UNAP
FINESI

IV Semestre

Modelos Discretos

Regresión Multinomial

Mtr. Alcides Ramos Calcina

Regresión Multinomial

Mtr. Alcides Ramos Calcina

Introducción

- En este tema se considera un modelo de regresión logística donde la variable dependiente tiene más de dos categorías.

- La respuesta puede bien ser nominal, la que trataremos en esta sección o puede ser ordinal y se tratara en la siguiente sección. A su vez, las variables explicativas pueden ser cuantitativas o categóricas.

- En los modelos de regresión multinomial se asume que los recuentos de las categorías de Y tienen una distribución multinomial. Esta distribución es, a su vez, una generalización de la distribución binomial.

- Por tanto, La regresión logística multinomial es una extensión de la regresión logística binomial.

1. Modelo Multinomial

En regresión logística binaria, se tomo una categoría,

$Y = 0$, como referencia y se definió la ventaja $P(Y = 1)/P(Y = 0)$ de la otra categoría de respuesta, $Y = 1$.

En el caso policotómico, como hay c categorías, tomando como referencia a una de ellas, por ejemplo, a la categoría 1, podemos definir $c - 1$ ventajas del tipo

$$\frac{P(Y=s)}{P(Y=1)} \qquad \text{con } s = 2, 3, \ \ldots, c.$$

por tanto, tenemos los $c - 1$ logits correspondientes.

Supongamos que codificamos las tres categorías de la variable respuesta como 0, 1 y 2. En el caso de regresión logística, el logit es $\log\left[\frac{P(Y=1)}{P(Y=0)}\right]$

1. Modelo Multinomial

Ahora el modelo necesita dos funciones logit ya que tenemos tres categorías, y necesitamos decidir que categorías queremos comparar.

Lo más general es utilizar Y = 0 como referencia y formar logits comparándola con Y = 1 y Y = 2.

Supongamos que tenemos *m* variables independientes o explicativas, entonces:

$$\log\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_{01} + \beta_{11}X_1 + \beta_{21}X_2 + \cdots + \beta_{m1}X_m$$

$$\log\left[\frac{P(Y=2)}{P(Y=0)}\right] = \beta_{02} + \beta_{12}X_1 + \beta_{22}X_2 + \cdots + \beta_{m2}X_m$$

1. Modelo Multinomial

con lo cual ahora tenemos el doble de coeficientes que en el caso de regresión logística. Las probabilidades se calcularían como:

$$P\left(Y=0 \mid X\right)=\frac{1}{1+e^{\beta_{01}+\beta_{11}X_1+\beta_{21}X_2+\cdots+\beta_{m1}X_m}+e^{\beta_{02}+\beta_{12}X_1+\beta_{22}X_2+\cdots+\beta_{m2}X_m}}$$

$$P\left(Y=1 \mid X\right)=\frac{e^{\beta_{01}+\beta_{11}X_1+\beta_{21}X_2+\cdots+\beta_{m1}X_m}}{1+e^{\beta_{01}+\beta_{11}X_1+\beta_{21}X_2+\cdots+\beta_{m1}X_m}+e^{\beta_{02}+\beta_{12}X_1+\beta_{22}X_2+\cdots+\beta_{m2}X_m}}$$

$$P\left(Y=2 \mid X\right)=\frac{e^{\beta_{02}+\beta_{12}X_1+\beta_{22}X_2+\cdots+\beta_{m2}X_m}}{1+e^{\beta_{01}+\beta_{11}X_1+\beta_{21}X_2+\cdots+\beta_{m1}X_m}+e^{\beta_{02}+\beta_{12}X_1+\beta_{22}X_2+\cdots+\beta_{m2}X_m}}$$

La estimación de los parámetros se hace nuevamente mediante el método de máxima verosimilitud.

