

Quinta entrega de Modelos Lineales

Fabrizio Camacho

2024-07-11

Ejercicio 1

Sea Y una variable con distribución perteneciente a la familia exponencial, esto es:

$$f_Y(y|\theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y; \phi)\right)$$

Empleando algunas propiedades del score $s(\theta)$ en clase se comprobó que $E(Y) = b'(\theta)$. Válgase de estas propiedades para demostrar el otro resultado que se planteó durante la clase, esto es:

$$\text{Var}(Y) = b''(\theta)a(\phi)$$

Se pide:

- 1) Demuestre este resultado.
- 2) Utilice estas propiedades para obtener la esperanza y Varianza de la distribución Poisson.
- 3) Utilice estas propiedades para obtener la esperanza y Varianza de la distribución Bernoulli.

1)

Definimos la función $\text{score}(\theta)$ como la derivada respecto de θ del logaritmo de la densidad $f_Y(y|\theta, \phi)$

$$s(\theta) = \frac{\partial(\log((y\theta - b(\theta))))}{\partial\theta} = \frac{\partial}{\partial\theta} \left(\frac{a(\phi)(y\theta - b(\theta))}{a(\phi)} + c(y; \phi) \right) = \frac{y - b'(\theta)}{a(\phi)}$$

Luego como:

$$E \left(\frac{y - b'(\theta)}{a(\phi)} \right) = \frac{E(y) - b'(\theta)}{a(\phi)} = \frac{\mu - b'(\theta)}{a(\phi)} = 0$$

Tenemos que:

$$b'(\theta) = \mu = E(Y)$$

Para la varianza de la función $s(\theta)$:

$$\text{Var}(s(\theta)) = E[s(\theta)^2] - (E[s(\theta)])^2 = E \left[\left(\frac{y - b'(\theta)}{a(\phi)} \right)^2 \right] - 0 = \frac{\text{Var}(Y)}{a(\phi)^2}$$

Dado que $\text{Var}(Y) = E(Y)^2 - E(Y^2)$:

$$\begin{aligned} \text{Var}(s(\theta)) &= E[s(\theta)^2] - E[s(\theta)]^2 = E(s(\theta)^2) = E \left[\left(\frac{(y - b'(\theta))^2}{a(\phi)^2} \right) \right] = \\ &= \left(\frac{E[(y - E(y))^2]}{a(\phi)^2} \right) = \frac{\text{Var}(Y)}{a(\phi)^2} = I(\theta) \end{aligned}$$

Donde:

$$I(\theta) = -E \left[\frac{\partial^2}{\partial\theta^2} \log f_Y(y, \theta) \right] = \frac{b''(\theta)}{a(\phi)}$$

$$I(\theta) = \frac{b''(\theta)}{a(\phi)} = \frac{\text{Var}(Y)}{a(\phi)^2}$$

Entonces:

$$\text{Var}(Y) = b''(\theta)a(\phi)$$

2)

Sea una variable Y con distribución $fy(y, \lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$:

$$fy(y, \lambda) = \exp(y \log(\lambda) - \lambda - \log(y!))$$

dado que $\theta = \log(\lambda)$, $a(\phi) = 1$, $b(\theta) = e^\theta$ y $c(y, \theta) = \log(y!)$:

$$E[Y] = b'(\theta) = e^\theta = \lambda$$

$$Var(Y) = b''(\theta)a(\phi) = e^\theta \cdot 1 = \lambda$$

3)

Sea una variable Y con distribución $fy(y, p) = p^y(1 - p)^{1-y}$:

$$fy(y, \theta) = \exp(y \log \left(\frac{p}{1-p} \right) + \log(1-p))$$

Dado que: $\theta = \log \frac{p}{1-p}$, $a(\phi) = 1$, $b(\theta) = \log(1 + e^\theta)$ y $c(y, \theta) = 0$

$$E[Y] = b'(\theta) = \frac{e^\theta}{1 + e^\theta} = p$$

$$Var(Y) = b''(\theta)a(\phi) = \frac{e^\theta}{(1 + e^\theta)^2} = p(1 - p)$$

Ejercicio 2

En el archivo `huracanes.xlsx` se dispone de la información de 337 huracanes registrados en la zona del Atlántico norte. A partir de este conjunto de datos se pide construir un modelo de regresión logística que permita determinar si la latitud donde se crea el huracán es capaz de predecir si este será de tipo tropical o no.

