Sesión 3

Contents

FACULTAD DE CIENCIAS SOCIALES - PUCP

Curso: POL 304 - Estadística para el análisis político 2 | Semestre 2024- 1

Jefas de Práctica: Karina Alcántara y Lizette Crispín

SESIÓN 3 - Regresión Logística Binaria

Diferencia de conceptos

	Urbana	Rural	Total
Sí	10992	1868	12860
No	5012	2675	7687
	16004	4543	20547

Los *odds* se interpretan como ratios, es decir, la cantidad de veces que algo pueda suceder sobre que no pueda suceder.

Probabilidad : qué tan posible es que ocurra un evento

$$\pi = \frac{N_{(X=1)}}{N}$$

La **probabilidad** de encontrar a una mujer sexualmente activa que use métodos anticonceptivos modernos es de:

Casos favorables / Casos posibles $\frac{12860}{20547} = 0.623$

"La probabilidad de que ocurra"

Odds : La probabilidad de un evento (p) sobre la probabilidad de que no ocurra (1-p)

$$\frac{\pi}{1-\pi}$$

Los **Odds** de encontrar a una mujer sexualmente activa que use métodos anticonceptivos modernos es de:

Prob que Sí ocurra / Prob que NO ocurra $\frac{12860/20547}{7687/20547} = 1.67$

"Veces más probable que ocurra a que no ocurra"

Odds Ratio: La posibilidad de que un evento ocurra según otra condición

$$\pi = \frac{OR}{OR+1}$$

El **Odds Ratio (OR)** de a posibilidad de que se usen métodos anticonceptivos modernos **según** área de residencia urbana :

$\frac{\text{odds urbana } 10992/5012}{\text{odds rural } 1868/2675} = 3.14$

"Veces más probable que ocurra el evento en zona urbana que en rural"

Ejemplo

			Total
Ganó	14	18	32
Perdió	3	4	7
	17	22	39

Un joven entrenador de ajedrez ha recorrido diversas regiones del Perú para concursar a nivel nacional junto a dos estudiantes . Luego de una extensa gira quiere conocer las probabilidades de ganar de su equipo para emplear estrategias de entrenamiento.

Probabilidad : qué tan posible es que ocurra un evento

$$\pi = \frac{N_{(X=1)}}{N}$$

Casos favorables / Casos posibles $\frac{32}{39} = 0.82$

"La probabilidad de que ocurra"

Odds : La probabilidad de un evento (p) sobre la probabilidad de que no ocurra (1-p)

$$\frac{\pi}{1-\pi}$$

$$\frac{\frac{32}{39}}{\frac{7}{39}} = 4.57$$

"Veces más probable que ocurra a que no ocurra"

Odds Ratio: La posibilidad de que un evento ocurra según otra condición

$$\frac{\frac{14}{17}}{\frac{3}{17}} \Big/ \frac{\frac{18}{22}}{\frac{4}{22}} = 1.022 \quad \text{OR}$$

"Veces más probable que ocurra el evento con Píkachu que con Squirtle"

Así podemos transformar el OR a probabilidad $\pi = \frac{OR}{OR+1}$

La revancha del odds

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_n X_n$$

Valor teórico de log(ods)