1. Modelo Multinomial

Generalizando lo anterior, sea Y una variable respuesta o dependiente nominal con c categorías con valores $Y = 1$, $Y = 2$, ..., $Y = c$, y sean X_1, X_2, …, X_m, un conjunto de m variables independientes o predictoras.

El modelo de *regresión logístico multionomial* o *policotómico* establece que:

$$\operatorname{logit}\left[P(Y=s)\right] = \log\left[\frac{P(Y=s)}{P(Y=1)}\right] = \beta_{0s} + \beta_{1s}X_1 + \beta_{2s}X_2 + \cdots + \beta_{ms}X_m$$

La definición del modelo multinomial conlleva por tanto que para cada predictora X_i tengamos que estimar c - 1 coeficientes.

Otra forma sería: $$p_s = P(Y=s) = \frac{e^{\beta_{0s}+\beta_{1s}X_1+\beta_{2s}X_2+\cdots+\beta_{ms}X_m}}{1+\sum_{s=2}^{c} e^{\beta_{0s}+\beta_{1s}X_1+\beta_{2s}X_2+\cdots+\beta_{ms}X_m}}$$

2. Interpretación de los Coeficientes

La interpretación de los coeficientes del modelo de regresión logística multinomial también guarda paralelismo con el caso binario.

Por ejemplo, si A es un sujeto expuesto y B uno no expuesto en relación a la variable X_i pero iguales valores en relación a las restantes independientes, el logit de la probabilidad de que A pertenezca a la categoría *s* es:

$$\log\left[\frac{P_A(Y=s)}{P_A(Y=1)}\right]=\beta_{0s}+\beta_{1s}X_1+\cdots+\beta_{is}1+\cdots+\beta_{ps}X_p$$

y, para el sujeto B

$$\log\left[\frac{P_B(Y=s)}{P_B(Y=1)}\right]=\beta_{0s}+\beta_{1s}X_1+\cdots+\beta_{is}0+\cdots+\beta_{ps}X_p$$

2. Interpretación de los Coeficientes

donde tenemos que el logaritmo de la razón de ventaja de estar en la categoría s respecto a la categoría de referencia, del sujeto A expuesto al factor representado por X_i respecto a uno B no expuesto, vienen dado precisamente por el coeficiente de tal variable, es decir,

$$OR_{is} = e^{\beta_{is}}$$

Donde OR_{is} es la razón de ventajas de estar en la categoría s respecto a la referencia, entre los dos sujetos que se diferencian en una unidad de la variable X_i.

De forma similar, si X_i es una variable cuantitativa con valores x_{iA}, x_{iB} en esos dos sujetos

$$OR_{is} = e^{\beta_{is}(x_{iA} - x_{iB})}$$

2. Interpretación de los Coeficientes

El modelo multinomial, además de proporcionar los riesgos de cada categoría a la diferencia, también permite comparar dos categorías cualesquiera; en efecto, de manera análoga que para la categoría s, para la categoría q.

$$\log\left[\frac{P(Y=q)}{P(Y=1)}\right]=\beta_{0q}+\beta_{1q}X_1+\cdots+\beta_{iq}0+\cdots+\beta_{mq}X_m$$

Restando las igualdades se tiene

$$\log\left[\frac{P(Y=s)}{P(Y=q)}\right]=\left(\beta_{0s}-\beta_{0q}\right)+\left(\beta_{1s}-\beta_{1q}\right)X_1+\cdots+\cdots+\left(\beta_{ms}-\beta_{mq}\right)X_m$$

Por lo que las diferencias $\beta_{is}-\beta_{iq}$ se pueden interpretar como:

$$e^{\beta_{is}-\beta_{iq}}$$

3. Estimación de los Coeficientes

La estimación se realiza a través del método de máxima verosimilitud. Como se dijo anteriormente, en este caso estimaremos (c-1)(m-1) parámetros β_{is} con s = 2, 3, …, c e i = 0, 1, …, m. Sean n_i el número de sujetos con un patrón de valores de las predictoras

