Actividad 6.- Regresión Poisson

José Carlos Sánchez Gómez

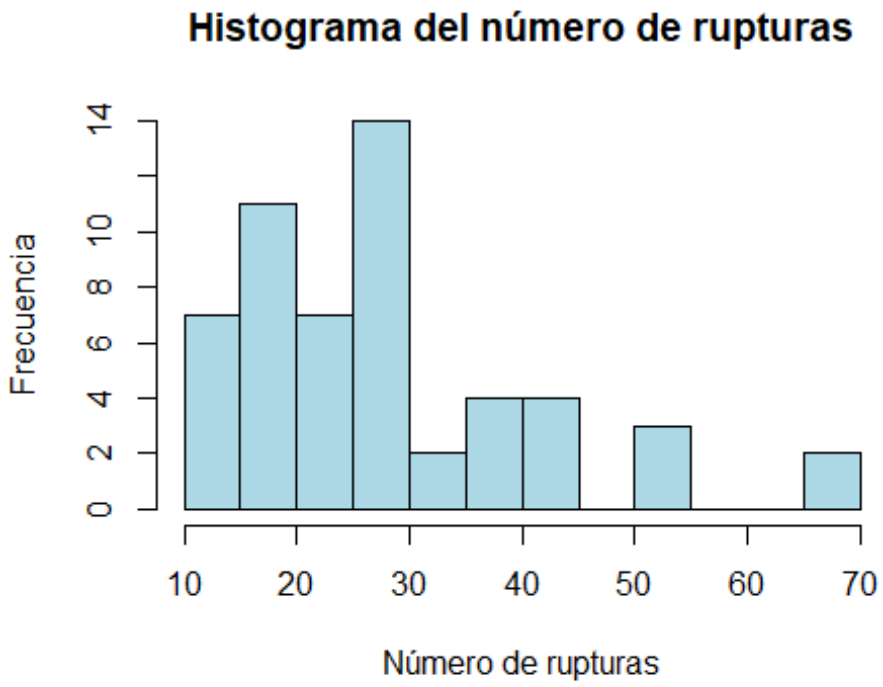
2024-11-04

```
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
```

Análisis Descriptivo

```
hist(data$breaks, main = "Histograma del número de rupturas", xlab =
"Número de rupturas", ylab = "Frecuencia", col = "lightblue", breaks =
10)
```



```r
mean_breaks = mean(data$breaks)
var_breaks = var(data$breaks)
cat("La media de rupturas es:", mean_breaks, "\n")

## La media de rupturas es: 28.14815

cat("La varianza de rupturas es:", var_breaks, "\n")

## La varianza de rupturas es: 174.2041
```

Por lo general, para que un modelo de Poisson sea efectivo, se requiere que los valores de la varianza y de la media sean similares. En este caso no lo son. Por lo que un modelo de Poisson no sea lo más acertado.

Ajuste de modelos de Poisson

Modelo sin interacción

```r
poisson_model1 <- glm(breaks ~ wool + tension, data = data, family =
poisson(link = "log"))
summary(poisson_model1)

##
## 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
```

Modelo:

$$\log(\text{breaks}) = 3.69196 - \text{woolB} \times (-0.2060) - \text{tensionM} \times (-0.3213) - \text{tensionH} \times 0.5185$$

Podemos observar que nuestras variables dummy tienen muy poca significancia, siendo asi la variable de woolB la que más significancia tiene. Este comportamiento no sucede en el modelo con interaccción.

###Modelo con interacción

```
poisson_model2 <- glm(breaks ~ wool * tension, data = data, family =
poisson(link = "log"))
summary(poisson_model2)
```

```
##
## Call:
## glm(formula = breaks ~ wool * tension, family = poisson(link = "log"),
##     data = data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     3.79674    0.04994  76.030  < 2e-16 ***
## woolB          -0.45663    0.08019  -5.694 1.24e-08 ***
## tensionM       -0.61868    0.08440  -7.330 2.30e-13 ***
## tensionH       -0.59580    0.08378  -7.112 1.15e-12 ***
## woolB:tensionM  0.63818    0.12215   5.224 1.75e-07 ***
## woolB:tensionH  0.18836    0.12990   1.450    0.147
## ---
## 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: 182.31  on 48  degrees of freedom
## AIC: 468.97
##
## Number of Fisher Scoring iterations: 4
```

Modelo:

$$\begin{aligned}\log(\text{breaks}) &= 3.79674 - \text{woolB} \times (-0.45663) - \text{tensionM} \times (-0.61886) \\ &- \text{tensionH} \times (-0.59580) + \text{woolB:tensionM} \times 0.63818 + \text{woolB:tensionH} \times 0.18836\end{aligned}$$

