Profesor: Héctor Bahamonde, PhD.

e:hector.bahamonde@uoh.cl

w:www.hectorbahamonde.com

Curso: MLE.

TA: Gonzalo Barría.

I. Outcomes Desordenados: Multinomial Logit/Probit

Muchas veces estamos interesados en variables dependientes que son nominales ("cualitativas"), pero donde no necesariamente cada valor representa un número o cantidad. Algunos ejemplos son el tipo de fruta que te guste mas ("pera", "manzana", "uva"). O por ejemplo, profesión ("profesor", "ingeniero", "médico").[1] No es posible ordenar frutas o profesiones porque estas categorías no representan cantidades. Es por esto que estos modelos son "desordenados".

Motivación El m-logit/probit sigue siendo una extensión del modelo logit/probit, y por extensión, muy parecido al o-logit/probit.

1. En primer lugar, si bien es cierto que el modelo o-logit/probit tiene un efecto estimado $\hat{\beta}$ constante para cada uno de los valores de y_i, en la especificación m-logit/probit los $\hat{\beta}$ pueden variar según los valores de y_i.
2. En segundo lugar, se puede pensar en el modelo m-logit/probit como una secuencia de modelos logits/probits: en base a combinatorias, se estiman modelos logit/probits individuales entre cada uno de los valores de y_i.

Retomemos el segundo punto. Supongamos que estamos estimando el consumo de tres tipos de frutas: "pera" (p), "manzana" (m), "uva" (u). Formalmente, el m-logit/probit estima lo siguiente,

[1] Aunque algunas veces el modelo m-logit/probit es usado cuando el supuesto de la regresión paralela no se cumple (Long 1997, p. 148).

$$\begin{aligned}
ln[\frac{\Pr(\mathrm{p}|\boldsymbol{x})}{\Pr(\mathrm{m}|\boldsymbol{x})}] &= \beta_{0,p|m} + \beta_{1,p|m}\boldsymbol{x} \\
ln[\frac{\Pr(\mathrm{m}|\boldsymbol{x})}{\Pr(\mathrm{u}|\boldsymbol{x})}] &= \beta_{0,m|u} + \beta_{1,m|u}\boldsymbol{x} \\
ln[\frac{\Pr(\mathrm{p}|\boldsymbol{x})}{\Pr(\mathrm{u}|\boldsymbol{x})}] &= \beta_{0,p|u} + \beta_{1,p|u}\boldsymbol{x}
\end{aligned} \tag{1}$$

Nota que cada β es distinto (no como en la parametrización) del o-logit/probit.

Ahora nota que debido a que,

$$ln[\frac{\Pr(\mathrm{p}|\boldsymbol{x})}{\Pr(\mathrm{m}|\boldsymbol{x})}] + ln[\frac{\Pr(\mathrm{m}|\boldsymbol{x})}{\Pr(\mathrm{u}|\boldsymbol{x})}] = ln[\frac{\Pr(\mathrm{p}|\boldsymbol{x})}{\Pr(\mathrm{u}|\boldsymbol{x})}] \tag{2}$$

y que debido a que $M = \{p, m, u\}$, el modelo m-logit/probit se reduce a $M - 1 = 2$ modelos logit (Ward and Ahlquist 2018, p. 163). Substantivamente, lo que esto requiere es que el analista especifique una **categoría de referencia**. Es decir, una categoría de M para poner en el denominador y realizar el modelo logit. **Ésta se escoge siguiendo criterios substantivos**. Matemáticamente, los resultados *no* se alteran (esto es claro en Equation 2). Pero substantivamente sí importa: cada coeficiente $\hat{\boldsymbol{\beta}}$ ahora representará el cambio de una de las categorías respecto de la categoría base.

