Regresiones ordinales con R

UNTREF

Modelos estadísticos

- Un modelo es una representación simplificada de un sistema (Bodo Winter, 2020)

- Podemos generar diferentes tipos de modelos según nuestros datos

Regresión lineal

- Tipo de variable respuesta: cuantitativa

- El modelo de regresión lineal (simple) es un modelo para el vínculo de dos variables aleatorias
  - X = variable predictora o covariable
  - Y = variable dependiente o de respuesta

- El modelo se denomina lineal pues propone que la Y depende linealmente de X

(Szretter, 2017)

Regresión lineal

- El modelo de regresión lineal (simple) es un modelo para el vínculo de dos variables aleatorias

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Donde

- β_0 = ordenada al origen
- β_1 = pendiente
- ε_i = error para el individuo i-ésimo

Alternativamente

$$E(Y \mid X) = \beta_0 + \beta_1 X$$

(Szretter, 2017)

Regresión lineal

- Supuestos del modelo lineal
  - los ε_i tiene media cero, $E(\varepsilon_i) = 0$
  - los ε_i tienen todos la misma varianza desconocida que llamaremos σ^2 y que es el otro parámetro del modelo, $Var(\varepsilon_i) = \sigma^2$ (homoscedasticidad)
  - 3. los ε_i tienen distribución normal
  - 4. los ε_i son independientes entre sí, y son no correlacionados con las X_i:

- Interpretación de coeficientes

(Szretter, 2017)

Regresión logística

- Tipo de variable respuesta: binaria
  - Dos valores posibles de la respuesta para cada ensayo → 0/1

- En el modelo de regresión logística, NO nos preguntamos por el valor de Y (1 o 0), sino:
  - Dado un valor de X, ¿cuál es la *probabilidad* de que Y tome valor 1?

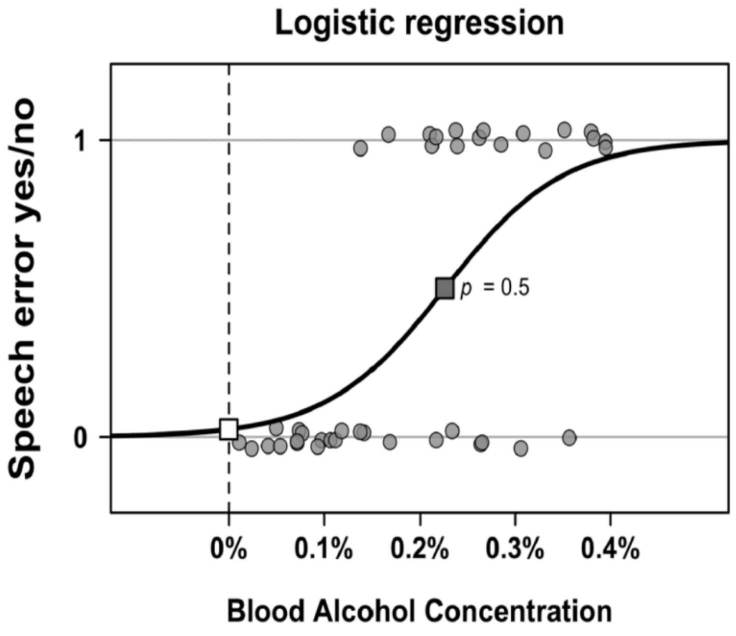


Figure 12.1. Speech errors as a function of blood alcohol concentration, treated as a binary categorical variable with superimposed logistic regression fit (bold curve); the white square indicates the intercept; the square in the middle indicates the point where making a speech error becomes more likely than not making a speech error

(Instituto de Cálculo UBA, 2017)

(Bodo Winter, 2020)

Regresión logística

- Modelo de regresión logística

$$\ln\left(\frac{\mathrm{p}(x)}{1 - \mathrm{p}(x)}\right) = \alpha + \beta\, x$$

- ODDS = proporción de la probabilidad de que algo ocurra sobre la probabilidad de que no ocurra

(Instituto de Cálculo UBA, 2017)

Regresión logística: Interpretación de coeficientes

El coeficiente β es el cambio en $\ln\left(\frac{\mathrm{p}(x)}{1 - \mathrm{p}(x)}\right) = \mathrm{logit}(\mathrm{p}(\mathrm{x}))$ **cuando la variable X aumenta en 1 unidad.**

Más específicamente

$$\ln\left(\frac{\mathrm{p}(x+1)}{1-\mathrm{p}(x+1)}\right) = \alpha + \beta \cdot (x+1)$$

Luego

$$\ln\left(\frac{\mathrm{p}(x+1)}{1-\mathrm{p}(x+1)}\right) - \ln\left(\frac{\mathrm{p}(x)}{1-\mathrm{p}(x)}\right) = \beta$$

y de manera equivalente

$$\ln\left(\frac{\frac{\mathrm{p}(x+1)}{1-\mathrm{p}(x+1)}}{\frac{\mathrm{p}(x)}{1-\mathrm{p}(x)}}\right) = \beta\,.$$