Por cada unidad, en cuanto aumenta el **log(ods)**

MODELO DE REGRESIÓN LOGÍSTICA

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1$$

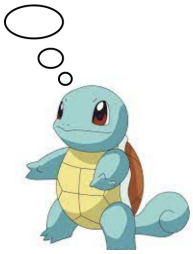
ECUACIÓN PARA HALLAR PROBABILIDAD

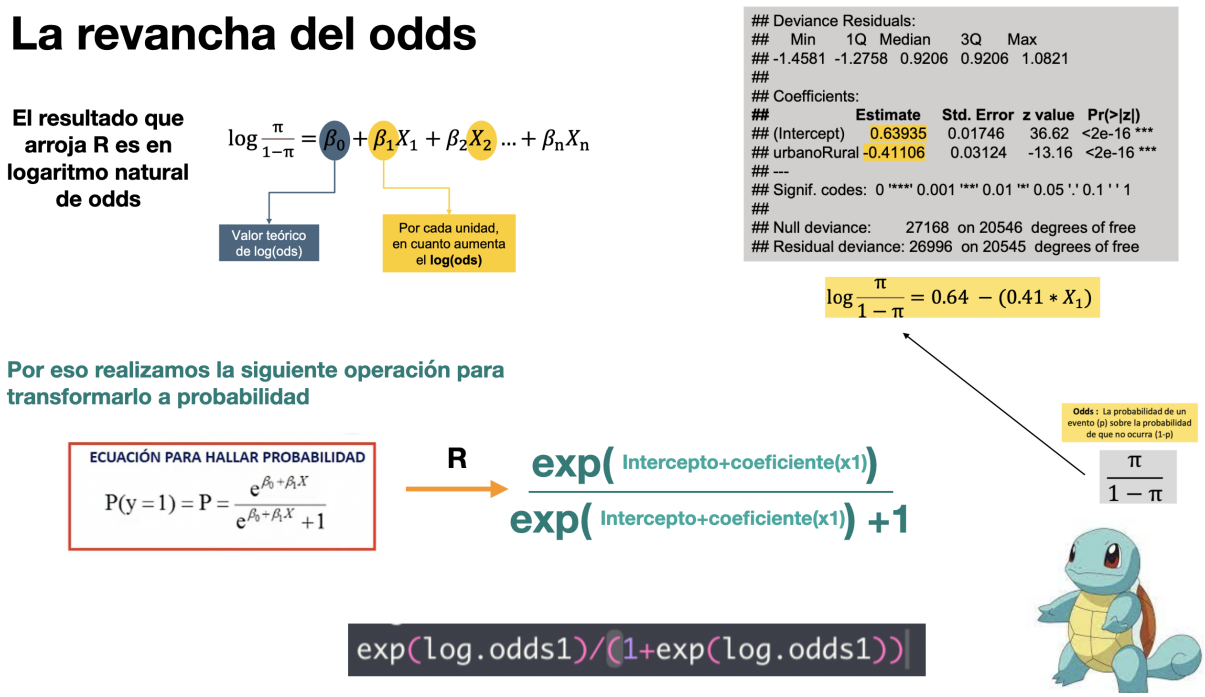
$$P(y=1) = P = \frac{e^{\beta_0+\beta_1 X}}{e^{\beta_0+\beta_1 X}+1}$$

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.4581  -1.2758  0.9206  0.9206  1.0821
##
## Coefficients:
##              Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept)   0.63935     0.01746    36.62  <2e-16 ***
## urbanoRural -0.41106     0.03124   -13.16  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Null deviance:      27168  on 20546  degrees of free
## Residual deviance: 26996  on 20545  degrees of free
```

$$\log \frac{\pi}{1-\pi} = 0.64 - (0.41 * X_1)$$ → Probabilidad

- Los *coefficients* obtenidos en esta regresión logística son el logaritmo natural de *odds*
- La función exponencial es la inversa del logaritmo
- Los *odds* son iguales a los exponenciales del coeficiente (porque este resultado viene en logaritmo natural)
- Si deseo el *odds* de un coeficiente entonces le aplico la función exponencial-> *exp(var)*





Modelo logístico binario

¿Qué factores pueden influenciar el que un docente quiera retornar a clases presenciales?

La base que usaremos hoy es la Encuesta Nacional a Docentes de Instituciones Educativas Públicas de Educación Básica Regular

Esta base de datos es del 2020, es decir, que hay que tomar en cuenta que se realizó en contexto de la pandemia. Entonces, hay diversas variables. Con respecto al cuidado de parientes, qué enfermedades ha tenido, satisfacción sobre temas personales o de la misma institución educativa.

