TEMA 3. Modelos de Elección Discreta

Profesor: Pedro Albarrán Pérez

Universidad de Alicante. Curso 2010/2011.

Contenido

Introducción

En ocasiones analizamos datos donde la variable *dependiente* de interés toma valores discretos:

1. *Variable dependientes binarias:* endeudarse o no
2. *Variables discretas sin ordenación:* modo de transporte (tren, autobús, etc...)
3. *Variables discretas con orden:* calificación/"rating" financiero
4. *Datos de conteo ("count data")* con variables discretas ordenadas que pueden tomar muchos valores diferentes: número de patentes de una empresa

Un modelo de regresión lineal puede no ser lo más adecuado en estos casos

- Los resultados son difíciles de interpretar: no se puede hablar de cambio continuo
- La variable dependiente sólo admite valores discretos, y puede que sólo no-negativos
- A veces, la variable dependiente debe entenderse cualitativa y no cuantitativamente.
- Podemos estar interesados en estimar la probabilidad de la ocurrencia de los distintos valores de la variable dependiente
  - no tanto en el valor esperado predicho

Empezaremos considerando el caso más sencillo: la variable dependiente sólo toma dos valores (es binaria).

- Una variable binaria (toma sólo dos valores):

$$Y_i = \begin{cases} 1 & \text{con probabilidad } p \\ 0 & \text{con probabilidad } 1-p \end{cases}$$

  - el valor 1 denota que el individuo ha tomado alguna acción
- sigue una distribución de Bernoulli

$$f(y) = \Pr(Y = y) = p^y (1-p)^{1-y}$$

  - $E(Y) = \Pr(Y = 1) = p$
  - $Var(Y) = p(1-p)$
- NO estamos interesados en esta distribución incondicional de Y, sino en su *distribución condicional*:
  - dadas sus características, cuál es la probabilidad de que el individuo i tome una acción $(Y_i = 1)$
- Generalizando el resultado anterior, la probabilidad de $Y = 1$ condicional en X es igual a la esperanza condicional de Y dado X

$$E(Y|X = x) = \Pr(Y = 1|X = x) = p(x)$$

  - $p(x)$ podría ser cualquier función

Modelo Lineal de Probabilidad

- El Modelo Lineal de Probabilidad simplemente supone que la esperanza condicional de la variable binaria Y es lineal.

$$E(Y|X = x) = \Pr(Y = 1|X = x) = p(x) = \beta_0 + \beta_1 x$$

- Todo lo que ya sabemos sobre el modelo de regresión lineal se puede aplicar directamente: estimación, contraste de hipótesis, interpretación de los parámetros, etc.
- Sólo debemos recordar que la esperanza condicional es, en este caso, una probabilidad.
- ¿Por qué no se utiliza frecuentemente el Modelo Lineal de Probabilidad?

MLP: limitaciones

1. *Heterocedasticidad*. El Modelo Lineal de Probabilidad es, por construcción, heterocedástico

$$Var\left(Y|X=x\right)=p\left(x\right)\left[1-p\left(x\right)\right]=\left(\beta_0+\beta_1 x\right)\left(1-\beta_0-\beta_1 x\right)$$

   - NO crucial: simplemente será necesario utilizar errores estándar robustos a la presencia de heterocedasticidad

2. En un modelo lineal, los valores de $E\left(Y|X=x\right)$ (que son probabilidades) NO están restringidos a estar comprendidos entre cero y uno
   - se pueden tener probabilidades mayores que uno o menores que cero
   - el cambio en la probabilidad esperada de $Y=1$ puede no tener sentido

- Debemos garantizar que los valores de $E(Y|X=x)=p(x)$ estén comprendidos entre cero y uno.
- Una elección natural es *una función de distribución acumulada* $F(\bullet)$, PERO sólo depende de una variable
  1. Primero utilizaremos una función $h(x_1,\ldots,x_k)$ que ofrezca un único valor (índice)
     - el valor puede estar fuera del intervalo cero y uno
  2. Posteriormente se aplica la función de distribución acumulada al índice
     $$E(Y|X=x)=p(x)=F(h(x_1,\ldots,x_k))$$
     y se obtienen valores entre el intervalo adecuado.
- La opción más simple consiste en utilizar un índice lineal
  $$h(x_1,\ldots,x_k)=\beta_0+\beta_1X_1+\cdots+\beta_kX_k$$
  - esto no supone una restricción ahora: $F(\bullet)$ transformará los valores al intervalo entre cero y uno
- Estos modelos se denominan *modelos de índice lineal*:
  $$E(Y|X=x)=\Pr(Y=1|X=x)=F(\beta_0+\beta_1X_1+\cdots+\beta_kX_k)$$

Aunque existen múltiples posibilidades para $F(\bullet)$, se suelen utilizar dos

1. La función de distribución acumulada de la normal estandarizada

$$\Phi(z) = \int_{-\infty}^{z} \phi(x)\, dx = \int_{-\infty}^{z} \frac{1}{2\pi} \exp\left(-\frac{1}{2}x^2\right) dx$$

donde $\phi(x)$ es la función de densidad de la normal estandarizada.

2. La función de distribución acumulada logística

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

- El modelo de índice lineal que utiliza esta función de distribución acumulada se denomina *modelo probit*:

$$E(Y|X=x) = \Pr(Y=1|X=x) = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$$

- El modelo de índice lineal que utiliza esta función de distribución acumulada se denomina *modelo logit*:

$$E(Y|X=x) = \Pr(Y=1|X=x) = \Lambda(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$$

Modelo de Elección Discreta

- Consideremos un individuo que se plantea comprar un bien (un coche, una casa, etc.).
  - Su decisión se basara en la utilidad que obtiene con la compra del bien frente a la que obtiene si no lo compra (con la mejor alternativa posible).
  - La utilidad del individuo de comprar el bien, V_{i1}, depende de las características del mismo, algunas de las cuales son observables (X) y otras no (ε).
- Supongamos que se puede expresar la *utilidad como un índice lineal* de las características

$$V_{i1} = \pi_0 + \pi_1 X_{1i}^1 + \cdots + \pi_k X_{ki}^1 + \varepsilon_{i1}$$

- Supongamos que la utilidad del individuo si no compra el bien depende de las características de la mejor alternativa posible:

$$V_{i0} = \delta_0 + \delta_1 X_{1i}^0 + \cdots + \delta_k X_{ki}^0 + \varepsilon_{i0}$$

- El individuo comprará el bien si la utilidad de comprarlo es mayor que la de no comprarlo $V_{i1} > V_{i0}$.