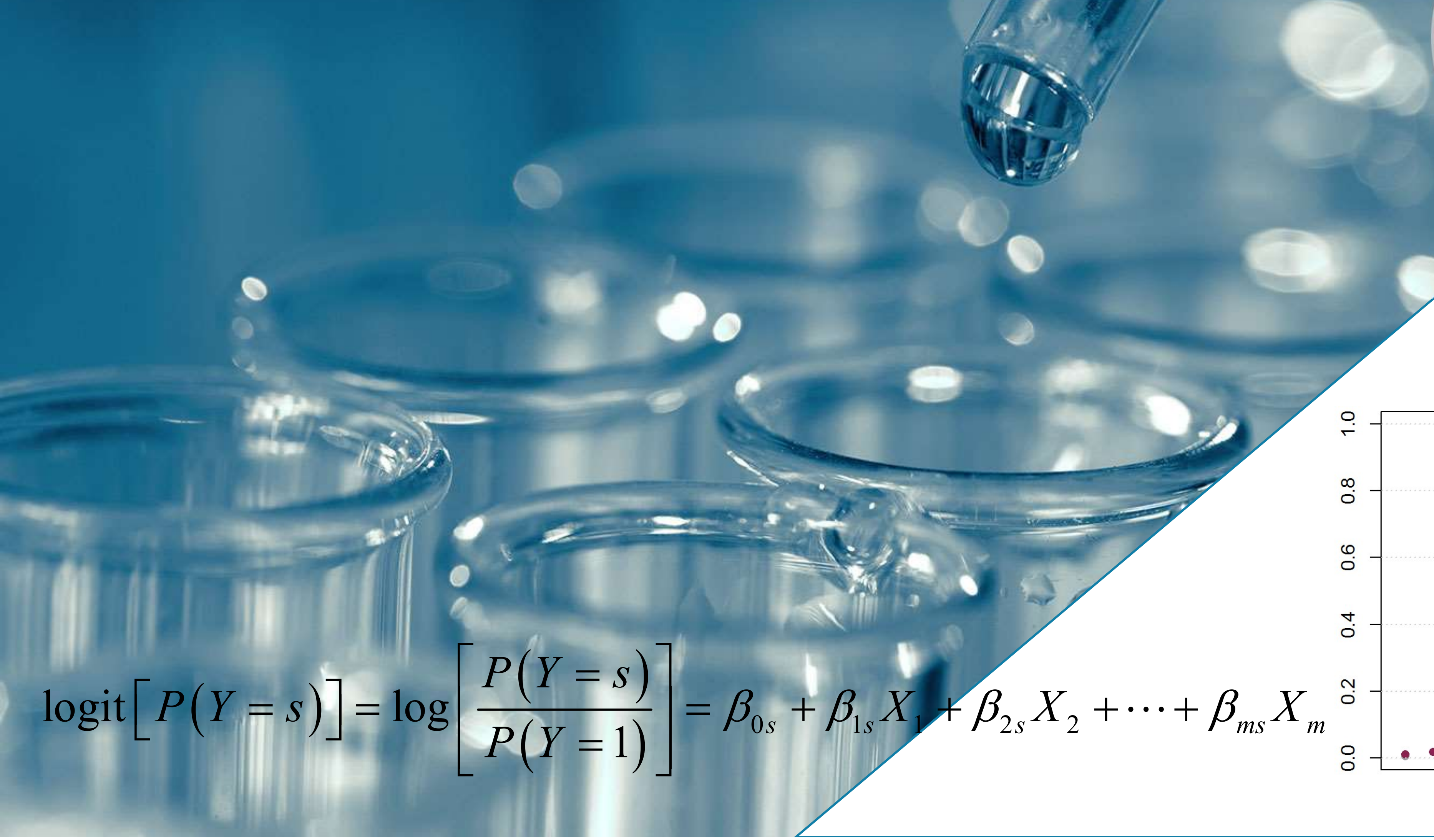
$$x_{1i}, x_{2i}, \cdots, x_{mi}$$

y sean

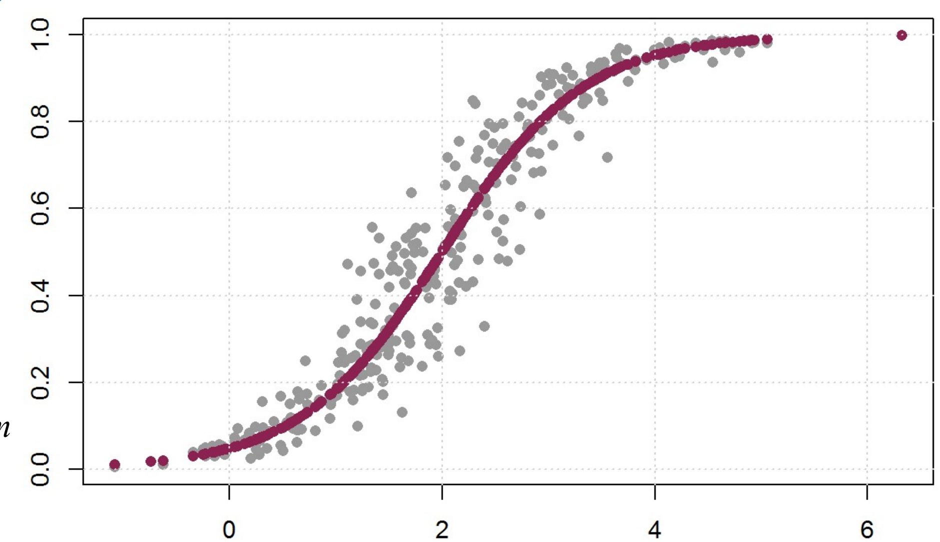
$$n_{i1}, n_{i2}, \cdots, n_{ic}$$

el número de ellos que pertenecen a las categorías 1, 2, 3, …, c, respectivamente.

EJEMPLOS



$$\operatorname{logit}\left[P\left(Y=s\right)\right]=\log\left[\frac{P\left(Y=s\right)}{P\left(Y=1\right)}\right]=\beta_{0s}+\beta_{1s}X_{1}+\beta_{2s}X_{2}+\cdots+\beta_{ms}X_{m}$$



Ejemplo 5

Los datos que vamos a utilizar en este caso provienen de un estudio cuyo objetivo es estudiarlos factores que influyen en el conocimiento, actitud y comportamiento de mujeres hacia las mamografías. Son datos sobre 412 mujeres y las variables son:

Descripción de las variables del estudio de mamografías.

Variable	Descripción	Valores
