# A6-Regresión Poisson

Facundo Colasurdo Caldironi

2024-10-29

## Regresión Poisson

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) Sigue el siguiente procedimiento de análisis:

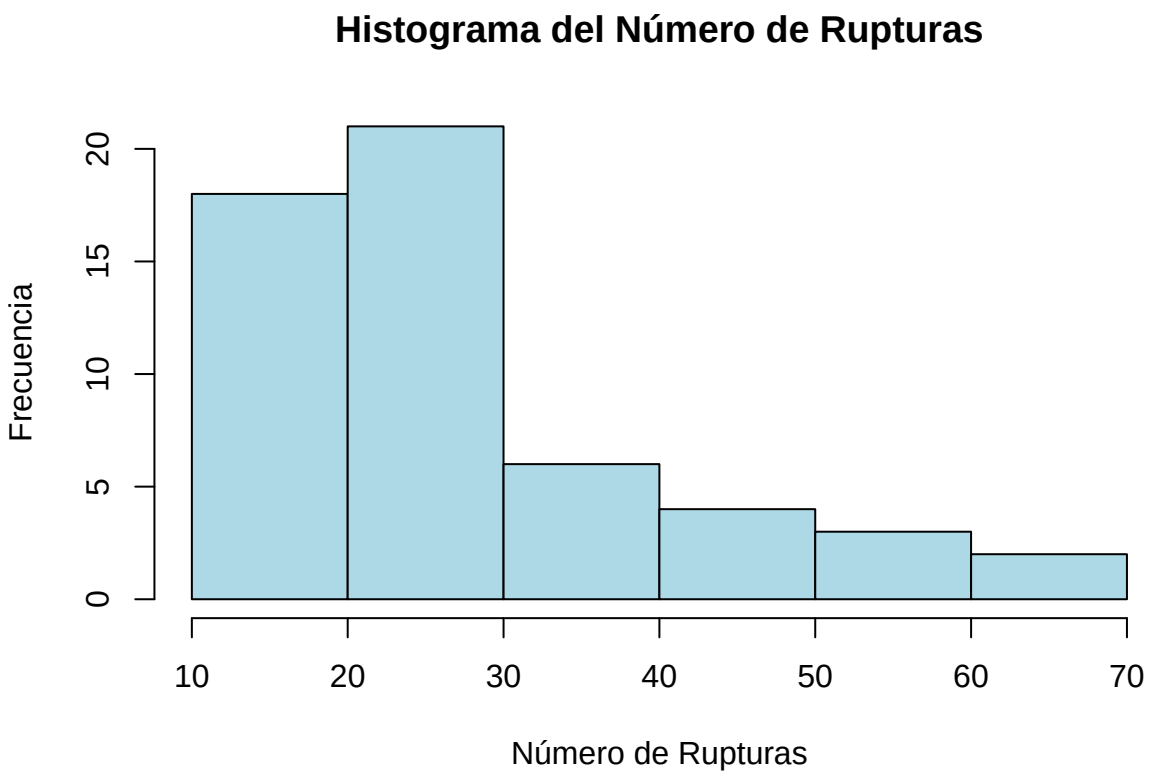
```
data<-warpbreaks
head(data,10)
```

```
##    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
```

##I. Análisis Descriptivo

Histograma del número de rupturas Obtén la media y la varianza de la variable dependiente Interpreta en el contexto de una Regresión Poisson

```
hist(data$breaks, main = "Histograma del Número de Rupturas", xlab = "Número de Rupturas", ylab = "Frecu
```




```
promedioRuptura <- mean(data$breaks)
varianzaRuptura <- var(data$breaks)
promedioRuptura
```

```
## [1] 28.14815
```

```
varianzaRuptura
```

```
## [1] 174.2041
```

Podemos ver como la varianza de 174.20 nos refleja la dispersion de los valores de numeros de ruptura, dentro de un modelo poisson, esta y la media deberian ser ser casi iguales, mas aqui podemos ver como la media 28.14 es significativamente menor

##II. Ajusta dos modelos de Regresión Poisson

Ajusta el modelo de regresión Poisson sin interacción Ajusta el modelo de regresión Poisson con interacción Usa los comandos: poisson_model<-glm(breaks ~ wool + tension, data, family = poisson(link = “log”)) S=summary(poisson_model) Interpreta los coeficientes de las variables Dummy. Escribe el modelo obtenido. Toma en cuenta que R genera variables Dummy para las variables categóricas. Para cada variable genera k-1 variables Dummy en k categorías.

