an unemployment spell or the probability of transition out of unemployment.

Introduction

- Main problems/characteristics:
  1. Several related distributional functions are of interest and either the duration or probability of transition may be modeled
  2. Many different sampling schemes are possible and statistical inference depends on both the duration model and the sampling scheme (**flow sampling vs. stock sampling**).
  3. Data on spell duration are often censored.
  4. Transition data can have several states, such as unemployment, part-time employment, full-time employment, and out-of-the labor force, and data for a given individual may be available on multiple transitions among these states.

Moving into Employment

- Suppose that we are interested in modelling the process of movement into employment for someone who is looking for a job.
- One way of building a model for this transition is to suppose that as a result of his search he receives offers for jobs from time to time.
- Of the possible offers he should get, some he would find worth accepting and some not.
- Then, whether he moves into employment on any day depends upon he receives an offer that day and whether, if he does, he deems it worth accepting.

Moving into Employment

- An economist could then develop this approach by asking, and solving, the questions:
  - What set of wage offers it is optimal for a given person to accept? and,
  - How much resources should optimally be devoted to search?
- Answers to these questions depend not only on the criterion of optimality chosen by the economist but also on the circumstances of the unemployed person, in particular, the resources he possesses and the constraints he faces.

Moving into Employment

- Suppose that the relevant circumstances of a person who has been looking for a job for t days are assembled in a vector $x(t)$.
- Elements of this vector might include, for example, the level of unemployment benefit available after t, or the average wage payable in jobs that might be offered to him.
- For a person described at t by $x(t)$ suppose his optimal amount of search produces a probability, say, $\lambda(x(t))\,dt$, that a job offer will be made to him in the interval of time from t to $t + dt$, and there is a probability $P(x(t))$ that if such offer is received it will be worth accepting.

Moving into Employment

- Then, the outcome of this model-building effort is a quantity

$$\lambda(x(t))P(x(t))\,dt = \theta(x(t))\,dt$$

  describing the probability of a transition out of unemployment in the time interval $(t,\ t + dt)$.
- The function $\theta(x(t))$ is called a hazard function, when, as in our example, there is only a single destination state (employment).
- There is an analogous set of functions when there are multiple destinations and in that context they are called transition intensities.
- Notice that, the choice variables in the transition model are the rate of search and the set of acceptable wage offers and these are not necessarily observable.

Moving into Employment

- Even if they are not observable we will be able to use the theory because of its implications for the hazard function.
- Even in the case where all relevant x were known to the investigator he still would not be able to say with certainty whether a transition will occur.
- Neither the econometrician nor the unemployed person can say for sure whether a transition will occur today. Thus the transition model is intrinsically stochastic.

Agenda

Definitions

- Let $T \geq 0$ denote the duration, which is the time elapsed until a certain event occurs, with some distribution in the population.
- t denotes a particular value for T.
- T is the time at which a person leaves the initial state. For example, if the initial state is unemployment, as in our motivation example, T would be the time, measured in, say, days, weeks or months, until a person becomes employed.
- The cumulative distribution function (cdf) of T is defined as

$$F(t) = P(T < t), \quad t \geq 0$$

Definitions

- The survivor function is defined as

$$S(t) \equiv 1 - F(t) = P(T \geq t), \quad t \geq 0$$

  and this is the probability of "surviving" past time t.
- Denote the density of T by

$$f(t) = \frac{dF}{dt}(t)$$

Definitions

- For $dt > 0$, $P(t \leq T < t + dt | T \geq t)$ is the probability of leaving the initial state in the interval $[t,\ t + dt)$ given survival up until time t.
- If we divide this probability by dt we get the average probability of leaving per unit time period over a short interval after t.
- Considering shorter and shorter intervals we formally define

$$\theta(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt} \tag{1}$$

  as the hazard function.
- The hazard function is the instantaneous rate of leaving per unit time period at t.

