Introducción a la regresión ordinal

Jose Barrera
jbarrera@mat.uab.cat

20 de mayo 2009

Introducción a la regresión ordinal

Introducción a la regresión ordinal

Introducción. Modelos de regresión para respuesta categórica

- Si queremos modelar una variable respuesta categórica, Y, de categorías $y_1, \ldots, y_g$, con un conjunto de variables explicativas (factores o covariables) $\mathbf{X} = (X_1, \ldots, X_m)$, mediante un modelo lineal general, podemos plantearnos las opciones siguientes:

g	¿Y ordinal?	Regresión	Modelamos
2	No importa	Logística	$f(P(Y = y_2 \mid \mathbf{X})) = \alpha + \beta' \mathbf{X}$
$\geqslant 3$	No	Multinomial	$f(P(Y = y_j \mid \mathbf{X}_i)) = \alpha_i + \beta_j' \mathbf{X}_i \ , \quad j = 2, \ldots, g; \ i = 1, \ldots, n$
$\geqslant 3$	Sí	Ordinal	$f(\gamma_j(\mathbf{X})) = f(P(Y \leqslant y_j \mid \mathbf{X})) = \alpha_j + \beta' \mathbf{X} \ , \quad j = 1, \ldots, g-1$

donde $f()$ es la llamada *función de enlace* (habitualmente Logit, Log-Log o Probit), $\alpha_j + \beta' \mathbf{X}$ es el predictor lineal y α_j y $\beta = (\beta_1, \ldots, \beta_m)'$ parámetros a estimar.

- Así, el modelo de regresión ordinal es adecuado para modelar una variable respuesta ordinal, Y, con categorías ordenadas $y_1, \ldots, y_g$, $g \geqslant 3$.

Formulación del modelo de regresión ordinal

- El modelo de regresión ordinal es

$$f(\gamma_j(\mathbf{X})) = \alpha_j + \beta'\mathbf{X}\,, \quad j = 1, \ldots, g-1 \quad , \quad \gamma_j(\mathbf{X}) = P(Y \leqslant y_j|\mathbf{X})$$

- Las funciones de enlace habituales son

$$\begin{array}{lcl} \text{Logit} & : & f(Y) = \text{logit}(P(Y)) = \log\left(\frac{P(Y)}{1-P(Y)}\right) = \log \text{Odds}(Y) \\ \text{Log-Log} & : & f(Y) = \log\left[-\log(1-P(Y))\right] \\ \text{Probit} & : & f(Y) = \Phi^{-1}(P(Y))\,,\ \Phi(x) = P(Z \sim N(0,1) \leqslant x) \end{array}$$

- La no inclusión de diferentes umbrales α_j implicaría

$$\widehat{P}(Y \leqslant y_j|\mathbf{X}) = \widehat{P}(Y \leqslant y_k|\mathbf{X}) \quad , \quad j \neq k.$$

- El enlace Logit es adecuado cuando la respuesta está uniformemente representada y el enlace Log-Log cuando predominan las categorías elevadas (según [3], citado en [1]). **En adelante consideraremos el enlace Logit**.
- El modelo proporciona las probabilidades

$$\begin{array}{lcl} \text{Acumuladas} & : & P(Y \leqslant y_j|\mathbf{X}) = \left[1 + e^{-\left(\alpha_j + \beta'\mathbf{x}\right)}\right]^{-1} \\ \text{Absolutas} & : & P(Y = y_j|\mathbf{X}) = P(Y \leqslant y_j|\mathbf{X}) - P(Y \leqslant y_{j-1}|\mathbf{X}) \end{array}$$

- Parametrización: Algunos autores y/o paquetes estadísticos consideran el predictor lineal $\alpha_j - \beta'\mathbf{X}$ y/o modelan $P(Y \geqslant y_j|\mathbf{X})$ en lugar de $P(Y \leqslant y_j|\mathbf{X})$.