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

```
##
## 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
```

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

```
##
## 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
```

Interpretacion de datos: Al analizar los valores, pudimos observar que la lana con el mejor resultado es la B, esto debido a que general, esta tiene un numero de rupturas menor a comparacion de la Lana A, la cual tiene una mayor cantidad de rupturas a tension baja, aunque tiene un numero de rupturas parecidas en tensiones medias y altas.

##III. Selección del modelo

Para seleccionar el modelo se toma en cuenta: Desviación residual: es la suma del cuadrado de los residuos estandarizados que se obtienen bajo el modelo. Con los grados de libertad se realiza una prueba de xˆ2 para significancia del modelo. AIC: Criterio de Aikaike Comparación entre los coeficientes y los errores estándar de de ambos modelos Desviación residual (Prueba de xˆ2) Si el modelo nulo explica a los datos, entonces la desviación nula será pequeña. Lo mismo ocurre con la Desviación residual . Puesto que es de suponer que el modelo contiene variables significativas, lo que importa que es la desviación residual del modelo sea suficientemente pequeño. La prueba de mide qué tan lejano está del cero la desviación residual del modelo. Entre más lejos esté del cero, el modelo será un buen modelo, entre más cerca, el modelo será un mal modelo que explicará poco la variabilidad de los datos. Su modelo supone: H0: Deviance = 0 H1: Deviance > 0 gl = gl_desviación residual (n-(p+1)) Usa los siguientes comandos: Valor frontera de la zona de rechazo (S es la variable que denota el summary del modelo): gl = S$null.deviance - S$df.residual qchisq(0.05,gl) Estadístico de prueba y valor p: dr = S$deviance cat("Estadístico de prueba =",dr) vp = 1-pchisq(dr,gl) cat("Valor p =",vp) Compara los AIC de cada modelo. Recuerda que un menor AIC indica un mejor modelo. Compara los coeficientes Compara los coeficientes de ambos modelos (haz una tabla para que se facilite la comparación) Compara el error estándar de cada estimador de de ambos modelos (haz una tabla para que se facilite la comparación) Interpreta los coeficientes de ambos modelos. Para interpretar mejor la interacción gráficala con el siguiente código: 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")) Define cuál de los dos es un mejor modelo.

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

gl = S$df.null-S$df.residual
cat("Grados de libertad =", gl, "\n")
```

```
## Grados de libertad = 3
```

```
valor_frontera <- qchisq(0.05, gl)
cat("Valor frontera de la zona de rechazo =", valor_frontera, "\n")
```

```
## Valor frontera de la zona de rechazo = 0.3518463
```

```
dr <- S$deviance
cat("Estadístico de prueba =", dr, "\n")
```

```
## Estadístico de prueba = 210.3919
```

```
vp <- 1 - pchisq(dr, gl)
cat("Valor p =", vp, "\n")
```

```
## Valor p = 0
```

```
AIC_sin_interaccion <- AIC(modeloPoisson)
cat("AIC del modelo sin interacción =", AIC_sin_interaccion, "\n")
```

```
## AIC del modelo sin interacción = 493.056
```

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

AIC_con_interaccion <- AIC(modeloPoissonInteraccion)
cat("AIC del modelo con interacción =", AIC_con_interaccion, "\n")
```

```
## AIC del modelo con interacción = 468.9692
```

```r
cat("Resumen del modelo sin interacción:\n")
```

```
## Resumen del modelo sin interacción:
```

```r
print(S)
```

```
##
## 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
```

```r
cat("Resumen del modelo con interacción:\n")
```

```
## Resumen del modelo con interacción:
```

```r
print(S_inter)
```

```
##
## 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
```