Definitions

- From equation (1) it follows that for "small" dt,

$$P(t \leq T < t + dt | T \geq t) \approx \theta(t)\, dt$$

- The rough interpretation of the function θ is that $\theta(t)\, dt$ is the probability of exit from a state in a short interval of length dt after t conditional on the state still being occupied at t.
- We can express the hazard function in terms of the density and cdf of T using the law of conditional probability,

$$P(t \leq T < t + dt | T \geq t) = \frac{P(t \leq T < t + dt)}{P(T \geq t)} = \frac{F(t + dt) - F(t)}{1 - F(t)}$$

Definitions

- Dividing by dt and letting dt goes to zero, we get

$$\begin{aligned}\theta(t) &= \lim_{dt\to 0} \frac{F(t+dt)-F(t)}{dt}\frac{1}{1-F(t)} = F'(t)\frac{1}{1-F(t)} \\ &= \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)}\end{aligned}$$

- Notice that the derivative of $S(t)$ is $-f(t)$, therefore we have

$$\theta(t) = -\frac{d\ log\ S(t)}{dt} \tag{2}$$

Definitions

- Equation (2) is a differential equation in t whose solution, subject to the initial condition $S(0) = 1$ (or $F(0) = 0$) is

$$S(t) = e^{-\int_0^t \theta(s)\, ds}, \tag{3}$$

- Equation (3) shows how one can calculate the probability distribution of duration of state occupancy given the hazard function.
- From the definition of the survivor function and equation (3) we have,

$$F(t) = 1 - e^{-\int_0^t \theta(s)\, ds}, \tag{4}$$

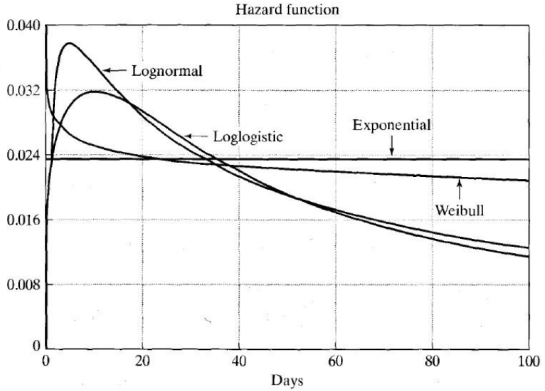
Definitions

- Straightforward differentiation of equation (4) gives the density of T as,

$$f(t) = \theta(t)\, e^{-\int_0^t \theta(s)\, ds}, \tag{5}$$

- Therefore, all probabilities can be computed using hazard functions.
- θ, f and S are alternative ways of describing the distribution of the probability of exit.
- Next figure shows some examples of hazard functions.

Definitions



Definitions

- The hazard function provides a convenient definition of duration dependence.
- Positive duration dependence exists at some point, say, t^* if $d\theta(t)/dt > 0$ at $t = t^*$.
- Positive duration dependence means that the probability of exiting the initial state increases the longer one is in the initial state.
- Negative duration dependence exists at some point, say, t^* if $d\theta(t)/dt < 0$ at $t = t^*$.
- Negative duration dependence means that the probability of exiting the initial state decreases the longer one is in the initial state.

Definitions

- The integrated hazard

$$\Theta(t) = \int_0^t \theta(s)\, ds \tag{6}$$

is also a useful function in practice. however, it does not have a convenient interpretation.

Agenda

Discrete data

- It is very common for a duration to be measured as an interval. For example, data may indicate that a transition occurred in a particular week, but the exact time in the week is not given.
- In such cases the transition times are said to be grouped and it is assumed that the hazard within the interval is constant.
- **Discrete-time hazard models** deal with such data.

Discrete data

- The starting point is to define the **discrete-time hazard function** as the probability of transition at discrete time $t_j, j = 1, 2, \ldots$, given survival to time t_j :

$$\begin{aligned} \lambda_j &= \Pr\left[T = t_j \mid T \geq t_j\right] \\ &= f^{\mathrm{d}}\left(t_j\right) / S^{\mathrm{d}}\left(t_{j_-}\right) \end{aligned} \tag{7}$$

  where the superscript d denotes discrete, and where $S^{\mathrm{d}}\left(a_-\right) = \lim_{t \to a_-} S^{\mathrm{d}}\left(t_j\right)$, an adjustment made because formally $S^{\mathrm{d}}(t)$ equals $\Pr[T > t]$ rather than $\Pr[T \geq t]$.