```
library(rio)
library(dplyr)
library(marginaleffects)
endo<-import("ENDO1.sav")
```

Estas son las variables que usaremos:

Variable dependiente: *(P2_2)* Retorno a clases

Variables independientes:

- **P1_24_E**: ¿Cuán satisfecho esta Ud. con los siguientes aspectos?: Su empleo en esta IE
- **P1_2**: EDAD
- **P1_4**: En su hogar, ¿vive usted con personas de la tercera edad?

- **P1_5**: En su hogar, ¿vive Ud. con personas que están en el grupo de riesgo ante COVID-19 por enfermedades preexistente

Durante el año 2020

- **P1_11_B**: ¿sufrió o sufre enfermedades respiratorias?
- **P1_11_F**: ¿sufrió o sufre ansiedad?
- **P1_11_G**: ¿sufrió o sufre depresión
- **P1_11_H**: ¿sufrió o sufre cancer?
- **P1_11_L**: ¿sufrió o sufre COVID-19?
- **P1_18**: ¿En este momento se encuentra pagando algún préstamo o crédito?

Limpieza de data

Selección de variables a usar

```
data <-endo%>%
       select( P2_2, P1_24_E, P1_2, P1_4, P1_5, P1_11_B, P1_11_F, P1_11_G,P1_11_H, P1_11_L, P1_18)
```

Tenemos variable de sexo, edad, si es que es area rural o urbana. También si es que el docente vive con personas de tercera edad, o con personas que tienen factores de riesgo de COVID, si en el 2020 han tenido depresión, ansiedad, enfermedades respiratorias, también hay otra variable sobre si regresarían a clases de manera presencial.

```
names(data)
```

Cambiamos los nombres para que sea más fácil identificar las variables

```
colnames(data)=c("Retorno" ,  "satIE","edad"  ,  "terEd" ,   "riesCov",  "resp","anx", "dep", "cancer",
```

```
data=as.data.frame(data[complete.cases(data),])
```

VARIABLE DEPENDIENTE: Retorno

```
table(data$Retorno)
```

```
##
##     0     1
##  1551 16484
```

```
data$Retorno <- as.factor(data$Retorno)
levels(data$Retorno)  <-  c("No", "Si")
table(data$Retorno) #confirmo el nuevo formato de la variable
```

```
##
##    No    Si
##  1551 16484
```

Ya teniendo lista la variable depediente vamos a realizar unos cuantos modelos y analizar el odds y la probabilidad.

MODELO 1 : Retorno y personas de la tercera edad

- VD: Retorno (variable dicotómica)
- VI: El docente vive con personas de la tercera edad terEd

```r
data$terEd <- ifelse(data$terEd == "1", "1","0")
data$terEd <- as.numeric(data$terEd)
table(data$terEd)
```

```
##
##     0     1
## 10660  7375
```

Creemos nuestro modelo (función glm). Recuerda que lo que se está modelando es el `logaritmo del odds` (p/1-p).

```r
modelo1 <- glm(Retorno ~ terEd,family= binomial,data)
summary(modelo1)
```

```
##
## Call:
## glm(formula = Retorno ~ terEd, family = binomial, data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.53621    0.03715  68.272  < 2e-16 ***
## terEd       -0.38557    0.05321  -7.246 4.31e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10575  on 18034  degrees of freedom
## Residual deviance: 10523  on 18033  degrees of freedom
## AIC: 10527
##
## Number of Fisher Scoring iterations: 5
```

Recordemos que

- *Los coefficientes obtenidos en esta regresión logística son el logaritmo natural de odds*

Es importante revisar el **signo** del coeficiente, ya que dependiendo de eso procederemos a interpretar. En este caso, el coeficiente es negativo; es decir, *la relación es inversa.* La interpretación de la probabilidad tendrá la siguiente forma:

> "cuando la VI aumenta o es 1 (para dicotómicas), la probabilidad de que la VD sea 1 (sea del caso de éxito) en promedio disminuye en [resultado de avg_slopes]…"

Ejemplo para VI numérica (edad)

- Cuando la edad **aumenta** en 1 año, la probabilidad de que la persona quiera emigrar a Europa **disminuye** en 20.5%.

Ejemplo para VI dicotómica:

- Cuando la persona **sí** tiene hijos **(es 1)** , la probabilidad de que la persona quiera emigrar a Europa disminuye en 31.3%

Ahora en el modelo1:

