Modelos multinomiales para variables categóricas

Gabriel V. Montes-Rojas

Modelo logit multinomial

Supongamos que la variable dependiente toma muchos valores, ej. $y = 0, 1, 2..., J$, aunque los valores de y no representan **ningún orden en particular**, $J + 1$ potenciales categorías. Éste es el modelo multinomial.

Ejemplo: Modelos de elección dicreta (discrete choice models). y podría ser la marca de un producto que el consumidor compra:
$y = 0$ A
$y = 1$ B
$y = 2$ C
$y = 3$ D
$y = 4$ E

Ejemplo: Participación en la fuerza laboral. y podría ser el status laboral de la persona:
$y = 0$ empleado
$y = 1$ desempleado
$y = 2$ fuera de la fuerza laboral

Modelo logit multinomial

- Seleccionemos sin pérdida de generalidad un **grupo base**. Por convención corresponde a $j = 0$.
- Cada valor de y contiene los parámetros β_j, $j = 1, 2, ..., J$, vectores $K_j \times 1$. (Si $J = 1$ tenemos el modelo logit.)
- En un modelo logit multinomial para modelar cada probabilidad tenemos

$$P[y = j|\boldsymbol{x}] = \frac{exp(\boldsymbol{x}\boldsymbol{\beta}_j)}{1 + \sum_{h=1}^{J} exp(\boldsymbol{x}\boldsymbol{\beta}_h)}, \ j = 1, 2, ..., J$$

- Notemos que la suma debe ser 1 (o sea $\sum_{h=0}^{J} P[y = h|\boldsymbol{x}] = 1$), entonces

$$P[y = 0|\boldsymbol{x}] = \frac{1}{1 + \sum_{h=1}^{J} exp(\boldsymbol{x}\boldsymbol{\beta}_h)}, \ j = 1, 2, ..., J$$

- Construir el logaritmo de la función de verosimilitud: $\ell_i(\boldsymbol{\beta}) = \sum_{j=0}^{J} 1[y_i = j] log(P[y = j|\boldsymbol{x}])$. McFadden (1984) muestra que esta función es cóncava global, entonces tiene un único máximo.

Modelo logit multinomial

- Los efectos parciales son

$$\frac{\partial P(y=j|\boldsymbol{x})}{\partial x_k} = P(y=j|\boldsymbol{x}) \left\{ \beta_{jk} - \left[\sum_{h=1}^{J} \beta_{hk} exp(\boldsymbol{x}\boldsymbol{\beta}_h) \right] / g(\boldsymbol{x}, \boldsymbol{\beta}) \right\},$$

donde β_{hk} es el elemento k de $\boldsymbol{\beta}_h$ y $g(\boldsymbol{x}, \boldsymbol{\beta}) = 1 + \sum_{h=1}^{J} exp(\boldsymbol{x}\boldsymbol{\beta}_h)$. Notemos que un cambio en x_k afecta todas las probabilidades simultáneamente.

- Una interpretación de $\boldsymbol{\beta}_j$ está dada por

$$p_j(\boldsymbol{x}, \boldsymbol{\beta}) / p_0(\boldsymbol{x}, \boldsymbol{\beta}) = exp(\boldsymbol{x}\boldsymbol{\beta}_j), j = 1, 2, ..., J.$$

- Entonces, el cambio en $p_j(\boldsymbol{x}, \boldsymbol{\beta}) / p_0(\boldsymbol{x}, \boldsymbol{\beta})$ ante un cambio en x_k (asumiendo es contínua) es $\beta_{jk} exp(\boldsymbol{x}\boldsymbol{\beta}_j)\Delta x_k$. O lo que es lo mismo, el log odds-ratio es lineal en $\boldsymbol{x}$: $log(p_j(\boldsymbol{x}, \boldsymbol{\beta}) / p_0(\boldsymbol{x}, \boldsymbol{\beta})) = \boldsymbol{x}\boldsymbol{\beta}_j$.
- Este resultado se extiende a comparaciones entre j y h: $log(p_j(\boldsymbol{x}, \boldsymbol{\beta}) / p_h(\boldsymbol{x}, \boldsymbol{\beta})) = \boldsymbol{x}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)$
- ¿Cómo se interpretaría el caso de x_k dummy? Hacer.

