A6-Regresión Poisson

Eryk Elizondo González A01284899

2024-10-30

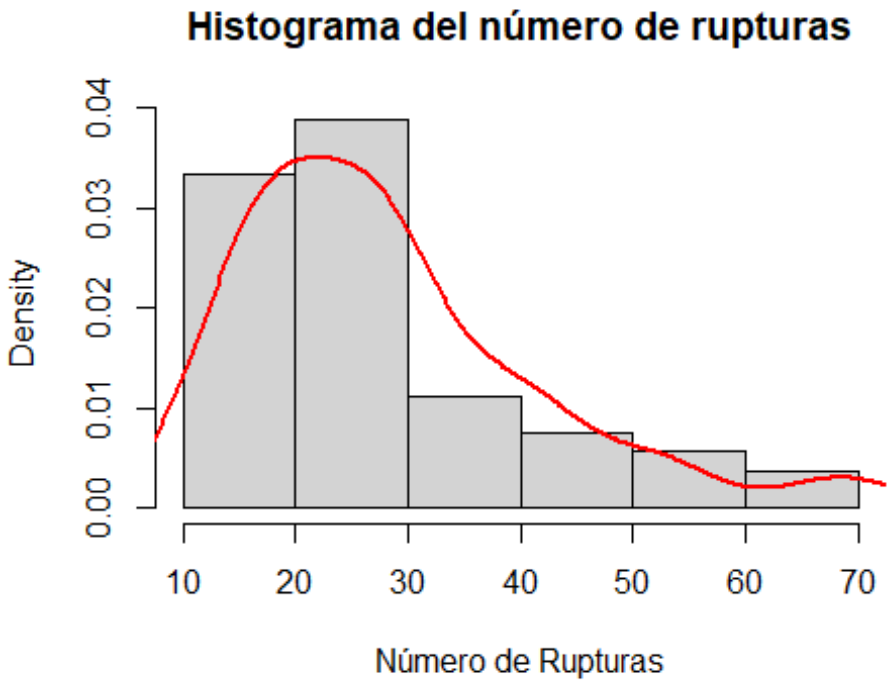
```
data<-warpbreaks
head(data,10)
```

```
##    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

```
M <- data$breaks
hist(M, xlab="Número de Rupturas", freq=FALSE,
     main="Histograma del número de rupturas")
lines(density(M), col="red", lwd=2)
```



Obtén la media y la varianza de la variable dependiente

```
M_media <- mean(M)
M_var <- var(M)
print(paste("Media: ", M_media))

## [1] "Media:  28.1481481481481"

print(paste("Varianza: ", M_var))

## [1] "Varianza:  174.204053109713"
```

Interpreta en el contexto de una Regresión Poisson

Observamos que la media y varianza son extremadamente diferentes lo cual no representa un comportamiento de distribución Poisson donde la media y varianza deben ser iguales.

II. Ajusta dos modelos de Regresión Poisson

Ajusta el modelo de regresión Poisson sin interacción

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

##
## 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 Obtenido

$$y_i = e^{3.69196 - 0.20599 woolB - 0.32132 tensionM - 0.51849 tensionH}$$

Con la lana B hay menos rupturas que con la lana A Con tensión alta hay menos rupturas, seguido de tensión media y finalmente tensión baja tiene mayor rupturas

Ajusta el modelo de regresión Poisson con interacción

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

```
##
## 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 Obtenido

$$y_i = e^{3.79674 - 0.45663 woolB - 0.6186 tensionM - 0.59580 tensionH + 0.63818 woolB * tension + 0.18836 woolB * tensionH}$$

Con la lana B hay menos rupturas que con la lana A Con tensión media hay menos rupturas, seguido de tensión alta y finalmente tensión baja tiene mayor rupturas. La intersección entre la lana b con tensión media o alta incrementa el numero de rupturas en comparación con la lana a o la lana b con tensión baja.

III. Selección del modelo

Desviación residual (Prueba de X^2)

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 X^2 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))

```
S_without <- summary(poisson_model_without)
S_with <- summary(poisson_model_with)

# Modelo sin interacción
gl_without <- S_without$df.null - S_without$df.residual
valor_critico_without <- qchisq(0.05, gl_without)
dr_without <- S_without$deviance
valor_p_without <- 1 - pchisq(dr_without, gl_without)

# Modelo con interacción
gl_with <- S_with$df.null - S_with$df.residual
valor_critico_with <- qchisq(0.05, gl_with)
dr_with <- S_with$deviance
valor_p_with <- 1 - pchisq(dr_with, gl_with)
```

```r
cat("Modelo sin interacción: \n")
```

```
## Modelo sin interacción:
```

```r
cat("  Estadístico de prueba =", dr_without, "\n")
```

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

```r
cat("  Valor frontera de la zona de rechazo =", valor_critico_without, "\n")
```

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

```r
cat("  Valor p =", valor_p_without, "\n\n")
```

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

```r
cat("Modelo con interacción: \n")
```

```
## Modelo con interacción:
```

```r
cat("  Estadístico de prueba =", dr_with, "\n")
```

```
##   Estadístico de prueba = 182.3051
```

```r
cat("  Valor frontera de la zona de rechazo =", valor_critico_with, "\n")
```

```
##   Valor frontera de la zona de rechazo = 1.145476
```

```r
cat("  Valor p =", valor_p_with, "\n")
```

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

La desviación residual del modelo con interacción es menor, lo que indica que este modelo se ajusta mejor los datos que el modelo sin interacción.