Supuestos Distribucionales: La distribución multinomial La distribución multinomial viene de la binomial que a su vez viene de la distribución Bernoulli. Formalmente, y siguiendo a Ward and Ahlquist (2018, p. 162), la distribución multinomial de la variable y_i,

$$y_i \sim (np_i, np_i(1 - p_i)) \tag{3}$$

y cuyo *probability mass function* (PMF) está definido por,[2]

$$y_i = (\sum_i^k x_i)! \prod_i^k \frac{p_i^{x_i}}{x_i!} \tag{4}$$

donde k es el número total de eventos ocurridos (*frutas*) x veces (distribución de *frutas*), y p_i denota la probabilidad de cada observación i. En simple, esta distribución clasifica un *data*

[2]Los PMF's son para variables nominales. Hasta el momento habíamos *probability density functions*, pero éstas son para variables continuas.

generating process en función de la probabilidad de que cada evento ocurra x veces, i.e. genera esa distribución.

Sin embargo, esta es la versión generalizada del modelo ologit. En el modelo estimado existen un par de detalles en el numerador que representan la inclusión de la categoria base. Por simpleza, omitiremos estos detalles.[3]

Finalmente, y tomando el modelo estructural,

$$x_i \beta_m + \epsilon_i \tag{5}$$

el analista puede asumir que ϵ_i está normalmente distribuido con $\phi(\epsilon_i) \sim (0, 1)$ llegando a la distribución multinomial probit, o está distribuido según la distribución logística con $\lambda(\epsilon_i) \sim (0, \frac{\pi^2}{3})$, llegando a la distribución multinomial logit.

Estimación: Probabilidades y Likelihood Continuando con el Equation 4, y tomando en cuenta el modelo estructural generalizado $x_i\beta_m + \epsilon_i$ en la Equation 5, es posible calcular probabilidades de la siguiente manera,

$$Pr(y_i = m) = \frac{exp(\boldsymbol{x}_i \boldsymbol{\beta}_m)}{\sum_i^M \boldsymbol{x}_i \boldsymbol{\beta}_m} \tag{6}$$

Es fácil ver entonces que existe un β_m por cada categoría m (fruta). No es como en el modelo o-logit/probit donde gracias al supuesto de la regresión paralela, existía un β para todas las categorías.

Si todas las observaciones i son independientes, el likelihood esta definido como sigue,

$$L(\boldsymbol{\beta}_m | y, \boldsymbol{x}) = \prod_m^M \prod_{y_i = m} \frac{exp(\boldsymbol{x}_i \boldsymbol{\beta}_m)}{\sum_i^M \boldsymbol{x}_i \boldsymbol{\beta}_m} \tag{7}$$

donde $\prod_{y_i=m}$ es el producto de todos los casos donde $y_i = m$.

IIA El modelo multinomial supone que "an individual's choice does not depend on the availability or characteristics of inaccessible alternatives" (Ward and Ahlquist 2018, p. 166). Esto es lo que se llama el supuesto de la *independence of irrelevant alternatives*. Lo que esto supone es que los respondentes tienen todas las alternativas M disponibles para responder, y que sacando una m de

[3] El estudiante interesado podrá referirse a Ward and Ahlquist (2018), en la ecuación 9.2.

ellas, esto no cambia la respuesta del respondente. Un ejemplo: los respondentes no escogen el mal menor.

Este supuesto es testable vía el test "H" de Hausman-McFadden, y de manera muy familiar, funciona comparando los coeficientes estimados $\hat{\boldsymbol{\beta}}_m$ cuando secuencialmente sacamos—con reemplazo—una categoría m a la vez. Esto se hace estimando un modelo "unrestricted" (u) con $M-1$ modelos "restricted" (r). Formalmente,

$$H = (\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_u)^T [\hat{\boldsymbol{V}}_r - \hat{\boldsymbol{V}}_u]^{-1} (\hat{\boldsymbol{\beta}}_r - \hat{\boldsymbol{\beta}}_u) \tag{8}$$

donde $\hat{\boldsymbol{\beta}}_r$ son los coeficientes del modelo "restricted" y $\hat{\boldsymbol{\beta}}_u$ son los coeficientes del modelo "unrestricted". Mientras que $\hat{\boldsymbol{V}}_r$ y $\hat{\boldsymbol{V}}_u$ son las varianzas-covarianzas de ambos coeficientes, respectivamente.