Modelo logit multinomial

- Una vez estimado podemos construir las probabilidades estimadas, $p_j(\boldsymbol{x}, \boldsymbol{\beta}), j = 0, 1, ..., J$.
- Para cada i podemos predecir el resultado usando la mayor probabilidad. O sea, $\hat{y}_i = max\{j = 0, 1, ..., J : p_j(\boldsymbol{x}_i, \hat{\boldsymbol{\beta}})\}$.
- Podríamos construir una medida de bondad del ajuste: $peudo - R^2 = \frac{\sum_{i=1}^{N} 1[\hat{y}_i = y_i]}{N}$.
- McFadden (1974) propone usar el likelihood ratio index:

$$LRI = 1 - \frac{\mathcal{L}(\hat{\boldsymbol{\beta}})}{\mathcal{L}(\boldsymbol{\beta} = 0)}$$

Modelo logit multinomial

- Una de las características de este modelo es que las **mismas** variables $\boldsymbol{x}$ se usan para todas las alternativas j.
- En este caso los controles afectan al individuo, pero no son específicas de las caratcerśticas $j = 0, 1, 2, ..., J$.
- El logit condicional (ver más abajo) permite que las variables de control sean específicas de cada opción.

Modelo logit multinomial: STATA

- `http://www.stata.com/manuals13/rmlogit.pdf`
- `mlogit y x1 x2 x3`
- `margins, dydx(*) predict(outcome(1))` (efectos marginales en el promedio de las densidades)
- `margins, dydx(*) predict(outcome(1)) atmeans` (efectos marginales para $y = 1$ sobre $\bar{x}$)

Modelo de elección probabilística: logit condicional

(McFadden, 1974, Wooldridge, 2012, cap.16)

- Supongamos que hay $j = 0, 1, ..., J$ variables latentes que representan la utilidad del individuo i
$$y^*_{ij} = \boldsymbol{x}_{ij}\boldsymbol{\beta}_j + a_{ij},$$
donde a_{ij} son variables no observadas que afectan los "gustos" de las personas. $\boldsymbol{x}_{ij}$ es un vector de $1 \times K$ que puede diferir entre alternativas e individuos (notar que depende de j, no sólo de i).
- Ejemplo: alternativas de transporte, $\boldsymbol{x}_{ij}$ puede contener el tiempo de viaje, o el costo del viaje.
- Ejemplo: alternativas de prepagas, $\boldsymbol{x}_{ij}$ puede contener el costo o las características del plan.
- Definamos $y_i = max\{y^*_{i0}, y^*_{i1}, ..., y^*_{iJ}\}$. Si $a_{ij}, j = 0, 1, ..., J$ son variables aleatorias independientes con distribución $F(a) = exp[-exp(-a)]$ (distribución de valores extremos de tipo I), entonces

$$P(y_i = j|\boldsymbol{x}_i) = \frac{exp(\boldsymbol{x}_{ij}\boldsymbol{\beta}_j)}{\sum_{h=0}^{J} exp(\boldsymbol{x}_{ih}\boldsymbol{\beta}_h)}, j = 0, 1, ..., J.$$

Modelo de elección probabilística: logit condicional

(McFadden, 1974, Wooldridge, 2012, cap.16)

- Los efectos marginales son

$$\partial p_j(\boldsymbol{x})/\partial x_{jk} = p_j(\boldsymbol{x})[1 - p_j(\boldsymbol{x})]\beta_{jk}, j = 0, 1, ..., J, k = 1, ..., K$$

$$\partial p_j(\boldsymbol{x})/\partial x_{hk} = p_j(\boldsymbol{x})p_h(\boldsymbol{x})\beta_{hk}, j = 0, 1, ..., J, k = 1, ..., K$$

