M5_A6

Sofia Cantu

2024-11-04

Regresión Poisson

```
if (!require(ggplot2)) install.packages("ggplot2")
```

```
## Loading required package: ggplot2
```

```
library(ggplot2)
if (!require(epiDisplay)) install.packages("epiDisplay")
```

```
## Loading required package: 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:ggplot2':
##
##     alpha
```

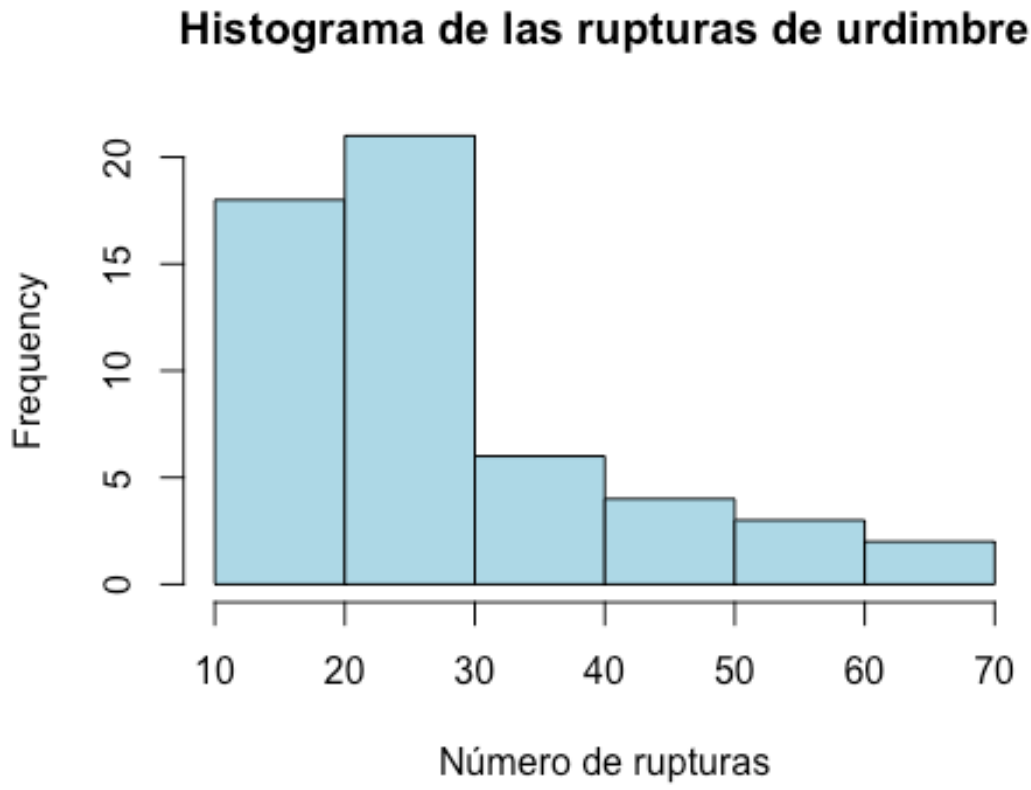
```
library(epiDisplay)
if (!require(MASS)) install.packages("MASS")
library(MASS)

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

```r
# 1.1 Histograma del número de rupturas
hist(data$breaks,
     main = "Histograma de las rupturas de urdimbre",
     xlab = "Número de rupturas",
     col = "lightblue",
     border = "black")
```




```r
# 1.2 Media y varianza de la variable dependiente (breaks)
mean_breaks <- mean(data$breaks)
var_breaks <- var(data$breaks)
cat("Media de las rupturas: ", mean_breaks, "\n")

## Media de las rupturas:  28.14815

cat("Varianza de las rupturas: ", var_breaks, "\n")

## Varianza de las rupturas:  174.2041

# 1.3 Interpretación en el contexto de una Regresión Poisson
cat("\nInterpretación:\n")
```

```
##
## Interpretación:

cat("La media de las rupturas es un indicador clave en una Regresión Poisson, ya que este tipo de regresión es adecuada para modelar variables de conteo. En el contexto del histograma de las rupturas de urdimbre, se observa que la mayoría de los eventos de ruptura se concentran entre 10 y 30, con una disminución progresiva a medida que el número de rupturas aumenta. \n")

## La media de las rupturas es un indicador clave en una Regresión Poisson, ya que este tipo de regresión es adecuada para modelar variables de conteo. En el contexto del histograma de las rupturas de urdimbre, se observa que la mayoría de los eventos de ruptura se concentran entre 10 y 30, con una disminución progresiva a medida que el número de rupturas aumenta.