Si los coeficientes $\hat{\boldsymbol{\beta}}_r$ y $\hat{\boldsymbol{\beta}}_u$ cambian mucho, y el cambio es estadísticamente significativo, habría evidencia de que al sacar una categoría m se estaría violando el supuesto *IIA*.

II. Programación

Carguemos los datos:

```
p_load(foreign)
dat <- read.dta("https://stats.idre.ucla.edu/stat/data/hsbdemo.dta")
```

Hagamos un resumen,

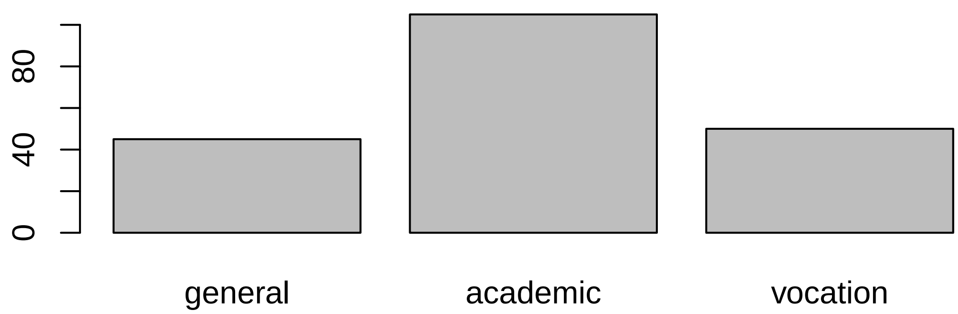
```
summary(dat)

##        id            female        ses          schtyp          prog
##  Min.   :  1.00   male  : 91   low   :47   public :168   general : 45
##  1st Qu.: 50.75   female:109   middle:95   private: 32   academic:105
##  Median :100.50                high  :58                 vocation: 50
##  Mean   :100.50
##  3rd Qu.:150.25
##  Max.   :200.00
##       read           write            math           science
```

```
##  Min.   :28.00   Min.   :31.00   Min.   :33.00   Min.   :26.00
##  1st Qu.:44.00   1st Qu.:45.75   1st Qu.:45.00   1st Qu.:44.00
##  Median :50.00   Median :54.00   Median :52.00   Median :53.00
##  Mean   :52.23   Mean   :52.77   Mean   :52.65   Mean   :51.85
##  3rd Qu.:60.00   3rd Qu.:60.00   3rd Qu.:59.00   3rd Qu.:58.00
##  Max.   :76.00   Max.   :67.00   Max.   :75.00   Max.   :74.00
##      socst                honors        awards          cid
##  Min.   :26.00   not enrolled:147   Min.   :0.00   Min.   : 1.00
##  1st Qu.:46.00   enrolled    : 53   1st Qu.:0.00   1st Qu.: 5.00
##  Median :52.00                      Median :1.00   Median :10.50
##  Mean   :52.41                      Mean   :1.67   Mean   :10.43
##  3rd Qu.:61.00                      3rd Qu.:2.00   3rd Qu.:15.00
##  Max.   :71.00                      Max.   :7.00   Max.   :20.00
```

En esta aplicación pensaremos en la variable `prog`: programa de estudio escogido. Veámos cómo se ve esta variable.

```
plot(dat$prog)
```



Expliquemos un poco de qué se tratan estos programas que se dan en Estados Unidos:

1. "**general**": programas universitarios que conducen hacia la primera mitad de una licenciatura (tipo bachillerato).
2. "**academic**": programas universitarios que conducen hacia el final una licenciatura (como tu programa, o el que estudié yo en pregrado).
3. "**vocational**": programas técnico.