- Se define la *variable latente* Y_i^* como la diferencia de las utilidades (no observadas)

$$Y_i^* = V_{i1} - V_{i0}$$

- Sí observamos la decisión del individuo que resulta de comparar ambas utilidades.

$$Y_i = \begin{cases} 1, & \text{si el individuo compra: } Y_i^* = V_{i1} - V_{i0} \geqslant 0 \\ 0, & \text{si el individuo NO compra: } Y_i^* = V_{i1} - V_{i0} < 0 \end{cases}$$

- La variable latente Y_i^* es un índice *lineal* de comparación de utilidades

$$Y_i^* = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i$$

  - donde X son diferencias en características de las dos opciones

- Si queremos calcular la esperanza condicional de la variable observable Y_i, tenemos

$$\begin{aligned} E\left(Y_i | X_{1i}, \ldots, X_{ki}\right) &= \Pr\left(Y_i = 1 | X_{1i}, \ldots, X_{ki}\right) = \Pr\left(Y_i^* > 0 | X_{1i}, \ldots, X_{ki}\right) \\ &= \Pr\left(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i > 0 | X_{1i}, \ldots, X_{ki}\right) \end{aligned}$$

- Sólo el término de error u_i es estocástico

$$\begin{aligned} E(Y_i|X_{1i},\ldots,X_{ki}) &= \Pr(\beta_0+\beta_1X_{1i}+\cdots+\beta_kX_{ki}+u_i>0|X_{1i},\ldots,X_{ki}) \\ &= \Pr(u_i>-\beta_0-\beta_1X_{1i}-\cdots-\beta_kX_{ki}|X_{1i},\ldots,X_{ki}) \end{aligned}$$

- El modelo se completa con un supuesto sobre la distribución del término de error u_i para calcular esta probabilidad
  - típicamente, se elige una distribución acumulada $F(z)=\Pr(U<z)$ que cumpla la propiedad de simetría:

$$1-F(-z)=F(z)$$

  - por tanto,

$$\begin{aligned} \Pr(u_i>-\beta_0-\beta_1X_{1i}-\cdots-\beta_kX_{ki}) &= 1-\Pr(u_i<-\beta_0-\beta_1X_{1i}-\cdots-\beta_kX_{ki}) \\ &= 1-F(-\beta_0-\beta_1X_{1i}-\cdots-\beta_kX_{ki}) \\ &= F(\beta_0+\beta_1X_{1i}+\cdots+\beta_kX_{ki}) \end{aligned}$$

- Si u_i sigue una distribución normal, se tiene un modelo probit para Y_i

$$E(Y_i|X_{1i},\ldots,X_{ki})=\Pr(Y_i=1|X_{1i},\ldots,X_{ki})=\Phi(\beta_0+\beta_1X_{1i}+\cdots+\beta_kX_{ki})$$

- Si u_i sigue una distribución logística, se tiene un modelo logit para Y_i

$$E(Y_i|X_{1i},\ldots,X_{ki})=\Pr(Y_i=1|X_{1i},\ldots,X_{ki})=\Lambda(\beta_0+\beta_1X_{1i}+\cdots+\beta_kX_{ki})$$

Ejemplo: la variable Y_i toma valor 1 si al lanzar una moneda sale cara y 0 en caso contrario

- Se tienen muestras con dos lanzamientos: los posibles valores son $\{0,0\}$, $\{0,1\}$, $\{1,0\}$ y $\{1,1\}$.
- Se tienen tres posibles monedas con distintas probabilidades de obtener cara: $p = \{0{,}1; 0{,}5; 0{,}9\}$
- La probabilidad de observar los distintos posibles valores en cada caso es:

	$p = 0{,}10$	$p = 0{,}50$	$p = 0{,}90$
$\{0,0\}$	0,81	0,25	0,01
$\{0,1\}$	0,09	0,25	0,09
$\{1,0\}$	0,09	0,25	0,09
$\{1,1\}$	0,01	0,25	0,81

- La probabilidad de observar cada uno de los cuatro casos varía según la moneda (parámetro p)
- Si observamos una muestra concreta, ¿cómo "predecir" la moneda que moneda se utilizó?

- Intuitivamente, la lógica que propone el método de máxima verosimilitud es:

  Dada un muestra observada, se elige como valor estimado aquél que maximiza la probabilidad (verosimilitud) de que precisamente esa muestra hubiera sido la observada.

- El método de máxima verosimilitud consta, pues, de dos pasos:
  1. Calcular la probabilidad de cada muestra como función de los parámetros del modelo.
     - Dada una muestra observada finalmente, la probabilidad de observar esa muestra varía sólo como función de los parámetros.
  2. Estimar el parámetro cómo el valor que hace máxima la probabilidad de observar una muestra concreta

$$y_i = \begin{cases} 1, & \text{si } y_i^* \geqslant 0 \\ 0, & \text{si } y_i^* < 0 \end{cases}$$

- Por simplicidad, la variable latente depende linealmente de una variable explicativa

$$\begin{aligned} y_i^* &= \beta_0 + \beta_1 X_1 + \varepsilon_i \\ \varepsilon_i | X_1 &\sim N(0,1) \end{aligned}$$

- Para estimar los parámetros β_0, β_1 de este modelo probit, escribiremos la función de verosimilitud
  - esto es, calcular la probabilidad de los parámetros en función de la muestra que observamos