Suposiciones sobre el modelo

- No se supone normalidad, homocedasticidad ni incorrelación de los residuos.
- El modelo de regresión ordinal sí supone una condición sobre los datos a modelar: *odds proporcionales* o *líneas paralelas*, ya que del modelo se deduce

$$\widehat{\mathrm{OR}}_{\Delta X_k = c}(Y \leqslant y_j) = e^{\beta_k c}\,, \quad \forall j = 1, \ldots, g-1$$

- Ejemplo: 540 profesores y profesoras de enseñanza secundaria clasificados según el grado de sensación de estrés laboral y el índice de diversidad sociocultural del alumnado del correspondiente centro de trabajo. (Datos ficticios).

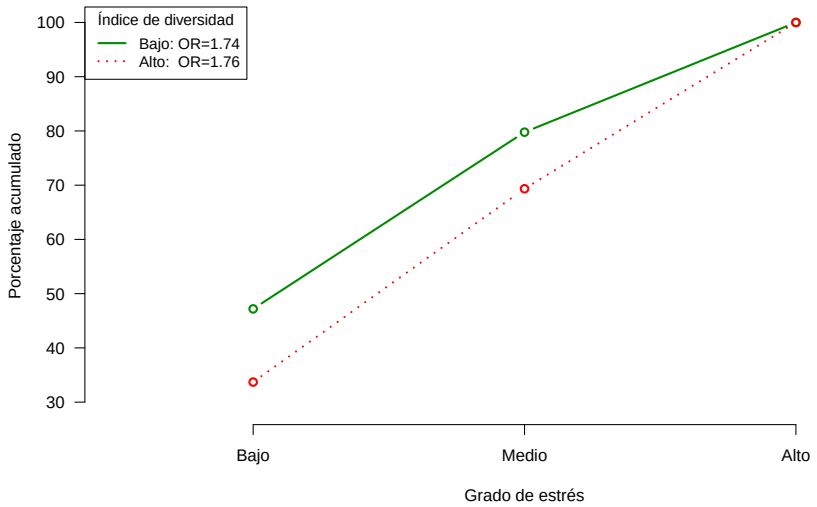
	Grado de estrés		
Índice de diversidad	Bajo	Medio	Alto
Bajo	84	58	36
Alto	122	129	111

($\widehat{\mathrm{OR}} \approx 1{,}74$)

	Grado de estrés	
Índice de diversidad	Bajo o Medio	Alto
Bajo	142	36
Alto	251	111

($\widehat{\mathrm{OR}} \approx 1{,}76$)

	Grado de estrés	
Índice de diversidad	Bajo	Medio o Alto
Bajo	84	94
Alto	122	240



Estimación y significación de los parámetros

- Podemos estimar los parámetros del modelo por MV maximizando (numéricamente) la función de verosimilitud

$$L(\alpha, \beta | Y, \mathbf{X}) = \text{✎} \cdots \text{✎} = \prod_{i=1}^{n} \prod_{j=2}^{g-1} \left[\frac{1}{1+e^{-(\alpha_1 + \boldsymbol{\beta}' \mathbf{x}_i)}} \right]^{\delta_{i1}} \left[\frac{1}{1+e^{-(\alpha_j + \boldsymbol{\beta}' \mathbf{x}_i)}} - \frac{1}{1+e^{-(\alpha_{j-1} + \boldsymbol{\beta}' \mathbf{x}_i)}} \right]^{\delta_{ij}}$$

donde

$$\delta_{ij} = \begin{cases} 1 & \text{si el } i\text{-ésimo individuo muestra } Y = y_j \\ 0 & \text{en caso contrario} \end{cases}$$

- $\hat{\theta}_{k,\text{MV}} \overset{\text{asint.}}{\sim} N\left(\theta_k, \sqrt{\hat{I}^{-1}_{kk}}\right)$ (I, matriz de información de Fisher) y podemos realizar la prueba de Wald para resolver el contraste de hipótesis

$$\text{H}_0 : \beta_k = 0 \text{ contra } \text{H}_1 : \beta_k \neq 0$$

según el estadístico

$$\frac{\hat{\beta}_k}{\sqrt{\hat{I}^{-1}_{kk}}} \overset{\text{H}_0}{\sim} N(0,1) \text{ o, equivalentemente, } \frac{\hat{\beta}^2_k}{\hat{I}^{-1}_{kk}} \overset{\text{H}_0}{\sim} \chi^2_1.$$

