Análisis de Datos Categóricos

Ayudantía 9

Felipe Olivares

Contenido

1. Regresión multinomial
2. Poisson

Regresión Multinomial

¿Qué es una regresión Multinomial?

Es una generalización de la regresión logística binomial, que busca hacerse cargo de la situación de una variable discreta que no necesariamente pueden estar ordenada y que tiene más de dos valores (cuestión que la regresión logística no puede cubrir). Un ejemplo clásico es cuándo se hace investigación sobre participación electorar. Tomemos el caso del voto por candidatos (cuándo hay más de dos candidatos, por ejemplo, en una primera vuelta electoral).

La regresión logística multinomial es una extensión simple de la regresión logística binaria que permite más de dos categorías de la variable dependiente o de resultado. Al igual que regresión logística, la regresión logística multinomial utiliza la estimación de máxima verosimilitud (MLE)

Regresión Multinomial

Para efecto de la regresión que vamos a utilizar, vamos a modelar la función de las formas de control de la protesta para distintas categorías de esta. Por lo tanto, es importante que definamos nuestra variable de referencia (o de contraste) respecto de las categorías de interés. En una regresión multinomial las estimaciones se obtienen respecto de una categoría

```
# agregamos la variable de participantes a otro subset de datos
df2 <- df1 %>% dplyr::select(apolicial,disruptiva,trabajadores, p11) %>%
  mutate( disruptiva = if_else(disruptiva=="Sí",1,0),
         trabajadores = if_else(trabajadores=="Sí",1,0),
         participantes = p11)  %>%
  mutate( participantes = as.numeric(participantes),
         disruptiva = as.numeric(disruptiva),
         trabajadores= as.numeric(trabajadores))  %>%
  na.omit(df2)
```

Regresión Multinomial

Revisemos los datos nuevamente. . .

```
head(df2)
```

```
# # A tibble: 6 x 5
#   apolicial               disruptiva trabajadores   p11 participantes
#   <fct>                        <dbl>        <dbl> <dbl>         <dbl>
# 1 Presencia policial               0            0    81            81
# 2 violencia policial               0            0    40            40
# 3 Enfrentamientos directos         1            0    20            20
# 4 Presencia policial               1            1  2000          2000
# 5 Enfrentamientos directos         0            0    60            60
# 6 Enfrentamientos directos         1            1   150           150
```

Regresión Multinomial

Formalmente:

$$\underbrace{\ln \frac{p_{ij}}{1 - p_{iJ}}}_{\text{logit}(p_{ij})} = \beta_{0j} + \beta_{j1}x_{i1} + \ldots + \beta_{jk}x_{ik}$$

Importante: Los coeficientes y sus transformaciones entregan información sobre las probabilidades relativas de los diferentes $j\'s$. Esto siempre respecto de una variable de referencia anteriormente seleccionada (para este caso presencia policial)

Regresión Multinomial

```r
mlogit1 <- multinom(apolicial ~ participantes + trabajadores, trace=F, data=df2)
summary(mlogit1)
```

```
# Call:
# multinom(formula = apolicial ~ participantes + trabajadores,
#     data = df2, trace = F)
#
# Coefficients:
#                          (Intercept) participantes trabajadores
# Acciones preventivas     -1.76407512  1.503145e-05   -0.4194244
# Enfrentamientos directos  0.07487663  5.924185e-06   -0.5319464
# violencia policial       -0.51631844  1.717192e-05   -0.1605411
#
# Std. Errors:
#                           (Intercept) participantes trabajadores
# Acciones preventivas     4.083983e-11  4.180938e-06 1.958634e-11
# Enfrentamientos directos 9.965112e-11  3.981581e-06 4.849998e-11
# violencia policial       8.978793e-11  3.558233e-06 5.598286e-11
#
# Residual Deviance: 6691.673
# AIC: 6709.673
```

Regresión Multinomial

Interpretación de los coeficientes de la variable trabajadores:

1. El logaritmo de un control con acciones preventivas versus un control de presencia policial aumenta en 0.0001 puntos de log odds por cada aumento que exista en el número de participantes de una protesta.
2. El logaritmo de un control con enfrentamientos directos versus un control de presencia policial disminuye en -0,53 puntos de log odds si existe presencia de trabajadores durante una manifestación.
3. Un aumento en 1 unidad de número de participantes se traduce en un aumento de 0.0001 unidades en el logit de un control violento de la protesta versus un control de presencial policial. El logaritmo de un control violento de la protesta versus un control de presencia policial aumenta en 0.0001 puntos de log odds por cada aumento que exista en el número de participantes.

