

Sesión 3

Contents

Modelo logístico binario	3
MODELO 1 : Retorno y personas de la tercera edad	5
MODELO 2 : Retorno, personas de la tercera edad y enfermedades	7
MODELO 3:	10



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:

$$\begin{array}{l} \text{Casos favorables} \quad 12860 \\ \text{Casos posibles} \quad 20547 \end{array} = 0.623$$

“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:

$$\begin{array}{l} \text{Prob que Sí ocurre} \\ 12860/20547 \\ \hline \text{Prob que NO ocurre} \\ 7687/20547 \end{array} = 1.67$$

“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 :

$$\begin{array}{l} \text{odds urbana} \\ 10992/5012 \\ \hline \text{odds rural} \\ 1868/2675 \end{array} = .3.14$$

“Veces más probable que ocurre el evento en zona urbana que en rural”

Ejemplo



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

Probabilidad : qué tan posible es que ocurra un evento

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

$$\begin{array}{l} \text{Casos favorables} \quad 32 \\ \text{Casos posibles} \quad 39 \end{array} = 0.82$$

“La probabilidad de que ocurra”

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

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

$$\begin{array}{l} \frac{32}{39} \\ \hline \frac{7}{39} \end{array} = 4.57$$

“Veces más probable que ocurra a que no ocurra”

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.

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

$$\begin{array}{l} \text{Pikachu} \\ \frac{14}{17} \\ \hline \text{Squirtle} \\ \frac{3}{17} \end{array} / \begin{array}{l} \text{Squirtle} \\ \frac{18}{22} \\ \hline \text{Pikachu} \\ \frac{4}{22} \end{array} = 1.022$$

“Veces más probable que ocurre el evento con Pikachu que con Squirtle”

Así podemos transformar el OR a probabilidad

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

La revancha del odds

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

Valor teórico de log(odds)

Por cada unidad, en cuanto aumenta el log(odds)

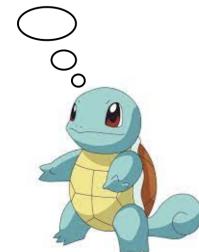
MODELO DE REGRESIÓN LOGÍSTICA

$$\log \left(\frac{p}{1-p} \right) = b_0 + b_1 X_1$$

ECUACIÓN PARA HALLAR PROBABILIDAD

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

- 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 exponentiales del coeficiente (porque este resultado viene en logaritmo natural)
- Si deseas el *odds* de un coeficiente entonces le aplico la función exponencial-> *exp(var)*



```
## 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
```

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

La revancha del odds

El resultado que arroja R es en logaritmo natural de odds

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

Valor teórico de log(odds)

Por cada unidad, en cuanto aumenta el log(odds)

```
## 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
```

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

Por eso realizamos la siguiente operación para transformarlo a probabilidad

ECUACIÓN PARA HALLAR PROBABILIDAD

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

R $\rightarrow \frac{\exp(\text{Intercepto}+\text{coeficiente}(x1))}{\exp(\text{Intercepto}+\text{coeficiente}(x1)) + 1}$

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

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



$$\exp(\text{log.odds1}) / (1 + \exp(\text{log.odds1}))$$

Modelo logístico binario

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)
endo=import("ENDO1.sav")
```

```
library(dplyr)
```

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?

```
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 área 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)

## [1] "P2_2"      "P1_24_E"    "P1_2"       "P1_4"       "P1_5"       "P1_11_B"    "P1_11_F"
## [8] "P1_11_G"   "P1_11_H"    "P1_11_L"    "P1_18"

colnames(data)=c("P2_2" , "satIE","edad" , "terEd" , "riesCov", "resp","anx", "dep", "cancer", "co
data=as.data.frame(data[complete.cases(data),])

```

Vamos a realizar diferentes modelos para calentar motores y volvemos expertas y expertos en la intrepretación de coeficientes.

Lo que queremos hacer es ver qué factores pueden influenciar en que un docente quiera retornar a clases presenciales

VARIABLE DEPENDIENTE: P2_2 Retorno a clases

```

table(data$P2_2)

##
##      0      1
## 1551 16484

data$retorno=as.factor(data$P2_2)
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

```

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).