- Sea un individuo i con valores observados $X_{1i} = x_{1i}$
- La probabilidad de que la variable dependiente para i tome valor 1 es:

$$\begin{aligned}
\Pr(Y_i = 1|X_{1i} = x_{1i}) &= \Pr(\beta_0 + \beta_1 x_1 + \varepsilon_i > 0) \\
&= \Pr(\varepsilon_i > -\beta_0 - \beta_1 x_1) \\
&= 1 - \Pr(\varepsilon_i < -\beta_0 - \beta_1 X_1) \\
&= 1 - \Phi(-\beta_0 - \beta_1 x_1) \\
&= \Phi(\beta_0 + \beta_1 x_1)
\end{aligned}$$

- Y la probabilidad de que tome valor 0 es:

$$\Pr(Y_i = 0|X_{1i} = x_{1i}) = 1 - \Phi(\beta_0 + \beta_1 x_1)$$

- Por tanto, la probabilidad de observar cada valor $y_i = \{0, 1\}$ para el individuo i es

$$\Pr(Y_i = y_i|x_{i1}; \beta_0, \beta_1) = [\Phi(\beta_0 + \beta_1 x_1)]^{y_i} [1 - \Phi(\beta_0 + \beta_1 x_1)]^{1-y_i}$$

- Si tenemos una muestra aleatoria, la probabilidad conjunta de observar a los N individuos de la muestra será:

$$\Pr(Y_1 = y_1, \ldots, Y_N = y_N) = \prod_{i=1}^{N} \Pr(Y_i = y_i | x_{i1}; \beta_0, \beta_1) =$$

$$L(\beta_0, \beta_1; y_1, \ldots, y_n) = \prod_{i=1}^{N} [\Phi(\beta_0 + \beta_1 x_1)]^{y_i} [1 - \Phi(\beta_0 + \beta_1 x_1)]^{1-y_i}$$

- La probabilidad conjunta $L(\beta_0, \beta_1; y_1, \ldots, y_n)$ como función de los parámetros se denomina *función de verosimilitud*
  - para cada valor de los parámetros, informa sobre cómo de verosímil (probable) resulta que se haya generado la muestra que observamos $(y_1, \ldots, y_n)$.
- En general, resulta más conveniente trabajar con la *función de log-verosimilitud*

$$\log L(\beta_0, \beta_1; y_1, \ldots, y_n) = \sum_{i=1}^{N} [y_i \log \Phi(\beta_0 + \beta_1 x_1)] (1 - y_i) \log [1 - \Phi(\beta_0 + \beta_1 x_1)]$$

- Nuestra estimación de los parámetros será aquella que haga máxima esta función

- La función de verosimilitud depende del supuesto distribucional del término de error

$$\varepsilon_i | X_1 \sim N(0,1)$$

- Si suponemos otra distribución (logística, por ejemplo) para el término de error, entonces
  - el mecanismo del método de máxima verosimilitud es el mismo
  - la probabilidad conjunta calculada sería diferente
- Por tanto, diferentes supuestos distribucionales implican distintas funciones de verosimilitud
- Asimismo, los estimadores pueden ser diferentes dependiendo del supuesto distribucional
  - el máximo de la función de (log-)verosimilitud puede ser distinto

- Los estimadores y sus propiedades dependen crucialmente de qué distribución se haya supuesto para el término de error.
  - el término de error es inobservable, no podemos conocer su distribución
- Si la distribución que suponemos resulta ser la verdadera distribución, el estimador máximo verosímil será
  1. consistente
     - cuando el tamaño muestral es grande, el valor estimado está próximo al verdadero valor de parámetro.
  2. eficiente (asintóticamente)
     - la varianza del estimador es la menor posible
- Si la distribución que suponemos no resulta ser la verdadera, no se puede garantizar esas buenas propiedades.
- En algunos casos, el estimador máximo verosímil sigue siendo consistente incluso si el supuesto distribucional no es cierto.
  - se habla de estimación por "pseudo" máxima verosimilitud.

En los modelos de elección discreta,

$$\begin{aligned} E\left(Y|X_1,\ldots,X_j,\ldots,X_k\right) &= \Pr\left(Y=1|X_1,\ldots,X_j,\ldots,X_k\right) \\ &= F\left(\beta_0+\beta_1X_1+\cdots+\beta_jX_j+\cdots+\beta_kX_k\right) \end{aligned}$$

- El efecto parcial o efecto marginal de una *variable continua* es:

$$\frac{\delta E\left(Y|X_1,\ldots,X_j,\ldots,X_k\right)}{\delta X_j} = \beta_j f\left(\beta_0+\beta_1X_1+\cdots+\beta_jX_j+\cdots+\beta_kX_k\right)$$

- El efecto parcial o efecto marginal de una *variable discreta* es:

$$F\left(\beta_0+\beta_1X_1+\cdots+\beta_j\left(c_j+1\right)+\cdots+\beta_kX_k\right)-$$

$$F\left(\beta_0+\beta_1X_1+\cdots+\beta_j\left(c_j\right)+\cdots+\beta_kX_k\right)$$

donde

- $F(\bullet)$ es una función de distribución acumulada
  - ($\Phi(\bullet)$ en el probit y $\Lambda(\bullet)$ en el logit)
- $f(z)=\frac{\delta F(z)}{\delta z}$ es su función de densidad.