Discrete data

- The **discrete-time survivor function** is obtained recursively from the hazard function as

$$\begin{aligned} S^{\mathrm{d}}(t) &= \Pr[T \geq t] \\ &= \prod_{j|t_j \leq t} (1 - \lambda_j) . \end{aligned} \tag{8}$$

- For example, $\Pr[T > t_2]$ equals the probability of no transition at time t_1 times the probability of no transition at time t_2 conditional on surviving to just before t_2, so that $\Pr[T > t_2] = (1 - \lambda_1) \times (1 - \lambda_2)$. The function $S^{\mathrm{d}}(t)$ is a decreasing step function with steps at $t_j, j = 1, 2, \ldots$

Discrete data

- The **discrete-time cumulative hazard function** is

$$\Lambda^{\mathrm{d}}(t) = \sum_{j|t_j \leq t} \lambda_j \tag{9}$$

Using (7), we have that the discrete probability that the spell ends at t_j is $\lambda_j S^{\mathrm{d}}(t_j)$

Agenda

Censoring mechanisms

- Survival data are usually censored, as some spells are incompletely observed.
- As an example, instead of observing the length of completed spell of unemployment, data may come from a survey of the currently unemployed, so that only the length of an incomplete spell of unemployment is observed.
- In practice data may be **right-censored, left-censored, or interval-censored**.
- For **right-censoring or censoring from above**, we observe spells from time 0 until a censoring time c. Some spells will have ended by this time anyway (completed spells), but others will be incomplete and all we know is that they will end some time in the interval (c, ∞).

Censoring mechanisms

- **Left-censoring or censoring from below** occurs when spells are known to end at some time in the interval $(0,\ c)$ but the exact time is unknown.
- The classical Tobit model is an example, where data on some spells are lost and the censoring time is unknown.
- **Interval-censoring** occurs when the completed spell length is observed but only in interval form such as in $[t_1^*,\ t_2^*]$.
- **Random censoring or exogenous censoring** means that each individual in the sample has a completed duration T_i^* and censoring time C_i^* that are independent of each other.
- We observe the completed duration T_i^* if the spell ends before the censoring time and the censoring time C_i^* if the spell ends after the censoring time.

Censoring mechanisms

- For standard survival analysis methods to be valid in the presence of censoring the censoring mechanism needs to be one with **independent (noninformative) censoring.**
- This means that parameters of the distribution of C_i^* are not informative about the parameters of the distribution of the duration T_i^*.
- Then one may treat the censoring indicator as exogenous, and it is then not necessary to model the censoring mechanism if interest lies in the duration model parameters.

Agenda

Nonparametric estimation

- Nonparametric estimation of survival functions are very useful for descriptive purposes.
- These methods are used to know the shape of the raw (unconditional) hazard or survival function before considering introducing regressors.
- Let $t_1 < t_2 < \cdots < t_j < \cdots < t_k$ denote the observed discrete failure times of the spells in a sample of size N, $N \geq k$.
- Define d_j to be the number of spells that end at time t_j (since the data are discrete d_j may exceed one).
- Define m_j to be the number of spells right-censored in the interval $[t_j, t_{j+1})$.
- The censoring mechanism is assumed to be independent censoring, so the only thing known about a spell censored in $[t_j, t_{j+1})$ is that the failure time is greater than t_j.

Nonparametric estimation

- Spells are **at risk** of failure if they have not yet failed or been censored.
- Define r_j to equal the number of spells at risk at time t_{j-}, that is, just before time t_j. Then $r_j = (d_j + m_j) + \cdots + (d_k + m_k) = \sum_{l|l \geq j} (d_l + m_l)$. Note that $r_1 = N$.
- In summary
  $d_j = \#$ spells ending at time t_j,
  $m_j = \#$ spells censored in $[t_j, t_{j+1})$,
  $r_j = \#$ spells at risk at time $t_{j-} = \sum_{l|l \geq j} (d_l + m_l)$.