```

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

```

Ahora vamos a calcular los coeficientes del modelo - [los *coefficients* obtenidos en esta regresión logística son el logaritmo natural de odds] - [La función exponencial es la inversa del logaritmo]

Revisemos los coeficientes.

```
coef(modelo1)
```

```

## (Intercept)      terEd
##     2.536211   -0.385568

```

Recuerda que 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*.

Damos el resultado en veces

```
1-exp(-0.385568)
```

```
## [1] 0.3199358
```

Si el docente vive con una persona de la tercera edad su deseo de retornar a la presencialidad disminuye en 0.32 veces.

Alternativo [Cuando calculamos el exponencial de los coeficientes, obtenemos el odds (# de veces de que ocurra)]. Sin embargo, como el coeficiente es negativo una de las alternativas es dividir el exponencial al 1 ($1/\exp(\log)$).

```
#Alternativo
1/exp(-0.385568)
```

```
## [1] 1.470449
```

Siempre que la relación es inversa, la interpretación es cuando la VI disminuye, la VD aumenta en xx veces. En este caso, 1.47 veces.

Tenemos dos maneras de poder analizar este resultado Recordemos que la VI era si el o la docente vive con personas de la tercera edad.

Otra manera analizarlo como una probabilidad, para ello es necesario realizar el siguiente cálculo $(1 - \exp(\text{coef})) * 100$

Recuerda que el coeficiente de la VI es -0.3860 es el

```
(1-exp(-0.3860))*100
```

```
## [1] 32.02295
```

En este caso podemos interpretar los resultados como “**si el docente vive con una persona de la tercera edad, la probabilidad de que quiera retornar a clases presenciales DISMINUYE en un 32.02%**”

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
```

```
coef(modelo2)
```

```
## (Intercept)      terEd      cancer       dep
##  2.5928268  -0.3837506  -0.9224211  -0.2627631
```

Ojo, los tres coeficientes son negativos. Calculemos el exponencial

Para calcular el odds.

```
exp(coef(modelo2))
```

```
## (Intercept)      terEd      cancer       dep
## 13.3675056   0.6813013   0.3975554   0.7689240
```

Nuevamente, el odds es menor que 1.

Como son menores que 1 entonces lo restamos y explicamos los resultados en base a la disminución de veces.

```
1-(exp(-0.3837506))
```

```
## [1] 0.3186987
```

```
1-(exp(-0.9224211))
```

```
## [1] 0.6024446
```

```
1-(exp(-0.2627631))
```

```
## [1] 0.231076
```

Análisis según n° de veces

- Si el docente vive con personas de la tercera edad el odds/probabilidad de que quiera retornar a clases presenciales disminuye en 0.31 veces
- Si el docente ha tenido o tiene cáncer el odds/probabilidad de que quiera retornar a clases presenciales disminuye en 0.6 veces
- Si el docente ha tenido o tiene depresión el odds/probabilidad de que quiera retornar a clases presenciales disminuye en 0.23 veces

Ahora analicemos las probabilidades:

```
#Cuando el odds es menor a 1  
(1-(exp(-0.38429)))*100
```

```
## [1] 31.90661
```

```
(1-(exp(-0.92276)))*100
```

```
## [1] 60.25794
```

```
(1-(exp(-0.26229)))*100
```

```
## [1] 23.07121
```

- Si el docente vive con personas de la tercera edad, la probabilidad de que quiera retornar a clases presenciales disminuye en 31%
- Si el docente ha tenido o tiene cáncer la probabilidad de que quiera retornar a clases presenciales disminuye en 60%
- Si el docente ha tenido o tiene depresión la probabilidad de que quiera retornar a clases presenciales disminuye en 23%

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))
```

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

```
exp(log.odds1)
```

```
## 1  
## 4.086311
```

```
exp(log.odds1)/(1+exp(log.odds1))
```

```
## 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.odds1 = predict(modelo2, data.frame(terEd = 0, cancer = 0, dep = 1))  
exp(log.odds1)/(1+exp(log.odds1)) #lo pasamos a probabilidades
```

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

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

MODELO 3:

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
```