Tomando exponencial de cada lado de la desigualdad queda:

$$\left. e^{\beta} = \frac{\frac{\mathrm{p}(x+1)}{1-\mathrm{p}(x+1)}}{\frac{\mathrm{p}(x)}{1-\mathrm{p}(x)}} \right\} \text{Odds ratio}$$

(Instituto de Cálculo UBA, 2017)

Regresión logística ordinal

Modelo de odds proporcionales / de logit acumulado

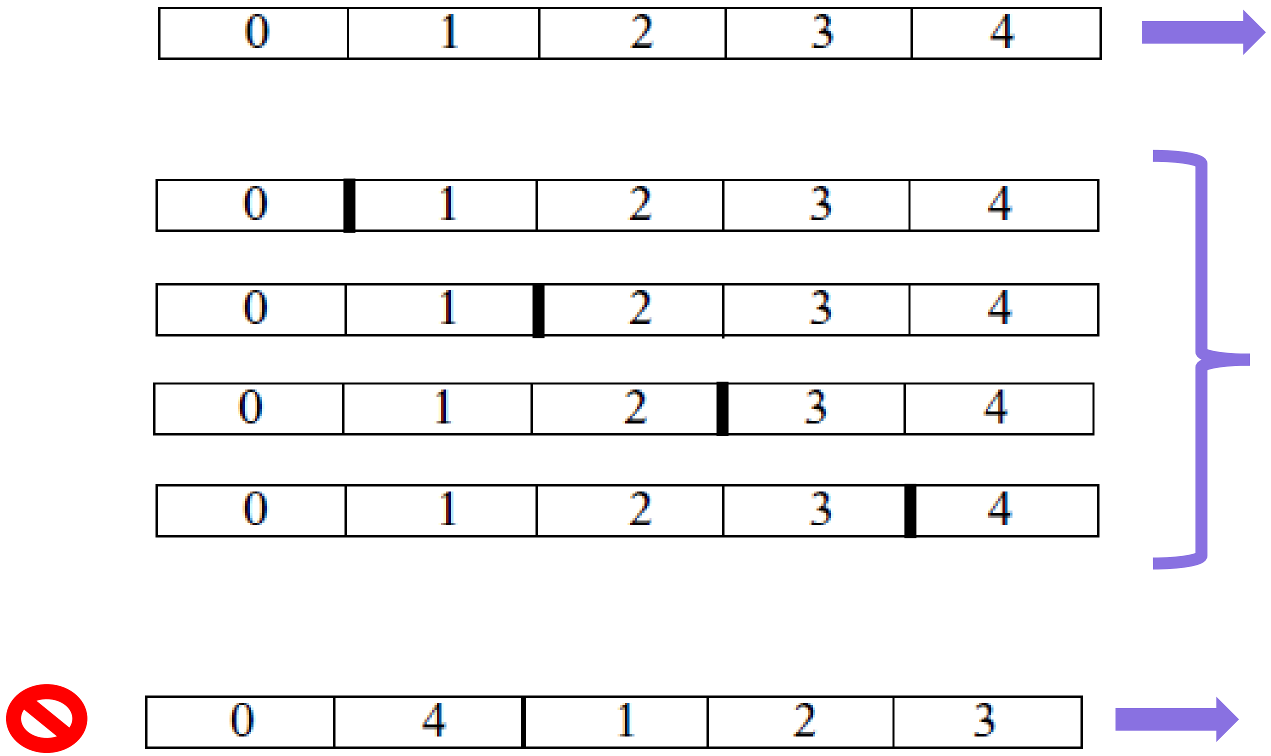
Regresión logística ordinal

- ¿Qué pasa cuando mi variable respuesta presenta más de dos categorías ordenadas?
  - Frecuencia de un síntoma, gravedad de un tumor, severidad de dolor

- No podemos usar regresión lineal ni regresión logística, pero sería interesante mantener la información del orden inherente de los niveles de la variable → podemos emplear una regresión logística ordinal usando el modelo de odds proporcionales (o modelo de logit acumulado)

Modelo de odds proporcionales

- El modelo de odds proporcionales requiere que colapsemos las categorías en dos:



Si nuestra variable respuesta Y tiene J categorías…

Hay J-1 formas de dicotomizar las categorías de la variable respuesta

Y hay formas no permitidas para mantener el orden natural

(Kleinbaum & Klein, 2002)

Modelo de odds proporcionales

- Entonces, si una variable respuesta variable respuesta ordinal D tiene G categorías (D = 0, 1, 2..., G – 1), entonces hay G – 1 formas de dicotomizar la respuesta.
- Así, las odds de que D ≥ g es igual a la probabilidad de D ≥ g dividido la probabilidad de que D < g donde (g = 1, 2, 3, ..., G – 1):

$$\text{odds}\,(D \geqslant g) = \frac{\mathrm{P}(D \geq g)}{\mathrm{P}(D < g)}$$