Independencia de alternativas irrelavantes

- Un gran problema de estos modelos (mlogit o clogit) es que la elección entre dos alternativas dadas no depende de una tercera. O sea, son comparaciones de a pares.
$$log(p_j(\boldsymbol{x},\boldsymbol{\beta})/p_h(\boldsymbol{x},\boldsymbol{\beta})) = \boldsymbol{x}(\boldsymbol{\beta}_j - \boldsymbol{\beta}_h)$$
Esto recibe el nombre de supuesto de la independencia de alternativas irrelevantes.
- Este supuesto viene del supuesto de independencia de los errores y homocedásticos.
- Se puede proponer un contraste de Hausman para ver la validez del modelo (Hausman y McFadden, 1984). Supongamos que la alternativa $j = h$ es irrelevante, entonces exlcuirla no afecta los resultados entre las restantes. Si es relevante, excluirla debería generar inconsistencias. Sin embargo el modelo con $j = 0, 1, ..., h-1, h+1, .., J$ es más eficiente que el modelo con $j = 0, 1, .., J$ (¿Por qué?)
Así,
$$(\hat{\boldsymbol{\beta}}_{all} - \hat{\boldsymbol{\beta}}_{-h})'[\hat{\boldsymbol{V}}_{-h} - \hat{\boldsymbol{V}}_{all})]^{-1}(\hat{\boldsymbol{\beta}}_{all} - \hat{\boldsymbol{\beta}}_{-h}) \xrightarrow{d} \chi^2_K$$
bajo la nula de alternativas irrelevantes.

Independencia de alternativas irrelavantes

- Existen otras alternativas que no tienen este supuesto.
- En este modelo $\boldsymbol{a}_i$ sigue una distribución multivariada normal con correlaciones arbitrarias entre a_{ij} y a_{ih}, para todo $j \neq h$.
- Sin embargo el modelo es mucho más complejo para estimar (problemas de convergencia).
- Ver en STATA: `http://www.stata.com/manuals13/rmprobit.pdf`

Modelo probit de orden (ordered probit model)

Supongamos que la variable dependiente toma muchos valores, ej. $y = 0, 1, 2..., J$, y los valores de y representan un orden en particular. Éste es el modelo de orden.

Ejemplo: y podría ser satisfacción con una compra de un bien; o evaluación de un candidato. En general aplica a evaluaciones subjetivas.
$y = 0$ "No me gusta"
$y = 1$ "Me es indiferente"
$y = 2$ "Me agrada"
$y = 3$ "Me gusta mucho"

¿Tiene sentido usar una regresión MCO?

Modelo probit de orden (ordered probit model)

- Asumamos una variable latente y^* dada por

$$y^* = \boldsymbol{x\beta} + e, \;\; e|\boldsymbol{x} \sim N(0,1)$$

- Consideremos un modelo con $J+1$ categorías indexadas por $j = 0, 1, 2..., J$y supongamos J umbrales o puntos de corte desconocidos $\alpha_1 < \alpha_2 < ... < \alpha_J$ que satisfacen
  $y = 0$ si $y^* \leq \alpha_1$
  $y = 1$ si $\alpha_1 < y^* \leq \alpha_2$
  $\vdots$
  $y = J$ si $y^* \geq \alpha_J$
- Entonces la distribución condicional es
  $P(y=0|\boldsymbol{x}) = P(y^* \leq \alpha_1|\boldsymbol{x}) = P(\boldsymbol{x\beta} + e \leq \alpha_1|\boldsymbol{x}) = \Phi(\alpha_1 - \boldsymbol{x\beta})$
  $P(y=1|\boldsymbol{x}) = P(\alpha_1 < y^* \leq \alpha_2|\boldsymbol{x}) = \Phi(\alpha_2 - \boldsymbol{x\beta}) - \Phi(\alpha_1 - \boldsymbol{x\beta})$
  $\vdots$
  $P(y=J-1|\boldsymbol{x}) = P(\alpha_{J-1} < y^* \leq \alpha_J|\boldsymbol{x}) = \Phi(\alpha_J - \boldsymbol{x\beta}) - \Phi(\alpha_{J-1} - \boldsymbol{x\beta})$
  $P(y=J|\boldsymbol{x}) = P(y^* > \alpha_J|\boldsymbol{x}) = 1 - \Phi(\alpha_J - \boldsymbol{x\beta})$

