Actividad 6. Regresión Poisson

Oscar Gutierrez

2024-10-29

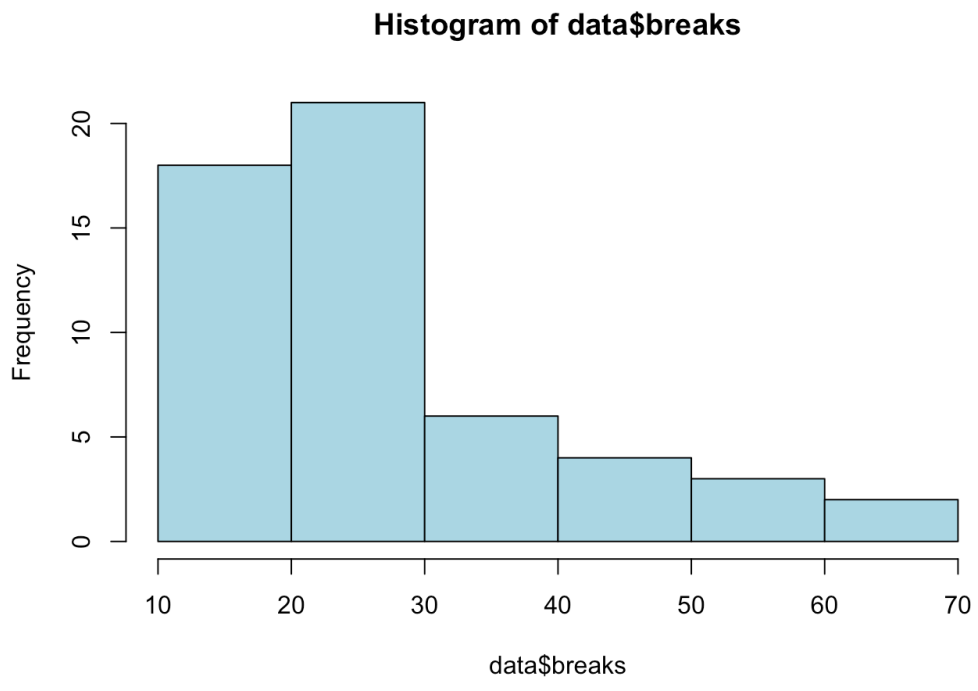
Regresión Poisson

```
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, col = 'lightblue')
```




```
summary(data)
```

```
##      breaks      wool   tension
##  Min.   :10.00   A:27   L:18
##  1st Qu.:18.25   B:27   M:18
##  Median :26.00          H:18
##  Mean   :28.15
##  3rd Qu.:34.00
##  Max.   :70.00
```

```
sd(data$breaks)
```

```
## [1] 13.19864
```

Se puede observar una distribución de los datos que está sesgada a la derecha, característica particular de las distribuciones de la familia exponencial. Las medidas importantes son la media = 26, desviación estandar = 13.2. La media es mayor a la mediana, lo cual también es característico de las distribuciones con sesgo a la derecha.

Una regresión Poisson puede ser útil para predecir el numero de rupturas puesto que se manejan únicamente valores enteros positivos, características de los procesos poisson.

Ajusta dos modelos de regresión Poisson

Modelo sin interacción

```
poisson_model_1 <-glm(breaks ~ wool + tension, data, family = poisson(link = "log"))
S1=summary(poisson_model_1)
S1
```

```
##
## 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
```

En el caso de este primer modelo, los resultados muestran que todas las variables son significativas para el modelo debido a que se obtienen valores-p menores a 0.05 (considerando una significancia del 5%). Adicionalmente, este modelo cuenta con residuos estandarizados de 210.39 y un indice de Aikaike de 493.06.

Modelo con interacción

```
poisson_model_2 <-glm(breaks ~ wool * tension, data, family = poisson(link = "log"))
S2=summary(poisson_model_2)
S2
```

```
##
## 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
```

En este segundo modelo, no todas las variables son significativas con una significancia del 5%, la interacción entre woolB y tensionH obtiene un valor mayor a 0.05, por lo que se considera no significativa, por lo que se puede descartar del modelo. Este segundo modelo cuenta con una desviación residual de 182.31, un poco menor que el modelo anterior, en índice de Aikaike de este segundo modelo es de 468.97.

Los modelos obtenidos fueron los siguientes: Sin interacción: $y = 3.69196 - 0.20599 \cdot \text{woolB} - 0.32132 \cdot \text{tensionM} - 0.51849 \cdot \text{tensionH}$

Con interacción:
$y = 3.79674 - 0.45663 \cdot \text{woolB} - 0.61868 \cdot \text{tensionM} - 0.59580 \cdot \text{tensionH} + 0.63818 \cdot (\text{woolB} \cdot \text{tensionM}) + 0.18836 \cdot (\text{woolB} \cdot \text{tensionH}$

Ambos modelos están compuestos por variables dummy, por lo que solamente se afecta la intersección con el eje y.