- Análogamente, el test de razón de verosimilitudes permite contrastar

$$\text{H}_0 : \beta_{k_1} = \cdots = \beta_{k_s} = 0 \text{ contra } \text{H}_1 : \exists r : \beta_{k_r} \neq 0.$$

Interpretación de los parámetros

- Consideraremos la parametrización

$$\log\left(\frac{P(Y\leqslant y_j|\mathbf{X})}{1-P(Y\leqslant y_j|\mathbf{X})}\right)=\alpha_j+\boldsymbol{\beta}'\mathbf{X}\,,\quad j=1,\ldots,g-1$$

- α_j:

$$\widehat{\log \text{Odds}}(Y\leqslant y_j|\mathbf{X}=0)=\alpha_j \iff e^{\alpha_j}=\widehat{\text{Odds}}(Y\leqslant y_j|\mathbf{X}=0)$$

- β_k:
  - Si X_k es un factor, $\widehat{\log \text{OR}}_{X_k}(Y\leqslant y_j)=\beta_k \iff e^{\beta_k}=\widehat{\text{OR}}_{X_k}(Y\leqslant y_j)$
  - Si X_k es covariable, $\widehat{\log \text{OR}}_{\Delta X_k=c}(Y\leqslant y_j)=\beta_k c \iff e^{\beta_k c}=\widehat{\text{OR}}_{\Delta X_k=c}(Y\leqslant y_j)$
  - Si existe interacción entre X_1 y X_2 (término $\beta_{12}X_1X_2$ en el modelo),
    - ★ Si X_1 y X_2 son factores: $\widehat{\text{OR}}_{X_1}(Y\leqslant y_j)=\begin{cases} e^{\beta_1} & , \quad X_2=0\\ e^{\beta_1+\beta_{12}} & , \quad X_2=1\end{cases}$
    - ★ Si X_1 es un factor y $X_2=x_2$ una covariable: $\widehat{\text{OR}}_{X_1}(Y\leqslant y_j)=\begin{cases} e^{\beta_1} & , \quad X_2=0\\ e^{\beta_1+\beta_{12}x_2} & , \quad X_2=1\end{cases}$
    - ★ Si X_1 es covariable y X_2 es un factor: $\widehat{\text{OR}}_{\Delta X_1=c}(Y\leqslant y_j)=\begin{cases} e^{\beta_1 c} & , \quad X_2=0\\ e^{(\beta_1+\beta_{12})c} & , \quad X_2=1\end{cases}$
    - ★ Si X_1 y $X_2=x_2$ son covariables: $\widehat{\text{OR}}_{\Delta X_1=c}(Y\leqslant y_j)=\begin{cases} e^{\beta_1 c} & , \quad X_2=0\\ e^{(\beta_1+\beta_{12}x_2)c} & , \quad X_2=1\end{cases}$
- X_k será un factor de riesgo (protección) si $\beta_k>0$ ($\beta_k<0$), si y_1 es *peor* que y_g.

Selección y validación del mejor modelo

- Cumplimiento de la asunción del modelo: test de *líneas paralelas*.
- Bondad de ajuste (R^2 ajustado).
- Estabilidad de los parámetros del modelo.
- Precisión en la clasificación. Podemos clasificar el individuo *i* de la muestra en aquella categoría *j* para la cual la probabilidad

$$\widehat{P}(Y = y_j|\mathbf{X} = \mathbf{X}_i) = \widehat{\gamma}_j(\mathbf{X} = \mathbf{X}_i) - \widehat{\gamma}_{j-1}(\mathbf{X} = \mathbf{X}_i)$$

  es máxima y crear y analizar la matriz de confusión (tabla de contingencia de las frecuencias observadas y esperadas).
- Parsimonia mediante, por ejemplo, el AIC (Criterio de Información de Akaike).

