A6-Regresión Poisson

Edgar Antonio Galarza López A00828688

2022-11-06

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)

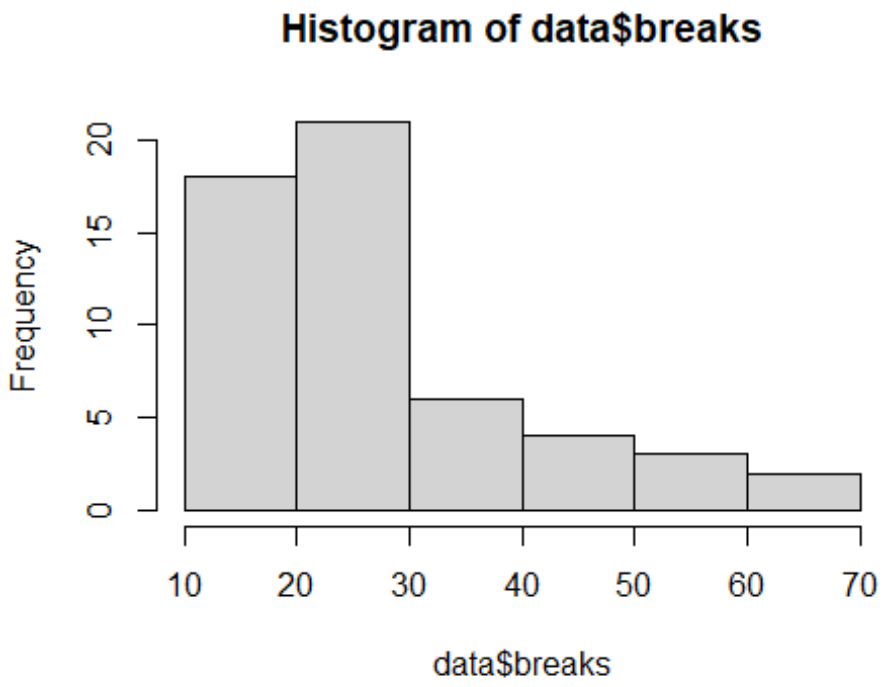
2. Analiza la base de datos:

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

Tenemos que para el conjunto de datos existen 3 variables diferentes. Breaks es una variable numérica que se refiere al número de pausas realizadas. Wool es un factor o variable dummy que describe el tipo de lana entre A y B. Tension de igual forma es una variable dummy que mide el nivel de tensión en el telar y pueden ser valores de L (low), M (medium), H (high). Tenemos un total de 54 entradas dentro del dataset.

Obtén y analiza el histograma del número de rupturas

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



Podemos analizar que los datos no tienen una distribución normal.

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
```

Según los resultados, tenemos una varianza mucho mayor a la media, lo que nos indica que habrá un exceso de dispersión en el modelo.

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).

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

*exp(α) =efecto sobre la media μ cuando $X = 0$

*exp(β) =cada incremento en X, la variable predictora muestra un efecto exp(β) sobre la media de Y, μ.

*Si $\beta = 0$ entonces exp(β) $= 1$ y el valor esperado es exp(α), por lo que Y y X no están relacionados.

*Si $\beta > 0$ entonces exp(β) > 1 y el valor esperado es exp(β) mayor que cuando X=0.

*Si $\beta < 0$ entonces exp(β) < 1 y el valor esperado es exp(β) menor que cuando X=0.

De igual forma podemos ver que todos los valores p son inferiores a 0.05, por lo que ambas variables (wool y tension) tienen efecto significativo sobre las roturas.

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.

Según la desviación obtenida en el análisis, se tiene que disminuyó su valor de 210.39 a 297.37. A medida que existe una mayor diferencia entre estos valores, quiere decir que existe un mal ajuste del modelo o una deviación excesiva.

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
```

Si se analiza ambos modelos, se tiene que los coeficientes son los mismos y lo único que cambia son los errores estándar. Debido a que en este modelo la media y varianza no son iguales, se está sobreestimación.