Compara los AIC de cada modelo. Recuerda que un menor AIC indica un mejor modelo.

- Modelo sin interacción: AIC = 493.06
- Modelo con interacción: AIC = 468.97

El AIC es más bajo para el modelo con interacción, lo que indica que este modelo es preferible en términos de balance entre ajuste y simplicidad

Compara los coeficientes

Coeficiente	Modelo sin Interacción (Est.)	Error Estándar sin Inter.	Modelo con Interacción (Est.)	Error Estándar con Inter.
(Intercept)	3.69196	0.04541	3.79674	0.04994
woolB	-0.20599	0.05157	-0.45663	0.08019

Coeficiente	Modelo sin Interacción (Est.)	Error Estándar sin Inter.	Modelo con Interacción (Est.)	Error Estándar con Inter.
tensionM	-0.32132	0.06027	-0.61868	0.08440
tensionH	-0.51849	0.06396	-0.59580	0.08378
woolB:tensionM	N/A	N/A	0.63818	0.12215
woolB:tensionH	N/A	N/A	0.18836	0.12990

Modelo sin interacción: - Con la lana B hay menos rupturas que con la lana A - Con tensión alta hay menos rupturas, seguido de tensión media y finalmente tensión baja tiene mayor rupturas

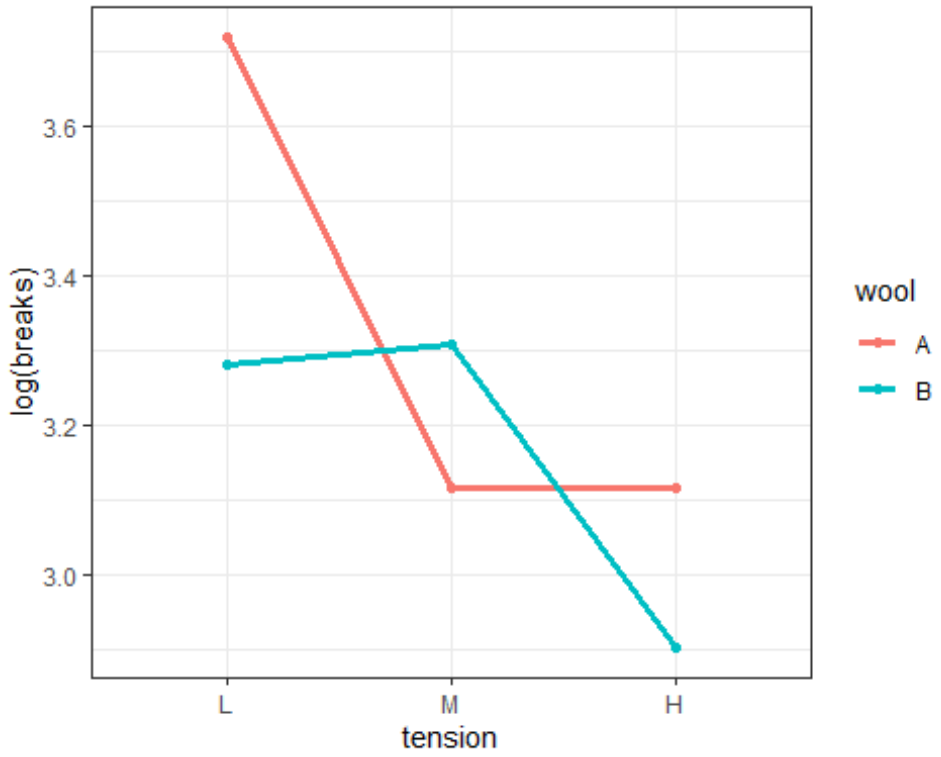
Modelo con interacción: - Con la lana B hay menos rupturas que con la lana A - Con tensión media hay menos rupturas, seguido de tensión alta y finalmente tensión baja tiene mayor rupturas. - La intersección entre la lana b con tensión media o alta incrementa el numero de rupturas en comparación con la lana a o la lana b con tensión baja.

```
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

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