- Notad que, a diferencia de los modelos lineales más simples,
  1. El efecto marginal no depende de un único parámetro
     - depende de todos los parámetros del modelo
     - depende de la forma funcional concreta de $f(\bullet)$
  2. El efecto marginal depende de los valores de las variables explicativas X.
     - individuos con valores diferentes en al menos una variable explicativa tienen distinto efecto marginal.
     - para cada combinación de valores de las X tendremos distintos efectos parciales
- En la práctica, se tiene un efecto marginal distinto para cada individuo de la muestra.
  - efectos *ceteris paribus* son efectos manteniendo constante el resto de variables en el valor de cada individuo
- Aunque el efecto marginal de X_j no sólo depende de β_j, el signo sí que coincide,
  - $f(\bullet) \geqslant 0$
  - $F(\bullet)$ es monótona no decreciente
- La significatividad del efecto marginal también coincide con la del parámetro, en general

Ejemplo 1

$$
\begin{aligned}
E(Y|X=x) &= \Pr(Y=1|X=x) = \delta_0 + \delta_1 X \\
E(Y|X=x) &= \Pr(Y=1|X=x) = \pi_0 + \pi_1 X + \pi_2 X^2 \\
E(Y|X=x) &= \Pr(Y=1|X=x) = F(\beta_0 + \beta_1 X)
\end{aligned}
$$

- Los efectos marginales son, respectivamente:

$$
\begin{aligned}
\frac{\delta E(Y|X=x)}{\delta X} &= \delta_1 \\
\frac{\delta E(Y|X=x)}{\delta X} &= \pi_1 + 2\pi_2 X \\
\frac{\delta E(Y|X=x)}{\delta X} &= \beta_1 f(\beta_0 + \beta_1 X)
\end{aligned}
$$

- Ejemplo: $Y = 1$ si un individuo decide retirarse y X es su edad
  - El efecto marginal debe ser diferente para $X = 20$ y $X = 60$
- El primer modelo lineal sólo ofrece un único efecto marginal
  - el promedio para los distintos valores de X
- El efecto marginal dependerá de X de forma diferente según
  - cómo se transforme X en el modelo lineal: $X^2, \log X, \frac{1}{X}, \ldots$
  - la forma funcional de $F(\bullet)$

Ejemplo

$$
\begin{aligned}
E(Y|X_1, X_2) &= \Pr(Y = 1|X_1, X_2) = \pi_0 + \pi_1 X_1 + \pi_2 X_2 + \pi_3 X_2^2 \\
E(Y|X_1, X_2) &= \Pr(Y = 1|X_1, X_2) = F(\beta_0 + \beta_1 X_1 + \beta_2 X_2)
\end{aligned}
$$

- Los efectos marginales de X_2 son, respectivamente:

$$
\begin{aligned}
\frac{\delta E(Y|X_1, X_2)}{\delta X_2} &= \pi_2 + 2\pi_3 X_2 \\
\frac{\delta E(Y|X_1, X_2)}{\delta X_2} &= \beta_2 f(\beta_0 + \beta_1 X_1 + \beta_2 X_2)
\end{aligned}
$$

- En ambos caso, el efecto marginal de X_2 es depende del valor inicial de X_2
- Pero cuando dos individuos tienen el mismo valor de X_2 y distinto valor de X_1, el efecto marginal d
  - en el modelo lineal, es el mismo para ambos individuos
  - en el modelo no lineal, es distinto porque sí depende de X_1
- Sería necesario incluir alguna interacción entre X_1 y X_2 en el modelo lineal para conseguir lo mismo

- A diferencia del modelo lineal, no existe un único efecto marginal
- Para cada combinación de valores en las variables explicativas, se tiene un efecto marginal diferente
  - si hay muchas variables explicativas y/o toman muchos valores diferentes, mayor número de posibles valores del efecto marginal.
- En la práctica, se tendrá un efecto marginal diferente para cada individuo en la muestra.
  - se tiene una distribución (muestral) de efectos parciales
- Inconveniente: *a priori*, no tenemos un valor resumen del efecto esperado del cambio en una variable

Efecto Marginal Evaluado en Valores Relevantes

- En ocasiones, queremos el efecto marginal para un conjunto de valores bien determinado
  - por ejemplo, para un individuo de un determinado nivel educativo, con una renta dada, etc.
  - para un único conjunto de valores dados para cada variable, se tendrá un único efecto marginal
- Para una variable continua, el efecto marginal evaluado en $\{X_1 = x_1^*, \ldots, X_k = x_k^*\}$ es:

$$em_*^j = \left.\frac{\delta E\left(Y|X_1,\ldots,X_j,\ldots,X_k\right)}{\delta X_j}\right|_{X_1=x_1^*,\ldots,X_k=x_k^*} = \beta_j f\left(\beta_0+\beta_1 x_1^*+\cdots+\beta_k x_k^*\right)$$

- Y para una variable discreta:

$$\begin{aligned} em_*^j &= F\left(\beta_0+\beta_1 x_1^*+\cdots+\beta_j\left(c_j+1\right)+\cdots+\beta_k x_k^*\right) \\ &- F\left(\beta_0+\beta_1 x_1^*+\cdots+\beta_j\left(c_j\right)+\cdots+\beta_k x_k^*\right) \end{aligned}$$

- No siempre tenemos claro en qué único conjunto de valores (para *todas* las variables) resulta interesante evaluar el efecto marginal.

Efecto Marginal Evaluado en la Media (o Mediana)

- Se puede utilizar la media (o mediana) de cada variable explicativa como valores representativos
- Para una variable continua, el efecto marginal evaluado en $\{X_1 = \overline{x}_1, \ldots, X_k = \overline{x}_k\}$ es:

$$em_{\bullet}^{j} = \left. \frac{\delta E\left(Y|X_1, \ldots, X_j, \ldots, X_k\right)}{\delta X_j} \right|_{X_1 = \overline{x}_1, \ldots, X_k = \overline{x}_k} = \beta_j f\left(\beta_0 + \beta_1 \overline{x}_1 + \cdots + \beta_k \overline{x}_k\right)$$

- Y para una variable discreta:

$$\begin{aligned} em_{\bullet}^{j} &= F\left(\beta_0 + \beta_1 \overline{x}_1 + \cdots + \beta_j\left(c_j + 1\right) + \cdots + \beta_k \overline{x}_k\right) \\ &- F\left(\beta_0 + \beta_1 \overline{x}_1 + \cdots + \beta_j\left(c_j\right) + \cdots + \beta_k \overline{x}_k\right) \end{aligned}$$

- Esta alternativa puede resultar inadecuada:
  1. la media no *es* siempre un valor representativo de una distribución
  2. cuando se eligen valores relevantes para varias variables, la elección debe ser conjunta (considerando relaciones entre ellas)
     - el valor medio de X_1 no tiene porque encontrarse asociado generalmente al valor medio de X_2.
- Se puede evaluar el efecto marginal en la media de algunas variables y en valores concretos de otras

Efecto Marginal Promedio (o Mediano, etc.)