Modelo probit de orden (ordered probit model)

- El modelo se puede estimar por MLE:
$\ell_i(\boldsymbol{\beta}) = 1[y_i = 0] log[\Phi(\alpha_1 - \boldsymbol{x\beta})] + 1[y_i = 1] log[\Phi(\alpha_2 - \boldsymbol{x\beta}) - \Phi(\alpha_1 - \boldsymbol{x\beta})] + ... + 1[y_i = J] log[1 - \Phi(\alpha_J - \boldsymbol{x\beta})]$.
- Los efectos marginales se pueden calcular como:

$$\partial p_0(\boldsymbol{x})/\partial x_k = -\beta_k \phi(\alpha_1 - \boldsymbol{x\beta}),$$

$$\partial p_J(\boldsymbol{x})/\partial x_k = \beta_k \phi(\alpha_J - \boldsymbol{x\beta}),$$

$$\partial p_j(\boldsymbol{x})/\partial x_k = \beta_k [\phi(\alpha_{j-1} - \boldsymbol{x\beta}) - \phi(\alpha_j - \boldsymbol{x\beta}), 0 < j < J$$

Modelo probit de orden (ordered probit model)

- Un supuesto que sale del modelo es el **supuesto de regresiones paralelas**. Es decir, todas las alternativas se basan en el mismo coeficiente $\boldsymbol{\beta}$ y sólo cambia la constante.
- Por ejemplo, notemos que $P(y \leq j|\boldsymbol{x}) = \sum_{h=0}^{j} P(y = h|\boldsymbol{x}) = \Phi(\alpha_1 - \boldsymbol{x\beta}) + (\Phi(\alpha_2 - \boldsymbol{x\beta}) - \Phi(\alpha_1 - \boldsymbol{x\beta})) + ... + (\Phi(\alpha_j - \boldsymbol{x\beta}) - \Phi(\alpha_{j-1} - \boldsymbol{x\beta})) = \Phi(\alpha_j - \boldsymbol{x\beta})$, (sólo cambia la constante).
- Podríamos armar el modelo de orden usando una serie de modelos probit. Por ejemplo, construir $w_{ij} = 1$ si $y_{ij} \leq j$, $w_{ij} = 0$ si $y_{ij} > j$ para $j = 0, 1, ..., J - 1$.
- Entonces tenemos,

$$P(y \leq j|\boldsymbol{x}) = P(y^* \leq \alpha_j|\boldsymbol{x}) = \Phi(\alpha_j - \boldsymbol{x\beta}_j)$$

- El modelo de orden asume que los parámetros $\boldsymbol{\beta}$ son los mismos en todos los probit bivariados, excepto por la constante (α). Si estimamos cada probit por separado entonces los betas son específicos de cada modelo y tenemos más flexibilidad.
- Un subproducto de este análisis es obtener

$$\frac{P(y \leq j|\boldsymbol{x})}{\partial x_h} = -\beta_{hj} g(\alpha_j - \boldsymbol{x\beta}_j)$$

Modelo probit de orden (ordered probit model): STATA

- `oprobit y x1 x2 x3`
- `mfx, predict(p outcome(1))` (efectos marginales para $y = 1$)
- `mfx, predict(p outcome(2))` (efectos marginales para $y = 2$)
- `http://www.stata.com/manuals13/rologit.pdf`
- `http://www.stata.com/manuals13/roprobit.pdf`

Interval regression

En el modelo anterior se asumía que no se conocían los puntos de corte. Si estos se conocieran podemos armar un modelo con más información.

Ejemplo: y podría ser niveles de ingreso
$y = 0$ "Ingreso 0"
$y = 1$ "$0 < Ingreso \leq \$1000$"
$y = 2$ "$\$1000 < Ingreso \leq \$2000$"
$y = 3$ "$\$2000 < Ingreso$"

¿Tiene sentido usar una regresión MCO?