El modelo de regresión ordinal en R

- Podemos ajustar un modelo de regresión ordinal en R con la función `polr` de la librería `MASS`.
- Esta función considera la parametrización

$$\log\left(\frac{P(Y \leqslant y_j|\mathbf{X})}{1-P(Y \leqslant y_j|\mathbf{X})}\right) = \alpha_j - \beta'\mathbf{X}$$

- Acepta cualquiera de los enlaces Logit, Log-Log o Probit.

Ejemplo de aplicación

- Estudio observacional transversal sobre estudiantes de un centro universitario de ciencias de la salud, con predominio de minorías, graduados entre los años 1999 y 2001 ([1]).
- Variable respuesta: nivel de satisfacción global que mostraron los estudiantes con su experiencia en la facultad ("muy insatisfecho", "insatisfecho", "satisfecho", "muy satisfecho").
- Variables explicativas: dos variables relacionadas con aspectos demográficos (sexo y etnia -Afroamericana o no-) y un cuestionario de 42 preguntas midiendo el grado de satisfacción con aspectos relacionados con el entorno de aprendizaje del estudiante.
- Criterios para la selección de los modelos: validación de la asunción del modelo, estabilidad, bondad de ajuste y parsimonia.
- Mejor modelo obtenido:
  - Enlace Logit.
  - Superó el test de *líneas paralelas*.
  - Resultaron significativas ($\alpha = 0{,}05$) tres variables explicativas (todas con asociación directa a los niveles altos de la variable respuesta): competencia del profesorado, relación alumnado-profesorado y promoción de hábitos saludables/prevención de enfermedades, además de los puntos de corte correspondientes a las categorías 2 ("satisfecho") y 3 ("muy satisfecho") de la variable respuesta.
  - La matriz de confusión clasificó correctamente el 75 % de los individuos.

Ejemplo de aplicación

- Coeficientes significativos en el modelo

Variable	Coeficiente	*p*-valor
Umbral ("Insatisfecho")	5,782	0,001
Umbral ("Satisfecho")	10,020	0,000
X_1 = Promoción salud/Prevención enfermedades	0,674	0,033
X_2 = Competencia profesorado	0,700	0,026
X_3 = Relaciones alumnado—profesorado	1,291	0,000

- Ejemplos de estimación de probabilidades (incorrecta/correcta)

$\alpha_j + \beta' \mathbf{X}$	Muy insatisfecho	Insatisfecho	Satisfecho	Muy satisfecho
$\mathbf{X} = 0$	0,50	0,50	0,00	0,00
$\mathbf{X} = X_1$	0,66	0,34	0,00	0,00
$\mathbf{X} = X_2$	0,67	0,33	0,00	0,00
$\mathbf{X} = X_3$	0,78	0,21	0,00	0,00
$\mathbf{X} = X_1 + X_2 + X_3$	0,93	0,06	0,00	0,00

$\alpha_j - \beta' \mathbf{X}$	Muy insatisfecho	Insatisfecho	Satisfecho	Muy satisfecho
$\mathbf{X} = 0$	0,50	0,50	0,00	0,00
$\mathbf{X} = X_1$	0,34	0,66	0,01	0,00
$\mathbf{X} = X_2$	0,33	0,66	0,01	0,00
$\mathbf{X} = X_3$	0,22	0,77	0,01	0,00
$\mathbf{X} = X_1 + X_2 + X_3$	0,07	0,89	0,04	0,00

Chen, C. y Hughes, J. (2004). Using Ordinal Regression Model to Analyze Student Satisfaction Questionnaires. *IR Applications*. Vol. 1. (Acceso al documento en `http://www.airweb.org/page.asp?page=554`).

Kleinbaum, D.G. y Klein, M. (2002). *Logistic Regression. A Self-Learning Text*. Springer.

SPSS, Inc. (2002). Ordinal Regression Analysis, *SPSS Advanced Models 10.0*. Chicago, IL.