Nonparametric estimation

- Using equation (7) an obvious estimator of the hazard function is the number of spells ending at time t_j divided by the number at risk of failure at time t_{j-}, or

$$\widehat{\lambda}_j = \frac{d_j}{r_j}$$

- The Kaplan-Meier estimator or product limit estimator of the survivor function is the sample analogue of equation (8)

$$\widehat{S}(t) = \prod_{j|t_j \leq t} \left(1 - \widehat{\lambda}_j\right) = \prod_{j|t_j \leq t} \frac{r_j - d_j}{r_j}. \tag{10}$$

- his is a decreasing step function with jump at each discrete failure time.

Nonparametric estimation

- In the case of no censoring $\widehat{S}(t)$ in (10) simplifies to $\widehat{S}(t) = r/N$, the number still at risk at time t divided by the sample size, which is one minus the empirical cdf.
- The discrete-time cumulative hazard function is defined in (9). The **Nelson-Aalen estimator** of the cumulative hazard function is the obvious sample analogue

$$\widehat{\Lambda}(t) = \sum_{j|t_j \leq t} \widehat{\lambda}_j = \sum_{j|t_j \leq t} \frac{d_j}{r_j}$$

Agenda

Some Distributions

Exponential Distribution

- For the exponential distribution with parameter $\theta > 0$ we have,

$$\begin{aligned} F(t) &= 1 - e^{-\theta t} \\ S(t) &= e^{-\theta t} \\ f(t) &= \theta e^{-\theta t} \\ \theta(t) &= \theta \\ \Theta(t) &= \theta t \end{aligned}$$

  - It is termed memoryless, because the hazard function is constant and so reflects no duration dependence.
  - The probability of exit from a state does not depends on how long it has been occupied.

Some Distributions

Weibull Distribution

- For the Weibull distribution with parameters $\alpha > 0$ and $\theta > 0$,

$$\begin{aligned} F(t) &= 1 - e^{-\theta t^{\alpha}} \\ S(t) &= e^{-\theta t^{\alpha}} \\ f(t) &= \theta \alpha t^{\alpha-1} e^{-\theta t^{\alpha}} \\ \theta(t) &= \theta \alpha t^{\alpha-1} \\ \Theta(t) &= \theta t^{\alpha} \end{aligned}$$

- It can be thought of as an exponential distribution on a re-scaled time axis (i.e. t^{α} has an exponential distribution with parameter θ).

Some Distributions

- Any hazard function can be transformed into the constant hazard by a transformation of the time scale.

Log-Logistic Distribution

- For the Log-Logistic distribution with parameters $\alpha > 0$ and $\theta > 0$,

$$\begin{aligned} F(t) &= 1 - [1/(1+\theta t^{\alpha})] \\ S(t) &= 1/(1+\theta t^{\alpha}) \\ f(t) &= \theta\alpha t^{\alpha-1}/(1+\theta t^{\alpha})^2 \\ \theta(t) &= \theta\alpha t^{\alpha-1}/(1+\theta t^{\alpha}) \\ \Theta(t) &= \ln(1+\theta t^{\alpha}) \end{aligned}$$

Agenda

Time-Invariant Covariates

- Usually in economics one is interested in hazard functions conditional on a set of covariates or regressors.
- When covariates do not change over time the hazard and all other features of T can be specified conditional on the covariates.
- The Conditional Hazard is

$$\theta(t;\mathbf{x}) = \lim_{dt\to 0} \frac{P(t \leq T < t+dt | T \geq t, \mathbf{x})}{dt} \tag{11}$$

  where $\mathbf{x}$ is a vector of explanatory variables.
- All of the formulas defined previously continue to hold provided the cdf and density are defined conditional on $\mathbf{x}$. For example,

$$\theta(t;\mathbf{x}) = \frac{f(t|\mathbf{x})}{1-F(t|\mathbf{x})}$$

  where $f(\cdot|\mathbf{x})$ is the density of T given $\mathbf{x}$.

Time-Invariant Covariates

- An important class of models with time-invariant regressors are the proportional hazard models.
- A proportional hazard can be written as

$$\theta(t;\mathbf{x}) = \kappa(\mathbf{x})\theta_0(t)$$

  where $\kappa(\cdot)$ is a nonnegative function of $\mathbf{x}$ and $\theta_0(t) > 0$ is called the baseline hazard.
- The baseline hazard is common to all units in the population.
- Individual hazard function differ proportionately based on a function $\kappa(\mathbf{x})$ of observed covariates.