Regresión Multinomial

Cálculemos efectos marginales de los logit de un control con enfrentamientos directos y de un control violento.

```
#Efecto 100 participantes
c(E= 0.07 + 0.001*100  -0.42*1,
  V= -0.52+ 0.001*100 -0.16*1)
```

```
#     E     V
# -0.25 -0.58
```

```
#Efecto 10000 participantes
c(E= 0.07 + 0.001*1000 -0.42*1,
  V= -0.52 + 0.001*1000 -0.16*1)
```

```
#    E    V
# 0.65 0.32
```

```
#Beta enfrentamientos directos
0.65 - 0.25
```

```
# [1] 0.4
```

```
#Beta Violencia policial
0.32 - 0.58
```

```
# [1] -0.26
```

Manteniendo todo constante, el aumento en 1000 participantes se traducen un aumento de 0.4 puntos de log odds para el control con enfrentamientos directos respecto de un conrol con presencia policial.

Manteniendo todo constante, el aumento en 1000 participantes se traduce en una disminución de 0.26 en el logaritmo de las odds de un control violento de la protesta versus un control de presencia policial.

Regresión Multinomial - Efectos multiplicativos

La trasformación de probabilidades a odds es monotónica, si la probabilidad aumenta también lo hacen los odds, y viceversa. El rango de valores que pueden tomar los odds es de $(0, \infty+)$. Dado que el valor de una probabilidad está acotado entre (0,1)

donde:

- $e^{\beta_{jk}}$ está restringido al rango $[0, \infty+)$. Es una constante que "comprime" o amplifica el ratio entre las probabilidades de j respecto de J.
- Si $\beta_{jk} < 0 \rightarrow (0 < e^{\beta_{jk}} < 1)$. Es decir, un aumento en x_k está asociado con una reducción (multiplicativa) del ratio entre las probabilidades de j versus J.
- Si $\beta_{jk} = 0 \rightarrow (e^{\beta_{jk}} = 1)$. Es decir, un cambio en x_k está asociado a un cambio nulo en el ratio entre las probabilidades de j versus J.
- Si $\beta_{jk} > 0 \rightarrow (e^{\beta_{jk}} > 1)$. Es decir, un aumento en x_k está asociado a aumento (multiplicativo) del ratio entre las probabilidades de j versus J.

Regresión Multinomial - Efectos multiplicativos

```
#Log odds
summary(mlogit1)$coefficients
```

```
#                          (Intercept) participantes trabajadores
# Acciones preventivas     -1.76407512  1.503145e-05   -0.4194244
# Enfrentamientos directos  0.07487663  5.924185e-06   -0.5319464
# violencia policial       -0.51631844  1.717192e-05   -0.1605411
```

```
#Odds ratio
exp(summary(mlogit1)$coefficients)
```

```
#                          (Intercept) participantes trabajadores
# Acciones preventivas       0.1713452      1.000015    0.6574251
# Enfrentamientos directos   1.0777512      1.000006    0.5874604
# violencia policial         0.5967133      1.000017    0.8516829
```

Regresión Multinomial

Precauciones en el uso de modelos multinomiales:

1. Efectos marginales son heterogéneos (multiplicidad de efectos)
2. Efectos heterogéneos se vuelven más complejo con la inclusión de predictores
3. Los efectos muchas veces no son monotónicos (pueden variar de signo)

Regresión Multinomial

Cambiando la categoría de referencia

```r
df2$apolicial2 <- relevel(df2$apolicial, ref = "Enfrentamientos directos") # 1º de cambiar valor de referencia

mlogit2 <- multinom(apolicial2 ~ participantes + trabajadores, trace=F, data=df2)
summary(mlogit2)
```

```
# Call:
# multinom(formula = apolicial2 ~ participantes + trabajadores,
#     data = df2, trace = F)
#
# Coefficients:
#                       (Intercept) participantes trabajadores
# Presencia policial   -0.07494316 -5.924776e-06    0.5314497
# Acciones preventivas -1.84033887  9.117559e-06    0.1127366
# violencia policial   -0.59112192  1.125851e-05    0.3712487
#
# Std. Errors:
#                        (Intercept) participantes trabajadores
# Presencia policial   9.494736e-11  3.982924e-06 4.923754e-11
# Acciones preventivas 3.224457e-11  3.773817e-06 1.350665e-11
# violencia policial   5.809218e-11  3.065192e-06 2.885183e-11
#
# Residual Deviance: 6691.673
# AIC: 6709.673
```

Regresión Poisson

Los modelos de regresión de Poisson se utilizan mejor para modelar variables de conteo o eventos en los que se cuentan los resultados. EStos datos deben ser valores discretos enteros no negativos que cuentan algo, como la cantidad de veces que ocurre un evento durante un período de tiempo determinado, por ejemplo, cantidad de protestas en un año.

