Tarea 3

Francisco Herrera Barajas

2025-05-05

Ejercicio 1

Importa la base de datos Datos_Tarea3.txt a R Studio, recodifica la variable volunfp para que la categoría SÍ esté representada por el valor 1 y la categoría NO por el 0, y elimina los valores perdidos de todas las variables. Después, crea una tabla de contingencia entre las variables género (gndr) y voluntariado recodificado (volunfp). En el documento Word, pega la tabla de contingencia y calcula las siguientes probabilidades y odds: -La probabilidad de ser voluntaria dado que se es mujer -La probabilidad de ser voluntario dado que se es hombre -La odds de ser voluntaria frente a no serlo en mujeres -La odds de ser voluntario frente a no serlo en hombres -La odds ratio entre las dos odds anteriores. Todas estas probabilidades, odds, y odds ratio tienen que estar debidamente interpretadas.

Cargamos las libreías necesarías

```
library(rio)
library(tidyverse)
library(ggplot2)
library(caret)
library(performance)
library(rcompanion)
library(gtsummary)
library(lmtest)
library(nnet)
library(dplyr)
library(flextable)
```

Importamos y visualizamos los datos

```
datos <- import("datos_Tarea3.csv")

head(datos)
```

```
##    idno volunfp lrscale gndr agea
## 1 10302       2       8    2   80
## 2 10684       1      88    1   57
## 3 10958       2       3    2   32
## 4 11128       2       5    1   68
## 5 11328       2       5    1   61
## 6 11691       1       7    2   31
```

Recodificamos la variable volunfp y la convertimos en factor

```
datos$volunfp <- ifelse(datos$volunfp == 2, 0,
                        ifelse(datos$volunfp == 1, 1, NA))

datos$volunfp <- factor(
  ifelse(datos$volunfp == 1, "Si",
         ifelse(datos$volunfp == 0, "No", NA)),
  levels = c("No", "Si")
)
```

Vemos la cantidad de NA que hemos generado.

```
nNA <- sum(is.na.data.frame(datos))
print(paste("Número de Na: ",nNA))
```

```
## [1] "Número de Na:  5"
```

Procedemos a ver el número de filas.

```
nrow(datos)
```

```
## [1] 936
```

Tenemos tan solo 5 filas con NA de 936, así que procedemos a borrar estas filas.

```
datos <- datos %>% drop_na()
```

Recodificamos y Convertimos en factor también la variable gndr

Primeramente, recodficamos la variable gndr entre 0 y 1 para mejorar la interpretación, quedando: hombre = 0 y mujer = 1. Continuamos convirtiendo la variable en factor, esto lo hacemos para que R entienda que esta variable es una variable categórica. También para mantener las etiquetas de los valores de la variable con su nombre, lo que facilita la interpretación del análisis.

```
datos$gndr <- ifelse(datos$gndr == 2, 1,
                        ifelse(datos$gndr == 1, 0, NA))

datos$gndr <- factor(
  ifelse(datos$gndr == 0, "Hombre",
         ifelse(datos$gndr == 1, "Mujer",NA)),
  levels = c("Hombre", "Mujer")
)
```

Tablas de contigencia

```r
datos %>%
  tbl_cross(row = gndr, col = volunfp, percent = "cell",
            label = list( gndr ~ "Género",volunfp ~ "Voluntariado")) %>%
  as_flex_table()
```

	Voluntariado		
	No	Si	Total
Género			
Hombre	361 (39%)	71 (7.6%)	432 (46%)
Mujer	409 (44%)	90 (9.7%)	499 (54%)
Total	770 (83%)	161 (17%)	931 (100%)

En esta primera tabla de contigencia vemos que el porcentaje del total es cada celda.

```r
datos %>%
  tbl_cross(row = gndr, col = volunfp, percent = "row",
            label = list( gndr ~ "Género",volunfp ~ "Voluntariado")) %>%
  as_flex_table()
```

	Voluntariado		
	No	Si	Total
Género			
Hombre	361 (84%)	71 (16%)	432 (100%)
Mujer	409 (82%)	90 (18%)	499 (100%)
Total	770 (83%)	161 (17%)	931 (100%)

En esta segunda tabla de contingencia vemos que el total se caclcula por fila lo que nos ayuda a ver las probabilidades condicionadas.