```r
avg_slopes(modelo1)[,c(1,3)]
```

```
##
##   Term Estimate
```

```
##  terEd  -0.0309
##
## Columns: term, estimate
```

Identificamos que la variación en la probabilidad es de -0.0309; es decir, disminuye un 3.09%. Esto quiere decir que cuando una persona sí vive con personas de tercera edad (es 1), la probabilidad de que quiera retornar a clases disminuye en un 0.0309 o en 3.09% (en promedio).

¿De donde sale este valor?

```
head(modelo1$fitted.values,10)
```

```
##         1         2         3         6         7        13        15        16
## 0.9266417 0.8957288 0.8957288 0.9266417 0.9266417 0.8957288 0.8957288 0.9266417
##        19        20
## 0.8957288 0.9266417
```

Solo existen dos posibles probabilidades: cuando no tiene personas de tercera edad, 0.926, y cuando sí tiene personas de tercera edad, 0.895. Entonces la reducción de la probabilidad será la diferencia entre ambos.

```
0.9266417 - 0.8957288
```

```
## [1] 0.0309129
```

MODELO 2 : Retorno, personas de la tercera edad y enfermedades

Agreguemos más variables:

- vive con personas de la tercera edad **(terEd)**
- ¿sufrió o sufre cancer? **(cancer)**
- ¿sufrió o sufre depresion? **(dep)**

Queremos saber si estas variables influyen en la probabilidad de que el docente quiera retornar o no a clases presenciales

```
modelo2 <- glm(Retorno ~ terEd+cancer+dep, family = binomial(link=logit),data = data)
summary(modelo2)
```

```
##
## Call:
## glm(formula = Retorno ~ terEd + cancer + dep, family = binomial(link = logit),
##     data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.59283    0.03943  65.750  < 2e-16 ***
## terEd       -0.38375    0.05328  -7.203 5.89e-13 ***
## cancer      -0.92242    0.21281  -4.334 1.46e-05 ***
## dep         -0.26276    0.06741  -3.898 9.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10575  on 18034  degrees of freedom
## Residual deviance: 10492  on 18031  degrees of freedom
## AIC: 10500
```

```
##
## Number of Fisher Scoring iterations: 5
```

Ojo, los tres coeficientes son negativos. Calculemos la probabilidad de que los docentes quieran retornar a las clases presenciales.

```
avg_slopes(modelo2)[,c(1,3)]
```

```
##
##    Term Estimate
##  cancer  -0.1038
##  dep     -0.0221
##  terEd   -0.0307
##
## Columns: term, estimate
```

Interpretemos **(Ojo: las variables aparecen en orden alfabético)**

- Si el docente ha tenido o tiene cáncer, la probabilidad de que quiera retornar a clases presenciales disminuye, en promedio, en 0.1038 o en 10.38%
- Si el docente ha tenido o tiene depresión, la probabilidad de que quiera retornar a clases presenciales disminuye, en promedio, en 0.0221 o en 2.21%
- Si el docente vive con personas de la tercera edad, la probabilidad de que quiera retornar a clases presenciales disminuye, en promedio, en 0.0307 o en 3.07%

Si queremos calcular datos determinados Ejemplo 1: Si el docente no vive con personas de la tercera edad, tiene cancer y tiene depresión

