

Regresión ordinal

Introducción

El test de Likert es una escala ordinal. Tratar las respuestas a un test de Likert como si fueran cuantitativas como se hizo en el análisis de la varianza del apartado anterior no es correcto por las siguientes razones:

- Los niveles de respuesta pueden no ser equidistantes: la distancia entre un par de opciones de respuesta puede no ser la misma para todos los pares de opciones de respuesta. Por ejemplo, la diferencia entre “Muy en desacuerdo” y “En desacuerdo” y la diferencia entre “De acuerdo” y “Muy de acuerdo” es de un nivel, pero psicológicamente puede ser percibida de forma diferente para cada sujeto.
- La distribución de las respuestas ordinales puede ser no normal. En particular esto sucederá si hay frecuencias altas de respuesta en los extremos del cuestionario.
- Las varianzas de las variables no observadas que subyacen a las variables ordinales observadas pueden diferir entre grupos, tratamientos, periodos, etc.

En Liddell y Kruschke (2018) se han analizado los problemas que puede ocasionar tratar datos ordinales como si fueran cuantitativos constatando que se pueden presentar las siguientes situaciones:

- Se pueden encontrar diferencias significativas entre grupos cuando no las hay: Error tipo I.
- Se pueden obviar diferencias cuando en realidad sí existen: Error tipo II.
- Incluso se pueden invertir los efectos de un tratamiento.
- También puede malinterpretarse la interacción entre factores.

Todos estos problemas pueden ser tratados con la regresión ordinal.

Variantes de la regresión ordinal.

Según Bürkner y Vuorre (2019) hay tres clases de regresión ordinal:

- Regresión ordinal acumulativa.
- Regresión ordinal secuencial.
- Regresión ordinal adyacente.

Nos centraremos en la primera ya que es el más habitual y adecuada para nuestro caso.

El modelo acumulativo, CM, presupone que la variable ordinal observada, Y , proviene de la categorización de una variable latente (no observada) continua \tilde{Y} . Hay K umbrales τ_k que particionan \tilde{Y} en $K + 1$ categorías ordenadas observables (ver Figura ??). Si asumimos que \tilde{Y}

tiene una cierta distribución (por ejemplo, normal) con distribución acumulada F , se puede calcular la probabilidad de que Y sea la categoría k de esta forma:

$$Pr(Y = k) = F(\tau_k) - F(\tau_{k-1})$$

Por ejemplo en la Figura ??,

$$Pr(Y = 2) = F(\tau_2) - F(\tau_1)$$

Si suponemos que, por ejemplo:

$$\tilde{Y} = \eta + \epsilon = b_1x_1 + b_2x_2 + \epsilon$$

Y que los errores son $N(0, \sigma^2)$.

Entonces:

$$Pr(\epsilon \leq z) = F(z)$$

Y:

$$Pr(Y \leq k \mid \eta) = Pr(\tilde{Y} \leq \tau_k \mid \eta) = Pr(\eta + \epsilon \leq \tau_k) = Pr(\epsilon \leq \tau_k - \eta) = F(\tau_k - \eta)$$

Por lo que:

$$Pr(Y = k) = \Phi(\tau_k - (b_1x_1 + b_2x_2)) - \Phi(\tau_{k-1} - (b_1x_1 + b_2x_2))$$

Donde hay que estimar los umbrales y los coeficientes de regresión.

Otra popular elección es suponer que la función acumulada se comporta como una logística. En ese caso, la interpretación de los coeficientes varía y se asemeja a la de la regresión logística. Se parte del supuesto de que:

$$\text{logit}(P(Y \leq k)) = \tau_k - \eta = \tau_k - (b_1x_1 + b_2x_2)$$

Se puede demostrar que, por ejemplo:

$$\frac{\frac{Pr(Y \leq k_1 \mid \eta)}{Pr(Y > k_1 \mid \eta)}}{\frac{Pr(Y \leq k_2 \mid \eta)}{Pr(Y > k_2 \mid \eta)}} = \exp(\tau_{k_1} - \tau_{k_2})$$

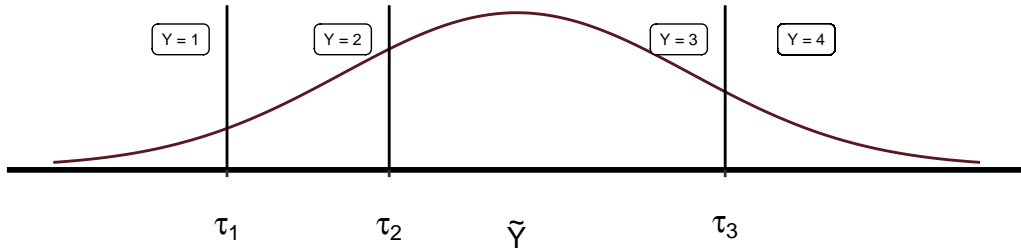


Figura 1: Regresión ordinal acumulativa.