Probabilidades condicionales

- P(Voluntaría | Mujer) = 0.18
- P(Voluntaría | Hombre) = 0.16

Podemos hacerlo prográmaticamente

```r
tabla <- datos %>%
  group_by(gndr, volunfp) %>%
  summarise(n = n(), .groups = "drop") %>%
  group_by(gndr) %>%
  mutate(probabilidad = n / sum(n))
```

```r
prob_mujer_voluntaria <- tabla %>%
  filter(gndr == "Mujer", volunfp == "Si") %>%
  pull(probabilidad)

prob_hombre_voluntario <- tabla %>%
  filter(gndr == "Hombre", volunfp == "Si") %>%
  pull(probabilidad)


print(paste("P(Voluntaria | Mujer) =", round(prob_mujer_voluntaria, 4)))
```

```
## [1] "P(Voluntaria | Mujer) = 0.1804"
```

```r
print(paste("P(Voluntario | Hombre) =", round(prob_hombre_voluntario, 4)))
```

```
## [1] "P(Voluntario | Hombre) = 0.1644"
```

Encajan con los resultados antes descritos para estas probabilidades condicionales.

La odds de ser voluntaria frente a no serlo en mujeres

```r
odd_m <- round(prob_mujer_voluntaria / (1 - prob_mujer_voluntaria), digits = 4)
print(paste("Odds mujer: ", odd_m))
```

```
## [1] "Odds mujer:  0.22"
```

Al ser la odd por debajo de 1, quiere decir que las mujeres tienen menor probabilidad relativa de ser voluntarias respecto así mismas.

La odds de ser voluntaria frente a no serlo en hombres

```r
odd_h <- round(prob_hombre_voluntario / (1 - prob_hombre_voluntario), digits = 4)
print(paste("Odds hombre: ", odd_h))
```

```
## [1] "Odds hombre:  0.1967"
```

Igual que las mujeres la odd esta por debajo de 1, los hombre tienen menor probabilidad relativa de ser voluntarias respecto así mismos.

La odds ratio entre las dos odds anteriores

```r
odd_ratio <- round(odd_m / odd_h, digits = 4)
print(paste("Odds ratio entre mujeres y hombres voluntarios: ", odd_ratio))
```

```
## [1] "Odds ratio entre mujeres y hombres voluntarios:  1.1185"
```

Este resultado quiere decir que las mujeres tienen un 1,11 más de posibilidades de ser volunarias que los hombres.

Ejercicio 2

Ajusta un modelo de regresión logística binaria, llamado modelo_1, en el que la variable dependiente sea si la persona ha hecho voluntariado o no (volunfp), y la variable independiente (o covariable) sea el enero (gndr). Interpreta todos los resultados: -Los coeficientes en escala logit -Los coeficientes en escala odds y odds ratio -La tabla de clasificación -El ajuste del modelo. -Obtén los pronósticos de la probabilidad de ser voluntario en ambos géneros

```
modelo_1 <- glm(volunfp ~ gndr, data = datos, family = binomial)
summary(modelo_1)
```

```
##
## Call:
## glm(formula = volunfp ~ gndr, family = binomial, data = datos)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.6262     0.1298 -12.526   <2e-16 ***
## gndrMujer     0.1123     0.1744   0.644     0.52
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 857.46  on 930  degrees of freedom
## Residual deviance: 857.05  on 929  degrees of freedom
## AIC: 861.05
##
## Number of Fisher Scoring iterations: 4
```

En esta salida tanto el coeficiente como el intercepto están em escala logit. Esta escala no es muy intuitiva para interpretar el modelo,esencialmente, es mejor transformar estos valores a odds y a probabilidades.

Sin embargo, si es interesante ver que el coeficiente para la variable género no es signficativo estadísticamente. Esto no esta diciendo que no hay una diferencia significativa entre géneros a la hora de participar en actividades de voluntariado.

Por otra parte, que el intercepto sí sea signicativo lo único que no esta diciendo es que la probabilidad de hacer voluntariados para el grupo de referencia en nuestro caso, los hombres, es distinta a 0,5, lo cual no es muy informativo.

