A6-Regresión Poisson

Oskar Arturo Gamboa Reyes

2024-10-29

Análisis Descriptivo

```
data<-warpbreaks
head(data,30)
```

```
##    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
## 11     21    A       M
## 12     29    A       M
## 13     17    A       M
## 14     12    A       M
## 15     18    A       M
## 16     35    A       M
## 17     30    A       M
## 18     36    A       M
## 19     36    A       H
## 20     21    A       H
## 21     24    A       H
## 22     18    A       H
## 23     10    A       H
## 24     43    A       H
## 25     28    A       H
## 26     15    A       H
## 27     26    A       H
## 28     27    B       L
## 29     14    B       L
## 30     29    B       L
```

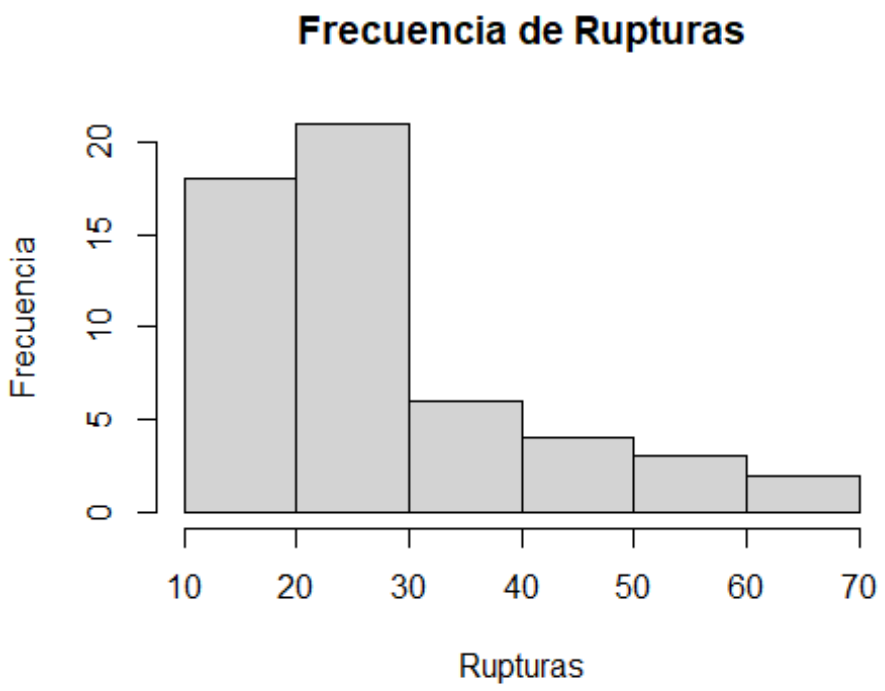
```
cat("Media: ", mean(data$breaks), "\n")
```

```
## Media:  28.14815
```

```
cat("Varianza: ", var(data$breaks), "\n")
```

```
## Varianza:  174.2041
```

```
hist(data$breaks, ylab="Frecuencia", xlab="Rupturas", main="Frecuencia de
Rupturas")
```



Podemos concluir que se va a usar un modelo Poisson ya que los datos de rupturas son enteros positivos. También podemos ver que la varianza es mayor que la media por lo que tendremos una gran dispersión en el modelo.

Ajusta dos modelos de Regresión Poisson

```
breaks = data$breaks
wool = data$wool
tension = data$tension

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

EL modelo sin interacción queda de la siguiente manera:

log(μ) = 3.69 - 0.20(woolB) - 0.32(tensionM) - 0.51(tensionH)

Intercepto: Representa el logaritmo de la cantidad esperada de rupturas para el grupo base (wool = “A” y tension = “L”).

woolB: es el cambio esperado de breaks cuando wool es “B” (comparado con “A”), manteniendo tension constante.

tensionM: es el cambio esperado de breaks cuando tension es “M” (comparado con “L”), manteniendo wool constante.

tensionH: es el cambio esperado de breaks cuando tension es “H” (comparado con “L”), manteniendo wool constante.

Tenemos dos conclusiones de estos coeficientes

Los coeficientes negativos indican una disminución en el número de rupturas (breaks) en comparación con la categoría de referencia. Por ejemplo, si wool = “B”, entonces el valor esperado de breaks disminuye en un factor de e^(−0.20) en comparación con wool = “A”.

Los coeficientes positivos tienen el efecto contrario, aumentan el numero de rupturas (breaks). Sin embargo podemos ver que las variables bases que tomó R hace que todas nuestras variables sean negativas.