- Para cada individuo i, se obtiene su efecto marginal evaluado en $\{X_1 = x_{1i}, \ldots, X_k = x_{ki}\}$

$$em_i^j = \left. \frac{\delta E\left(Y | X_1, \ldots, X_j, \ldots, X_k\right)}{\delta X_j} \right|_{X_1 = x_{1i}, \ldots, X_k = x_{ki}} = \beta_j f\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}\right)$$

- Y para una variable discreta:

$$\begin{aligned} em_*^j &= F\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_j\left(c_j + 1\right) + \cdots + \beta_k x_{ki}\right) \\ &- F\left(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_j\left(c_j\right) + \cdots + + \beta_k x_{ki}\right) \end{aligned}$$

- Se tiene una *distribución (muestral) de efectos marginales*
  - puede resumirse con el Efecto Marginal Promedio

$$\overline{em}^j = \frac{1}{N} \sum_{i=1}^{N} em_i^j$$

- Este enfoque sí controla por la relación entre variables explicativas.
- Como disponemos de toda la distribución, podemos interesarnos
  - por percentiles representativos de la parte alta o de la parte baja
  - por medidas de dispersión

Nota sobre los Errores Estándar

- La varianza de un efecto marginal *depende de la varianza de todos los coeficientes estimados y de una forma no lineal:*

$$Var\left(\widehat{em}^{j}_{*}|X\right) = Var\left(\widehat{\beta}_j f\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \cdots + \widehat{\beta}_k x_k^*\right)\right)$$

  - resulta complicado calcular la varianza exacta (en muestra finitas)
- Se puede utilizar el llamado *método delta* para calcular la varianza asintótica

$$AVar\left(\widehat{em}^{j}_{*}|X\right) = J' V J$$

como una función de

  - la matriz de varianzas y covarianzas de los coeficientes estimados $V = Var\left(\widehat{\beta}\right)$
  - el jacobiano de la función $f(\bullet)$

$$J = \frac{\delta}{\delta \beta'}\left[\widehat{\beta}_j f\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_1^* + \cdots + \widehat{\beta}_k x_k^*\right)\right]$$

Predicción de Probabilidades

- Los modelos no lineales también permiten predecir los valores de la variable dependiente
- PERO para variables dependientes discretas, resulta más conveniente predecir probabilidades de los valores
- El modelo binario se refiere directamente a la probabilidad predicha de $Y = 1$

$$\widehat{E}[Y|X_1, X_2] = \widehat{\Pr}(Y = 1|X_1, X_2) = F\left(\widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2\right)$$

  - en función de los parámetros estimados
  - y de los valores de las variables explicativas en los que condicionemos
- La probabilidad de $Y = 0$ se obtiene automáticamente como su complementaria

- Se tienen diversas formas de predecir probabilidades
  1. Evaluada en valores concretos o en la media (mediana)
     - predicciones puntuales para unos valores no disponibles en la muestra (contrafactual)
     - puede contradecir la relación real entre variables explicativas
  2. Evaluado en los valores observados para cada individuo
     - Como en el caso lineal, la media de la variable dependiente coincidirá con la media de la probabilidad predicha
- Se pueden combinar ambos enfoques: algunos valores fijos y otros observados
  - se obtiene una distribución de probabilidades predichas para un contrafactual (valores fijados artificialmente)
- Debe calcularse el error estándar de las probabilidades predichas
  - puesto que dependen de estimaciones de parámetros
- Se utiliza el método delta para obtener esos errores estándar

$$AVar\left(\widehat{\Pr}\left(Y=1|X_1=x_1^*,X_2=x_2^*\right)\right)=G'VG$$

donde $G=\frac{\delta}{\delta\beta'}\left[F\left(\widehat{\beta}_0+\widehat{\beta}_1x_1^*+\widehat{\beta}_2x_2^*\right)\right]$

- Los coeficientes estimados en modelos Probit, Logit y de Probabilidad Lineal tienen una escala muy diferentes

$$\begin{aligned} \widehat{\beta}_{logit} &\simeq 4\widehat{\beta}_{lineal} \\ \widehat{\beta}_{probit} &\simeq 2{,}5\widehat{\beta}_{lineal} \\ \widehat{\beta}_{logit} &\simeq 1{,}6\widehat{\beta}_{probit} \end{aligned}$$

- Esto NO supone que las implicaciones de ambos modelos sea diferentes
- Los efectos marginales y las probabilidades predichas son muy similares entre los tres modelos
  - sólo existen diferencias de cierta importancia en las colas entre el probit y el logit.
- Hay que recordar que el modelo lineal de probabilidad sólo estima el efecto marginal promedio y la probabilidad predicha promedio
  - el efecto marginal promedio del logit y del probit serán similares al coeficiente del modelo de probabilidad lineal

Bondad de Ajuste

- El valor de la *(log-)verosimilitud*.
  - A mayor (log-)verosimilitud, mejor será el modelo
  - Sólo puede compararse entre modelos de la misma clase
- El *pseudo-R^2 de McFadden*:

$$\widetilde{R}^2 = 1 - \frac{L_N\left(\widehat{\beta}\right)}{L_N\left(\overline{y}\right)}$$

  - Es una medida de la mejora relativa en la log-verosimilitud
    - del modelo que incluye variables explicativas, $L_N\left(\widehat{\beta}\right)$
    - respecto al modelo sólo con constante, $L_N\left(\overline{y}\right)$ (probabilidad media incondicional)
    - Esta medida está entre cero y uno, PERO NO representa proporción de varianza explicada por el modelo