Coficientes en escala odds y odd ratio

```
nuevos_datos <- data.frame(gndr = 0)
nuevos_datos <- data.frame(gndr = factor("Hombre", levels = levels(datos$gndr)))

my_log_odds1 <- round(predict(modelo_1, newdata = nuevos_datos), digits = 4)
my_log_odds1
```

```
##       1
## -1.6262
```

```r
odd_1 <- round(exp(my_log_odds1), digits = 4)
print(paste("Odds para el grupo referencia (hombre): ", odd_1))
```

```
## [1] "Odds para el grupo referencia (hombre):  0.1967"
```

El resultado encaja con la odd de hombre obtenida en el ejercicio anterior. Sacamos la odd de mujer para poder conseguir la odds ratio.

```r
nuevos_datos1 <- data.frame(gndr = 1)
nuevos_datos1 <- data.frame(gndr = factor("Mujer", levels = levels(datos$gndr)))

my_log_odds2 <- round(predict(modelo_1, newdata = nuevos_datos1), digits = 4)
my_log_odds2
```

```
##       1
## -1.5139
```

```r
odd_2 <- round(exp(my_log_odds2), digits = 4)
odd_2
```

```
##      1
## 0.2201
```

```r
# Ponemos de numerador las odds de mujer para que sea más facil de interpretar
odds_ratio <- round(odd_2/ odd_1, digits = 4)
print(paste("Odds ratio: ", odds_ratio))
```

```
## [1] "Odds ratio:  1.119"
```

Las odds ratios es exactamente igual que hemos calculado en el ejercicio anterior.

Tabla de clasificación

Utilizamos como umbral el valor 0,17 por estar entre las dos probabilidades condicionales que predice el modelo:

- P(Voluntaría | Mujer) = 0.18
- P(Voluntaría | Hombre) = 0.16

De este modo, el umbral permite separar ambos grupos.

```r
glm.probs <- predict(modelo_1, newdata = datos, type = "response")
glm.predict <- ifelse(glm.probs > 0.17, "Si", "No")
glm.predict <- factor(glm.predict, levels = c("No", "Si"))
confusionMatrix(data = glm.predict, reference = datos$volunfp)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No  Si
##         No 361  71
##         Si 409  90
##
##                Accuracy : 0.4844
##                  95% CI : (0.4519, 0.5171)
##     No Information Rate : 0.8271
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : 0.0152
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.4688
##             Specificity : 0.5590
##          Pos Pred Value : 0.8356
##          Neg Pred Value : 0.1804
##              Prevalence : 0.8271
##          Detection Rate : 0.3878
##    Detection Prevalence : 0.4640
##       Balanced Accuracy : 0.5139
##
##        'Positive' Class : No
##
```

En nuestro caso, con una única variable predictora categórica binaria, la matriz de confusión no aporta información adicional útil, ya que la predicción se reduce a replicar el valor del predictor.

El ajuste del modelo.

```
nagelkerke(modelo_1, null = NULL, restrictNobs = FALSE)
```

```
## $Models
##
## Model: "glm, volunfp ~ gndr, binomial, datos"
## Null:  "glm, volunfp ~ 1, binomial, datos"
##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                          0.000484863
## Cox and Snell (ML)                0.000446464
## Nagelkerke (Cragg and Uhler)      0.000741778
##
## $Likelihood.ratio.test
##  Df.diff LogLik.diff   Chisq p.value
##       -1    -0.20788 0.41575 0.51906
##
## $Number.of.observations
```

```
##
## Model: 931
## Null:  931
##
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitting with ML"
##
## $Warnings
## [1] "None"
```

Las distintas formas de medir el R^2 son prácticamente cero lo que indica que la variable género no es un buen predictor de la participación en actividades de voluntariado en esta muestra.

Además, por el test de razón de similitudes, al no ser significativo este nos esta indicando que el modelo no ajusta mejor que un modeo nulo.

Obtén los pronósticos de la probabilidad de ser voluntario en ambos géneros.

```
prob <- odd_1 / (1+odd_1)
print(paste("Probabilidad de ser volubtario siendo hombre: ", prob))
```

```
## [1] "Probabilidad de ser volubtario siendo hombre:  0.164368680538147"
```

```
prob2 <- odd_2 / (1+odd_2)
print(paste("Probabilidad de voluntario siendo mujer: ", prob2))
```

```
## [1] "Probabilidad de voluntario siendo mujer:  0.1803950495861"
```

Son las mismas proabilidades del ejercicio anterior.

Ejercicio 3

Ajusta un nuevo modelo (llamado modelo_2), en el que incluyas como variable independiente la edad (agea) e ideología política (lrscale), además de la variable género (gndr). Recuerda centrar previamente ambas variables para facilitar su interpretación. Interpreta todos los resultados: -Los coeficientes en escala logit -Los coeficientes en escala odds y odds ratio -La tabla de clasificación -El ajuste del modelo. -Obtén los pronósticos de la probabilidad de ser voluntario para un hombre completamente de izquierdas con una edad media.