```
poisson_model2<-glm(breaks ~ wool * tension, data = data, family =
poisson(link = "log"))
S2 = summary(poisson_model2)
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
```

El modelo con interacción queda de la siguiente manera:

log(μ) = 3.79 - 0.45(woolB) - 0.61(tensionM) - 0.59(tensionH) + 0.63(woolB:tensionM) + 0.18(woolB:tensionH)

woolB:tensionM: Representa el efecto combinado de wool = “B” y tension = “M”. Muestra cuánto varía el efecto de wool = “B” cuando también está en tension = “M”, comparado con la categoría de referencia.

woolB:tensionH: Representa el efecto combinado de wool = “B” y tension = “H”, en comparación con la referencia.

Si wool = “B” y tension = “M”, el número de rupturas aumenta en un factor de e^0.63 en comparación con el caso sin interacción. De la misma manera aumenta en un factor de e^0.18 si wool = “B” y tension = “L”.

Selección del modelo

Desviación Residual

Modelo sin interacción

```
gl <- S1$df.null - S1$df.residual
cat("Grados de libertad =", gl, "\n")

## Grados de libertad = 3

valor_rechazo <- qchisq(0.05, gl)
cat("Valor de corte para la zona de rechazo =", valor_rechazo, "\n")
```

```
## Valor de corte para la zona de rechazo = 0.3518463

dr <- S1$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
```

Modelo con interacción

```
gl <- S2$df.null - S2$df.residual
cat("Grados de libertad =", gl, "\n")

## Grados de libertad = 5

valor_rechazo <- qchisq(0.05, gl)
cat("Valor de corte para la zona de rechazo =", valor_rechazo, "\n")

## Valor de corte para la zona de rechazo = 1.145476

dr <- S2$deviance
cat("Estadístico de prueba =", dr, "\n")

## Estadístico de prueba = 182.3051

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

## Valor p = 0
```

Los resultados de esta prueba nos muestran que los modelos no tienen significancia con los datos. Podemos ver que los dos modelos tienen un corte bastante pequeño para los para ser rechazados (0.35 y 1.14) y el estadístico de prueba en ambos es extremadamente alto (210 y 182), además el valor p es igual a 0 en ambos modelos por loq ue indica que la probabilidad de tener esta distribución de datos para ambos modelos es nula.

AIC

Primer modelo AIC = 493.06 Segundo modelo AIC = 468.97

En terminos relativos es mejor el modelo con interacción ya que el AIC es menor, sin embargo ninguno de los dos modelos describen bien los datos proporcionados.

Coeficientes

```
coef_sin_interaccion <- S1$coefficients[, 1]
se_sin_interaccion <- S1$coefficients[, 2]
```

```r
coef_con_interaccion <- S2$coefficients[, 1]
se_con_interaccion <- S2$coefficients[, 2]


tabla_coeficientes_sin <- data.frame(Estimador = rownames(S1$coefficients),
Sin_Interaccion = coef_sin_interaccion)
tabla_coeficientes_con <- data.frame(Estimador = rownames(S2$coefficients),
Con_Interaccion = coef_con_interaccion)


tabla_coeficientes <- merge(tabla_coeficientes_sin, tabla_coeficientes_con,
by = "Estimador", all = TRUE)