La distribución de Poisson se usa regularmente para encontrar la probabilidad de ocurrencia de un evento dentro de un intervalo de tiempo dado (por ejemplo, cantidad de protestas en el año 2019). Dado que estamos hablando de un recuento, con la distribución de Poisson, el resultado debe ser 0 o superior; no es posible que un evento ocurra un número negativo de veces.

Propiedades –> Ocurrencia e independencia: La ocurrencia de un delito en un barrio no condiciona la tasa promedio de delitos en el barrio. Es decir, un "exito" no condiciona la ocurrencia de otro "exito". La ocurrencia de eventos no tienen un límite superior (infinito positivo)

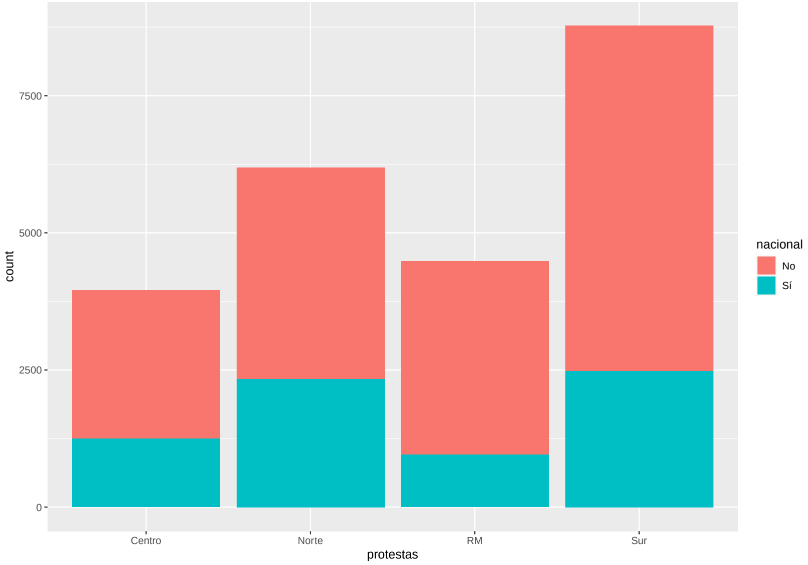
Regresión Poisson

```r
# agregamos la variable de participantes a otro subset de datos



load(url("https://github.com/mebucca/cda_soc3070/blob/master/ta/ta_3/data_OCS.Rdata?raw=true"))


df3<-df1 %>%  dplyr::select(ano,macrozona,nacional) %>%
               group_by(ano,macrozona) %>% # por año
            mutate(protestas = row_number())

head(df3)
```

```
# # A tibble: 6 x 4
# # Groups:   ano, macrozona [4]
#   ano   macrozona nacional protestas
#   <fct> <chr>     <fct>        <int>
# 1 2009  RM        No               1
# 2 2009  Norte     No               1
# 3 2009  Norte     No               2
# 4 2009  Centro    Sí               1
# 5 2009  Sur       No               1
# 6 2009  Sur       No               2
```

Regresión Poisson

```
# ejemplo de conteo de protestas según macrozona y protesta nacional Chile 2009-2019

ggplot(df3, aes(protestas)) +  geom_bar(aes(macrozona, fill=nacional))
```



Regresión Poisson

```
poisson1 <- glm(protestas ~ nacional + macrozona, family=poisson(link="log"),data=df3)
summary (poisson1)
```

```
#
# Call:
# glm(formula = protestas ~ nacional + macrozona, family = poisson(link = "log"),
#     data = df3)
#
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -33.291  -12.423   -3.388    6.142   43.681
#
# Coefficients:
#                 Estimate Std. Error z value Pr(>|z|)
# (Intercept)    5.3396210  0.0010732  4975.2   <2e-16 ***
# nacionalSí     0.3524846  0.0007197   489.7   <2e-16 ***
# macrozonaNorte 0.3369208  0.0012453   270.6   <2e-16 ***
# macrozonaRM    0.2387811  0.0013605   175.5   <2e-16 ***
# macrozonaSur   0.6759749  0.0011489   588.4   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
#     Null deviance: 5492689  on 23397  degrees of freedom
# Residual deviance: 4774197  on 23393  degrees of freedom
# AIC: 4944351
#
# Number of Fisher Scoring iterations: 5
```