```
log.odds1<-predict(modelo2, data.frame(terEd = 0, cancer = 1, dep = 1))

exp(log.odds1)/(1+exp(log.odds1))#lo pasamos a probabilidades
```

```
##         1
## 0.8033939
```

La **probabilidad** estimada de que quiera retornara a clases presenciales es de 0.80

Ejemplo 2: Si el docente **no** vive con personas de la tercera edad, no tiene cancer y tiene depresión

```
log.odds2<-predict(modelo2, data.frame(terEd = 0, cancer = 0, dep = 1))
exp(log.odds2)/(1+exp(log.odds2)) #lo pasamos a probabilidades
```

```
##         1
## 0.9113365
```

La probabilidad de que quiera retornara a clases presenciales es de 0.91

MODELO 3: Retorno, personas de la tercera edad, ansiedad, edad

Nuestras explicativas serán si la persona vive o no con personas de la tercera edad, tiene o ha tenido ansiedad y la variable edad.

```
modelo3<-glm(Retorno ~ terEd+anx+edad, family = binomial, data = data)
summary(modelo3)
```

```
##
## Call:
## glm(formula = Retorno ~ terEd + anx + edad, family = binomial,
```

```
##     data = data)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.547717   0.137744  25.756  < 2e-16 ***
## terEd       -0.354417   0.053486  -6.626 3.44e-11 ***
## anx         -0.437741   0.055822  -7.842 4.44e-15 ***
## edad        -0.019287   0.002844  -6.782 1.19e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10575  on 18034  degrees of freedom
## Residual deviance: 10414  on 18031  degrees of freedom
## AIC: 10422
##
## Number of Fisher Scoring iterations: 5
```

Analicemos las probabilidades con los efectos marginales:

```
avg_slopes(modelo3)
```

```
##
##   Term Contrast Estimate Std. Error     z Pr(>|z|)    S    2.5 %   97.5 %
##  anx      1 - 0  -0.0367   0.005015 -7.31   <0.001 41.8 -0.04650 -0.02684
##  edad     dY/dX  -0.0015   0.000223 -6.75   <0.001 35.9 -0.00194 -0.00107
##  terEd    1 - 0  -0.0282   0.004333 -6.50   <0.001 33.5 -0.03664 -0.01966
##
## Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
## Type:  response
```

Interpretemos:

- Si el docente ha tenido o tiene ansiedad la probabilidad de que quiera retornar a clases presenciales disminuye en 0.0367 o en 3.67%
- Si el docente aumenta en 1 año su edad la probabilidad de que quiera retornar a clases presenciales disminuye en 0.0015 o en 0.15%
- Si el docente vive con personas de la tercera edad la probabilidad de que quiera retornar a clases presenciales disminuye en 0.0282 o en 2.82%

Ahora obtengamos la probabilidad de que quieran retornar a clases presenciales según ciertas condiciones:

¿Qué sucede cuando el docente NO vive con personas de tercera edad (es 0), ha tenido o tiene ansiedad (es 1) y su edad es de 50 años?

```
log.odds3<-predict(modelo3, data.frame(terEd = 0, anx = 1, edad = 50))
exp(log.odds3)/(1+exp(log.odds3)) #para pasarlo a probabilidad
```

```
##         1
## 0.8952608
```

Cuando un o una docente vive con personas de tercera edad, tiene ansiedad y tenga 50 años, la probabilidad de que quiera retornar a clases presenciales es de 0.89 o de 89.34%.

Ahora obtengamos la probabilidad con menos edad.

```r
log.odds4<-predict(modelo3, data.frame(terEd = 0, anx = 1, edad = 25))
exp(log.odds4)/(1+exp(log.odds4))
```

```
##       1
## 0.93263
```

Cuando un o una docente vive con personas de tercera edad, tiene ansiedad y tenga 25 años, la probabilidad de que quiera retornar a clases presenciales es de 0.93 o de 93.27%