Selección del modelo

H0: Deviance = 0 H1: Deviance > 0

```
gl1 = S1$df.null-S1$df.residual
qchisq(0.05,gl1)
```

```
## [1] 0.3518463
```

```
dr1 = S1$deviance
cat("Estadístico de prueba =",dr1, "\n")
```

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

```
vp1 = 1-pchisq(dr1,gl1)
cat("Valor p =",vp1)
```

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

```
gl2 = S2$df.null-S2$df.residual
qchisq(0.05,gl2)
```

```
## [1] 1.145476
```

```
dr2 = S2$deviance
cat("Estadístico de prueba =",dr2, "\n")
```

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

```
vp2 = 1-pchisq(dr2,gl2)
cat("Valor p =",vp2)
```

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

En ambos modelos el valor p es de 0, por lo que se rechaza la hipótesis nula y se considera la aternativa, la desviación es mayor a 0.

AIC

Los valores para el criterio de Aikaike de ambos modelos son: modelo1 AIC = 493.06, modelo2 AIC = 468.97. Un menor valor para el AIC indica un mejor modelo, por lo que con este criterio debería seleccionarse el segundo modelo.

Comparación de coeficientes

Term	Estimate (Model 1)	Std. Error (Model 1)	Pr(>|z|) (Model 1)	Estimate (Model 2)	Std. Error (Model 2)	Pr(>|z|) (Model 2)
(Intercept)	3.69196	0.04541	< 2e-16 ***	3.79674	0.04994	< 2e-16 ***
woolB	-0.20599	0.05157	6.49e-05 ***	-0.45663	0.08019	1.24e-08 ***
tensionM	-0.32132	0.06027	9.73e-08 ***	-0.61868	0.08440	2.30e-13 ***
tensionH	-0.51849	0.06396	5.21e-16 ***	-0.59580	0.08378	1.15e-12 ***
woolB:tensionM	-	-	-	0.63818	0.12215	1.75e-07 ***
woolB:tensionH	-	-	-	0.18836	0.12990	0.147

El intercept en el modelo 1 es de 3.69 mientras que para el modelo 2 es de 3.8, este valor representa la intersección con el eje y cuando se tienen tension L y Wool A.

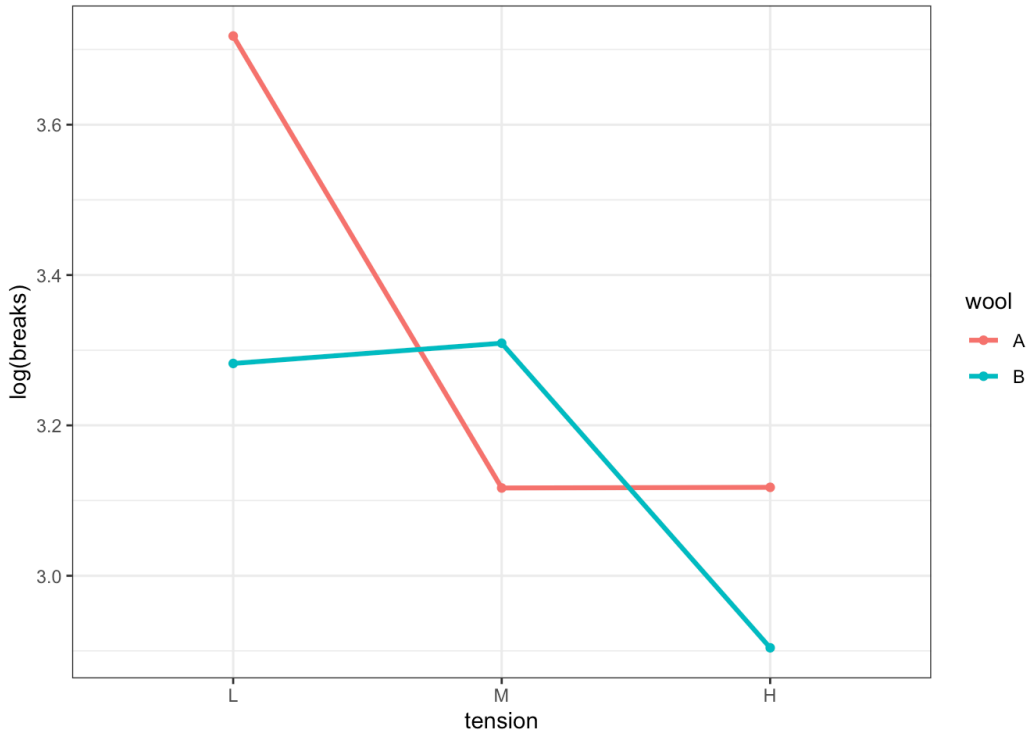
WoolB tiene un valor negativo para ambos modelos, esto significa que este tipo de lana reduce la cantidad de breaks.

tensionM y tensionH también tienen valores negativos, por lo que cuando se tienen estas tensiones, se reducen los breaks.

La interacción entre woolB con las tensiones M y H tienen valores positivos, por lo que estos términos aumentan la cantidad de breaks.