El paquete de `R` que usaremos se llama `nnet`—éste especifica que la variable dependiente debe ser `factor`. Además, debemos escoger la categoría de referencia:

```
is(dat$prog)[1] # ok, es factor

## [1] "factor"

dat$prog2 <- relevel(dat$prog, ref = "academic") # ref cat
```

Ahora estimemos el modelo,

```
p_load(nnet)
modelo <- multinom(prog2 ~ ses + write, data = dat)

## # weights:  15 (8 variable)
## initial  value 219.722458
## iter  10 value 179.982880
## final  value 179.981726
## converged
```

Hagamos una tabla.

```
p_load(texreg)
texreg(modelo) # usa "screenreg" no "texreg".
```

Lo que acabamos de estimar tiene un correlato en la parte teórica que acabamos de discutir en Equation 1.

	Model 1
general: (Intercept)	2.85^{*}
	(1.17)
general: sesmiddle	-0.53
	(0.44)
general: seshigh	-1.16^{*}
	(0.51)
general: write	-0.06^{**}
	(0.02)
vocation: (Intercept)	5.22^{***}
	(1.16)
vocation: sesmiddle	0.29
	(0.48)
vocation: seshigh	-0.98
	(0.60)
vocation: write	-0.11^{***}
	(0.02)
AIC	375.96
BIC	402.35
