Regresion Poisson

Andrea Piñeiro

2022-11-07

1. Paquete dataset

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
```

2. Analiza la base de datos

- Describe las variables y el número de datos.

```
cat('Número de columnas ', length(data), '\n')
```

```
## Número de columnas  3
```

```
cat('Numéro de filas ', nrow(data))
```

```
## Numéro de filas  54
```

Como se puede observar, en la base de datos se cuenta con 3 columnas. Y 54 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
```

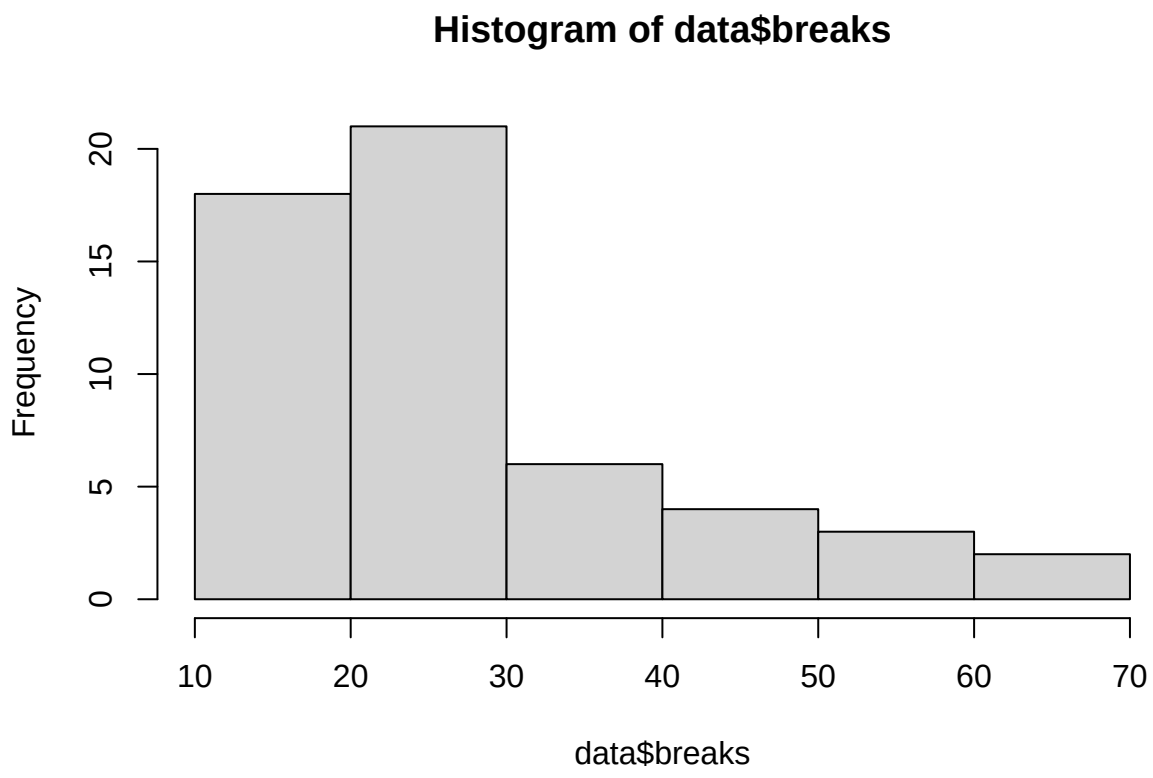
```
ls.str(data)
```

```
## breaks :  num [1:54] 26 30 54 25 70 52 51 26 67 18 ...
## tension :  Factor w/ 3 levels "L","M","H": 1 1 1 1 1 1 1 1 1 2 ...
## wool :  Factor w/ 2 levels "A","B": 1 1 1 1 1 1 1 1 1 1 ...
```

Como se puede observar la variables *breaks* es una variable numércia que tiene una media de 28 y el rango va de 10 a 70. La variable *wool* puede tener dos valores (A y B); y los datos se encuentran distribuidos con estas variables. La mitad tienen A y la otra mitad B. La variable *tension* puede tener uno de 3 valores (L, M, H). Y de igual manera cada tercio de los datos cuenta con cada una de ellas.

- Obtén y analiza el histograma del número de rupturas

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



Como se puede observar los datos de rupturas no tienen una distribución normal. Hay muchos datos en un rango de 10 a 30, y después comienzan a disminuir cada vez más.

Los datos de ruptura son asimétricos hacia la derecha.

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

```
cat('Media: ', mean(data$breaks), '\n')
```

```
## Media:  28.14815
```

```
cat('Varianza: ', var(data$breaks))
```

```
## Varianza:  174.2041
```

La media y la varianza no son iguales. En este caso la varianza es mucho mayor que la media por lo que tendremos sobredispersión en el modelo.

3. Modelo de Regresión Poisson

```
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.

Tomando un *valor de significancia* de 0.05; sabemos que cierta variable influye en el modelo si el valor p es menor al definido anteriormente.

Como se puede observar todas las variables influyen en el modelo, por lo que tanto el tipo de tela como la tención influyen en el modelo.

De igual manera podemos ver por las estrellas a la derecha de cada variable que todas tienen la misma significancia. Algo importante también a notar es que cuando la tensión toma un valor de H el efecto en el número de rupturas es inversamente mayor que el de las demás variables.

- La desviación residual debe ser menor que los grados de libertad.

Como se puede observar la desviación residual que tenemos es de 210.39 en 50 grados de libertad. Como la desviación residual es mayor significa que los errores estándar son incorrectos y no están siendo tomados en cuenta por el modelo.

- La desviación excesiva nula muestra que tan bien se predice la variable de respuesta mediante un modelo que incluye solo el intercepto (gran media).

Como se puede observar entre ambas desviaciones tenemos una diferencia de 3 grados de libertad. Para la desviación nula tenemos un valor de 297.37.

$297.37 - 210.39 = 86.98$

Como podemos ver es una gran diferencia entre ambos valores por lo que hay mal ajuste en el modelo.

- Si hay un mal modelo, recurre a usar un modelo cuasi Poisson.

Para tener un error mejor, se usará el modelo cuasi Poisson.

```
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
```

Con estos resultados podríamos afirmar que el modelo es bueno si los coeficientes son iguales.

```
cmodel1 = coef(poisson.model)
cmodel2 = coef(poisson.model2)

cmodel1
```

```
## (Intercept)       woolB    tensionM    tensionH
##   3.6919631  -0.2059884  -0.3213204  -0.5184885
```

```
cmodel2
```

```
## (Intercept)       woolB    tensionM    tensionH
##   3.6919631  -0.2059884  -0.3213204  -0.5184885
```

Como se observa los coeficientes son iguales por lo que el modelo es bueno.