Predicciones y Bondad de Ajuste

- La bondad de ajuste puede medirse como capacidad para *predecir adecuadamente* los datos observados
- Las probabilidad predichas serán valores entre cero y uno, PERO los datos observados son exactamente cero o uno

1. Comparación de probabilidades predichas con frecuencias muestrales
   - Se divide la muestra en g subgrupos
   - Calculamos la diferencia entre las probabilidad media predicha y la observada en cada subgrupo
2. Convertir la predicción en valores binarios y calcular el *porcentaje de observaciones correctamente clasificadas*
   - El resultado predicho de un individuo será 1 si su probabilidad predicha supera un cierto umbral (p.e., 0.5 o la media de la variable dependiente)
   - La medida de bondad de ajuste es el porcentaje de resultados predichos que coinciden con los observados

1. Test de Wald:
   - estima el modelo bajo la hipótesis alternativa (modelo sin restringir)
   - comprueba si las estimaciones satisfacen las restricciones: poca "distancia" a lo especificado por el modelo restringido (H_0)
   - Los test de Wald tienen una distribución asintótica conocida bajo H_0
   - Cuando la probabilidad de observar valores muy alejados es pequeña de acuerdo con la distribución normal, se rechaza H_0
2. Test del Ratio de Verosimilitudes
   - se estiman el modelo no restringido (H_a) y el restringido (H_0)
     - el modelo restringido siempre tendrá menor verosimilitud (por imponer restricciones)
   - Una "gran" diferencia en las verosimilitudes es poco probable bajo H_0
   - El estadístico de contraste es

$$LR = -2\left[\ln L\left(\widehat{\theta}_r\right) - \ln L\left(\widehat{\theta}_u\right)\right] \overset{a}{\sim} \chi^2_{(q)} \text{ bajo } H_0$$

   - Ambos tests son *asintóticamente equivalentes:* se obtendrán resultados similares (rechazo o no de H_0, p-valor, etc.)

Introducción

- Los modelos binarios pueden generalizarse de varias formas:
  - modelos univariantes multinomiales: una variable dependiente con múltiples categorías mutuamente exclusivas
    - las categorías pueden ser ordenadas o no
    - las variables explicativas pueden ser especificas de cada alternativa, etc.
  - modelos multivariantes para variables discretas
    - útil para varias categorías no mutuamente exclusivas
- Existen varios modelos diferentes para adecuarse a cada situación
  - algunos modelos son consistentes con maximación de utilidad
- Los parámetros NO son, en general, directamente interpretables
  - un coeficiente positivo NO implica necesariamente un aumento de la probabilidad
- Los efectos marginales informan del cambio en la probabilidad de observar cada una de las distintas categorías
  - no en una única probabilidad

Modelos Multinomiales

- El valor de la variable dependiente Y_i para el individuo i es una de m alternativas

$$Y_i = r, \qquad r = 1, 2, \ldots, m$$

  - los valores son arbitrarios (no afectan a resultados)
  - salvo en los modelos ordenados
- La probabilidad de la alternativa r para el individuo i, condicional en las variables explicativas X_i

$$p_{ir} = \Pr\left(Y_i = r | X_i\right) = F_r\left(X_i; \theta\right)$$

  donde $F_r(\bullet)$ depende del modelo multinomial concreto
- El efecto marginal de la variable explicativa j sobre la probabilidad de la alternativa r para el individuo i es

$$em^{irj} = \frac{\delta \Pr\left(Y_i = r | X_i\right)}{\delta x_{ij}} = \frac{\delta F_r\left(X_i; \theta\right)}{\delta x_{ij}}$$

- Las probabilidades y los efectos marginales dependen de los valores concretos de las variables explicativas en que se evalúen

Estimación por Máxima Verosimilitud

- Para un individuo i

$$\Pr\left(Y_i = r|X_i\right) = p_{i1}^{y_{i1}} \times \cdots \times p_{im}^{y_{im}} = \prod_{s=1}^{m} p_{is}^{y_{is}}$$

donde

$$y_{ir} = \begin{cases} 1, & \text{si } Y_i = r \\ 0, & \text{en caso contrario} \end{cases}$$

- Para una muestra aleatoria de individuos, la función de verosimilitud es

$$\Pr\left(Y_1 = y_1, \ldots, Y_N = y_N\right) = \prod_{i=1}^{N} \Pr\left(Y_i = r|X_i\right) = \prod_{i=1}^{N}\prod_{s=1}^{m} p_{is}^{y_{is}}$$

$$L\left(\theta\right) = \prod_{i=1}^{N}\prod_{s=1}^{m} \left[F_s\left(X_i;\theta\right)\right]^{y_{is}}$$

y, por tanto, la log-verosimilitud es

$$\ln L\left(\theta\right) = \sum_{i=1}^{N}\sum_{s=1}^{m} y_{is} \ln F_s\left(X_i;\theta\right)$$

- Ya conocemos las propiedades generales del método de máxima verosimilitud
  - por lo visto para modelos binarios
- Sabemos cómo obtener los errores estándar y realizar contrastes
- Los estimadores máximo verosímiles tienen buenas propiedades, siempre que $F_r(\bullet)$ esté correctamente especificado:
  - supuesto distribucional correcto
  - modelo multinomial adecuado a los datos
- Se puede obtener el pseudo-R^2 como medida de bondad de ajuste