tabla_errores_sin <- data.frame(Estimador = rownames(S1$coefficients),
SE_Sin_Interaccion = se_sin_interaccion)
tabla_errores_con <- data.frame(Estimador = rownames(S2$coefficients),
SE_Con_Interaccion = se_con_interaccion)


tabla_errores <- merge(tabla_errores_sin, tabla_errores_con, by =
"Estimador", all = TRUE)


print("Tabla de comparación de coeficientes:")
```

```
## [1] "Tabla de comparación de coeficientes:"
```

```r
print(tabla_coeficientes)
```

```
##        Estimador Sin_Interaccion Con_Interaccion
## 1    (Intercept)       3.6919631       3.7967368
## 2       tensionH      -0.5184885      -0.5957987
## 3       tensionM      -0.3213204      -0.6186830
## 4          woolB      -0.2059884      -0.4566272
## 5 woolB:tensionH              NA       0.1883632
## 6 woolB:tensionM              NA       0.6381768
```

```r
print("Tabla de comparación de errores:")
```

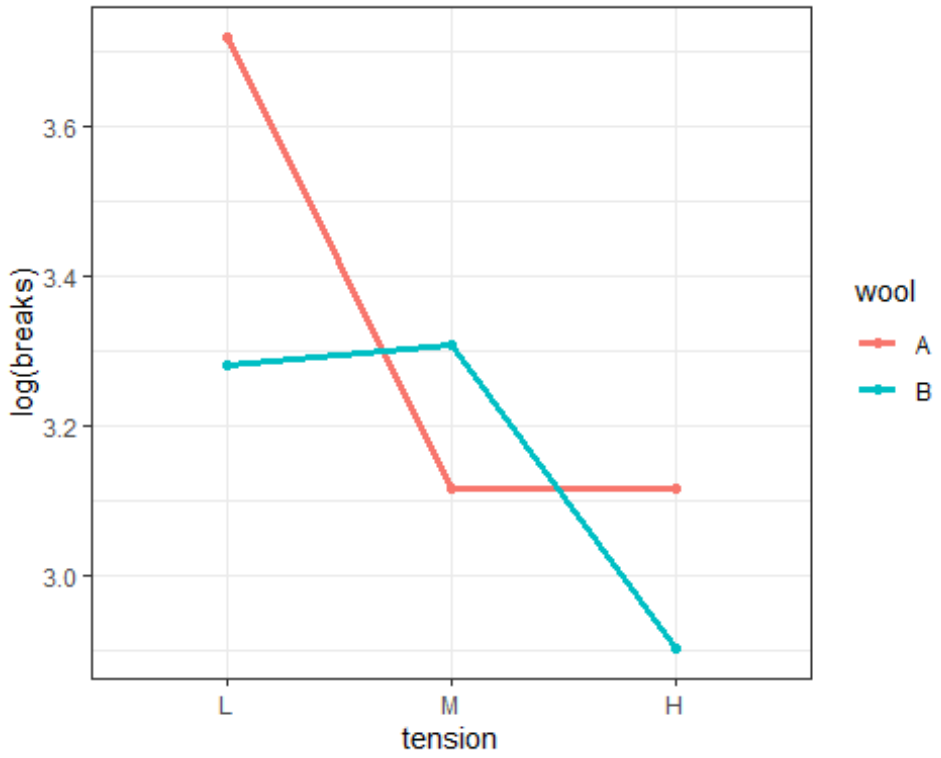
```
## [1] "Tabla de comparación de errores:"
```

```r
print(tabla_errores)
```

```
##        Estimador SE_Sin_Interaccion SE_Con_Interaccion
## 1    (Intercept)         0.04541069         0.04993753
## 2       tensionH         0.06395944         0.08377723
## 3       tensionM         0.06026580         0.08440012
## 4          woolB         0.05157117         0.08019202
```

```
## 5 woolB:tensionH                  NA         0.12989529
## 6 woolB:tensionM                  NA         0.12215312
```