Se pide:

- 1) Ajuste un GLM adecuado a la variable de respuesta que se indica (tropical vs no tropical).
- 2) Indique el poder predictivo de este modelo en base a los indicadores *pseudo-R2* vistos en clase.
- 3) Interprete el valor del coeficiente asociado a la variable *Latitud de creacion* (variable `lat_creacion` en el conjunto de datos).
- 4) A partir de los resultados del modelo realice un gráfico que permita visualizar la probabilidad de que un huracán sea tropical en función de su latitud de creación.

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr    1.5.1
## v ggplot2    3.5.1      v tibble     3.2.1
## v lubridate  1.9.3      v tidyr      1.3.1
## v purrr      1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
## here() starts at C:/Users/fabri/OneDrive/Escritorio/Cosas/Facultad/1-Modelos lineales/Modelos-Lineales
```

1)

```
mod <- glm(trop2 ~ lat_creacion, data = huracanes, family = binomial) #Modelo
```

2)

```
Mcfaden <- 1 - (mod$deviance / mod$null.deviance)
```

```
cat("Pseudo R2=",Mcfaden)
```

```
## Pseudo R2= 0.4989746
```

3)

```
summary(mod)
```

```
##
## Call:
## glm(formula = trop2 ~ lat_creacion, family = binomial, data = huracanes)
##
## Coefficients:
```

```
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept)   9.08263    0.96148   9.446  <2e-16 ***
## lat_creacion -0.37283    0.03947  -9.447  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 463.11  on 336  degrees of freedom
## Residual deviance: 232.03  on 335  degrees of freedom
## AIC: 236.03
##
## Number of Fisher Scoring iterations: 6
```

El modelo explica el **49.9%** de la varianza al explicar si el huracan es tropical o no según la latitud según *Mc Faden*

La variable explicativa *lat_creación* es significativa para el modelo con un *p-valor* < 0.001

La relación entre la latitud y el tipo de huracan esta dada por:

A medida que si la latitud aumenta una unidad la probabilidad de que no sea tropical es de **68.87%**, siendo este valor e^{β} .

Mientras que la latitud aumenta una unidad la probabilidad de que el huracan sea tropical son del **31,13%**, siendo este valor el complemento de la interpretación anterior.

```
ggplot(huracanes, aes(x = lat_creacion, y = trop2)) +
  geom_point() +
  #xlab('Latitud') +
  #ylab('Prob Tropical') +
  labs(title = "Probabilidad de que un huracán sea tropical en función de su latitud de creación",
        x = "Latitud",
        y = "Prob tropical") +
  theme_bw() -> p1
#p1

