Regresión Poisson

Jacobo Hirsch Rodriguez

2024-10-29

Cargamos la base de datos y vemos los primeros 10 elementos

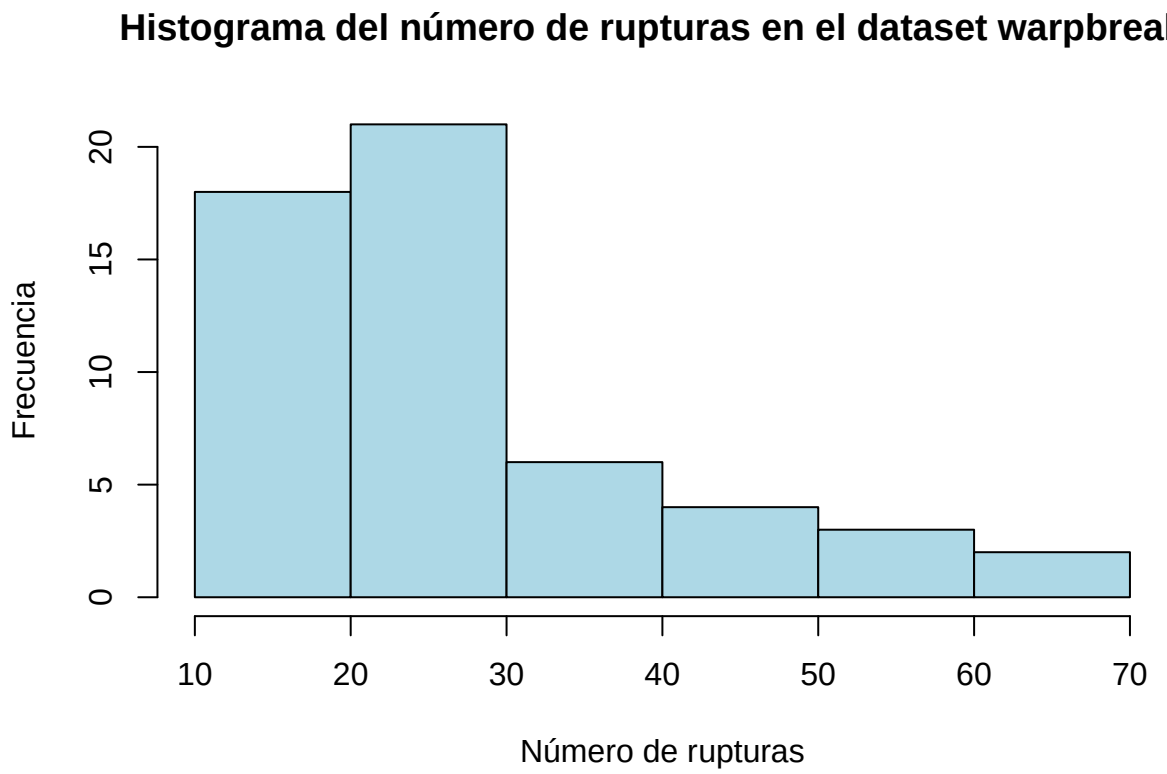
```
data<-warpbreaks
head(data,10)
```

```
##    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
```

la base de datos se puede describir con los siguientes: breaks: número de rupturas wool: tipo de lana (A o B) tensión: el nivel de tensión (L, M, H)

#I. Análisis Descriptivo

```
# Crear un histograma del número de rupturas
hist(data$breaks,
     main = "Histograma del número de rupturas en el dataset warpbreaks",
     xlab = "Número de rupturas",
     ylab = "Frecuencia",
     col = "lightblue",
     border = "black")
```



también quería calcular la moda , pero no es necesario para el analisis.

```
# Calcular la tabla de frecuencia
tabla_frecuencia <- table(data$breaks)

# Encontrar el valor o los valores con la frecuencia máxima (moda)
moda <- names(tabla_frecuencia[tabla_frecuencia == max(tabla_frecuencia)])

print(moda)
```

```
## [1] "21" "26" "29"
```

Obtén la media y la varianza de la variable dependiente

```
media_breaks <- mean(data$breaks)
print(media_breaks)
```

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

```
varianza_breaks <- var(data$breaks)
print(varianza_breaks)
```

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

en una regresión poisson se asume que la media y la varianza son iguales, se puede observar que existe una diferencia muy grande entre estas dos para nuestra variable dependiente, esto indica sobredispersión, que significa que la variabilidad en los datos es mucho mayor de lo que el modelo de Poisson espera.