```
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"))
```



Si comparamos estos dos modelos podemos ver que con interacción es un mejor modelo, ya que toma en cuenta la relación de woolB y tensionM y tensionH, lo que permite que exista una variación más similar a los datos. Sin embargo ninguno de los dos modelos es significativo y las predicciones van a ser bastante erroneas si decidimos usar uno de estos modelos, lo mejor sería cambiar de estrategia.

Evaluación de los supuestos

Independencia

```
library(zoo)
```

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

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

```
library(lmtest)
dwtest(poisson_model)
```

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

```
dwtest(poisson_model2)
```

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

Ambos modelos cumplen con el supuesto de idependencia, ya que los valores DW y p-values altos indican que no hay autocorrelación entre los residuos.

Prueba de Sobredispersión

```
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)
```

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

```
## $chisq
## [1] 210.3919
##
## $df
## [1] 50
##
## $p.value
## [1] 1.44606e-21
```

```
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
```

Hipótesis nula

H0: No hay sobredispersión. H1: Hay sobredispersión.

Al tener valores de Chi-Cuadrada (210 y 182) tan altos comparada con los grados de libertad (50 y 48), además de tener un valor p demasiado pequeño (prácticamente 0 para los dos) podemos rechazar la hipótesis nula y asumir que hay una sobredipersión, ya que ningún modelo fue útil voy a intentar con otros modelos.

Nuevos Modelos

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
```

Independencia

```
library(zoo)
library(lmtest)
dwtest(poisson_model3)

##
##  Durbin-Watson test
##
## data:  poisson_model3
## DW = 2.2376, p-value = 0.575
## alternative hypothesis: true autocorrelation is greater than 0
```

Prueba de Sobredispersión

```
pearson_chisq <- sum(residuals(poisson_model3, type = "pearson")^2)


df_residual <- poisson_model3$df.residual


sobredispersion_ratio <- pearson_chisq / df_residual


cat("Estadística de Pearson:", pearson_chisq, "\n")

## Estadística de Pearson: 180.6663

cat("Grados de libertad residuales:", df_residual, "\n")

## Grados de libertad residuales: 48

cat("Razón de sobredispersión:", sobredispersion_ratio, "\n")

## Razón de sobredispersión: 3.763881
```

Modelo Binomial Negativa

```
poisson_model4 = model.nb = glm.nb(breaks ~ wool * tension, data, control =
glm.control(maxit=1000))
summary(poisson_model4)
```

```
##
## 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
```

Independencia

```
library(zoo)
library(lmtest)
dwtest(poisson_model4)
```

```
##
##  Durbin-Watson test
##
## data:  poisson_model4
## DW = 2.2376, p-value = 0.575
## alternative hypothesis: true autocorrelation is greater than 0
```

Prueba de Sobredispersión

```r
library(epiDisplay)

poisgof(poisson_model4)
```

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

Al analizar los dos modelos nuevos, Cuasi Poisson (con interacción) y Binomial Negativo (con interacción), podemos ver que los resultados son mejores para la Binomial. Los resultados de la independencia son parecidos, el valor estadistico de los dos modelos es muy cercano a 2 (2.23 y 2.23) y con un p-value alto (0.573 y 0.575), las dos demuestran una independencia entre los residuos. Sin embargo donde realmente destaca el modelo Binomial es en la sobredispersión, obtiene un p-value de 0.27 mejor que cualquier modelo, lo que indica que no hay una sobredispersión en los residuos, mientras que el modelo Cuasi Poisson obtiene una razón de dispersión de casi 4, cualquier razón mayor a 1 es bastante mala.

Mejor modelo

$\log(E[Y]) = 3.79 - 0.45(\text{wool B}) - 0.61(\text{tensionM}) - 0.59(\text{tensionH}) + 0.63(\text{woolB}\times\text{tensionM}) + 0.18(\text{woolB}\times\text{tensionH})$

Podemos concluir que el mejor modelo es Binomial Negativo ya que tiene un mejor ajuste de sobredispersión.