cat("Este patrón sugiere una distribución sesgada hacia la izquierda, lo cual es común en datos de conteo, donde los eventos más bajos en frecuencia son más comunes que los valores extremos. Un modelo de Regresión Poisson asume que la media es aproximadamente igual a la varianza. Sin embargo, si la varianza de los datos es significativamente mayor que la media, esto indicaría la presencia de sobredispersión. La sobredispersión puede afectar la bondad de ajuste del modelo y provocar errores de estimación. \n")

## Este patrón sugiere una distribución sesgada hacia la izquierda, lo cual es común en datos de conteo, donde los eventos más bajos en frecuencia son más comunes que los valores extremos. Un modelo de Regresión Poisson asume que la media es aproximadamente igual a la varianza. Sin embargo, si la varianza de los datos es significativamente mayor que la media, esto indicaría la presencia de sobredispersión. La sobredispersión puede afectar la bondad de ajuste del modelo y provocar errores de estimación.
```

II. Ajuste de modelos de Regresión Poisson

```
# 2.1 Modelo de regresión Poisson sin interacción
poisson_model <- glm(breaks ~ wool + tension, data = data, family = poisson(link = "log"))
S <- summary(poisson_model)
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
```

```
# 2.2 Modelo de regresión Poisson con interacción
poisson_model_interaction <- glm(breaks ~ wool * tension, data = data, family
= poisson(link = "log"))
S_interaction <- summary(poisson_model_interaction)
print(S_interaction)
```

```
##
## 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
```

```
# 2.3 Interpretación de los coeficientes de las variables Dummy
cat("\nInterpretación de los coeficientes:\n")
```

```
##
## Interpretación de los coeficientes:
```

```
cat("Los coeficientes obtenidos en el modelo de regresión Poisson representan
el logaritmo natural del cambio esperado en el número de rupturas cuando la v
ariable predictora cambia, manteniendo las demás constantes. Las variables Du
mmy generadas por R representan comparaciones con la categoría de referencia
```

```
para cada variable categórica. Para las variables wool y tension, se generan
k-1 variables Dummy, donde k es el número de categorías.\n")

## Los coeficientes obtenidos en el modelo de regresión Poisson representan e
l logaritmo natural del cambio esperado en el número de rupturas cuando la va
riable predictora cambia, manteniendo las demás constantes. Las variables Dum
my generadas por R representan comparaciones con la categoría de referencia p
ara cada variable categórica. Para las variables wool y tension, se generan k
-1 variables Dummy, donde k es el número de categorías.

# Modelo sin interacción obtenido
cat("\nModelo sin interacción:\n")

##
## Modelo sin interacción:

cat("breaks = exp(\u03b20 + \u03b21*woolB + \u03b22*tensionM + \u03b23*tensio
nH)\n")

## breaks = exp(β0 + β1*woolB + β2*tensionM + β3*tensionH)

# Modelo con interacción obtenido
cat("\nModelo con interacción:\n")

##
## Modelo con interacción:

cat("breaks = exp(\u03b20 + \u03b21*woolB + \u03b22*tensionM + \u03b23*tensio
nH + \u03b24*woolB*tensionM + \u03b25*woolB*tensionH)\n")

## breaks = exp(β0 + β1*woolB + β2*tensionM + β3*tensionH + β4*woolB*tensionM
+ β5*woolB*tensionH)
```

III. Selección del modelo

```
# 3.1 Calcular y comparar la desviación residual y realizar la prueba de chi-
cuadrado

# Cálculo de los grados de libertad y valor crítico
cat("Desviación residual y prueba de significancia:\n")

## Desviación residual y prueba de significancia:

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

## Grados de libertad:  3

valor_critico <- qchisq(0.95, gl)
cat("Valor frontera de la zona de rechazo: ", qchisq(0.05, gl), "\n")

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

```r
#Estadístico de prueba y valor p
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

# 3.2 Comparar los AIC de ambos modelos
cat("\nComparación del AIC:\n")

##
## Comparación del AIC:

cat("AIC del modelo sin interacción: ", S$aic, "\n")

## AIC del modelo sin interacción:  493.056

cat("AIC del modelo con interacción: ", S_interaction$aic, "\n")

## AIC del modelo con interacción:  468.9692

AIC_sin_interaccion <- AIC(poisson_model) # me dalo mismo que S$aic = 493.056
AIC_con_interaccion <- AIC(poisson_model_interaction) # me dalo mismo que S_i
nteraction$aic = 468.9692

# 3.3 Comparar los coeficientes y errores estándar de ambos modelos y crear t
ablas para facilitar la comparación.
cat("\nComparación de los coeficientes y errores estándar: (en el data.frame)
\n")

##
## Comparación de los coeficientes y errores estándar: (en el data.frame)

coef_comparison <- data.frame(
  Modelo = c("Sin Interacción", "Con Interacción"),
  Coeficientes = c(coef(poisson_model), coef(poisson_model_interaction)),
  Errores_Estandar = c(S$coefficients[, 2], S_interaction$coefficients[, 2])
)
print(coef_comparison)

##             Modelo Coeficientes Errores_Estandar
## 1  Sin Interacción    3.6919631       0.04541069
## 2  Con Interacción   -0.2059884       0.05157117
## 3  Sin Interacción   -0.3213204       0.06026580
## 4  Con Interacción   -0.5184885       0.06395944
## 5  Sin Interacción    3.7967368       0.04993753
## 6  Con Interacción   -0.4566272       0.08019202
## 7  Sin Interacción   -0.6186830       0.08440012
## 8  Con Interacción   -0.5957987       0.08377723
```

```
## 9  Sin Interacción    0.6381768       0.12215312
## 10 Con Interacción    0.1883632       0.12989529
```