Regresión Poisson

$$\frac{\partial \ln(\mu)}{\partial x_k} = \beta_k$$

"Un cambio (infinitesimal) en Δ unidades de x_k se traduce en un cambio en $\Delta\beta_k$ unidades en $\ln(\mu)$"

Regresión Poisson

```r
#Efectos multiplicativos sobre la tasa (sin offset)

summary(poisson1) # exponenciado- Tasa sobre las protestas
```

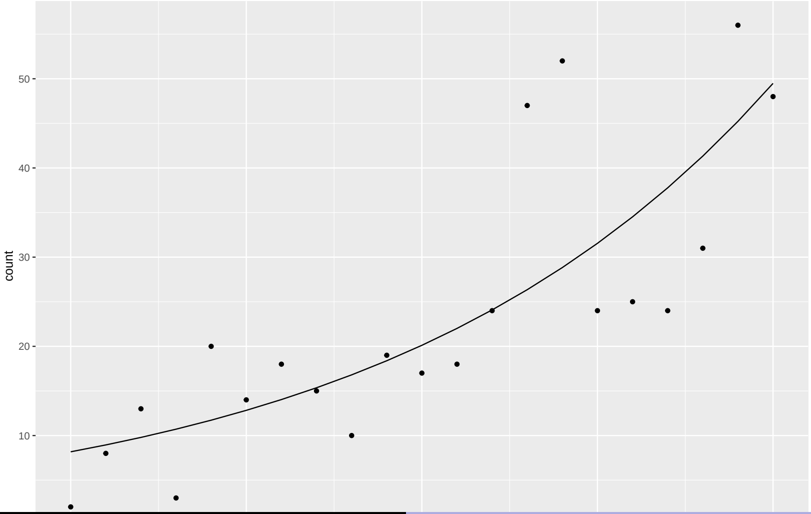
```
#
# Call:
# glm(formula = protestas ~ nacional + macrozona, family = poisson(link = "log"),
#     data = df3)
#
# Deviance Residuals:
#     Min       1Q   Median       3Q      Max
# -33.291  -12.423   -3.388    6.142   43.681
#
# Coefficients:
#                 Estimate Std. Error z value Pr(>|z|)
# (Intercept)    5.3396210  0.0010732  4975.2   <2e-16 ***
# nacionalSí     0.3524846  0.0007197   489.7   <2e-16 ***
# macrozonaNorte 0.3369208  0.0012453   270.6   <2e-16 ***
# macrozonaRM    0.2387811  0.0013605   175.5   <2e-16 ***
# macrozonaSur   0.6759749  0.0011489   588.4   <2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
#     Null deviance: 5492689  on 23397  degrees of freedom
# Residual deviance: 4774197  on 23393  degrees of freedom
# AIC: 4944351
#
# Number of Fisher Scoring iterations: 5
```

```r
tasa1= exp(5.3396210 +0.3369208 + 0.3524846*0 ) # sin protesta nacional para la macrozona norte
tasa2= exp(5.3396210 +0.3369208 + 0.3524846*1 ) # con protesta nacional protesta nacional para la macrozona norte

#efecto multiplicativo de la protesta nacional
tasa1/tasa2
```

```
# [1] 0.7029394
```

Regresión Poisson

```r
# ejemplo creado
year <- 1990:2010 # años observados
count <- c(2, 8, 13, 3, 20, 14, 18, 15, 10, 19, 17, 18, 24, 47, 52, 24, 25, 24, 31, 56, 48) # cantidad de protestas
df <- data.frame(year, count)
poisson2 <- glm(count ~ year, family = "poisson", data = df)
poisson2$model$fitted <- predict(poisson2, type = "response")
ggplot(poisson2$model) + geom_point(aes(year, count)) + geom_line(aes(year, fitted))
```



Regresión Poisson

Veamos otros ejemplos..