A pesar de que las variables de este modelo igual tienen poca significancia, es mejor al anterior, puesto que este posee valores no tan bajos como el anterior. Como es el caso de la interacción entre woolB y tensionH, en dónde obtenemos que el valor p de la interacción es de 0.147. ## Selección del modelo

Desviación residual

```
S1 = summary(poisson_model1)
S2 = summary(poisson_model2)

gl1 = S1$df.null - S1$df.residual
gl2 = S2$df.null - S2$df.residual

cat("Modelo Sin Interacción", "\n")
```

```
## Modelo Sin Interacción
```

```r
# Valor frontera para el modelo sin interacción
cat("Valor frontera:", qchisq(0.05, gl1), "\n")
```

```
## Valor frontera: 0.3518463
```

```r
# Estadístico de prueba y valor p para el modelo sin interacción
dr1 = S1$deviance
vp1 = 1 - pchisq(dr1, gl1)
cat("Estadístico de prueba: ", dr1, "\n")
```

```
## Estadístico de prueba:  210.3919
```

```r
cat("Valor p: ", vp1, "\n", "\n")
```

```
## Valor p:  0
##
```

```r
# Valor frontera para el modelo con interacción
cat("Modelo Con Interacción", "\n")
```

```
## Modelo Con Interacción
```

```r
cat("Valor frontera:", qchisq(0.05, gl2), "\n")
```

```
## Valor frontera: 1.145476
```

```r
# Estadístico de prueba y valor p para el modelo con interacción
dr2 = S2$deviance
vp2 = 1 - pchisq(dr2, gl2)
cat("Estadístico de prueba: ", dr2, "\n")
```

```
## Estadístico de prueba:  182.3051
```

```r
cat("Valor p: ", vp2, "\n")
```

```
## Valor p:  0
```

De estos datos entendemos que el modelo sin interacción tiene una mejor variabilidad en sus datos, puesto que su valor de deviance es mayor. Un valor mayor en este rubro indica mayor variabilidad.

AIC: Criterio Aike

```r
aic_no_interaccion = AIC(poisson_model1)
aic_interaccion = AIC(poisson_model2)

cat("AIC Modelo Sin Interacción: ", aic_no_interaccion, "\n")
```

```
## AIC Modelo Sin Interacción:  493.056
```

```r
cat("AIC Modelo Con Interacción: ", aic_interaccion, "\n")
```

```
## AIC Modelo Con Interacción:  468.9692
```

Ambos cuentan con un valor de AIC muy elevado, sin embargo, el que posee un valor menor es el modelo con interacción. Suponiendo que es un mejor modelo que el de sin interacción.

Comparación de coeficientes

```
coef_model_1 = data.frame(
  Variable = rownames(summary(poisson_model1)$coefficients),
  Coeficiente = summary(poisson_model1)$coefficients[, "Estimate"]
)
print(coef_model_1)
```

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

```
coef_model_2 = data.frame(
  Variable = rownames(summary(poisson_model2)$coefficients),
  Coeficiente = summary(poisson_model2)$coefficients[, "Estimate"]
)

print(coef_model_2)
```

```
##                      Variable Coeficiente
## (Intercept)       (Intercept)   3.7967368
## woolB                   woolB  -0.4566272
## tensionM             tensionM  -0.6186830
## tensionH             tensionH  -0.5957987
## woolB:tensionM woolB:tensionM   0.6381768
## woolB:tensionH woolB:tensionH   0.1883632
```