me	Frecuencia de chequeo.	0: Nunca 1: El último año 2: Más de un año
symp	La mamografía no es necesaria a menos que tengas algún síntoma.	1: Muy de acuerdo 2: De acuerdo 3: En desacuerdo 4: Muy en desacuerdo
pb	Beneficio percibido	5 - 20
hist	Madre y hermana con historia de cancer de mama.	0: No 1: Si
bse	¿Te han enseñado a explorarte?	0: No 1: Si
detc	¿Es probable que una mamografía detecte un nuevo caso de cancer?	0: No 1: Algo 2: Muy probable

Ejemplo 5

La base de datos se encuentran el el archivo mamexp.txt, como se muestra a continuación:

mamexp: Bloc de notas

Archivo Editar Ver

me	symp	pb	hist	bse	dect
0	3	7	0	1	1
0	2	11	0	1	2
0	3	8	1	1	2
1	3	11	0	1	2
2	4	7	0	1	2
0	3	7	0	1	2
2	4	6	0	1	1
0	4	6	0	1	2
0	2	6	0	1	2
1	4	6	0	1	2
0	4	8	0	1	1
0	3	6	1	0	2
0	4	6	0	1	2
0	1	5	1	1	2
0	2	8	0	0	1
0	4	11	0	1	2
0	3	6	0	1	2
1	4	5	0	1	2