Preparación

Rows: 2,980

Columns: 6

```
$ Group    <fct> AB, AB, AB, AB, AB, AB, AB, AB, AB, AB, AB, AB, AB, AB, AB, A~
$ Period   <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2~
$ Treat    <fct> A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, B, B, B~
$ Subject  <fct> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4~
$ Question <fct> Q01, Q02, Q03, Q04, Q05, Q06, Q07, Q08, Q09, Q10, Q11, Q12, Q~
$ Response <ord> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3~
```

Tabla 1: Resumen de frecuencias de respuesta.

Group	Period	Treat	Response				
			1	2	3	4	5
AB	1	A	2	25	71	203	434
AB	2	B	87	185	121	172	166
BA	1	B	76	174	127	237	138
BA	2	A	2	30	64	345	321

Modelo de enlace logit acumulado

Vamos a ajustar el modelo con la función de enlace logit:

$$\text{logit}(P(y_i \leq k)) = \log \frac{P(y_i \leq k)}{1 - P(y_i \leq k)} \quad (1)$$

La función de enlace logit acumulada (Ecuación ??) no está definida para $k = K$, ya que $1 - P(Y_i \leq K) = 1 - 1 = 0$.

En nuestra escala de Likert tenemos $K = 5$ niveles, el modelo mixto que vamos a plantear es el siguiente:

$$\begin{aligned} \text{logit}(p(y_i \leq k)) &= \tau_k - \beta_1 \text{Period}_i - \beta_2 \text{Treat}_i - \beta_3 \text{Question}_i - u(\text{Subject}_i) \\ i &= 1, \dots, n \quad k = 1, \dots, K - 1 \end{aligned}$$

donde τ_k es el umbral de la categoría k y son $K - 1 = 4$ interceptores. Los coeficientes de los efectos fijos, β_1 , β_2 y β_3 , son independientes, por lo que cada β tiene el mismo efecto en los $K - 1$ logits acumulados. El efecto aleatorio, Subject, se presupone que sigue una distribución normal: $u(\text{Subject}_i) \sim N(0, \sigma_u^2)$.

En esencia lo que estamos haciendo es un modelo en cadena de regresiones logísticas donde la respuesta binaria se corresponde con “menor o igual que cierto nivel frente a mayor que ese nivel”.

En el caso particular de $K = 5$, los umbrales τ_k se interpretan como:

- $k = 1$: log-odds del nivel = 1 vs. 2-5
- $k = 2$: log-odds del nivel = 1-2 vs. 3-5
- $k = 3$: log-odds del nivel = 1-3 vs. 4-5
- $k = 4$: log-odds del nivel = 1-4 vs. 5

Ajuste del modelo.

Vamos a usar las funciones `ordinal::clm` y `ordinal::clmm` que respectivamente permiten ajustar modelos únicamente con efectos fijos o modelos mixtos (efectos fijos y aleatorios).

Comenzamos con un modelo simple que tiene un único predictor:

$$\text{logit}(p(y_i \leq k)) = \tau_k - \beta_2 \text{Treat}_i$$

```

clm_likert_treat <-
  clm(
    Response ~ Treat,
    data = df, link = "logit"
  )
summary(clm_likert_treat)

```

```

formula: Response ~ Treat
data:    df

```

```

link threshold nobs logLik   AIC      niter max.grad cond.H
logit flexible  2980 -3966.11 7942.21 5(0)   1.64e-10 3.1e+01

```

Coefficients:

```

      Estimate Std. Error z value Pr(>|z|)
TreatB  -1.7206     0.0731  -23.54  <2e-16 ***
---

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Threshold coefficients:

```

      Estimate Std. Error z value
1|2 -3.97230     0.09678 -41.045
2|3 -2.45446     0.06812 -36.029
3|4 -1.66453     0.05936 -28.042
4|5 -0.10547     0.04946  -2.132

```

El método `summary` muestra la información resumen. Para su interpretación vamos a seguir Christensen (2018). La sección de coeficientes es la más importante. Se muestra la estimación de parámetros, el error “standard” y la significación estadística de acuerdo al test de Wald ¹. Comprobamos que el valor es claramente significativo. Es decir, que los estudiantes han valorado de forma diferente la calidad del subtítulo en ambos vídeos. El estimador de máxima verosimilitud del coeficiente `TreatB` es -1.72. La interpretación que podemos hacer es la siguiente:

- El signo es negativo. Eso quiere decir que el subtítulo que hemos denominado como tratamiento B es valorado como de peor calidad por los estudiantes ².
- Los umbrales que cortan la función latente se han desplazado β_2 hacia la izquierda entre el subtítulo A y B.

¹El test de Wald es un contraste de hipótesis estadístico en el que se evalúa si el valor estimado es cero suponiendo que $W = \left(\frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \right)^2 \sim \chi^2$.

²Recuérdese que el parámetro β_2 tiene signo negativo en la ecuación de regresión.