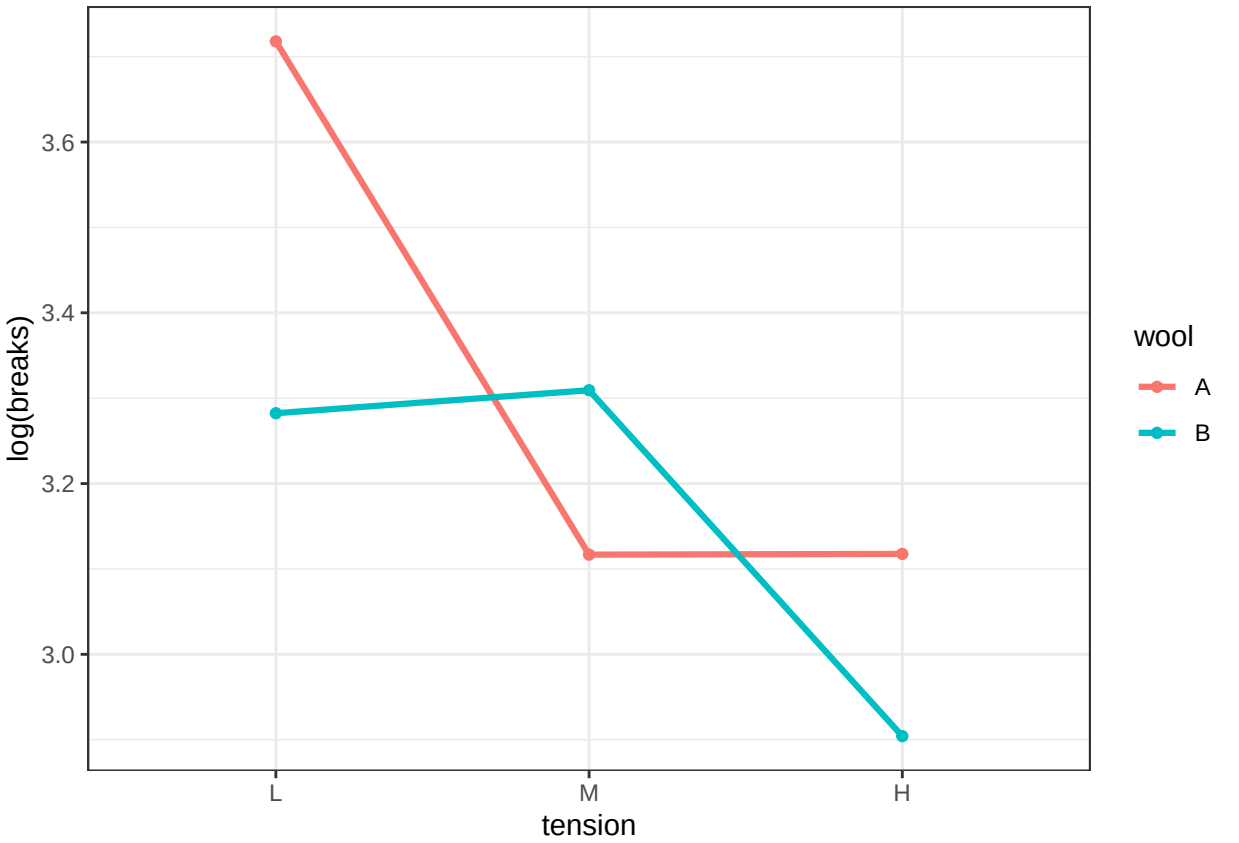
Al comparar el AIC, podemos ver que el modelo con interaccion tiene un menor AIC, este siendo de 468.97 a comparacion del Aic del que no tiene interaccion, el cual da un valor mas alto de 493.06

Compara los coeficientes de ambos modelos Al comprarar los dos modelos, fue posible ver que wool B, tension M y tension H son los que tienen los coeficientes mas grandes en con interaccion.