Ln 1, Col 1 100% Windows (CRLF) UTF-8

A continuación, realizamos la lectura de los datos

```
setwd("C:/…")
datos <- read.table("mamexp.txt", header = TRUE)
View(datos)
```

Ejemplo 5

Las variables categóricas tienen valores numéricos que identifique cada categoría, en este caso R piensa por defecto que es una variable continua que toma valores 0 y 1, no sabe que es una variable categórica.

Indicaremos a R que esa variable no es numérica, que es categórica, con la función factor.

```
datos$symp <- factor(datos$symp)
datos$hist <- factor(datos$hist)
datos$bse <- factor(datos$bse)
datos$dect <- factor(datos$dect)
attach(datos)
```

Ejemplo 5

Ahora revisaremos las frecuencias de las categorías de la variable “me”.

```
table(me)
me
  0   1   2
234 104  74
```

Tenemos 234 casos que nunca se hicieron el chequeo, 104 el último año y 74 con más de un año.

Iniciaremos con el ajuste de modelos multinomiales univariantes (con 2 de las variables con hist y dect). La función **multinom** nos da las funciones logit de las variables que ajustamos, referenciadas al cero (nunca).

Ejemplo 5

- **Modelo 1:** $$\log\left[\frac{P(Y=s)}{P(Y=0)}\right]=\beta_{0s}+\beta_{1s}hist$$

Estimación del modelo

```
# install.packages("nnet")
# library(nnet)
modelo1 <- multinom(me ~ hist, data = datos)
# weights:  9 (4 variable)
initial  value 452.628263
iter  10 value 396.169969
iter  10 value 396.169969
final  value 396.169969
converged
summary(modelo1)
Call:
multinom(formula = me ~ hist, data = datos)
```

```
Coefficients:
  (Intercept)    hist1
1  -0.9509794 1.256531
2  -1.2504803 1.009384

Std. Errors:
  (Intercept)     hist1
1   0.1277115 0.3746633
2   0.1428926 0.4275097

Residual Deviance: 792.3399
AIC: 800.3399
```

Mtr. Alcides Ramos Calcina

Ejemplo 5

Se tiene los dos modelos logísticos correspondientes a la comparación de cada categoría con la referencia:

$$\log\left[\frac{P(Y=1)}{P(Y=0)}\right] = \log\left[\frac{P(\acute{u}ltimo\_a\tilde{n}o)}{P(nunca)}\right] = -0.95098 + 1.25653hist$$

$$\log\left[\frac{P(Y=2)}{P(Y=0)}\right] = \log\left[\frac{P(m\acute{a}s\_de\_a\tilde{n}o)}{P(nunca)}\right] = -1.25048 + 1.00938hist$$

Estimación de OR

Entonces, el logaritmo de la razón de ventaja de estar en la categoría 1 y 2 respecto a la categoría de referencia 0, es:

$$\widehat{OR}_1 = e^{\beta_{11}} = e^{1.25653} = 3.51321$$

Las mujeres con historia familiar de cáncer tienen un odds de haberse hecho una mamografía en el ultimo año que es 3.51 veces mayor a las mujeres sin historia familiar.

Ejemplo 5

$$\widehat{OR}_2 = e^{\beta_{12}} = e^{1.00938} = 2.74391$$

Las mujeres con historia familiar tienen un odds ratio 2.74 mayor de haberse hecho una mamografía hace mas de un año que las que no tienen historial familiar.