Modelos aditivos de utilidad aleatoria

Algunos modelos pueden interpretarse como resultado de maximización de utilidad

- La utilidad de la alternativa r para el individuo i resulta de la suma de
  - un componente determinístico V_{ir}
  - un componente aleatorio inobservado ε_{ir}

$$U_{ir} = V_{ir} + \varepsilon_{ir}$$

- Se observa que el individuo i elige la alternativa r, $Y_i = r$, si obtiene la mayor utilidad entre alternativas

$$\begin{aligned} \Pr(Y_i = r) &= \Pr(U_{ir} \geqslant U_{is}), \quad \text{para todo } s \\ &= \Pr(U_{is} - U_{ir} \leqslant 0), \quad \text{para todo } s \\ &= \Pr(\varepsilon_{is} - \varepsilon_{ir} \leqslant V_{ir} - V_{is}), \quad \text{para todo } s \end{aligned}$$

Un modelo multinomial concreto especifica típicamente:

- $V_{ir} = x'_{ir}\beta + z'_i\gamma_r$
  - los regresores x_{ir} son variables específicas para cada alternativa
    - (ej. características de la opción)
  - los regresores z_i son variables invariantes a la alternativa, aunque potencialmente con impacto diferente en cada alternativa
    - (ej. características del individuo)
- Una distribución conjunta de $\varepsilon_{i1}, \ldots, \varepsilon_{im}$
  - distintos supuestos llevan a diferentes formas de $F_r(\bullet)$

Modelo Logit Multinomial

- Todas las variables explicativas son invariantes a la alternativa
- Los errores siguen una distribución conjunta logística; por tanto,

$$p_{ir} = \frac{\exp\left(X_i'\beta_r\right)}{\sum_{s=1}^{m}\exp\left(X_i'\beta_s\right)} \quad r = 1, \ldots, m$$

  - el vector β_s se fija a cero en una categoría (base)
  - los coeficientes se interpretan con respecto a la categoría base
- Puede interpretarse como una serie de modelos logit para pares de alternativas

$$\Pr\left(Y_i = r | Y_i = r \text{ o } Y_i = 1\right) = \frac{\Pr\left(Y_i = r\right)}{\Pr\left(Y_i = r\right) + \Pr\left(Y_i = 1\right)} = \frac{\exp\left(X_i'\beta_r\right)}{1 + \exp\left(X_i'\beta_r\right)}$$

  - siendo $s = 1$ la categoría base
- Se pueden definir los "odd-ratios" o ratios de riesgos relativos

$$\frac{\Pr\left(Y_i = r\right)}{\Pr\left(Y_i = 1\right)} = \exp\left(X_i'\beta_r\right)$$

  - $\exp\left(\beta_{rj}\right)$ representa el cambio relativo en la probabilidad de la alternativa j frente a la alternativa 1 cuando x_{irj} aumenta en una unidad.

- En lugar de una esperanza condicional, el logit multinomial ofrece la probabilidad de cada alternativa
- Y los correspondientes efectos marginales de los regresores sobre cada probabilidad

$$\frac{\delta p_{ir}}{\delta X_i} = p_{ir}\left(\beta_r - \overline{\beta}_i\right)$$

  - $\overline{\beta}_i$ es una media ponderada (por las probabilidades) para el individuo i del vector de coeficientes en todas las alternativas

$$\overline{\beta}_i = \sum_{s=1}^{m} p_{is}\beta_s$$

- El signo de los coeficientes NO coincide con el del efecto marginal: para una variable x_{ij} el efecto marginal es positivo si $\beta_{j,r} > \overline{\beta}_{j,i}$
- Tanto las probabilidades predichas como los efectos marginales dependen de los valores en que se evalúen

Modelo Logit Condicional

- Es una extensión del modelo Logit Multinomial que permite regresores específicos para cada alternativa

$$p_{ir} = \frac{\exp\left(X_{ir}'\beta + Z_i'\gamma_r\right)}{\sum_{s=1}^{m}\exp\left(X_{is}'\beta + Z_i'\gamma_s\right)} \qquad r = 1, \ldots, m$$

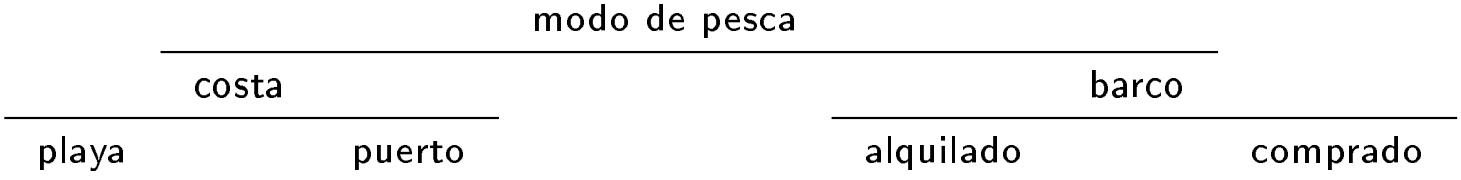
- Los coeficientes de los regresores específicos de cada alternativa son fácilmente interpretables:

$$\frac{\delta p_{ir}}{\delta X_{jis}} = \begin{cases} p_{ir}\left(1 - p_{ir}\right)\beta_j, & r = s \\ -p_{ir}p_{is}\beta_j, & r \neq s \end{cases}$$

  - si $\beta_j > 0$, un incremento de una variable en una alternativa supone mayor probabilidad de elegir esa categoría y menor de elegir el resto

Logit Anidado

- Los modelos logit multinomial/condicional imponen una restricción: la elección entre pares de alternativas es un modelo logit binario
  - este supuesto, conocido como de **independencia de alternativas irrelevantes**, puede ser muy restrictivo
  - ej.: problema del elección "coche/bus rojo/bus azul"
- Es decir, suponen que los errores de cada alternativa ε_{ir} son *independientes* e idénticamente distribuidos como valor extremo
- El modelo Logit Anidado requiere una estructura "anidada"
  - las alternativas se reparten en grupos
  - los errores ε_{ir} están correlacionados dentro del grupo, pero incorrelacionados fuera del grupo
- Por ejemplo,



Modelo Logit Anidado (cont.)