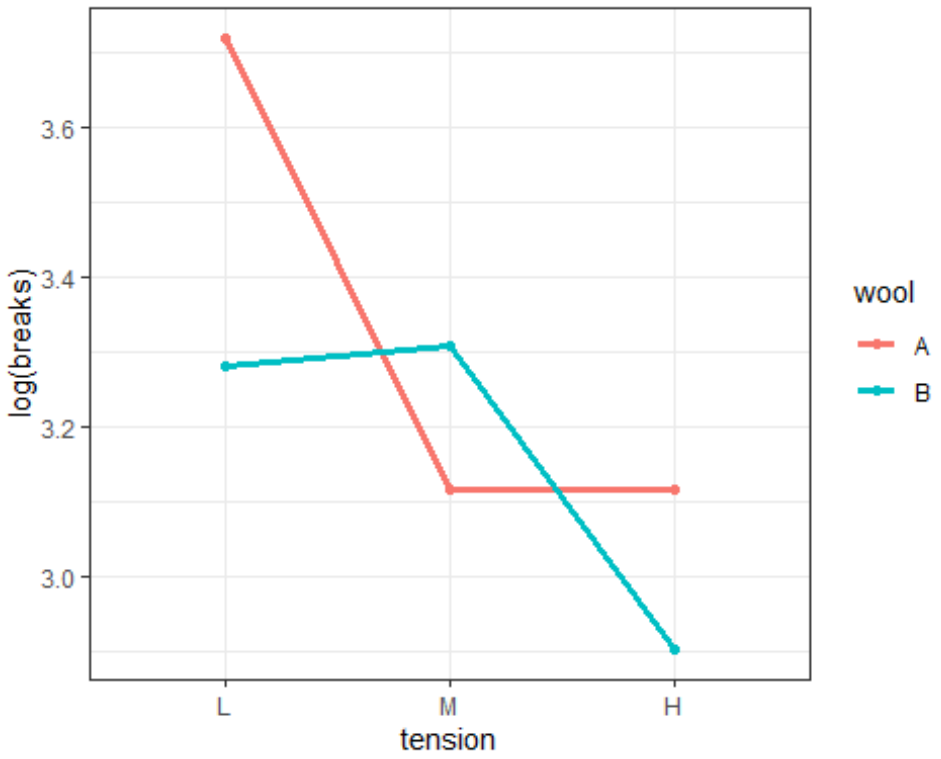
```
error_model_1 = data.frame(
  Variable = rownames(summary(poisson_model1)$coefficients),
  Error = summary(poisson_model1)$coefficients[, "Std. Error"]
)

error_model_2 = data.frame(
  Variable = rownames(summary(poisson_model2)$coefficients),
  Error = summary(poisson_model2)$coefficients[, "Std. Error"]
)

print(error_model_1)
```

```
##                Variable      Error
## (Intercept) (Intercept) 0.04541069
## woolB             woolB 0.05157117
## tensionM       tensionM 0.06026580
## tensionH       tensionH 0.06395944
```

```
print(error_model_2)
```

```
##                       Variable      Error
## (Intercept)       (Intercept) 0.04993753
## woolB                   woolB 0.08019202
## tensionM             tensionM 0.08440012
## tensionH             tensionH 0.08377723
## woolB:tensionM woolB:tensionM 0.12215312
## woolB:tensionH woolB:tensionH 0.12989529
```

```
library(ggplot2)
ggplot(data, aes(x = tension, y = log(breaks), group = wool, color =
wool)) +
  stat_summary(fun = mean, geom = "point") +
  stat_summary(fun = mean, geom = "line",lwd=1.1) +
  theme_bw() +
  theme(panel.border = element_rect(fil="transparent"))
```



Dentro de los dos modelos, considero que el mejor es el modelo con interacción. La razón de esto es que en la prueba de AIC tuvo un mejor puntaje que su contrincante. En las demás pruebas se desempeñaron bastante similar. En la prueba de desviación residual, el modelo con interacción lo supero por unos puntos, sin embargo, esto no significa que el modelo con interacción no cuente con una variabilidad en sus datos. Además de que en los coeficientes que ambos compartian, tenian un valor similar de error. Dado que en la única prueba que ganó fue la de AIC, considero que el mejor modelo es el de con interacción. Recomiendo buscar un mejor modelo que se acople a los datos.

Evaluación de los supuestos

```
library("lmtest")
```

```
## Cargando paquete requerido: zoo
```

```
##
## Adjuntando el paquete: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
dwtest(poisson_model2)
```

```
##
##  Durbin-Watson test
##
## data:  poisson_model2
## DW = 2.2376, p-value = 0.575
## alternative hypothesis: true autocorrelation is greater than 0
```

Al tener un valor p muy elevado, podemos decir que el modelo con interacción no cuenta con independencia.

```
library(epiDisplay)
```

```
## Cargando paquete requerido: foreign
```

```
## Cargando paquete requerido: survival
```

```
## Cargando paquete requerido: MASS
```

```
## Cargando paquete requerido: nnet
```

```
##
## Adjuntando el paquete: 'epiDisplay'
```

```
## The following object is masked from 'package:lmtest':
##
##     lrtest
```

```
## The following object is masked from 'package:ggplot2':
##
##     alpha
```

```
poisgof(poisson_model2)
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 182.3051
##
```

```
## $df
## [1] 48
##
## $p.value
## [1] 1.582538e-17
```

Dado al valor bajo de p, hay evidencia de una sobredispersion dentro del modelo. Y lo que indica que el modelo de Poisson no se ajusta bien a los datos. Recomiendo buscar otro tipo de modelo.

