Regresión de Poisson

Héctor San Román Caraza

6/11/2022

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)
- tensión: el nivel de tensión (L, M, H)

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.

```
summary(data)
```

```
##      breaks      wool   tension
##  Min.   :10.00   A:27   L:18
##  1st Qu.:18.25   B:27   M:18
##  Median :26.00          H:18
##  Mean   :28.15
##  3rd Qu.:34.00
##  Max.   :70.00
```

```
cat("Dimensión de los datos y su tipo: \n")
```

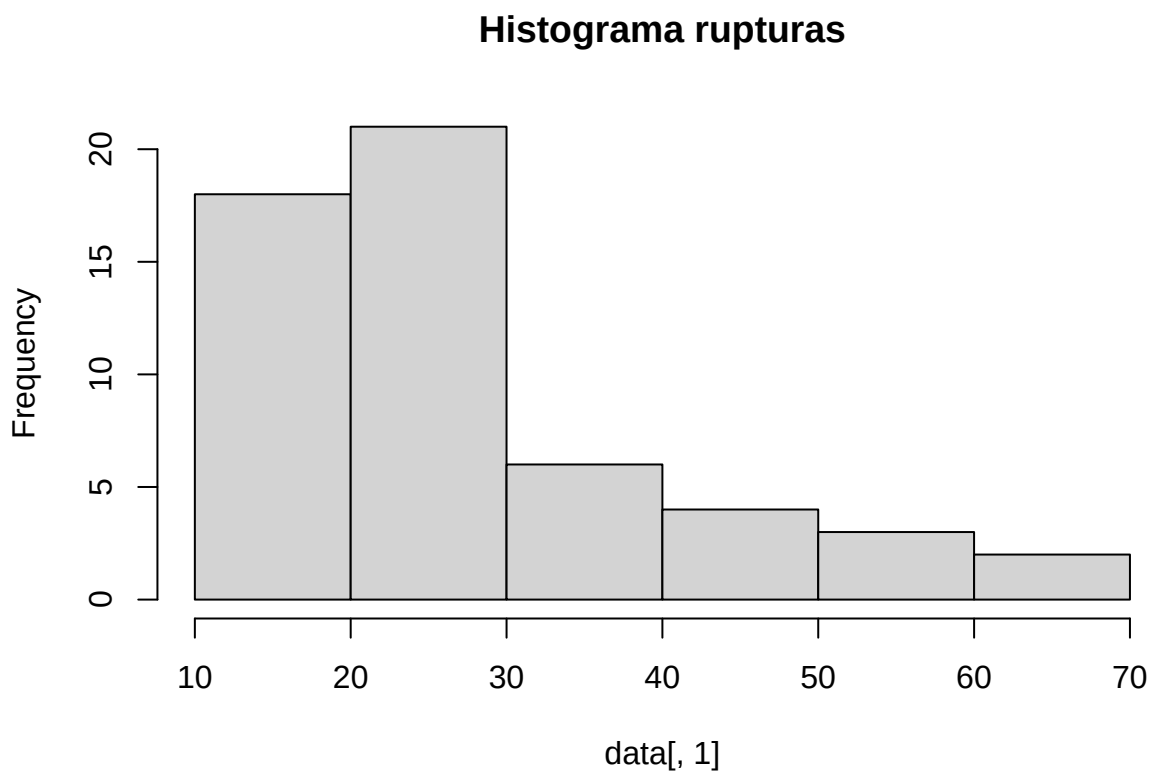
```
## Dimensión de los datos y su tipo:
```

```
str(data)
```

```
## 'data.frame':    54 obs. of  3 variables:
##  $ breaks : num  26 30 54 25 70 52 51 26 67 18 ...
##  $ wool   : Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
##  $ tension: Factor w/ 3 levels "L","M","H": 1 1 1 1 1 1 1 1 2 ...
```

Obtén y analiza el histograma del número de rupturas

```
hist(data[,1], main = "Histograma rupturas")
```



Vemos que encontramos una mayor frecuencia en rupturas de entre 10 a 30, estas hacen la mayoría de los datos.