Time-Invariant Covariates

- Typically, $\kappa(\cdot)$ is parameterized as $\kappa(\mathbf{x}) = e^{\mathbf{x}\beta}$, where β is a vector of parameters.
- Then,

$$\ln \theta(t; \mathbf{x}) = \mathbf{x}\beta + \ln \theta_0(t)$$

  and $\beta_{\mathbf{j}}$ measures the semi-elasticity of the hazard with respect to $\mathbf{x}_{\mathbf{j}}$.
- One of the most used proportional hazard (PH) model is the Cox PH model
  - is a semiparametric model
  - makes no assumptions about the form of $\theta_0(t)$ (nonparametric part of model)
  - assumes parametric form for the effect of the predictors on the hazard
  - In most situations, we are more interested in the parameter estimates than the shape of the hazard

Agenda

Time-Varying Covariates

- Studying hazard functions is more complicated when we wish to model the effects of time-varying covariates on the hazard.
- It makes no sense to specify the distribution of the duration T conditional on the covariates at only one time period.
- Let $\mathbf{x}(\mathbf{t})$ denote the vector of regressors at time t.
- For $t \geq 0$, let $\mathbf{X}(\mathbf{t})$ denote the covariate path up through time t: $\mathbf{X}(\mathbf{t}) \equiv \{\mathbf{x}(\mathbf{s}) : 0 \leq s \leq t\}$.
- Following Lancaster (1990, chapter 2), we define the conditional hazard function at time t by

$$\theta(t; \mathbf{X}(\mathbf{t})) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t, \mathbf{X}(\mathbf{t} + \mathbf{dt}))}{dt} \tag{12}$$

assuming that this limit exists.

Time-Varying Covariates

- By definition, the covariates are sequentially exogenous because, by specifying $\theta(t; \mathbf{X}(\mathbf{t}))$ we are conditioning on current and past covariates.
- One case where this limit exists very generally occurs when T is continuous and for each t, $\mathbf{x}(\mathbf{t} + \mathbf{dt})$ is constant for all $dt \in [0, \eta(t)]$ for some function $\eta(t) > 0$.
- In this case we can replace $\mathbf{X}(\mathbf{t} + \mathbf{dt})$ with $\mathbf{X}(\mathbf{t})$ in equation (12) because $\mathbf{X}(\mathbf{t} + \mathbf{dt}) = \mathbf{X}(\mathbf{t})$ for dt sufficiently small.

Time-Varying Covariates

- In practice, since the interval of observation is discrete we will have to assume that time varying covariates are constant over this interval in which case there is no problem in defining equation (12).
- It is important to know if time-varying covariates are strictly exogenous. Lancaster (1990) provides a definition that rules out feedback from the duration to future values of the covariates.
- If $\mathbf{X}(\mathbf{t}, \mathbf{t}+\mathbf{dt})$ denotes the covariate path from time t to $t+dt$, then Lancaster's exogeneity condition is

$$P[\mathbf{X}(\mathbf{t}, \mathbf{t}+\mathbf{dt}) | T \geq t+dt, \mathbf{X}(\mathbf{t})] = P[\mathbf{X}(\mathbf{t}, \mathbf{t}+\mathbf{dt}) | \mathbf{X}(\mathbf{t})] \tag{13}$$

for all $t \geq 0$, $dt \geq 0$.

Time-Varying Covariates

- The definition of strict exogeneity applies to covariates whose entire path is well-defined whether or not the person is in the initial state.
- One example are the so called external covariates (Kalbfleisch and Prentice, 1980) having the feature that the covariance path is independent of whether any person has or has not left the initial state. The city labor force participation in the moving into employment case.
- Other covariates are not external to each individual but have paths that are still defined after the person leaves the initial state. For example, marital status is well defined after someone becomes employed, but it is possibly related to whether someone has been unemployed. Whether marital status satisfy condition (13) is an empirical issue.