Los errores estándar de las variables en el modelo 1 son menores, esto significa que se tiene menos variabilidad si se obtienen estos coeficientes a partir de diferentes muestras de la misma población.

```
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"))
```



En esta gráfica se puede observar la interacción entre los tipos de wool y la tension, donde la woolB tiene menores breaks para las tensiones L y H, mientras que woolA tiene menos breaks en tension M.

El mejor modelo

El mejor modelo seria el segundo, puesto que logra capturar relaciones más complejas ya que considera la interacción entre las variables, además de que el valor para el AIC es menor en este modelo, lo que también indica que es mejor al modelo 1. Cabe recalcar que la interacción entre woolB y tensionH puede no ser tan relevante ya que el valor p obtenido es relativamente alto.

Evaluación de los supuestos

Independencia

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

```
library(lmtest)
```

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

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

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

```
dwtest(poisson_model_1)
```

```
##
##  Durbin-Watson test
##
## data:  poisson_model_1
## DW = 2.0332, p-value = 0.3896
## alternative hypothesis: true autocorrelation is greater than 0
```

```
dwtest(poisson_model_2)
```

```
##
##  Durbin-Watson test
##
## data:  poisson_model_2
## DW = 2.2376, p-value = 0.575
## alternative hypothesis: true autocorrelation is greater than 0
```

De acuerdo con la prueba de Durbin-Watson, ambos modelos cumplen con independencia puesto que el valor-p obtenido es menor a alpha = 0.05.

Sobredispersión

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

```
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(poisson_model_1)
```

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

```
poisgof(poisson_model_2)
```

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

La evaluación de esta prueba para ambos modelos resulta con un valor p menor a 0.05, por lo que se tiene una sobredispersión.

Modelos Cuasi Poisson y Binomial

Debido a la presencia de sobredispersión en el modelo anterior, se harán otros dos modelos, uno Cuasi Poisson y otro Binomial, ambos considerarán la interacción entre las variables.

```
poisson_model_3<-glm(breaks ~ wool * tension, data = data, family = quasipoisson(link = "log"))
S3 = summary(poisson_model_3)
S3
```

```
##
## 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
```

```
bnm  = glm.nb(breaks ~ wool * tension, data, control = glm.control(maxit=1000))
S4 = summary(bnm)
S4
```

```
##
## 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
```

La interacción entre woolB y tensionH no es relevante para ninguno de los modelos, todas las demás variables sí lo son si se considera un alpha de 0.05. Las estimaciones para todos los coeficientes es similar para ambos modelos, al igual que el error estándar.

Los valores para las desviaciones difieren entre los dos modelos, en general son menores para el modelo binomial.

Seleccionar modelo

H0: Deviance = 0 H1: Deviance > 0

```
gl3 = S3$df.null-S3$df.residual
qchisq(0.05,gl3)
```

```
## [1] 1.145476
```

```
dr3 = S3$deviance
cat("Estadístico de prueba =",dr3, "\n")
```

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

```
vp3 = 1-pchisq(dr3,gl3)
cat("Valor p =",vp3)
```

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

```
gl4 = S4$df.null-S4$df.residual
qchisq(0.05,gl2)
```

```
## [1] 1.145476
```

```
dr4 = S4$deviance
cat("Estadístico de prueba =",dr4, "\n")
```

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

```
vp4 = 1-pchisq(dr4,gl4)
cat("Valor p =",vp4)
```

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

En ambos modelos se rechaza la hipótesis nula, por lo que la desviación es mayor a 0.

El mejor modelo

Ambos de estos modelos obtienen resultados muy similares para las estimaciones de los coeficientes de cada una de las variables, sin embargo, una diferencia importante a considerar son los valores de las desviaciones.

Para el modelo Cuasi Poisson :

Null deviance: 297.37 Residual deviance: 182.31

Mientras que para el modelo binomial:

Null deviance: 86.759
Residual deviance: 53.506

Menos desviación es mejor, ya que esto significa que los valores predichos se aproximan más a los reales, por lo que el modelo binomial es la mejor opción.

Evaluación de los supuestos

Independencia

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

```
dwtest(bnm)
```

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

No hay autocorrelación entre errores para el modelo binomial con un alpha de 0.05.

Sobredispersión

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

```
poisgof(bnm)
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 53.50616
##
## $df
## [1] 48
##
## $p.value
## [1] 0.2711637
```

La prueba de sobredispersión obtiene un valor-p mayor a alpha = 0.05, por lo que no se rechaza la hipótesis nula, esto quiere decir que no hay sobredispersión.

Conclusión

En este ejercicio, el modelo binomial resulta ser el mejor modelo puesto que la mayoría de las variables son significativas, además de que obtiene los valores para desviación residual y nula más bajos de todas las opciones. También cabe mencionar que cumple con todos los supuestos de la regresión, hay independencia en los errores y el modelo no cuenta con sobredispersión.

Por estas razones, este es el modelo apropiado para predecir el número de breaks.