# Act6 Regresión Poisson C_IACD_Estadística

A01750164 | Paul Martín García Morfín

2022-11-06

## Regresión Poisson

### Punto 1

Trabajaremos con el paquete dataset, que incluye la base de datos warpbreaks, que contiene datos del hilo (yarn) para identificar cuáles variables predictoras afectan la ruptura de urdimbre.

```
data = warpbreaks
head(data,10)
```

```
##    breaks wool tension
## 1      26    A       L
## 2      30    A       L
## 3      54    A       L
## 4      25    A       L
## 5      70    A       L
## 6      52    A       L
## 7      51    A       L
## 8      26    A       L
## 9      67    A       L
## 10     18    A       M
```

Este conjunto de datos indica cuántas roturas de urdimbre ocurrieron para diferentes tipos de telares por telar, por longitud fija de hilo:

- breaks: número de rupturas
- wool: tipo de lana (A o B)
- tension: el nivel de tensión (L, M, H)

### Punto 2

Analiza la base de datos:

- Describe las variables y el número de datos. Describe los valores que toma y qué tipo de variable son.

```
dim(data)
```

```
## [1] 54  3
```

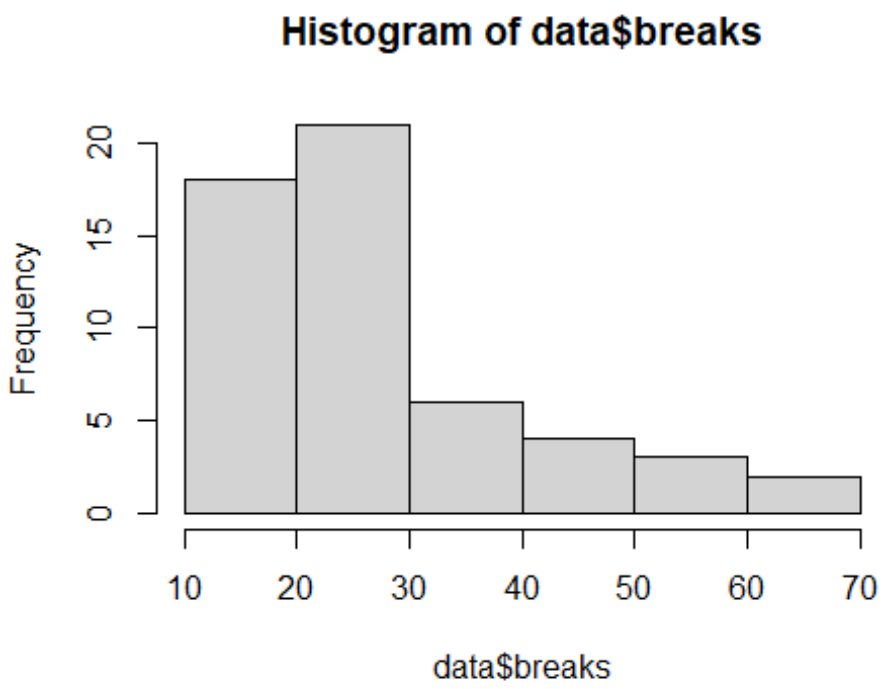
```
library(dplyr);
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

glimpse(data)

## Rows: 54
## Columns: 3
## $ breaks  <dbl> 26, 30, 54, 25, 70, 52, 51, 26, 67, 18, 21, 29, 17,
12, 18, 35…
## $ wool    <fct> A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A, A,
A, A, A,…
## $ tension <fct> L, L, L, L, L, L, L, L, L, M, M, M, M, M, M, M, M, M,
H, H, H,…
```

Finalmente, tenemos 3 variables con 54 observaciones, de las cuales hay:

- 1 variable numérica (dbl): breaks (número de rupturas)
- 2 variables categóricas (fct): wool (tipo de lana: A o B) y tension (nivel de tensión: L, M o H)
- Obtén y analiza el histograma del número de rupturas

```
hist(data$breaks)
```




Podemos observar que los valores más comúnes se encuentran entre 10 y 30 roturas de urdimbre. La dispersión de datos es de los 10 hasta los 70. Parece que los datos muestran una asimetría a la derecha, lo que es un indicio de que no están distribuidos normalmente. Parece que no hay valores atípicos.