Log Likelihood	-179.98
Deviance	359.96
Num. obs.	200
K	3

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1: *Statistical models*

$$\begin{aligned} ln[\frac{\text{Pr(program: general)}}{Pr(\text{program: academic})}] &= \beta_{10} + \beta 11(\text{ses} = \text{middle}) + \beta_{12}(\text{ses} = \text{high}) + \beta_{13}(\text{write}) \\ ln[\frac{\text{Pr(program: vocation)}}{Pr(\text{program: academic})}] &= \beta_{20} + \beta 22(\text{ses} = \text{middle}) + \beta_{22}(\text{ses} = \text{high}) + \beta_{23}(\text{write}) \end{aligned} \tag{9}$$

O lo que en números es (tomando los resultados de la tabla),

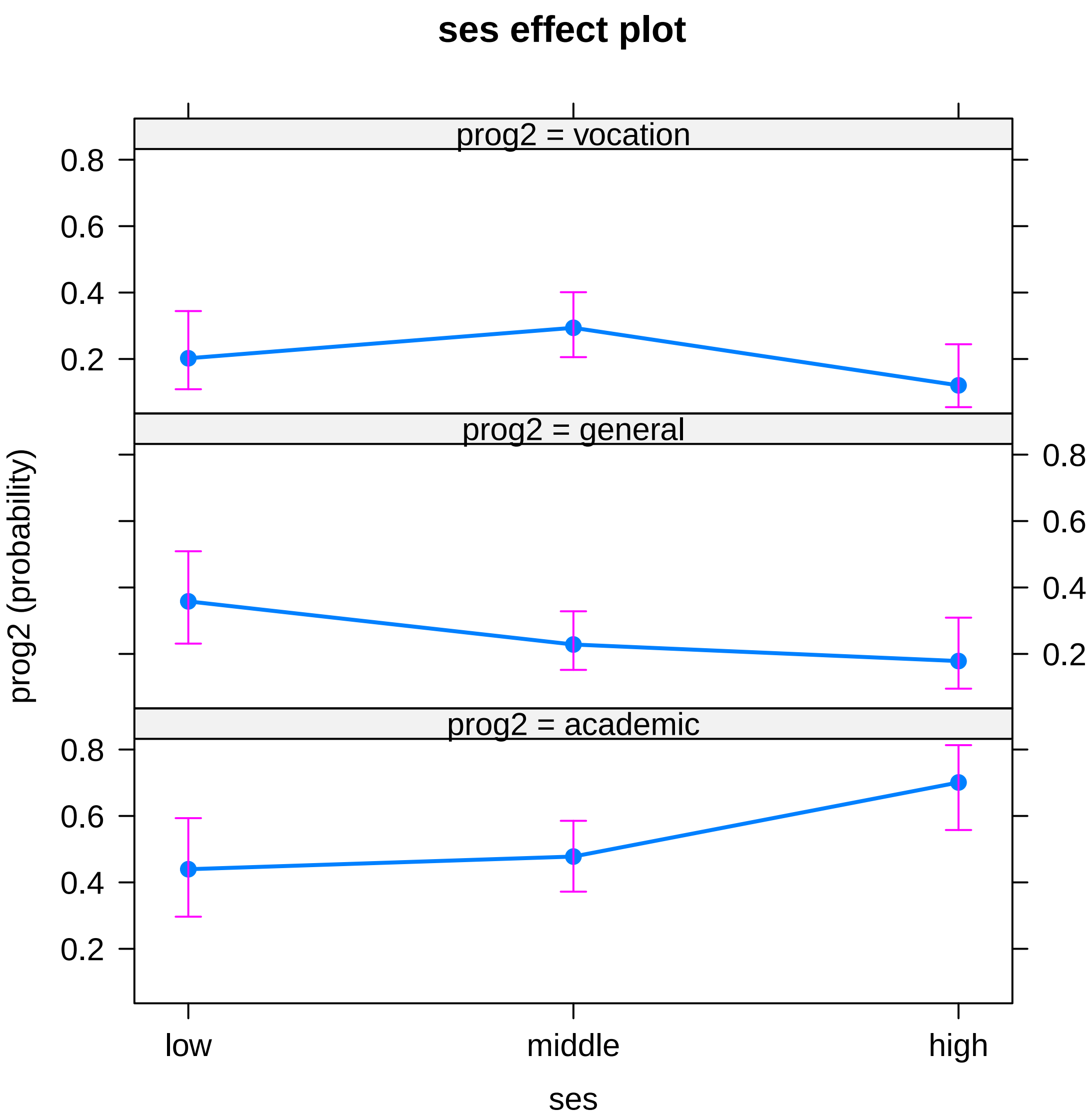
$$\begin{aligned} ln[\frac{\text{Pr(program: general)}}{Pr(\text{program: academic})}] &= 2.85 - 0.53(\text{ses} = \text{middle}) - 1.16(\text{ses} = \text{high}) - 0.06(\text{write}) \\ ln[\frac{\text{Pr(program: vocation)}}{Pr(\text{program: academic})}] &= 5.22 + 0.29(\text{ses} = \text{middle}) - 0.98(\text{ses} = \text{high}) - 0.11(\text{write}) \end{aligned} \tag{10}$$