# la interpretacion de b1
betas <- coef(mod)
exp(betas[2])
```

```
## lat_creacion
##      0.6887826
```

```
#La 'CHANCE' de un diagnostico de CHD aumenta 12% x cada año

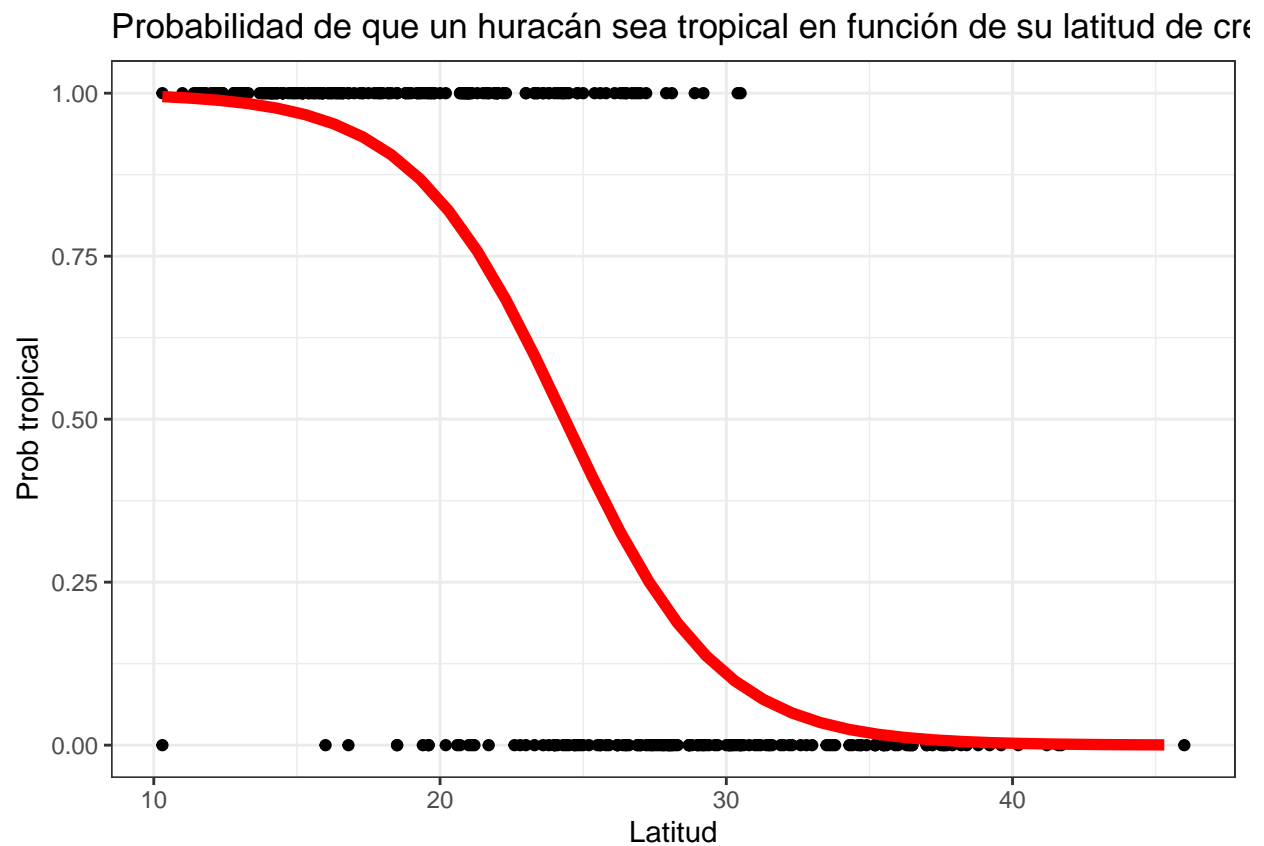
# Agregamos la curva pi(edad) al grafico
xp <- min(huracanes$lat_creacion):max(huracanes$lat_creacion)
b0 <- betas[1]
b1 <- betas[2]
curva <- data.frame(lat_creacion = xp, trop2 = exp(b0 + b1*xp)/(1 + exp(b0 + b1*xp)))

#p1 +
# geom_line(data = curva, col = 'red', size = 2)
```

```
# que habria pasado si hubiesemos ajustado el modelo de RLS

mod2 <- lm(trop2 ~ lat_creacion, data = huracanes)
b0 <- coef(mod2)[1]
b1 <- coef(mod2)[2]
recta <- data.frame(lat_creacion = xp, trop2 = b0 + b1*xp)

# grafiquemos la curva Y la recta
p1 +
  geom_line(data = curva, col = 'red', size = 2)
```



```
# geom_line(data = recta, col = 'blue', size = 2)
```