Ahora, estimamos los OR en R.

```
exp(coef(modelo1))
  (Intercept)    hist1
1   0.3863624 3.513213
2   0.2863672 2.743911
```

Ejemplo 5

- **Modelo 2:** $\log\left[\dfrac{P(Y=s)}{P(Y=0)}\right]=\beta_{0s}+\beta_{1s}dect$

Estimación del modelo

```
modelo2 <- multinom(me ~ dect, data = datos)
# weights:  12 (6 variable)
initial  value 452.628263
iter  10 value 389.200621
final  value 389.200537
converged
summary(modelo2)
Call:
multinom(formula = me ~ dect, data = datos)
```

```
Coefficients:
  (Intercept)      dect1     dect2
1   -2.565201  0.7062589 2.1062453
2   -1.178617 -0.3926042 0.1977986

Std. Errors:
  (Intercept)    dect1     dect2
1    1.037867 1.083278 1.0464712
2    0.571762 0.634349 0.5936115

Residual Deviance: 778.4011
AIC: 790.401
```

Ejemplo 5

Los dos modelos logísticos correspondientes son:

$$\log\left[\frac{P(ultimo\_año)}{P(nunca)}\right] = -2.56520 + 0.70626dect1 + 2.10625dect2$$

$$\log\left[\frac{P(más\_de\_año)}{P(nunca)}\right] = -1.17862 - 0.39260dect1 + 0.19780dect2$$

Estimación de OR

$$\widehat{OR}_{1\_\text{dect1}} = e^{\beta_{11\_\text{dect1}}} = e^{0.70626} = 2.02640 \quad \text{y} \quad \widehat{OR}_{1\_\text{dect2}} = e^{\beta_{11\_\text{dect2}}} = e^{2.10625} = 8.21733$$

En el primer OR indica que, las mujeres con ***algo*** *de probabilidad de detectar nuevo cáncer tienen un odds de haberse hecho una mamografía en el* ***ultimo año*** *que es 2.03 veces mayor a las mujeres que no tienen probabilidad de contraer nuevo cáncer y en el segundo, las mujeres que es* ***muy probable*** *detectar un nuevo cáncer su odds es de 8.22 veces más.*

Ejemplo 5

$$\widehat{OR}_{2\_dect1} = e^{\beta_{12\_dect1}} = e^{-0.39260} = 0.67530 \quad \text{y} \quad \widehat{OR}_{2\_dect2} = e^{\beta_{12\_dect2}} = e^{0.19780} = 1.21872$$

Con respecto a la segunda ecuación, las mujeres con algo de probabilidad de detectar nuevo cáncer tienen un odds ratio 0.68 mayor de haberse hecho una mamografía hace más de un año que las que no tienen probabilidad de contraer nuevo cáncer y para las mujeres que es muy probable detectar un nuevo cáncer su odds es de 1.22 veces más.

En R se tiene.

```
exp(coef(modelo2))
  (Intercept)      dect1    dect2
1  0.07690374 2.0263961 8.217330
2  0.30770391 0.6752959 1.218717
```

Ejemplo 5

Ahora estimamos los siguientes modelos:

- Modelo 3: $$\log\left[\frac{P(Y=s)}{P(Y=0)}\right] = \beta_{0s} + \beta_{1s} symp$$

- Modelo 4: $$\log\left[\frac{P(Y=s)}{P(Y=0)}\right] = \beta_{0s} + \beta_{1s} pb$$

- Modelo 5: $$\log\left[\frac{P(Y=s)}{P(Y=0)}\right] = \beta_{0s} + \beta_{1s} bse$$

Ejemplo 5

Se tiene:

```
modelo3 <- multinom(me ~ symp, data = datos)
summary(modelo3)
exp(coef(modelo3))

modelo4 <- multinom(me ~ pb, data = datos)
summary(modelo4)
exp(coef(modelo4))

modelo5 <- multinom(me ~ bse, data = datos)
summary(modelo5)
```

La interpretación de los coeficientes se deja a consideración de los estudiantes.

FINESI

Modelos Discretos

IV Semestre

GRACIAS

https://aulavirtual2.unap.edu.pe/