```
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"))
```



La gráfica muestra que el efecto de la tensión en el número de rupturas depende del tipo de lana. Con la lana a, la tensión alta reduce significativamente las rupturas, mientras que con la lana B, la tensión baja y media tienen efectos similares, pero la tensión alta disminuye aún más las rupturas.

Con esto, podemos confirmar que el modelo con interacción es el mejor para capturar estas diferencias en el comportamiento entre tipos de lana y niveles de tensión

Define cuál de los dos es un mejor modelo.

Dado los análisis anteriores, se concluye que el modelo con interacción es mejor tanto con una desviación residual menor, un AIC menor y principalmente que este modelo representa y permite visualizar la interacción entre los tipos de lana y los niveles de tensión, evidente del análisis de los coeficiente sy el gráfico previo.

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.

Hipótesis

- H_0: Los errores no están autocorrelacionados
- H_1:: Los errores están autocorrelacionados

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor p < α

```
bgtest(poisson_model_with)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  poisson_model_with
## LM test = 1.3701, df = 1, p-value = 0.2418
```

Conslusión

No rechazamos la hipótesis nula (H_0) a favor de la alternativa (H_1). Esto significa que los errores no están autocorrelacionados.

Sobredispersión de los residuos.

Hipótesis

- H_0: No hay una sobredispersión del modelo
- H_1:: Hay una sobredispersión del modelo

Valor frontera

$$\alpha = 0.03$$

Regla de decisión

- Se rechaza H_0 si valor p < α

```
poisgof(poisson_model_with)
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 182.3051
##
## $df
## [1] 48
##
```

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

Conslusión

Rechazamos la hipótesis nula (H_0) a favor de la alternativa (H_1). Esto significa que hay una sobredispersión del modelo. Esto nos indica que el modelo no cumple con el supuesto de que la media es igual a la varianza de los residuos.

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.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
```

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
```

Define el mejor modelo usando las mismas pruebas y crtierios que usaste en los modelos Poisson

Coeficiente	Quasi-Poisson (Est.)	Error Estándar Quasi-Poisson	Binomial Negativo (Est.)	Error Estándar Binomial Negativo
(Intercept)	3.79674	0.09688	3.7967	0.1081
woolB	-0.45663	0.15558	-0.4566	0.1576
tensionM	-0.61868	0.16374	-0.6187	0.1597
tensionH	-0.59580	0.16253	-0.5958	0.1594
woolB:tensionM	0.63818	0.23699	0.6382	0.2274
woolB:tensionH	0.18836	0.25201	0.1884	0.2316

```
S_quasi <- summary(poisson.model3)
S_bnm <- summary(bnm)
```

```r
# Modelo quasi-Poisson
gl_quasi <- S_quasi$df.null - S_quasi$df.residual
valor_critico_quasi <- qchisq(0.05, gl_quasi)
dr_quasi <- S_quasi$deviance
valor_p_quasi <- 1 - pchisq(dr_quasi, gl_quasi)

# Modelo binomial negativo
gl_bnm <- S_bnm$df.null - S_bnm$df.residual
valor_critico_bnm <- qchisq(0.05, gl_bnm)
dr_bnm <- S_bnm$deviance
valor_p_bnm <- 1 - pchisq(dr_bnm, gl_bnm)

# Resultados
cat("Modelo quasi-Poisson: \n")
```

```
## Modelo quasi-Poisson:
```

```r
cat("  Estadístico de prueba =", dr_quasi, "\n")
```

```
##   Estadístico de prueba = 182.3051
```

```r
cat("  Valor frontera de la zona de rechazo =", valor_critico_quasi, "\n")
```

```
##   Valor frontera de la zona de rechazo = 1.145476
```

```r
cat("  Valor p =", valor_p_quasi, "\n\n")
```

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

```r
cat("Modelo binomial negativo: \n")
```

```
## Modelo binomial negativo:
```

```r
cat("  Estadístico de prueba =", dr_bnm, "\n")
```

```
##   Estadístico de prueba = 53.50616
```

```r
cat("  Valor frontera de la zona de rechazo =", valor_critico_bnm, "\n")
```

```
##   Valor frontera de la zona de rechazo = 1.145476
```

```r
cat("  Valor p =", valor_p_bnm, "\n")
```

```
##   Valor p = 2.647427e-10
```

- Modelo quasi-Poisson: AIC = NA
- Modelo binomial: AIC = 405.12

V. Define cuál es tu mejor modelo

Tomando en consideración el modelo inicial, el quasi-Poisson y Binomial, el mejor modelo es el Binomial negativo. Esto se debe a que tiene un AIC de 405.12 que es notablemente

menor al modelo con interacción con 468.97 y un AIC más bajo indica un mejor ajuste a los datos. Además, el modelo binomial negativo presenta una desviación residual menor de 53.506 a comparacion de los demas modelos con mayor desviación, de la misma forma que el AIC, una menor desviación indica mejor ajuste a los datos. Finalmente, el modelo binomial asume una dispersión cercana a 1, apropiado para datos con sobredispersión, lo cual es más robusto que el quasi-Poisson con una dispersión de 3.76389 que representa un peor ajuste.