Compara el error estándar de cada estimador de Bi de ambos modelos Los errores estándar en el modelo con interacción son generalmente mayores, esto debido a que busca ajustar un mayor numero de datos posibles

```r
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(fill = "transparent"))
```




Define cuál de los dos es un mejor modelo. De los dos modelos podemos que el modelo con interaccion es el que mejor logro adapatarse esto debido a que tiene los mayores coeficientes de ambos modelos, al mismo tiempo que tiene el menor AIC, a comparacion de la Sin interaccion AIC, lo cual nos respalda que tiene un mejor resultado.

##IV. Evaluación de los supuestos

Los supuestos principales que se deben cumplir son:

Independencia: haz la misma prueba de independencia que usaste en los modelos lineales. Sobredispersión de los residuos. La sobredispersión de los residuos indicará que el modelo no cumple con el supuesto de que la media es igual a la varianza de los residuos. Para probarla se usa la prueba posgof, que es una prueba con gl = grados de libertad residual. La desviación estándar se compara con los grados de libertad de la desviación residual, no deben ser muy diferentes. Esto indicará una sobredispersión de los residuos: H0: No hay una sobredispersión del modelo H1: Hay una sobredispersión del modelo Usa el comando: library(epiDisplay) poisgof(pm) Si hay un mal modelo, recurre a usar: Modelo cuasi Poisson: poisson.model3<-glm(breaks ~ wool + tension, data = data, family = quasipoisson(link = "log")) summary(poisson.model2) Modelo Binomial Negativa (intenta imaginar qué es lo que cambia en este modelo con respecto al Poisson): bnm = model.nb = glm.nb(breaks ~ wool * tension, data, control = glm.control(maxit=1000)) summary(bnm) Define si usas defines tus modelos con interacción o sin interacción (no hagas los dos) Define el mejor modelo usando las mismas pruebas y crtierios que usaste en los modelos Poisson

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
dwtest(modeloPoissonInteraccion)
```

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

```
library(epiDisplay)
```

```
## Loading required package: foreign

## Loading required package: survival

## Loading required package: MASS

## Loading required package: nnet

##
## Attaching package: 'epiDisplay'

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

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

```
poisgof(modeloPoissonInteraccion)
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 182.3051
##
## $df
## [1] 48
##
## $p.value
## [1] 1.582538e-17
```

```r
#Modelo cuasi 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)  3.69196    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
```

```r
#Modelo Binomial Negativa (intenta imaginar qué es lo que cambia en este modelo con respecto al Poisson
bnm = model.nb = glm.nb(breaks ~ wool * tension, data, control = glm.control(maxit=1000))
summary(bnm)
```

```
##
## Call:
## glm.nb(formula = breaks ~ wool * tension, data = data, control = glm.control(maxit = 1000),
##     init.theta = 12.08216462, link = log)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      3.7967     0.1081  35.116  < 2e-16 ***
## woolB           -0.4566     0.1576  -2.898 0.003753 **
## tensionM        -0.6187     0.1597  -3.873 0.000107 ***
## tensionH        -0.5958     0.1594  -3.738 0.000186 ***
## woolB:tensionM   0.6382     0.2274   2.807 0.005008 **
## woolB:tensionH   0.1884     0.2316   0.813 0.416123
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(12.0822) family taken to be 1)
##
##     Null deviance: 86.759  on 53  degrees of freedom
## Residual deviance: 53.506  on 48  degrees of freedom
## AIC: 405.12
##
```

```
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  12.08
##           Std. Err.:  3.30
##
##  2 x log-likelihood:  -391.125
```

```
library(lmtest)
dwtest(poisson.model3)
```

```
##
##  Durbin-Watson test
##
## data:  poisson.model3
## DW = 2.0332, p-value = 0.3896
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(bnm)
```

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

##V. Define cuál es tu mejor modelo

El mejor modelo fue el Modelo Binomial Negativa, esto debido a que originalmente, el modelo poison con interaccion fallaba ya que nos daba la hipotesis uno la cual nos dice que existe una sobredispersión del modelo, por eso tuvimos que probar con otros metodos, y de estos, fue el Modelo Binomial Negativa el cual fue el tomo en mas cuenta la significancia de los coeficientes y un menor valor de P.