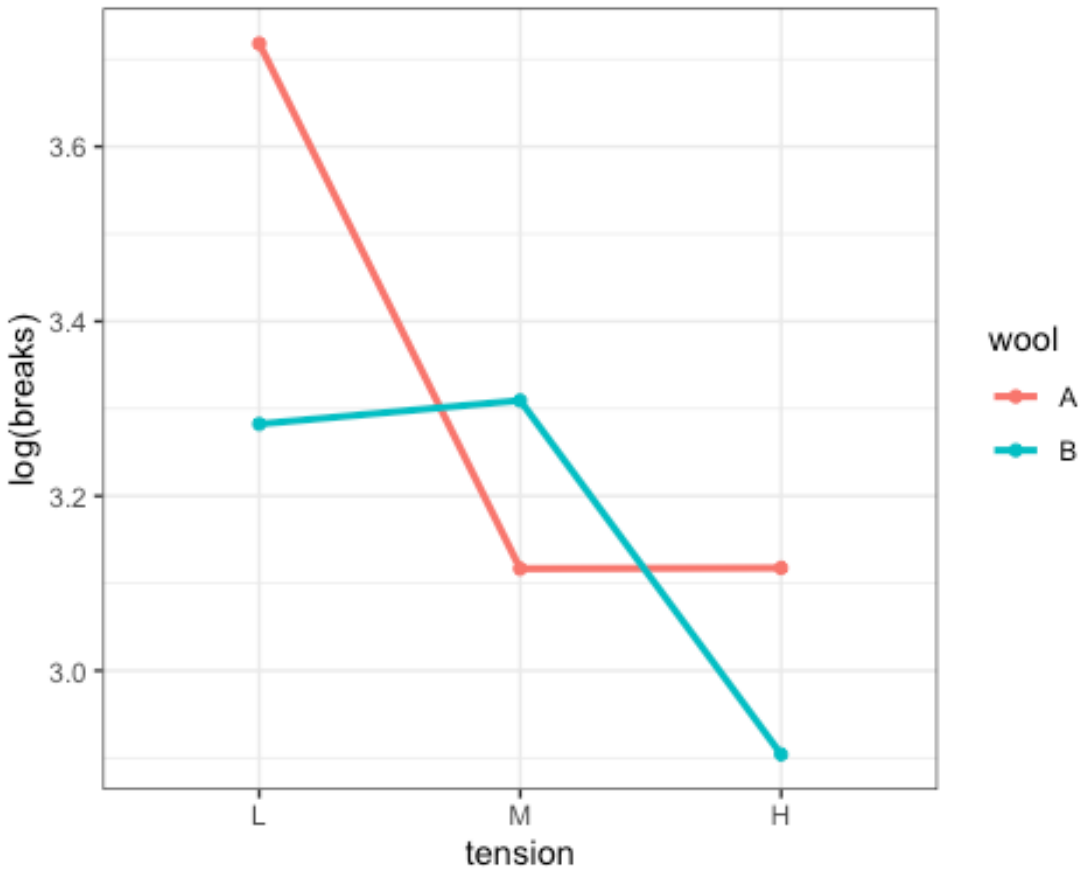
```
# 3.4 Interpretación de los coeficientes
cat("\nInterpretación de los coeficientes de ambos modelos:\n")
```

```
##
## Interpretación de los coeficientes de ambos modelos:
```

```
cat("Los coeficientes se interpretan de forma similar para ambos modelos, per
o el modelo con interacción incluye términos adicionales que muestran el efec
to combinado de 'wool' y 'tension'.\n")
```

```
## Los coeficientes se interpretan de forma similar para ambos modelos, pero
el modelo con interacción incluye términos adicionales que muestran el efecto
combinado de 'wool' y 'tension'.
```

```
# 3.5 Interpretar los coeficientes y graficar la interacción
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"))
```



```r
cat("\nConclusión:\n")
```

```
##
## Conclusión:
```

```r
cat("El modelo con el menor AIC es el preferido, y la interpretación de la de
sviación residual sugiere qué tan bien ajustado está el modelo. La comparació
n de los coeficientes y sus errores estándar ayuda a evaluar la estabilidad d
e las estimaciones.\n")
```

```
## El modelo con el menor AIC es el preferido, y la interpretación de la desv
iación residual sugiere qué tan bien ajustado está el modelo. La comparación
de los coeficientes y sus errores estándar ayuda a evaluar la estabilidad de
las estimaciones.
```