Recuerda revisar los signos, para poder identificar el tipo de relación.

```
coef(modelo3)

## (Intercept)      terEd          anx          edad
## 3.54771740 -0.35441669 -0.43774133 -0.01928669
```

Interpretemos según n° de veces

```
1-(exp(-0.35441669))
```

```
## [1] 0.2984174
```

```
1-(exp(-0.43774133))
```

```
## [1] 0.3545073
```

```
1-(exp(-0.01928669 ))
```

```
## [1] 0.01910189
```

Análisis

- Si el docente vive con personas de la tercera edad el odds de que quiera retornar a clases presenciales disminuye en 0.29 veces
- Si el docente ha tenido o tiene ansiedad el odds de que quiera retornar a clases presenciales disminuye en 0.35 veces
- Si el docente aumenta en un 1 su edad el odds de que quiera retornar a clases presenciales disminuye en 0.02 veces

```
(1-(exp(-0.35441669))) *100
```

Ahora en porcentaje (probabilidad)

```
## [1] 29.84174
```

```
(1-(exp(-0.43774133)))*100
```

```
## [1] 35.45073
```

```
(1-(exp(-0.01928669 )))*100
```

```
## [1] 1.910189
```

- Si el docente vive con personas de la tercera edad la probabilidad de que quiera retornar a clases presenciales disminuye en 29%
- Si el docente ha tenido o tiene ansiedad la probabilidad de que quiera retornar a clases presenciales disminuye en 35.5%
- Si el docente aumenta en 1 su edad la probabilidad de que quiera retornar a clases presenciales disminuye en 1.9%

Ahora obtengamos la probabilidad según casos

```
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.

Ahora obtengamos la probabilidad con menor edad.

```
log.odds3 = predict(modelo3, data.frame(terEd = 0, anx = 1, edad = 25))  
exp(log.odds3)/(1+exp(log.odds3))
```

```
## 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.

Este fue un capítulo de...

- No me gusta las **Matemáticas**
- y ¿Qué quieres estudiar?
- Ciencias sociales
- ...



Extra: Ayuda divina

disclaimer : Material opcional y autodidacta para lxs alumnxs que desean ahondar más en el tema de funciones y usarlo para el presente tema.

Les proponemos esta función (Agradezcanle a su profesor) que facilita la interpretación de los resultados. La función se llama Divine.Help, para poder usarla solo necesitas indicar como argumento al nombre de tu modelo: Divine.Help(modelo). Recuerda que esta función solo podrá ejecutarse si previamente has ejecutado el código que crea la función.

```
Divine.Help <- function(model) {  
  # Extraer los coeficientes del modelo  
  coeficients <- coef(model)[-1] # Excluye el intercepto  
  # Inicializar un vector para almacenar los efectos  
  effects <- numeric(length(coeficients))  
  
  for (i in 1:length(coeficients)) {  
    if (coeficients[i] < 1) {  
      effects[i] <- round(1 - exp(coeficients[i]), 2)  
    } else {  
      effects[i] <- round(exp(coeficients[i] - 1), 2)  
    }  
  }  
  
  # Generar la interpretación en lenguaje natural  
  interpretation <- paste0("Un aumento de una unidad en la variable ", names(coeficients),  
    " está asociado con un cambio de ",  
    abs(effects)*100, "% en la probabilidad de éxito.")
```

```

# Devolver los coeficientes, efectos y la interpretación en un dataframe
result <- data.frame(Coefficient = coeficients,
                      Exp = exp(coeficients),
                      Probability = effects,
                      Interpretation = interpretation)
return(result)
}

Divine.Help(modelo2)

```

```

##           Coefficient      Exp Probability
## terEd    -0.3837506 0.6813013      0.32
## cancer   -0.9224211 0.3975554      0.60
## dep      -0.2627631 0.7689240      0.23
##
## terEd    Un aumento de una unidad en la variable terEd está asociado con un cambio de 32% en la probal
## cancer  Un aumento de una unidad en la variable cancer está asociado con un cambio de 60% en la probal
## dep      Un aumento de una unidad en la variable dep está asociado con un cambio de 23% en la probal

```