- El modelo Logit Anidado se puede derivar un problema de maximización de utilidad
  - suponiendo que los errores siguen una distribución multivariante (de valor extremo de Gumbel)
- El modelo Logit Anidado se puede definir para múltiples niveles
  - aunque típicamente se tienes dos
- Las probabilidades del modelo Logit Anidado son "similares" a las de Logit Condicional
  - aunque con una estructura anidada específica
- La interpretación de coeficientes y efectos marginales es igual a la discutida antes
  - según se tengan regresores específicos de cada alternativa o no
  - notad que si se tienen regresores no específicos a la alternativa, debe haber una categoría base con su vector de parámetros igual a cero

Probit Multinomial

- Permite relajar el supuesto de independencia de alternativas irrelevantes
  - permite un patrón de correlaciones en los errores muy flexible
  - no es necesario definir una estructura anidada
- Dado un modelo de utilidad aleatoria, donde la utilidad de la alternativa r es

$$U_{ir} = x_{ir}'\beta + z_i'\gamma_r + \varepsilon_{ir}$$

- Se supone que los errores siguen una distribución conjunta normal

$$\varepsilon \sim N(0, \Sigma)$$

donde $\varepsilon = (\varepsilon_{i1}, \ldots, \varepsilon_{im})$

- La probabilidad de elegir la alternativa r es

$$p_{ir} = \Pr(Y_i = r) = \Pr\left(\varepsilon_{is} - \varepsilon_{ir} \leqslant (x_{ir} - x_{is})'\beta + z_i'(\gamma_r - \gamma_s)\right), \quad \text{para todo } s$$

  - implica resolver una integral de dimensión $m-1$, sin solución cerrada y difícil de resolver

Probit Multinomial (cont.)

- La estimación del modelo necesita
  - imponer restricciones sobre Σ
  - obtener las probabilidades integrando numéricamente o utilizar el método de máxima verosimilitud simulada
- La interpretación de probabilidades predichas y de efectos marginales es "similar" a lo discutido anteriormente

Modelos para variable discreta ordenada

- En algunos casos, las variables categóricas están ordenadas de manera natural
- Sea una variable latente

$$Y_i^* = X_i'\beta + u_i$$

- La variable categórica se observa según Y_i^* cruza secuencialmente determinados umbrales

$$Y_i = r, \quad \text{si } \alpha_{r-1} < Y_i^* \leqslant \alpha_r, \quad r = 1, \ldots, m$$

donde $\alpha_0 = -\infty$ y $\alpha_m = \infty$

- La probabilidad de cada alternativa es

$$\begin{aligned} \Pr(Y_i = r) &= \Pr(\alpha_{r-1} < Y_i^* \leqslant \alpha_r) \\ &= \Pr(\alpha_{r-1} < X_i'\beta + u_i \leqslant \alpha_r) \\ &= \Pr(\alpha_{r-1} - X_i'\beta < u_i \leqslant \alpha_r - X_i'\beta) \\ &= F(\alpha_r - X_i'\beta) - F(\alpha_{r-1} - X_i'\beta) \end{aligned}$$

$$p_{ir} = \Pr(Y_i = r) = F(\alpha_r - X_i'\beta) - F(\alpha_{r-1} - X_i'\beta)$$

- La función de distribución acumulada $F(\bullet)$ depende del supuesto sobre el término de error
  - si u_i sigue una distribución normal estándar, se tiene el probit ordenado $F(\bullet) = \Phi(\bullet)$
  - si u_i sigue una distribución logística, se tiene el logit ordenado $F(\bullet) = \Lambda(\bullet)$
- Se pueden identificar y estimar
  - bien $m-1$ umbrales α y el vector β si el modelo no incluye constante
  - o bien $m-2$ umbrales α y el vector β incluyendo constante
- Las probabilidades predichas y los efectos marginales se calculan de las formas habituales
- El efecto marginal de una variable continua es

$$\frac{\delta \Pr(Y_i = r)}{\delta X_{ij}} = [F'(\alpha_r - X_i'\beta) - F'(\alpha_{r-1} - X_i'\beta)]\beta_j$$

  - el signo del efecto marginal y del coeficiente β_j coinciden

Modelos Multivariantes

- Dos o más variables categóricas se tienen que analizar conjuntamente si
  - existe simultaneidad: las variables categóricas dependen de las otras
  - no existe simultaneidad, pero los errores están correlacionados
- En el caso más sencillo tenemos dos variables binarias relacionadas sólo por la correlación de los errores: **modelo bi-probit**
- Se tienen dos variables latentes

$$\begin{aligned} Y_{1i}^* &= X_{1i}'\beta + \varepsilon_{1i} \\ Y_{2i}^* &= X_{2i}'\beta + \varepsilon_{2i} \end{aligned}$$

  donde ε_{1i} y ε_{2i} siguen una distribución conjunta normal
  - con esperanzas cero,
  - varianzas 1 y correlación ρ
- Se observan dos variables binarias

$$Y_{1i} = \begin{cases} 1, & \text{si } Y_{1i}^* > 0 \\ 0, & \text{si } Y_{1i}^* \leqslant 0 \end{cases} \quad \text{y} \quad Y_{2i} = \begin{cases} 1, & \text{si } Y_{2i}^* > 0 \\ 0, & \text{si } Y_{2i}^* \leqslant 0 \end{cases}$$

  - existen cuatro resultados mutuamente excluyentes

- SÓLO cuando $\rho = 0$, se pueden estimar *separadamente dos probits*
- El modelo conjunto se estima por máxima verosimilitud, a partir de las expresiones de las probabilidades para los cuatro casos
  - aunque no existe una expresión analítica cerrada
- Estimando adecuadamente (de forma conjunta) se obtienen resultados sustancialmente diferentes de los probits separados
  - tanto en las probabilidades $\Pr(Y_{1i} = 1|X_i)$ y $\Pr(Y_{2i} = 1|X_i)$, y en las cuatro conjuntas
  - como en los efectos marginales
- La generalización del modelo, especialmente si una variable depende de la otra, complica sustancialmente la estimación