IV. Evaluación de los supuestos

```r
# 4.1 Prueba de independencia
cat("\nPrueba de independencia:\n")
```

```
##
## Prueba de independencia:
```

```r
# Esta prueba puede incluir el cálculo de la autocorrelación o métodos simila
res usados en modelos lineales.


# 4.2 Prueba de sobredispersión
cat("\nPrueba de sobredispersión:\n")
```

```
##
## Prueba de sobredispersión:
```

```r
poisgof(poisson_model)
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 210.3919
##
## $df
## [1] 50
##
## $p.value
## [1] 1.44606e-21
```

```r
# 4.3 Si el modelo muestra sobredispersión, ajustar con modelos alternativos
cat("\nAjuste de modelos alternativos en caso de sobredispersión:\n")
```

```
##
## Ajuste de modelos alternativos en caso de sobredispersión:
```

```r
# Modelo Quasi-Poisson
poisson_model_quasi <- glm(breaks ~ wool + tension, data = data, family = qua
sipoisson(link = "log"))
summary(poisson_model_quasi)
```

```
##
## 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
bnm <- glm.nb(breaks ~ wool * tension, data = data)
summary(bnm)
```

```
##
## Call:
## glm.nb(formula = breaks ~ wool * tension, data = data, init.theta = 12.082
16462,
##     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
```

V. Definir cuál es el mejor modelo

```
cat("5.1 Comparación de AIC:\n El modelo con interacción tiene un AIC más baj
o (468.97) comparado con el modelo sin interacción (493.06). Un menor AIC ind
ica un mejor ajuste del modelo, lo que sugiere que el modelo con interacción
es preferible en términos de información.\n")
```

```
## 5.1 Comparación de AIC:
##  El modelo con interacción tiene un AIC más bajo (468.97) comparado con el
modelo sin interacción (493.06). Un menor AIC indica un mejor ajuste del mode
lo, lo que sugiere que el modelo con interacción es preferible en términos de
información.
```

```
cat("5.2 Sobredispersión\n La prueba de bondad de ajuste muestra una p-value
extremadamente bajo, lo que indica que hay sobredispersión en el modelo de Po
isson. Esto sugiere que un modelo de Poisson puede no ser adecuado, y se debe
n considerar modelos alternativos como el modelo Quasi-Poisson o el modelo Bi
nomial Negativa. \n")
```

```
## 5.2 Sobredispersión
##  La prueba de bondad de ajuste muestra una p-value extremadamente bajo, lo
que indica que hay sobredispersión en el modelo de Poisson. Esto sugiere que
un modelo de Poisson puede no ser adecuado, y se deben considerar modelos alt
ernativos como el modelo Quasi-Poisson o el modelo Binomial Negativa.
```

```
cat("5.3 Modelos alternativos\n El modelo Binomial Negativa presenta un AIC d
e 405.12, lo cual es significativamente más bajo que el AIC de los modelos Po
isson, indicando un ajuste mejor y más apropiado para los datos, especialment
e en casos de sobredispersión. \n")
```

```
## 5.3 Modelos alternativos
##  El modelo Binomial Negativa presenta un AIC de 405.12, lo cual es signifi
cativamente más bajo que el AIC de los modelos Poisson, indicando un ajuste m
ejor y más apropiado para los datos, especialmente en casos de sobredispersió
n.
```

```
cat("5.4 Errores estándar y significancia\n El modelo con interacción muestra
errores estándar más altos en los coeficientes, lo cual es común cuando se ag
```

```
regan términos de interacción. Sin embargo, los coeficientes principales y las interacciones en el modelo son significativos, a excepción del término woolB:tensionH.\n")
```

```
## 5.4 Errores estándar y significancia
##  El modelo con interacción muestra errores estándar más altos en los coeficientes, lo cual es común cuando se agregan términos de interacción. Sin embargo, los coeficientes principales y las interacciones en el modelo son significativos, a excepción del término woolB:tensionH.
```

```
cat("\n5.5 Conclusión final: \n El modelo Binomial Negativa es el más adecuado, ya que aborda la sobredispersión presente y tiene el AIC más bajo.\n")
```

```
##
## 5.5 Conclusión final:
##  El modelo Binomial Negativa es el más adecuado, ya que aborda la sobredispersión presente y tiene el AIC más bajo.
```