#II. Ajusta dos modelos de Regresión Poisson

Ajusta el modelo de regresión Poisson sin interacción

```
# Ajustar el modelo de regresión Poisson sin interacción
modelo_poisson_sin_interaccion <- glm(breaks ~ wool + tension, data = warpbreaks, family = poisson(link

# Resumen del modelo
S1 <- summary(modelo_poisson_sin_interaccion)
S1
```

```
##
## Call:
## glm(formula = breaks ~ wool + tension, family = poisson(link = "log"),
##     data = warpbreaks)
##
## 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 toma la lana de tipo “A” y la tensión baja (“L”) como categorías de referencia, prediciendo una tasa de rupturas de aproximadamente 40.11 para esta combinación.

Tipo de Lana: Cambiar de lana “A” a “B” reduce la tasa de rupturas en un 18.6%, sugiriendo que la lana “B” es menos propensa a rupturas bajo la misma tensión. Tensión: Tensión Media (“M”) reduce las rupturas en un 27.5% en comparación con la tensión baja. Tensión Alta (“H”) reduce las rupturas en un 40.5%, siendo la opción más eficaz para disminuir rupturas. Todos los coeficientes son estadísticamente significativos ($p < 0.001$), confirmando que estos efectos son reales y no atribuibles al azar.

Ajusta el modelo de regresión Poisson con interacción

```
# Ajustar el modelo de regresión Poisson con interacción
modelo_poisson_con_interaccion <- glm(breaks ~ wool * tension, data = warpbreaks, family = poisson(link

# Resumen del modelo
S2 <- summary(modelo_poisson_con_interaccion)
S2
```

```
##
## Call:
## glm(formula = breaks ~ wool * tension, family = poisson(link = "log"),
##     data = warpbreaks)
##
## 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 utiliza lana de tipo “A” y tensión baja (“L”) como referencia, prediciendo una tasa de 44.59 rupturas para esta combinación.

Efecto de la Lana: Cambiar a lana “B” reduce las rupturas en un 36.6% en comparación con lana “A” bajo tensión baja. Efecto de la Tensión: Tensión Media (“M”) reduce las rupturas en un 46.1%. Tensión Alta (“H”) reduce las rupturas en un 44.9%. Interacciones: Lana “B” con tensión media (“M”) aumenta la tasa de rupturas en un 89% respecto a lo esperado si los efectos fueran aditivos, indicando que esta combinación no es óptima para minimizar rupturas. Lana “B” con tensión alta (“H”) no presenta un efecto interactivo significativo, sugiriendo efectos aditivos de estos factores. Conclusiones: Para minimizar rupturas, se recomienda utilizar lana “B” con tensión baja o alta, mientras que lana “B” con tensión media podría incrementar las rupturas. Este modelo con interacciones proporciona una visión detallada para la selección de materiales y condiciones óptimas de tensión.

#III. Selección del modelo

##Vamos a comparar las desviaciones residuales

```r
# Modelo sin interacción
gl1 <- S1$df.null - S1$df.residual
qchisq_val1 <- qchisq(0.05, gl1)
dr1 <- S1$deviance
vp1 <- 1 - pchisq(dr1, gl1)

cat("Modelo sin interacción:\n")
```

```
## Modelo sin interacción:
```

```r
cat("Grados de libertad =", gl1, "\n")
```

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

```r
cat("Valor frontera de la zona de rechazo =", qchisq_val1, "\n")
```

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

```r
cat("Estadístico de prueba =", dr1, "\n")
```

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

```r
cat("Valor p =", vp1, "\n\n")
```

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

```r
# Modelo con interacción
gl2 <- S2$df.null - S2$df.residual
qchisq_val2 <- qchisq(0.05, gl2)
dr2 <- S2$deviance
vp2 <- 1 - pchisq(dr2, gl2)

cat("Modelo con interacción:\n")
```

```
## Modelo con interacción:
```

```r
cat("Grados de libertad =", gl2, "\n")
```

```
## Grados de libertad = 5
```

```r
cat("Valor frontera de la zona de rechazo =", qchisq_val2, "\n")
```

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

```r
cat("Estadístico de prueba =", dr2, "\n")
```

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

```r
cat("Valor p =", vp2, "\n\n")
```

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

##Modelo sin interaccion: el modelo sin interacción parece ser que no explica bien los datos ya que hay mucha variabilidad no explicada en el modelo, esto se concluyo dado que el estadistico de prueba es muy superior al valor frontera de rechazo y el valor p nos indico que es estadisticamente significativo.