(Kleinbaum & Klein, 2002)

Modelo de odds proporcionales

- **Supuesto de odds poporcionales**
  - **El odds ratio del efecto de la variable de exposición es invariante a dónde se dicotomizaron las categorías de la variable respuesta.**
  - En otras palabras, el OR que evalúa el efecto de una variable predictora en cualquiera de las comparaciones va a ser el mismo más allá de donde se haga el punto de corte.
  - Si tenemos una variable respuesta con 5 niveles y una variable predictora categórica dicotómica (E = 1, E = 0), …
    - …bajo este supuesto, el OR que compara las categorías mayores o iguales a 1 con las menores o iguales a 1 es IGUAL al OR que compara las categorías mayores o iguales a 4 con las menores o iguales a 4

(Kleinbaum & Klein, 2002)

Modelo de odds proporcionales

- **Supuesto de odds poporcionales**

  Esto implica que:
  - sólo hay un parámetro (β) por cada variable predictora
  - Hay una intercepta separada para cada una de las G – 1 comparaciones

- Esto NO implica que el odds de un patrón de exposición sea invariante
  - Por ejemplo, para el valor de exposición 0 las odds que comparan categorías mayores o iguales a 1 con las menores a 1 no son equivalentes a las odds que comparan categorías mayores o iguales a 4 con las menores a 4

(Kleinbaum & Klein, 2002)

Modelo de odds proporcionales

- El modelo para un predictor

D = variable respuesta ordinal
G = categorías
g = 1, 2, 3, ..., G - 1

$$\text{odds} = \frac{P(D \geq g \mid X_1)}{1 - P(D \geq g \mid X_1)} = \frac{P(D \geq g \mid X_1)}{P(D < g \mid X_1)}$$

$$= \frac{\dfrac{1}{1+\exp[-(\alpha_g + \beta_1 X_1)]}}{\dfrac{\exp[-(\alpha_g + \beta_1 X_1)]}{1+\exp[-(\alpha_g + \beta_1 X_1)]}} = \exp(\alpha_g + \beta_1 X_1)$$

Odds de una *inequidad*

$$\text{odds} = \frac{P(D^* \boxed{\leq} g \mid X_1)}{P(D^* \boxed{>} g \mid X_1)} = \exp(\alpha_g^* \boxed{-} \beta_1^* X_1)$$

Versión alternativa
Mismos β pero distintos α

(Kleinbaum & Klein, 2002)

Regresión ordinal con el paquete de R **ordinal**

Christensen (2018, 2022)

Paquete **ordinal**

- Existen diversos paquetes en R que permiten implementar modelos de regresión ordinal (e.g. MASS, VGAM), pero nos vamos a enfocar en el paquete ordinal (Christensen, 2022)

- **Ordinal** nos permite ajustar modelos con la función `clm` (de *cumulative link models*)

Ajustando un modelo con `clm` en R

- Usamos el dataset `wine` que viene con este paquete
- Son datos de un experimento (Randall, 1989) sobre factores que afectan la amargura percibida en el vino:
  - Rating: amargura del vino (de 1 = menos amargo a 5 = más amargo)
  - Temperatura: Factor de tratamiento del vino (frío o caliente)
  - Contacto: Factor de tratamiento del vino, si hubo contacto entre el jugo y la piel de las uvas al aplastarlas durante la producción (sí o no)
  - Hay 72 observaciones hechas por 9 jueces que evaluaron vinos de dos botellas de cada una de las cuatro condiciones de tratamiento

(Christensen, 2018)

Ajustando un modelo con `clm` en R

- Ajustamos el siguiente modelo:

$$\text{logit}(P(Y_i \leq j)) = \theta_j - \beta_1(\text{temp}_i)$$

$$i = 1, \ldots, n, \; j = 1, \ldots, J - 1$$

- Donde
- $\beta_1(\text{temp}_i)$ toma los valores $\beta_1(\text{frío})$ y $\beta_1(\text{caliente})$
- Es un modelo para la probabilidad acumulada del rating *i*ésimo cayendo en la *j*ésima categoría o más baja, donde *i* indiza todas las observaciones (n = 72), $j = 1, \ldots, J$ indiza las categorías de respuesta ($J = 5$) y θ_j es la intercepta del *j*ésimo logit acumulado: $\text{logit}(P(Y_i \leq j))$.

(Christensen, 2018)

¡Pasemos al código!

Bibliografía sobre regresión ordinal

- Christensen, R. H. B. (2018). Cumulative link models for ordinal regression with the R package ordinal. *Submitted in J. Stat. Software*, *35*.

- Kleinbaum, D. G., Klein, M., & Pryor, E. R. (2002). Ordinal Logistic Regression. En *Logistic regression: a self-learning text* (Vol. 94). New York: Springer.