- Obtén la media y la varianza del número de rupturas, ¿puedes decir que son iguales o diferentes?

```
mean(data$breaks)

## [1] 28.14815

var(data$breaks)

## [1] 174.2041
```

Podemos observar que la media y la varianza son diferentes.

## Punto 3

Ajusta el modelo de regresión Poisson. Usa el mando:

```
poisson.model = glm(breaks ~ wool + tension, data, family = poisson(link
= "log"))
summary(poisson.model)

##
## Call:
## glm(formula = breaks ~ wool + tension, family = poisson(link = "log"),
```

```
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6871  -1.6503  -0.4269   1.1902   4.2616
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.69196    0.04541  81.302  < 2e-16 ***
## woolB       -0.20599    0.05157  -3.994 6.49e-05 ***
## tensionM    -0.32132    0.06027  -5.332 9.73e-08 ***
## tensionH    -0.51849    0.06396  -8.107 5.21e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 297.37  on 53  degrees of freedom
## Residual deviance: 210.39  on 50  degrees of freedom
## AIC: 493.06
##
## Number of Fisher Scoring iterations: 4
```

- Interpreta la información obtenida. Toma en cuenta que R genera variables Dummy para las variables categóricas. Para cada variable genera k-1 variables Dummy en k categorías (recuerda qué es una variable Dummy).

La desviación residual es la difrencia entre lo que observas y lo que estimas a través del modelo. La variable dependiente es *breaks*, mientras que *wool* y *tension* son variables predictoras (se agregan las variables dummy debido a que *tension* y *wool* son categóricas). Se busca que la desviación residual sea menor a los grados de libertad, lo cual no ocurre en este caso ya que es mayor. Se comparan la media y la varianza de la variable *breaks*, pero no son iguales y el modelo obtenido es malo, ya que nos indica que hay una desviación excesiva (las varianzas no son iguales y la desviación residual es mucho mayor a los grados de libertad)

La desviación nula es la que compara el modelo sin ningún predictor, únicamente con la media general. Tiene que ser mayor que la desviación residual porque ahí sí tenemos predictores (la desviación nula no los tiene) que nos expliquen la variación.

- La desviación residual debe ser menor que los grados de libertad para asegurarse que no exista una dispersión excesiva. Una diferencia mayor, significará que aunque las estimaciones son correctas, los errores estándar son incorrectos y el modelo no lo toma en cuenta.
- La desviación excesiva nula muestra que tan bien se predice la variable de respuesta mediante un modelo que incluye solo el intercepto (gran media). Una diferencia en los valores significa un mal ajuste.
- Si hay un mal modelo, recurre a usar un modelo cuasi Poisson, si los coeficientes son los mismos, el modelo es bueno:

```
poisson.model2 = glm(breaks ~ wool + tension, data = data, family =
quasipoisson(link = "log"))
summary(poisson.model2)
```

```
##
## Call:
## glm(formula = breaks ~ wool + tension, family = quasipoisson(link =
"log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6871  -1.6503  -0.4269   1.1902   4.2616
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.69196    0.09374  39.384  < 2e-16 ***
## woolB       -0.20599    0.10646  -1.935 0.058673 .
## tensionM    -0.32132    0.12441  -2.583 0.012775 *
## tensionH    -0.51849    0.13203  -3.927 0.000264 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 4.261537)
##
##     Null deviance: 297.37  on 53  degrees of freedom
## Residual deviance: 210.39  on 50  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

Cuando la distribución de Poisson no es totalmente correcta, se usa el método de cuasi Poisson. Esto ocurre cuando la media y la varianza no son iguales y por lo tanto el modelo sobre estima o subestima la dispersión de los datos. Al comparar los modelos nos damos cuenta de que los coeficientes son los mismos entonces la conclusión final es que el modelo es bueno.

Nota: EL AIC no se puede comparar porque da NA en el seguno modelo, ya que no es posible calcularlo usando el método de quasipoisson (por algunas suposiciones que se hacen en éste). Se debe quitar uno de los predictores para poder calcularlo.