Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey

Actividad 4. Regresión Poisson

TC3007C.501 Inteligencia artificial avanzada para la ciencia de datos II

Profesores:

Iván Mauricio Amaya Contreras

Blanca Rosa Ruiz Hernández

Félix Ricardo Botello Urrutia

Edgar Covantes Osuna

Felipe Castillo Rendón

Hugo Terashima Marín

Alumno:

Alberto H Orozco Ramos - A00831719

13 de Octubre de 2023

Actividad 4: Regresión Poisson

Instrucciones

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)
```

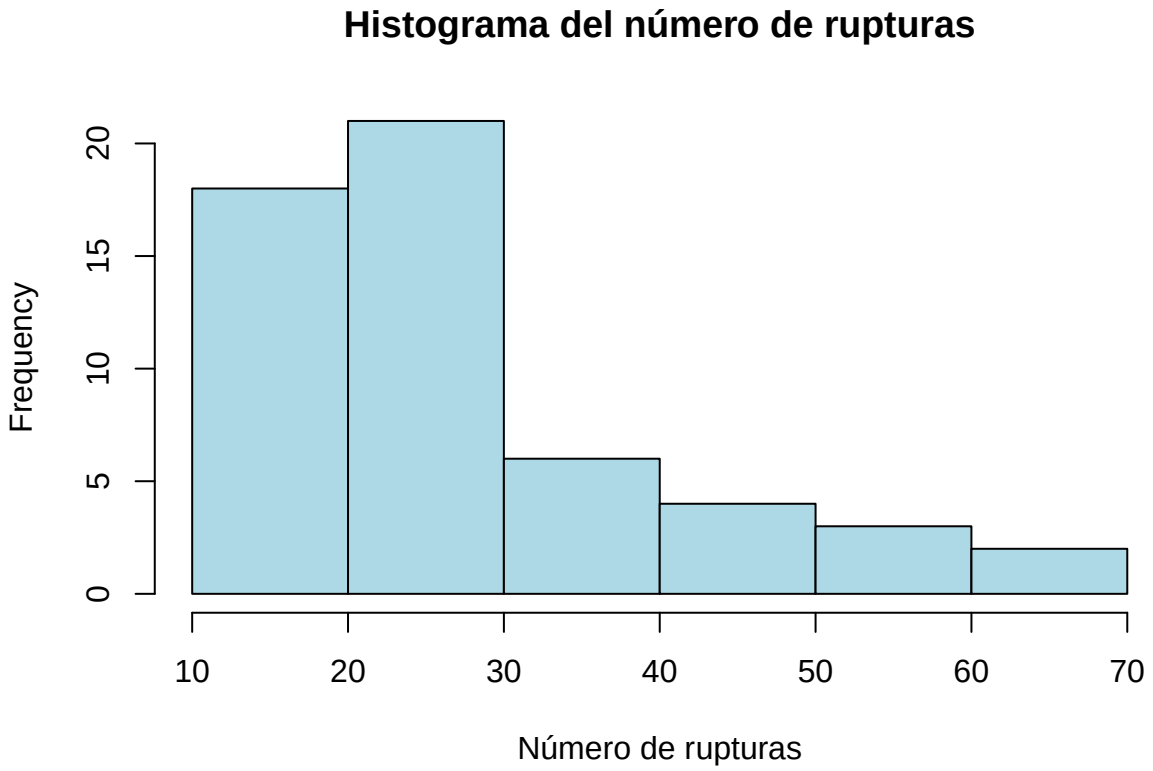
```
##    breaks wool tension
## 1      26    A       L
## 2      30    A       L
## 3      54    A       L
## 4      25    A       L
## 5      70    A       L
## 6      52    A       L
## 7      51    A       L
## 8      26    A       L
## 9      67    A       L
## 10     18    A       M
```

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)

Obtén: * Histograma del número de rupturas

```
hist(data$breaks, main = "Histograma del número de rupturas", xlab = "Número de rupturas", col = "lightb
```



- Obtén la media y la varianza

```
mean_breaks <- mean(data$breaks)
var_breaks <- var(data$breaks)

cat("Media del número de rupturas:", mean_breaks, "\n")
```

```
## Media del número de rupturas: 28.14815
```

```
cat("Varianza del número de rupturas:", var_breaks, "\n")
```

```
## Varianza del número de rupturas: 174.2041
```

- 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)
##
## 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.*

  El coeficiente para **woolB** es negativo (-0.20599), lo que indica que el tipo de lana B se encuentra asociado con menos rupturas en comparación con el tipo de lana A. Además, los coeficientes negativos para 'tensionM' y 'tensionH' indican que niveles más altos de tensión están asociados con menos rupturas.

- *La desviación residual debe ser mayor que los grados de libertad para asegurarse que no exista una dispersión excesiva. Una diferencia menor, significará que aunque las estimaciones son correctas, los errores estándar son incorrectos y el modelo no los toma en cuenta.*

La desviación residual es de 210.39, y tiene 50 grados de libertad. Es menor que los grados de libertad, lo que sugiere que el modelo se encuentra bien ajustado.

- *La desviación excesiva nula muestra que tan bien se predice la variable de respuesta mediante un modelo que incluye solo el intercepto (gran media) mientras que el residual con la inclusión de variables. Una diferencia en los valores significa un mal ajuste.*

La diferencia entre la desviación nula (297.37) y la deviación residual (210.39) indica cuánto mejor se ajusta eñ modelo en comparación con un modelo nulo (solo intercepto). La diferencia entre ambos es de 86.98, lo cual es una diferencia bastante marcada y algo siginificativa para determinar que esta se trata de un buen 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)
##
## 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
```

De los resultados proporcionados por el *summary()* del model Cuasi Poisson, podemos destacar:

1. Intersección: La intersección (3.69196) es el número medio logarítmico de ruptruas cuando todos los predictores se encuentran en sus niveles de referencia.
2. WoolB: El coeficiente (-0.20599) para woolB indica la relación logarítmica del el número medio de rupturas entre la lana B y la referente lana A. El p-value (0.058673) sugiere una significancia marginal.
3. TensionM: El coeficiente (-0.32132) para tensionM indica la relación logarítmica del número medio de rupturas entre la tensión nivel M y la referente tensión nivel L. El p-value (0.012775) es significante.
4. TensionH: El coeficiente (-0.51849) para la tensionH indica la relación logarítmica del número medio de rupturas entre la tension nivel H y el referente de tensión nivel L. El p-valur (0.000264) es altamente significante.
5. El parámetro de dispersión (4.261537) representa la escala de parámetro de la distribución Cuasi-Poisson. Este es mayor que 1, por lo que indica sobredispersión comparado con una distribución poisson ideal.