III. Interpretación

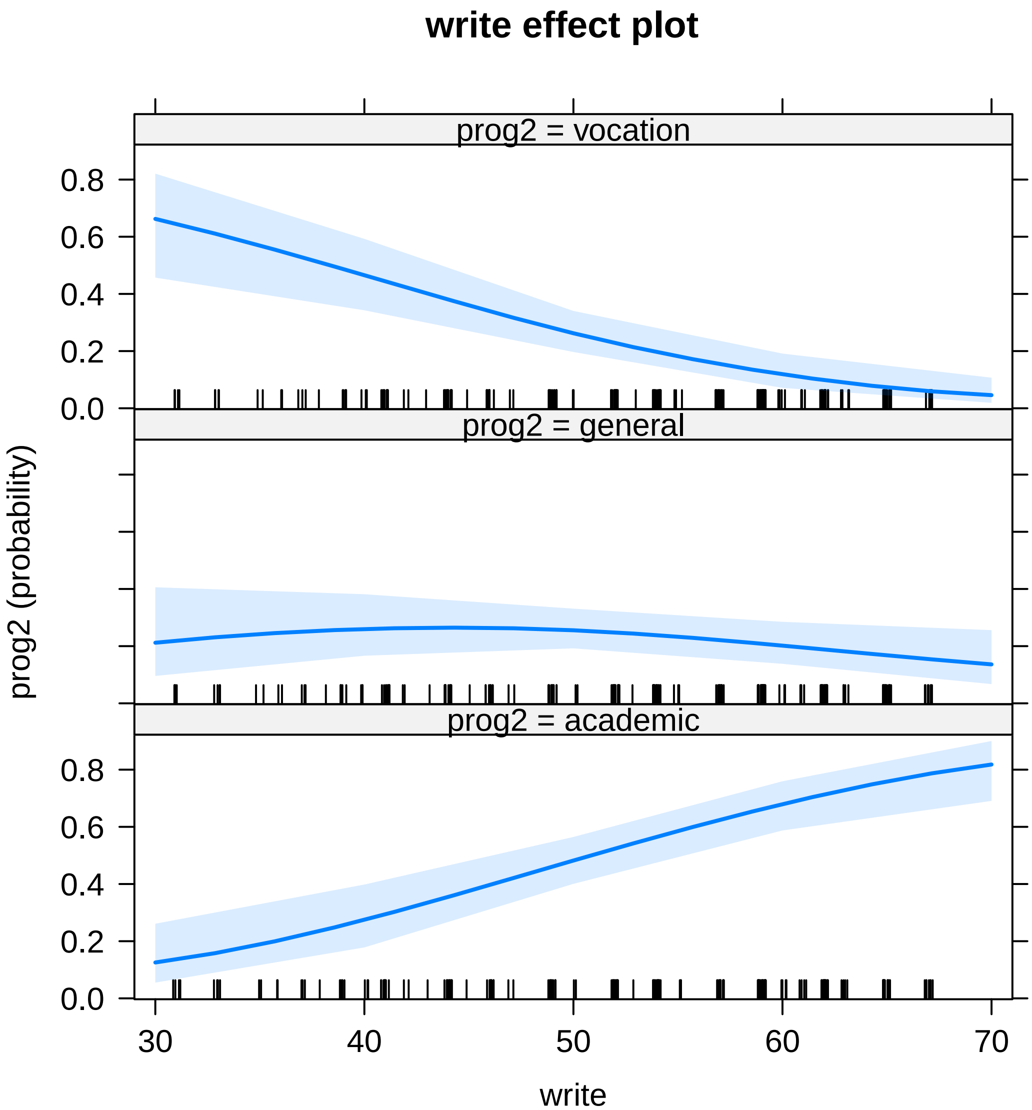
Ahora interpretaremos el modelo. Tomando la Equation 10, se interpreta como siempre. Por ejemplo, el β_{13} significa, cuando `write` sube una unidad, los **log-odds** de estar en un programa `general` **versus** un programa `academic` decrecen -0.06. Otro ejemplo es el β_{11}: cuando subo en una unidad `ses` (es decir de `low` a `middle`), los **log-odds** de estar en un programa `general` **versus** un programa `academic` decrecen -0.53.

Este modelo es bastante complejo. No solo en su parametrización (Equation 4), sino que también en su interpretación (Equation 10). Es por esto que nos saltaremos los *odds rations* y los cambios marginales. Todos ellos están en escala de log-ods, dificultando aun más la interpretación substantiva de la tabla de regresión. Es por esto que saltaremos directo a las *predicted probabilities*. La ventaja es que podremos calcular todas las probabilidades, *incluso de la categoría base*.

```
p_load(effects)
plot(effect("ses", modelo))
```




```
plot(effect("write", modelo))
```



Algunos ejemplos de interpretación:

1. Estudiantes con *background* socio-económico (`ses`) **alto** tienden a preferir programas `academic`.
2. Estudiantes con *background* socio-económico (`ses`) **bajo** tienden a preferir programas `general`.
3. Estudiantes que logran puntajes **altos** en la prueba de escritura (`write`) tienden a preferir programas `academic`.

4. Estudiantes que logran puntajes **bajos** en la prueba de escritura (`write`) tienden a preferir programas `vocation`.

```
knitr::purl('Multinomial.Rnw')

## Error in parse_block(g[-1], g[1], params.src, markdown_mode):  Duplicate chunk label
'setup', which has been used for the chunk:
## if (!require("pacman")) install.packages("pacman"); library(pacman)
## p_load(knitr)
## set.seed(2020)
## options(scipen=9999999)

Stangle('Multinomial.Rnw')

## Writing to file Multinomial.R

## Error in match.arg(options$results, c("verbatim", "tex", "hide")):  'arg' should
be one of "verbatim", "tex", "hide"
```