Time-Varying Covariates

- The definition of strict exogeneity cannot be applied to time-varying covariates whose path is not defined once the person leaves the initial state.
- These are internal covariates. For example consider an example of job tenure duration where a time-varying covariate is wage paid in that job. It makes no sense to define the future wage path in that job.
- It is clear that internal covariates cannot satisfy any reasonable strict exogeneity assumption.
- This fact is important when estimating duration models with unobserved heterogeneity.

Agenda

Single-Spell data

- Assume the population of interest is individuals entering the initial state during a given interval of time, say $[0, b]$, where $b > 0$ is a known constant.
- By convention, let zero denote the earliest calendar date that an individual can enter the initial state, and b is the last possible date.
- For example, if we are interested in the population of workers who became unemployed at any time during 2005, and unemployment duration is measured in years (with 0.5 meaning half a year) then $b = 1$. If duration is measured in weeks, then $b = 52$; if duration is measured in days, then $b = 365$.

Single-Spell data

- We restrict attention to single-spell data. That is, we use, at most, one completed spell per individual. If after leaving the initial state, an individual subsequently reenters the initial state in the interval $[0, b]$, we ignore this information.
- Covariates in this analysis are time invariant, meaning that we collect covariates on individuals at a given point in time -usually, at the beginning of the spell- and we do not re-collect data on the covariates during the course of the spell.
- With flow sampling we sample individuals who enter the state at some point during the interval $[0, b]$, and we record the length of time each individual is in the initial state.

Single-Spell data and Flow sampling

- We collect data on covariates known at the time the individual entered the initial state.
- Suppose we are interested in the population of workers who became unemployed at any time during 2005. Then, we randomly sample from the, say, population of male workers who became unemployed during 2005.
- At the beginning of the unemployment spell we might obtain information on tenure in last job, wage on last job, gender, marital status, unemployment benefits etc.

Single-Spell data and Flow sampling

- There are two common ways to collect data on unemployment spells.
- First: we may randomly sample individuals from a large population, say, all working-age individuals in the country for a given year, say, 2005.
- Some fraction of these people will be in the labor force and will become unemployed during 2005 -and this group of people who become unemployed is our random sample of all workers who become unemployed during 2005.
- Second: retrospective sampling.
- Suppose that, for a given province in the country we have access to unemployment records for 2005. We can obtain a random sample of all workers who become unemployed during 2005.

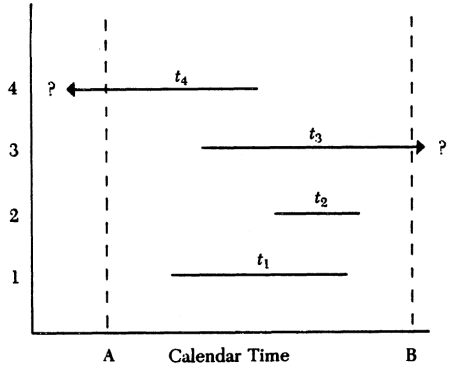
Single-Spell data and Flow sampling

- Flow data are usually subject to right censoring.
- That is, after a certain amount of time, we stop following individuals in the sample.
- For individuals who have completed their spells in the initial state, we observe the exact duration. But for those still in the initial state, we only know that the duration lasted as long as the tracking period.
- In the unemployment duration example, we might follow each individual for a fixed length of time, say, two years. If unemployment spells are measured in weeks, we would have censoring at 104 weeks.

Single-Spell data and Flow sampling

- Alternatively, we might stop tracking individuals at a fixed calendar date, say the last week of 2006. Because individuals can become unemployed at any time during 2005, calendar-date censoring results in censoring times that differ across individuals.

Flow sampling



A: Beginning of study period (e.g., March survey)
B: End of study period (e.g., April survey)
t_1, t_2: Completed spells
t_3: Right-censored spell
t_4: Left-censored spell

Single-Spell data and Flow sampling

- For a random draw i from the population, let $a_i \in [0, b]$ denote the time at which individual i enters the initial state (the "starting time")
- Let t_i^* denote the length of time in the initial state (the duration)
- Let $\mathbf{x_i}$ denote the vector of observed covariates.
- We assume that t_i^* has a continuous conditional density $f(t|\mathbf{x_i}; \beta)$, $t \geq 0$, where β is a vector of unknown parameters.
- Without right censoring we would observe a random sample on $(a_i, t_i^*, \mathbf{x_i})$, and estimation would be a standard exercise on conditional maximum likelihood.