##Modelo con interaccion

aunque el valor de desviación residual sigue siendo alto, el modelo con interacción muestra una desviación residual más baja (182.31) en comparación con el modelo sin interacción (210.39), lo que sugiere que el modelo con interacción ofrece un mejor ajuste que el modelo sin interacción, explicando mejor la variabilidad de los datos. El valor p igualmente indica que es estadisticamente significativo. X$

Comparamos los valores del AIC

```r
AIC_sin_interaccion <- AIC(modelo_poisson_sin_interaccion)
AIC_con_interaccion <- AIC(modelo_poisson_con_interaccion)

cat("AIC del modelo sin interacción:", AIC_sin_interaccion, "\n")
```

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

```r
cat("AIC del modelo con interacción:", AIC_con_interaccion, "\n")
```

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

el AIC es una medida que evalúa el equilibrio entre la calidad de ajuste del modelo y su complejidad (número de parámetros). En general, un AIC más bajo indica un mejor modelo, ya que sugiere un buen ajuste con menos complejidad. En este escenario el modelo que tiene un AIC más bajo es el modelo con interacción.

#Comparacion de los coeficientes y sus errores estándar

```r
# dataframe para los coeficientes del modelo sin interacción
coef_sin_interaccion <- data.frame(
  Term = rownames(S1$coefficients),
  Coef_Sin_Interaccion = S1$coefficients[, "Estimate"],
  StdErr_Sin_Interaccion = S1$coefficients[, "Std. Error"]
)

# dataframe para los coeficientes del modelo con interacción
coef_con_interaccion <- data.frame(
  Term = rownames(S2$coefficients),
  Coef_Con_Interaccion = S2$coefficients[, "Estimate"],
  StdErr_Con_Interaccion = S2$coefficients[, "Std. Error"]
)

# unimos ambas tablas por el término para que coincidan los nombres de los coeficientes
coef_comparison <- merge(coef_sin_interaccion, coef_con_interaccion, by = "Term", all = TRUE)
print(coef_comparison)
```

```
##             Term Coef_Sin_Interaccion StdErr_Sin_Interaccion
## 1    (Intercept)            3.6919631             0.04541069
## 2       tensionH           -0.5184885             0.06395944
## 3       tensionM           -0.3213204             0.06026580
## 4          woolB           -0.2059884             0.05157117
## 5 woolB:tensionH                   NA                     NA
## 6 woolB:tensionM                   NA                     NA
##   Coef_Con_Interaccion StdErr_Con_Interaccion
## 1            3.7967368             0.04993753
## 2           -0.5957987             0.08377723
## 3           -0.6186830             0.08440012
## 4           -0.4566272             0.08019202
## 5            0.1883632             0.12989529
## 6            0.6381768             0.12215312
```

observamos que el tipo de lana (wool) y el nivel de tensión (tension) afectan significativamente la tasa de rupturas en los hilos. El modelo sin interacción muestra que tanto el uso de lana tipo “B” como los niveles

de tensión media (“M”) y alta (“H”) reducen la tasa de rupturas en comparación con la lana tipo “A” y la tensión baja (“L”). Sin embargo, el modelo con interacción revela que la combinación de lana “B” con tensión media aumenta la tasa de rupturas en comparación con los efectos individuales de cada factor. Esto indica que hay un efecto combinado entre wool y tension que el modelo sin interacción no puede capturar.

##Mejor modelo

Escogería el modelo con interacción debido a su menor AIC y a su capacidad para capturar efectos combinados significativos entre el tipo de lana y el nivel de tensión, especialmente en la interacción woolB:tensionM. Este modelo no solo proporciona un mejor ajuste a los datos, sino que también permite un entendimiento más detallado de cómo las variables afectan conjuntamente la tasa de rupturas, con la única desventaja a comparación del modelo sin interacción de que este es más complejo.