Intento con otro modelo de poisson

```
poisson_model3 = glm(breaks ~ wool * tension, data = data, family =
quasipoisson(link = "log"))
summary(poisson_model3)

##
## Call:
## glm(formula = breaks ~ wool * tension, family = quasipoisson(link =
"log"),
##     data = data)
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.79674    0.09688  39.189  < 2e-16 ***
## woolB          -0.45663    0.15558  -2.935 0.005105 **
## tensionM       -0.61868    0.16374  -3.778 0.000436 ***
## tensionH       -0.59580    0.16253  -3.666 0.000616 ***
## woolB:tensionM  0.63818    0.23699   2.693 0.009727 **
## woolB:tensionH  0.18836    0.25201   0.747 0.458436
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 3.76389)
##
##     Null deviance: 297.37  on 53  degrees of freedom
## Residual deviance: 182.31  on 48  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

A primera vista parece ser que este modelo es superior a los anteriores, puesto que demuestra tener una mayor significancia en sus coeficientes, a diferencia de los anteriores. Esto se puede observar con el incremento en el valor p de cada coeficiente.

Pruebas

```
S3 = summary(poisson_model3)

gl3 = S3$df.null - S3$df.residual

cat("Modelo Con Interacción 2", "\n")
```

```r
## Modelo Con Interacción 2

cat("Valor frontera:", qchisq(0.05, gl3), "\n")

## Valor frontera: 1.145476

dr3 = S3$deviance
vp3 = 1 - pchisq(dr3, gl3)
cat("Estadístico de prueba: ", dr3, "\n")

## Estadístico de prueba:  182.3051

cat("Valor p: ", vp3, "\n")

## Valor p:  0
```

Este modelo cuenta con un buen valor estadistico (deviance) lo cual indica que existe una mayor variabilidad en sus datos.

AIC: Criterio Aike

```r
aic_modelo_3 = AIC(poisson_model3)
cat("AIC Modelo Con Interacción 2: ", aic_modelo_3, "\n")

## AIC Modelo Con Interacción 2:  NA
```

Debido a que estamos usando un modelo que ocupa la familia Quasipoisson, no es posible obtener el resultado de AIC, ya que este no esta definido para modelos ajustados con esta familia.

Comparación de coeficientes

```r
coef_model_3 = data.frame(
  Variable = rownames(summary(poisson_model3)$coefficients),
  Coeficiente = summary(poisson_model3)$coefficients[, "Estimate"]
)
print(coef_model_3)

##                       Variable Coeficiente
## (Intercept)       (Intercept)   3.7967368
## woolB                   woolB  -0.4566272
## tensionM             tensionM  -0.6186830
## tensionH             tensionH  -0.5957987
## woolB:tensionM woolB:tensionM   0.6381768
## woolB:tensionH woolB:tensionH   0.1883632

error_model_3 = data.frame(
  Variable = rownames(summary(poisson_model3)$coefficients),
  Error = summary(poisson_model3)$coefficients[, "Std. Error"]
)

print(error_model_3)
```

```
##                       Variable      Error
## (Intercept)       (Intercept) 0.09688254
## woolB                   woolB 0.15557852
## tensionM             tensionM 0.16374255
## tensionH             tensionH 0.16253410
## woolB:tensionM woolB:tensionM 0.23698620
## woolB:tensionH woolB:tensionH 0.25200659
```

Conclusión

Finalmente, después de analizar los primeros dos modelos, compararlos, y ponerlos a prueba. Decidi que no eran modelos optimos para como estan distribuidos sus datos. Los resultados que obtuvieron en las pruebas no fueron satisfactorios, por lo que opte a hacer otro modelo. Este último se desempeño mejor que sus predecesores. Sus valores p de cada coeficiente eran mayor, lo que representaban una mayor significancia en cada variable. De igual manera, los errores de estos mismo son bajos. En las otras pruebas se desempeño de buena manera, logrando tener una buena desviación residual. La función final otorgada por el último modelo es esta:

Modelo:

$$\begin{aligned}\log(\text{breaks}) \\ = 3.79674 - \text{woolB} \times 0.45663 - \text{tensionM} \times 0.61886 - \text{tensionH} \times 0.59580 \\ + \text{woolB:tensionM} \times 0.63818 + \text{woolB:tensionH} \times 0.18836\end{aligned}$$