Single-Spell data and Flow sampling

- To account for right censoring, we assume that the observed duration, t_i, is obtained as

$$t_i = \min(t_i^*,\ c_i) \tag{14}$$

where c_i is the censoring time for individual i. Notice that in some cases c_i is constant across i.

- We assume that, conditional on the covariates, the true duration is independent of the starting point a_i, and the censoring time c_i:

$$D(t_i^*|\mathbf{x_i}, a_i, c_i) = D(t_i^*|\mathbf{x_i}) \tag{15}$$

where $D(\cdot|\cdot)$ denotes the conditional distribution.

Single-Spell data and Flow sampling

- Assumption (15) clearly holds when a_i and c_i are constant for all i.
- Under assumption (15) the distribution of t_i^* given $(\mathbf{x_i}, a_i, c_i)$ does not depend on (a_i, c_i).
- Therefore, if the duration is not censored, the density of $t_i = t_i^*$ given $(\mathbf{x_i}, a_i, c_i)$ is simply $f(t|\mathbf{x_i}, \beta)$.
- The probability that t_i is censored is

$$P(t_i^* \geq c_i|\mathbf{x_i}) = 1 - F(c_i|\mathbf{x_i}, \beta) \tag{16}$$

where $F(t|\mathbf{x_i}, \beta)$ is the conditional cdf of t_i^* given $\mathbf{x_i}$.

Single-Spell data and Flow sampling

- Let d_i be a censoring indicator ($d_i = 1$ if uncensored, $d_i = 0$ if censored).
- The conditional likelihood for observation i can be written as

$$f(t_i|\mathbf{x_i}, \beta)^{d_i}[1 - F(t_i|\mathbf{x_i}, \beta)]^{1-d_i} \tag{17}$$

- Given data on (t_i, d_i, $\mathbf{x_i}$) for a random sample of size N, the maximum likelihood estimator of β is obtained by maximizing

$$\sum_{i=1}^{N}\{d_i \ln f(t_i|\mathbf{x_i}, \beta) + (1 - d_i) \ln [1 - F(t_i|\mathbf{x_i}, \beta)]\} \tag{18}$$

- With the maximum likelihood estimations we can compute the hazard functions with the formulas above.

Single-Spell data and Stock sampling

- With stock sampling we randomly sample from individuals that are in the initial state at a given point in time.
- The population is again individuals who enter the initial state during a specified interval $[0, b]$.
- Rather than observe a random sample of people flowing into the initial state, we can only obtain a random sample of individuals that are in the initial state at time b.
- In addition to the possibility of right censoring, we may face the problem of left censoring, which occurs when some or all of the starting times a_i are not observed.

Single-Spell data and Stock sampling

- For now assume: (1) we observe the starting times a_i for all individuals we sample at time b; and (2) we can follow sampled individuals for certain lenght of time after we observed them at time b.
- In the unemployment duration example, where the population comprises workers who became unemployed at some point during 2005, stock sampling would occur if we randomly sampled from workers who were unemployed during the last week of 2005.
- This kind of sampling causes a clear sample selection problem: we necessarily exclude from our sample any individual whose unemployment spell ended before the last week of 2005.

Single-Spell data and Stock sampling

- Because these spells were shorter than a year, we cannot just assume that the missing observations are randomly missing.
- The sample selection problem caused by stock sampling is just the same as the left truncation problem.
- Under the assumptions that we observe the a_i and can observe some spells past sampling date b, left truncation is fairly easy to deal with.