Adecuar la base de datos

Primeramente, procedemos a dejar la variable lrscale(ideología política) unicamente con valores entre 0 y 10.

```
datos <- datos[datos$lrscale >= 0 & datos$lrscale <=10,]
```

Vamos a tratar a esta variable como una variable continua (cuantitativa ordinal) para nuestro análisis.

Centramos las variables escala política y edad

```
datos$agea_c <- scale(datos$agea, center = TRUE, scale = FALSE)
datos$lrscale_c <- scale(datos$lrscale, center = TRUE, scale = FALSE)
```

Reconvertimos las columnas en vectores porque trás centrarlas se convierten en matriz.

```
datos$agea_c <- as.numeric(datos$agea_c)
datos$lrscale_c <- as.numeric(datos$lrscale_c)
```

Coeficientes en escala Logit

```
modelo_2 <- glm(volunfp ~ gndr + agea_c + lrscale_c, data = datos, family = binomial)
summary(modelo_2)
```

```
##
## Call:
## glm(formula = volunfp ~ gndr + agea_c + lrscale_c, family = binomial,
##     data = datos)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.599367   0.137916 -11.597   <2e-16 ***
## gndrMujer    0.117173   0.184460   0.635    0.525
## agea_c      -0.001941   0.002536  -0.765    0.444
## lrscale_c   -0.010188   0.038894  -0.262    0.793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 773.62  on 825  degrees of freedom
## Residual deviance: 772.15  on 822  degrees of freedom
## AIC: 780.15
##
## Number of Fisher Scoring iterations: 5
```

Como dijimos anteriormente, los coeficientes están en escala logit, lo que los hace poco intuitivos. Ninguno de los predictores (edad, ideología ni género) es estadísticamente significativo, lo cual sugiere que no son buenos predictores de la participación en actividades de voluntariado.

Aunque el intercepto sí es significativo, esto simplemente indica que la probabilidad de hacer voluntariado para una persona media (en edad e ideología) y de género masculino es distinta de 0,5.

Coeficientes en escala odds y odds ratio

Vamos a sacar los ocoeficientes en escala odds. No obstante, ya no podemos sacarlos como en el anterior modelo ya que con las variables continuas como edad y escala política hay que elegir valores concretos para el resto de predictores.

Odds para distintos valores de edad, siendo hombre y teniendo una ideología de centro

```
new_edad <- data.frame(
  gndr = factor("Hombre", levels = levels(datos$gndr)),
  agea_c = c(-10, 0, 10),
  lrscale_c = 0
)

logit_edad <- round(predict(modelo_2, newdata = new_edad), 4)

odds_edad <- round(exp(logit_edad), 4)

print(paste("Distintas odds para edad:", paste(odds_edad, collapse = ", ")))
```

```
## [1] "Distintas odds para edad: 0.206, 0.202, 0.1981"
```

La interpretación de esta salida es que al parecer cuanto más edad se reduce muy levemente las odds de participar en actividades de voluntariado.

Odds para distintos valores de ideología manteniendo constante los valores de edad y siendo hombre.

```
news_ideo <- data.frame(
  gndr = factor("Hombre", levels = levels(datos$gndr)),
  agea_c = 0,
  lrscale_c = c(-5, 0, 5)
)

logit_ideo <- round(predict(modelo_2, newdata = news_ideo), 4)

odds_ideo <- round(exp(logit_ideo), 4)

print(paste("Distintas odds para la escala política: ", paste(odds_ideo, collapse = ", ")))
```

```
## [1] "Distintas odds para la escala política:  0.2126, 0.202, 0.192"
```

Esto lo podemos interpretar como que a medida que la ideología tiende más a la de derecha las odds de participar en actividades de voluntariado se reducen levemente.

Odds para distintos valores de género manteniendo lso valores de idiología en el centro.

```
nuevos_gndr <- data.frame(
  gndr = factor(c("Hombre", "Mujer"), levels = levels(datos$gndr)),
  agea_c = 0,
  lrscale_c = 0
```

```
)

logit_gndr <- round(predict(modelo_2, newdata = nuevos_gndr), 4)
odds_gndr <- round(exp(logit_gndr), 4)

print(paste("Odds para hombre y mujer: ", paste(odds_gndr, collapse = ", ")))
```

```
## [1] "Odds para hombre y mujer:  0.202, 0.2271"
```

Estos datos nos muestran que las mújeres tienden un poco más que los hombres a participar en actividades de voluntariado.

