A6_Regresión Poisson

Eduardo Alvarado Gómez A01251534

2022-11-06

```
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
```

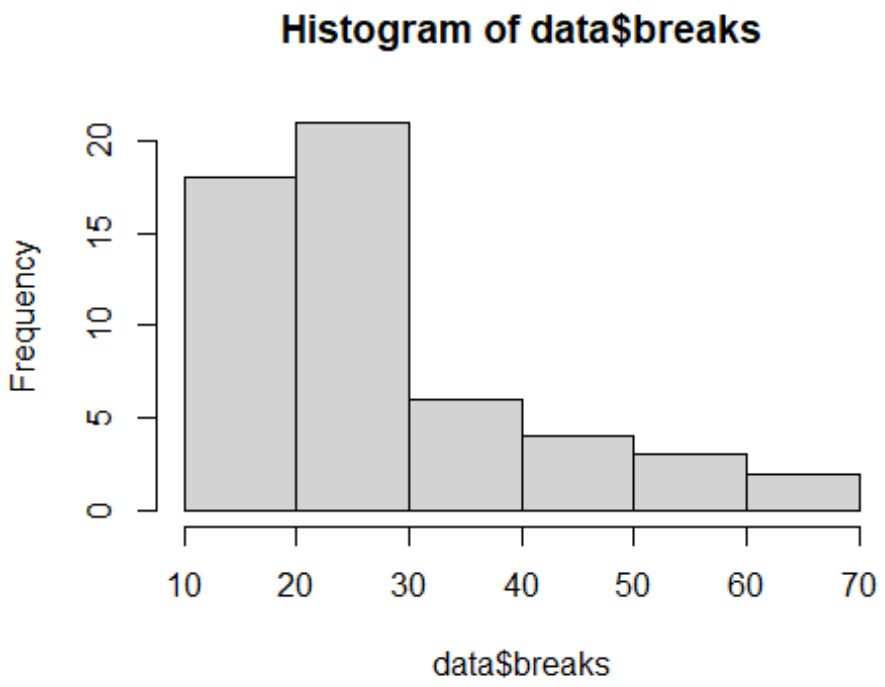
2. Analiza la base de datos:

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

Hay 3 variables diferentes: Breaks es una variable numérica, que indica el número de pausas realizadas, Wool describe el tipo de lana entre A y B y Tension el nivel de tensión en el telar. (pueden ser valores de L (low), M (medium), H (high)). Hay 54 registros.

Obtén y analiza el histograma del número de rupturas

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



Los datos no muestran normalidad.

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
```

Se observa una varianza mucho mayor a la media, es decir alta dispersión.

3. Ajusta el modelo de regresión Poisson. Usa el mando:

```
poisson.model<-glm(breaks ~ wool + tension, data, family = poisson(link = "lo
g"))
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).

Realizando un análisis del modelo, tenemos que según los valores de los coeficientes:

- $\exp(\alpha) =$ efecto sobre la media μ cuando X=0
- $\exp(\beta) =$ incremento en X, la predictora muestra un efecto $\exp(\beta)$ sobre la media de Y.
- Si $\beta = 0$, $\exp(\beta) = 1$ y el valor esperado es $\exp(\alpha)$, por lo que Y y X no están relacionados.
- Si $\beta > 0$, $\exp(\beta) > 1$ y el valor esperado es $\exp(\beta)$ mayor que cuando X=0.
- Si $\beta < 0$, $\exp(\beta) < 1$ y el valor esperado es $\exp(\beta)$ menor que cuando X=0.
- Los valores p son inferiores a 0.05, por lo que ambas variables tienen efecto significativo.

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 se predice la variable de respuesta con un modelo que incluye solo gran media y una diferencia en los valores es un mal ajuste.

Según el análisis y la desviación, disminuyó de 210.39 a 297.37. Una mayor diferencia entre estos valores, indica un mal ajuste del modelo o una deviación muy grande.

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 = quasipoiss
on(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
```

Los coeficientes son los mismos y lo único que cambia son los errores estándar. Debido a que la media y varianza no son iguales, hay sobreestimación.