En la segunda parte de este taller trabajaremos Vamos a trabajar con la base de datos "nomel", la cual contabiliza el número de personas que sufrieron de cáncer a la piel durante un año en dos ciudades de Estados Unidos para distintos grupos etarios. Los datos incluyen las siguientes variables:

1. cases: que contabiliza el número de personas a las que se le detectaron melanomas para cada grupo etario de cada ciudad,
2. age.range: una variable categórica que indica el grupo etario de las personas (15-24, 25-34, 35-44, 46-54, 55-64, 64-74, 75-84, 85 o más),
3. n: variable que contabiliza el número total de personas que pertenecen a cada grupo etario en cada ciudad, y
4. city: una variable binaria que indica la ciudad de residencia de las personas.

```
#Exploración base de datos
head(nonmel)
```

```
#   cases      n        city age.range
# 1     1 172675 Minneapolis     15_24
# 2    16 123065 Minneapolis     25_34
# 3    30  96216 Minneapolis     35_44
# 4    71  92051 Minneapolis     45_54
# 5   102  72159 Minneapolis     55_64
# 6   130  54722 Minneapolis     65_74
```

Regresión Poisson

Con la base de datos nonmel, estimaremos dos modelos poisson donde el número de personas a las que se le detectaron melanomas es la variable dependiente, y ciudad y grupos etarios son las variables independientes. Use la ciudad de Minneapolis y el grupo etario mayor como categorías de referencia. Así mismo, estimaremos un segundo modelo con la variable *n* como variable de exposición o offset (controlando por el logaritmo de la población).

```
model1 = glm(cases~city+age.range, data=nonmel, family="poisson")
## se ajustan los resultados al tamaño de los grupos etarios. Para ello se utiliza el Offset.
model2 = glm(cases~city+age.range, data=nonmel, offset=log(n), family="poisson")
```

Regresión Poisson

Table 1: Modelos poisson

	Model 1	Model 2
(Intercept)	3.44^{***}	-5.48^{***}
	(0.10)	(0.10)
cityDallas	0.86^{***}	0.80^{***}
	(0.05)	(0.05)
age.range15_24	-3.04^{***}	-6.17^{***}
	(0.46)	(0.46)
age.range25_34	-0.66^{***}	-3.54^{***}
	(0.17)	(0.17)
age.range35_44	0.35^{**}	-2.33^{***}
	(0.13)	(0.13)
age.range45_54	1.02^{***}	-1.58^{***}
	(0.11)	(0.11)
age.range55_64	1.23^{***}	-1.09^{***}
	(0.11)	(0.11)
age.range65_74	1.43^{***}	-0.53^{***}
	(0.11)	(0.11)
age.range75_84	1.23^{***}	-0.12
	(0.11)	(0.11)
AIC	136.40	120.50
BIC	143.35	127.46
Log Likelihood	-59.20	-51.25
Deviance	24.15	8.26
Num. obs.	16	16

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

Regresión Poisson

¿Cuál de los dos modelos es preferible?

Al incorporar un offset se estima el número esperado de melanómas por grupo etario considerando el número de personas que hay en cada grupo etario, captando la heterogeneidad al interior de dichos grupos. Sin el offset, los coeficientes indican el número esperado de personas con melanóma sin considerar que al incluir más personas, hay mayores oportunidades de ver personas con melanóma dado la mayor cantidad de personas.

Regresión Poisson

Estimación del cambio porcentual en $E(y_i \mid X_{ik})$, utilizamos $exp(\beta_i \delta)$.

En base al modelo 2 calculamos el cambio porcentual en el numero de melonómas entre las personas de Dallas y Minneapolis?

```
exp(0.8039)
```

```
## [1] 2.234237
```

```
(2.234237-1)*100
```

```
## [1] 123.4237
```

El cambio porcentual para las personas que habitan en Dallas es igual a 123%, es decir, para las personas de Dallas, la tasa de casos con melanoma aumenta en un 123% respecto de quienes viven en Minneapolis, manteniendo todas las demas variables constantes, utilizando un 99% de confianza.

Regresión Poisson

Para estimar el cambio discreto en $E(y_i \mid X_{ik})$, utilizamos $exp(\beta_0 + \beta_1 + \sum_{k=2}^{K-1} \beta_k X_{ik}) - exp(\beta_0 + \sum_{k=2}^{K-1} \beta_k X_{ik})$.

```
X1=c(1,0,1,0,0,0,0,0,0)
X2=c(1,0,0,0,0,0,1,0,0)

## para tener en consideración la variable offset es necesario a agregarla a este cambio discreto que nos piden.
n.1 = nonmel$n[nonmel$age.range=="15_24" & nonmel$city=="Minneapolis"]
n.2 = nonmel$n[nonmel$age.range=="55_64" & nonmel$city=="Minneapolis"]

exp(sum(X1*model2$coef)) * n.1 - exp(sum(X2*model2$coef)) * n.2
```

```
## [1] -99.62563
```

Entre los habitantes de Minneapolis se espera observar 100 personas con melanóma más entre las personas de ente 55 y 64 años, respecto a aquellas con edades de entre 15 y 24 años.