Odds ratio.

```
odds_ratios_modelo_2 <- exp(coef(modelo_2))
odds_ratios_modelo_2
```

```
## (Intercept)   gndrMujer      agea_c   lrscale_c
##   0.2020243   1.1243143   0.9980607   0.9898635
```

Estos coeficientes los podemos interpretar:

1) La mujeres tienen una odds ratio de 1,12 más de participar en acitividades voluntarias que los hombres.
2) Por cada año adicional se reduce las odds de de ser voluntanrio.
3) Por cada punto en la escala del espectro político más, es decir pensamiento político más de derechas, disminuyen las odds de participar.

Tabla de clasificación.

```
probabilidades <- predict(modelo_2, type = "response")
predicciones <- ifelse(probabilidades >= 0.17, 1, 0)
table(Predicho = predicciones, Real = datos$volunfp)
```

```
##         Real
## Predicho  No  Si
##        0 178  33
##        1 501 114
```

```
acurracy_2 <- (178+114) / (178+501+33+114)
print(paste("Precisión: ",acurracy_2))
```

```
## [1] "Precisión:  0.353510895883777"
```

La precisión del modelo sigue siendo muy baja, aun bajando el umbral como el ejercicio anterior.

Ajuste del modelo 2

```r
nagelkerke(modelo_1, null = NULL, restrictNobs = FALSE)
```

```
## $Models
##
## Model: "glm, volunfp ~ gndr, binomial, datos"
## Null:  "glm, volunfp ~ 1, binomial, datos"
##
## $Pseudo.R.squared.for.model.vs.null
##                              Pseudo.R.squared
## McFadden                           -0.1078320
## Cox and Snell (ML)                 -0.0937413
## Nagelkerke (Cragg and Uhler)       -0.1660990
##
## $Likelihood.ratio.test
##  Df.diff LogLik.diff   Chisq p.value
##       -1      41.711 -83.422       1
##
## $Number.of.observations
##
## Model: 931
## Null:  826
##
## $Messages
## [1] "Note: For models fit with REML, these statistics are based on refitting with ML"
##
## $Warnings
## [1] "WARNING: Fitted and null models have different numbers of observations"
```

Los pseudo R^2 siguen siendo muy bajos para para este modelo. A pesar de añadir nuevas variables estas no consiguen explicar bien si una persona realiza actividades de voluntariado o no.

Obtén los pronósticos de la probabilidad de ser voluntario para un hombre completamente de izquierdas con una edad media

Ya que el modelo fue hecho con una variable centrada hay que centrar el valor del cual se hace la predicción.

```r
media_lrscale <- mean(datos$lrscal)
```

```
## Warning in mean.default(datos$lrscal): argument is not numeric or logical:
## returning NA
```

```r
lr_izquierda_c <- 0 - media_lrscale

perfil_izquierda <- data.frame(
  gndr = factor("Hombre", levels = levels(datos$gndr)),
  agea_c = 0,
  lrscale_c = lr_izquierda_c
)

probabilidad <- predict(modelo_2, newdata = perfil_izquierda, type = "response")

print(paste("Probabiliad de que un hombre sea completamente de izquierdas con edad media y que realize
```

```
## [1] "Probabiliad de que un hombre sea completamente de izquierdas con edad media y que realize volun
```

Ejercicio 4.

Compara el ajuste del modelo_2 con el ajuste del modelo_1.

Como al recodificar la variable lrscal para el segundo modelo hemos eliminado ciertas filas los dos modelos estan desbalanceados. Para subsanar este hecho vamos a volver a realizar el modelo 1 sin estas filas

```
modelo_1_plus <- glm(volunfp ~ gndr, data = datos, family = binomial)
summary(modelo_1_plus)
```

```
##
## Call:
## glm(formula = volunfp ~ gndr, family = binomial, data = datos)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5969     0.1371 -11.650   <2e-16 ***
## gndrMujer     0.1213     0.1833   0.662    0.508
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 773.62  on 825  degrees of freedom
## Residual deviance: 773.18  on 824  degrees of freedom
## AIC: 777.18
##
## Number of Fisher Scoring iterations: 4
```

Los resultados son bastante parecidos al modelo 1.

```
anova(modelo_1_plus,modelo_2)
```

```
## Analysis of Deviance Table
##
## Model 1: volunfp ~ gndr
## Model 2: volunfp ~ gndr + agea_c + lrscale_c
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       824     773.18
## 2       822     772.15  2    1.031   0.5972
```

Según esta salida añadir las variables de edad y escala ideológica no mejora el ajuste del modelo más complejo.