Single-Spell data and Stock sampling

- To account for the truncated sampling, we must modify the density in equation (17) to reflect the fact that part of the population is systematically omitted from the sample.
- Let $(a_i, c_i, \mathbf{x_i}, t_i)$ denote a random draw from the population of all spells starting in $[0, b]$.
- We observe this vector if and only if the person is still in the initial state at time b, that is, if and only if $a_i + t_i^* \geq b$ or $t_i^* \geq b - a_i$, where t_i^* is the true duration.
- Under the conditional independence assumption (15),

$$P(t_i^* \geq b - a_i | a_i, c_i, \mathbf{x_i}) = 1 - F(b - a_i | \mathbf{x_i}; \beta) \tag{19}$$

where $F(\cdot|\mathbf{x_i}; \beta)$ is the cdf of t_i^* given $\mathbf{x_i}$ as before.

Single-Spell data and Stock sampling

- The log-likelihood function can be written as

$$\sum_{i=1}^{N}\{d_i \ln f(t_i|\mathbf{x_i},\beta) + (1-d_i)\ln\left[1-F(t_i|\mathbf{x_i},\beta)\right] - \ln\left[1-F(b-a_i|\mathbf{x_i};\beta)\right]\} \quad (20)$$

  where again, $t_i = c_i$ when $d_i = 0$.
- Unlike the case of flow sampling, with stock sampling both the starting dates, a_i, and the length of the sampling interval, b, appear in the conditional likelihood function.
- Their presence makes it clear that specifying the interval $[0,\ b]$ is important for analyzing stock data.

Agenda

Single-Spell data and Unobserved Heterogeneity

- It may be the case that some elements of $\mathbf{x_i}$ are unknown to the investigator and must be supposed to vary over the population. Example: reservation wage in a job search model.
- This gives a second source of stochastic variation in the model in the form of unobserved heterogeneity.
- The key assumptions used in most models that incorporate unobserved heterogeneity are:
  (1) the heterogeneity is independent of the observed covariates, as well as starting times and censoring times.
  (2) the heterogeneity has a distribution known up to a finite number of parameters.
  (3) the heterogeneity enters the hazard function multiplicatively.

Single-Spell data and Unobserved Heterogeneity

- For a random draw i from the population, a Weibull hazard function conditional on observed covariates $\mathbf{x_i}$ and unobserved heterogeneity ν_i is

$$\theta(t; \mathbf{x_i}; \nu_i) = \nu_i \alpha t^{\alpha-1} e^{\mathbf{x_i}\beta} \tag{21}$$

  where $x_{i1} \equiv 1$ and $\nu_i > 0$.
- To identify the parameters α and β we need a normalization on the distribution of ν_i. The most common is $E(\nu_i) = 1$.
- This implies that, for a given vector $\mathbf{x}$, the average hazard is $\alpha t^{\alpha-1} e^{\mathbf{x}\beta}$.
- In the general case, where the cdf of t_i^* given $(\mathbf{x_i}, \nu_i)$ is $F(t|\mathbf{x_i}, \nu_i; \beta)$, we can obtain the distribution of t_i^* given $\mathbf{x_i}$ by integrating out the unobserved effect.

Single-Spell data and Unobserved Heterogeneity

- Because ν_i and $\mathbf{x_i}$ are independent, the cdf of t_i^* given $\mathbf{x_i}$ is

$$G(t|\mathbf{x_i};\beta,\rho) = \int_0^\infty F(t|\mathbf{x_i},\nu_i;\beta)h(\nu;\rho)\,d\nu \tag{22}$$

  where, for concreteness, the density of ν_i, $h(\cdot;\rho)$, is assumed continuous and depends on the unknown parameters ρ.
- For flow data, the log-likelihood function is

$$\sum_{i=1}^{N}\{d_i \ln g(t_i|\mathbf{x_i},\beta,\rho) + (1-d_i)\ln\left[1-G(t_i|\mathbf{x_i},\beta,\rho)\right]\} \tag{23}$$

  where $g(t|\mathbf{x_i},\beta,\rho)$ is the density of t_i^* given $\mathbf{x_i}$.
- We should assume that $D(t_i^*|\mathbf{x_i},\nu_i,a_i,c_i) = D(t_i^*|\mathbf{x_i},\nu_i)$ and $D(\nu_i|\mathbf{x_i},a_i,c_i) = D(\nu_i)$. These assumptions ensure that the conditional independence condition (15) holds.