Modelos Lineales para Clasificación

Modelos para Respuesta Ordinal

Juan Zamora O.

INSTITUTO DE ESTADÍSTICA

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Modelos con respuestas ordinales

En esta parte del Curso, revisaremos problemas en que las categorias de la respuesta en el problema de regresión están ordenadas

Modelos en donde Las respuestas tienden a tomar cateogrías más altas o más bajas a medida que el valor de una covariable se incrementa

Es posible explotar información del ordenamiento natural de las categorías y así obtener un modelo más parsimonioso que el multinomial

Modelos con respuestas ordinales

Errores de modelamiento comúnes

Uso de modelos clásicos ignorando que la respuesta es ordinal. Esto genera una interpretación deficiente de los resultados.

Otra manera de abordar este tipo de problemas es mediante el uso de regresión binaria sobre una varsión colapsada de las categorías. El problema en este caso es la pérdida de información.

Modelos con respuestas ordinales

Problemas de ejemplo

- Recuperación de pacientes según calidad de vida (Excelente, Buena, Suficiente, Pobre)
- Dolor (Ninguno, Leve, Considerable, Severo)
- Inclinación política (Extrema izquierda, Izquierda moderada, Moderado, Derecha moderada, Extrema derecha)
- Gasto de una Institución (Muy bajo, Medio, Muy alto)
- Categorización de una variable ordinal. Indice de masa corporal (BMI=peso/altura): $(< 18.5, 18.5 - 25, 25 - 30, > 30)$ para (Bajo peso, Peso normal, Sobrepeso)

El problema ordinal

Sea c el número de categorías de la variable respuesta Y. A su vez las probabilidades de la respuesta satisfacen $\prod_{i=1}^{c} \pi_i = 1$

El foco está en modelo como $P(y = j),\ j = 1, 2, \ldots, c$, depende de las variables explicativas x

Cuando todas las covariables son discretas, el conjunto de datos puede ser agrupado en forma de conteo en las c categorias de Y para cada configuración de las variables explicativas

Estrategia de modelamiento

Una forma de modelar respuestas binarias es comenzar desde modelos binarios. Es decir, las categorías ordenadas $1 \ldots c$ son transformadas a una respuesta binaria

Modelo Acumulativo Simple

Una forma simple y muy usada es aquella que se deriva del supuesto que las categorías observadas en Y representan una versión discreta de un modelo de regresión continua

$$\widetilde{Y} = \boldsymbol{x}^T\beta + \epsilon$$

Por lo tanto $Y = r \Leftrightarrow \tau_{r-1} < \widetilde{Y} \leq \tau_r$. Siendo $(\tau_0, \tau_1, \ldots, \tau_c)$ puntos de corte en la escala latente. Por lo tanto,

$$P(Y \leq r|\boldsymbol{x}) = P(\widetilde{Y} \leq \tau_r|\boldsymbol{x}) = F(\tau_r - \boldsymbol{x}^T\beta)$$

Luego el modelo queda especificado como

$$F^{-1}(P(Y \leq r|\boldsymbol{x})) = \tau_r - \boldsymbol{x}^T\beta$$

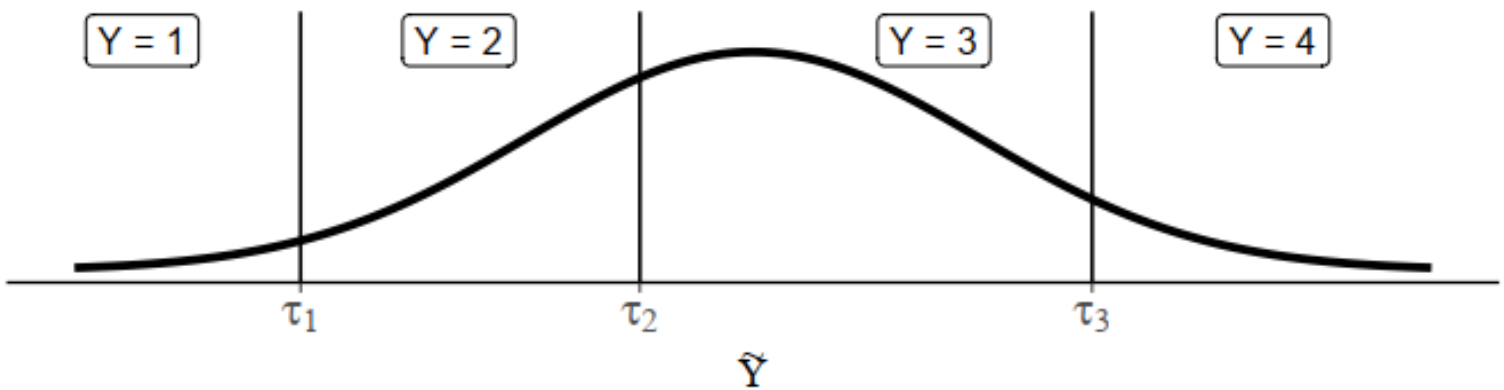
Por ejemplo, cuando F es la función acumulada de la distribución logística, su inversa F^{-1} es la función *logit*. Por lo tanto, el **modelo acumulativo logit** consiste en

$$logit(P(Y \leq r|\boldsymbol{x})) = \log\left(\frac{P(Y \leq r|\boldsymbol{x})}{P(Y > r|\boldsymbol{x})}\right) = \tau_r - \boldsymbol{x}^T\beta$$

Este modelo funciona bien cuando el modelo de regresión se aplica a una respuesta logística.

Otro modelo similar se genera cuando F es la función de densidad acumulada de una normal standard. Este modelo se denomina modelo acumulativo probit y funciona bien cuando el modelo de regresión subyacente tiene una $\widetilde{Y} \sim \mathcal{N}(0, 1)$

Por ejemplo, para una respuesta con 4 niveles cuyo modelo de regresión subyacente con una covariable que además supone una distribución normal standard para $\widetilde{Y}$:



Conclusiones generales

Usualmente, un coeficiente $\beta_j > 0$ se interpreta como un efecto positivo de esa covariable en la probabilidad de que Y tome determinada categoría.

A menudo se hace una interpretación de la variable latente (no observada). Sin embargo, el modelo resultante puede ser usado sin referencia alguna a la variable latente.

Debido a la derivación del modelo de regresión latente, el efecto de $\boldsymbol{x}$ no depende de la categoría, es decir es un **efecto global**.

El modelo más usado es el logit acumulado o también llamado modelo de probabilidades proporcionales (*proportional odds*)