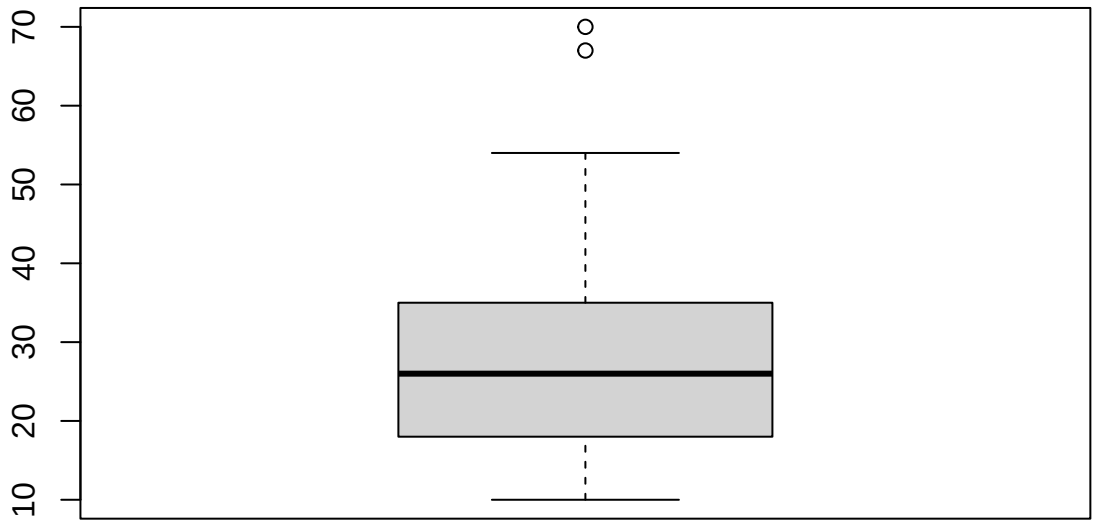
Obtén la media y la varianza del número de rupturas, ¿puedes decir que son iguales o diferentes?

```
cat("Media: " ,mean(data[,1]), "\n", "Varianza: ", var(data[,1]))
```

```
## Media:  28.14815
##  Varianza:  174.2041
```

Vemos que existe una varianza muy alta, hablandonos de que los datos se encuentran muy dispersos. Checaremos ahora los outliers para revisar mejor esta dispersión.

```
boxplot(data[,1])
```



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 ca

- La desviación residual debe ser menor que los grados de libertad para asegurarse que no exista una di

- La desviación excesiva nula muestra que tan bien se predice la variable de respuesta mediante un model

Si hay un mal modelo, recurre a usar un modelo cuasi Poisson, si los coeficientes son los mismos, el mo

```
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
```

Vemos que la desviación residual es mucho mayor a los grados de libertad. Lo que nos indica que las predicciones son correctas; sin embargo, los errores son incorrectos y existe una gran varición en nuestros datos. Podemos ver que con el modelo cuassi poisson los coeficientes son los mismos, ayudandonos aver que el modelo es correcto. La variabilidad de los datos que vimos al sacar su varianza está representada en los modelos.

```
library(arm)
```

```
## Warning: package 'arm' was built under R version 4.1.3
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.13-1, built: 2022-8-25)
```

```
## Working directory is C:/Users/hsrc1/Downloads
```

```
se.coef2 = coef(poisson.model2)
exp(se.coef2)
```

```
## (Intercept)       woolB    tensionM    tensionH
##  40.1235380   0.8138425   0.7251908   0.5954198
```

```
cat("Variación de cambiar tipo de lana A a B: ", 1-0.8138425)
```

```
## Variación de cambiar tipo de lana A a B:  0.1861575
```

Sacamos el exponente para analizar la variación en el tipo de tela que usemos.Vemos que las rupturas bajan un 18% cuando cambias de tipo de tela A a la B.