#IV. Evaluación de los supuestos

se escogio el modelo con interacción para hacer las siguientes pruebas

Los supuestos principales que se deben cumplir son:

##Independencia: haz la misma prueba de independencia que usaste en los modelos lineales.

```
library(lmtest)
```

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

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

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

```
# Prueba de Durbin-Watson para independencia en el modelo con interacción
dwtest(modelo_poisson_con_interaccion, alternative = "two.sided")
```

```
##
##  Durbin-Watson test
##
## data:  modelo_poisson_con_interaccion
## DW = 2.2376, p-value = 0.8499
## alternative hypothesis: true autocorrelation is not 0
```

el valor dw cercano a 2 sugiere que no hay autocorrelación en los residuos y el valor-p indica que no tenemos suficiente evidencia para rechazar la hipótesis nula de que no hay autocorrelación. Esto apoya aún más la independencia de los residuos en este modelo.

##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

```r
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
```

```r
# Prueba de sobredispersión para el modelo con interacción
poisgof(modelo_poisson_con_interaccion)
```

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

Dado el valor p extremadamente bajo, podemos concluir que el modelo de Poisson no cumple con el supuesto de que la varianza es igual a la media (sobredispersión). Esto sugiere que el modelo de Poisson no es adecuado y que la variabilidad en los datos es mayor de lo que el modelo de Poisson puede explicar.

##Nuevos modelos Modelo cuasi Poisson:

```r
# Ajuste del modelo Cuasi-Poisson con interacción
modelo_quasi_poisson <- glm(breaks ~ wool * tension, data = warpbreaks, family = quasipoisson(link = "l

# Resumen del modelo
summary(modelo_quasi_poisson)
```

```
##
## Call:
## glm(formula = breaks ~ wool * tension, family = quasipoisson(link = "log"),
##     data = warpbreaks)
##
## 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):

```r
library(MASS)

# Ajuste del modelo Binomial Negativa con interacción
modelo_binomial_negativa <- glm.nb(breaks ~ wool * tension, data = warpbreaks, control = glm.control(ma

# Resumen del modelo
summary(modelo_binomial_negativa)
```

```
##
## Call:
## glm.nb(formula = breaks ~ wool * tension, data = warpbreaks,
##     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
```

vamos a hacer dos pruebas rápidas para ver cual de estos dos nuevos modelos es el mejor:

##Prueba AIC

```
cat("AIC del modelo Poisson:", AIC(modelo_poisson_con_interaccion), "\n")
```

```
## AIC del modelo Poisson: 468.9692
```

```
cat("AIC del modelo Binomial Negativa:", AIC(modelo_binomial_negativa), "\n")
```

```
## AIC del modelo Binomial Negativa: 405.1248
```

##Sobredispersión de los residuos.

```
# Verificar sobredispersión en el modelo Cuasi-Poisson
S_qp <- summary(modelo_quasi_poisson)
deviance_qp <- S_qp$deviance
df_residual_qp <- S_qp$df.residual
dispersion_ratio_qp <- deviance_qp / df_residual_qp

cat("Modelo Cuasi-Poisson:\n")
```

```
## Modelo Cuasi-Poisson:
```

```
cat("Desviación residual =", deviance_qp, "\n")
```

```
## Desviación residual = 182.3051
```

```
cat("Grados de libertad residuales =", df_residual_qp, "\n")
```

```
## Grados de libertad residuales = 48
```

```
cat("Relación de sobredispersión =", dispersion_ratio_qp, "\n\n")
```

```
## Relación de sobredispersión = 3.798024
```

```
# Verificar sobredispersión en el modelo Binomial Negativa
S_bn <- summary(modelo_binomial_negativa)
deviance_bn <- S_bn$deviance
df_residual_bn <- S_bn$df.residual
dispersion_ratio_bn <- deviance_bn / df_residual_bn

cat("Modelo Binomial Negativa:\n")
```

```
## Modelo Binomial Negativa:
```

```r
cat("Desviación residual =", deviance_bn, "\n")
```

```
## Desviación residual = 53.50616
```

```r
cat("Grados de libertad residuales =", df_residual_bn, "\n")
```

```
## Grados de libertad residuales = 48
```

```r
cat("Relación de sobredispersión =", dispersion_ratio_bn, "\n")
```

```
## Relación de sobredispersión = 1.114712
```

se obtiene la conclusión de los nuevos modelos en el último apartado

#V. Define cual es tu mejor modelo

Con base en estos resultados, el modelo de Binomial Negativa es claramente el mejor de los dos por las siguientes razones:

Tiene una relación de sobredispersión cercana a 1, lo que sugiere que captura adecuadamente la variabilidad de los datos sin problemas de sobredispersión. Tiene un AIC considerablemente menor que el modelo Poisson, lo que indica un mejor ajuste global.