A6-Regresión Poisson_fer

Fernanda Pérez

2024-10-29

Regresión Poisson

Trabajaremos con el paquete dataset, que incluye la base de datos warpbreaks, que contiene datos del hilo (yarn) para identificar cuáles variables predictoras afectan la ruptura de urdimbre.

Este conjunto de datos indica cuántas roturas de urdimbre ocurrieron para diferentes tipos de telares por telar, por longitud fija de hilo:

breaks: número de rupturas wool: tipo de lana (A o B) *tensión: el nivel de tensión (L, M, H)

Sigue el siguiente procedimiento de análisis:

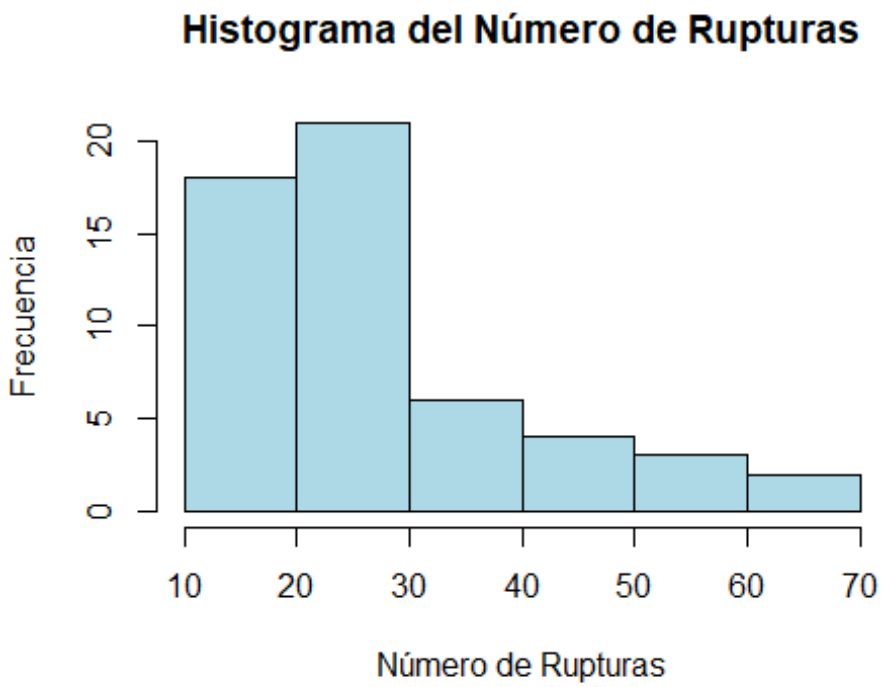
I. Análisis Descriptivo

Histograma del número de rupturas

```
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
```

```
hist(warpbreaks$breaks,
     main = "Histograma del Número de Rupturas",
     xlab = "Número de Rupturas",
     ylab = "Frecuencia",
     col = "lightblue",
     border = "black")
```



El histograma del número de rupturas nos muestra una distribución asimétrica a la derecha. La mayoria de rupturas se concentra en los valores más bajos en el rango de entre 10 y 30 rupturas, con una frecuencia máxima alrededor de los 20-30. La frecuencia disminuye poco a poco conforme el número de rupturas aumenta, indicando que las rupturas más altas son menos frecuentes.

Obtén la media y la varianza de la variable dependiente

```
# Media
media_breaks <- mean(warpbreaks$breaks)
cat("media breaks:", media_breaks, "        ")

## media breaks: 28.14815

# Varianza
varianza_breaks <- var(warpbreaks$breaks)
cat("varianza breaks:", varianza_breaks)

## varianza breaks: 174.2041
```

La media del número de rupturas es 28.14815 y la varianza es 174.2041. En el contexto de una regresión de Poisson, uno de los supuestos es que la media y la varianza de la variable de conteo sean aproximadamente iguales, pero en este caso, la varianza es considerablemente mayor que la media, lo cual indica una posible sobre-dispersión en los datos.

Interpreta en el contexto de una Regresión Poisson

El histograma del número de rupturas en el contexto de poisson, lo que esperamos es que la media y la varianza de los datos de conteo sean parecidos y/o muy cercanos, la forma del grafico del histograma, es decir la distribución que presenta nos da indicios de que es una posible adecuación para un modelo de Poisson,, sin embargo tenemos que confirmar con los valores de media breaks y varianza breaks y ver que los valores sean similares. Si la diferencia entre esos valores es grande podria indicar una sobre-dispersión.

Ya que análisamos media breaks y varianza breaks vemos que la varianza es considerablemente mayor que la media, lo cual indica una posible sobre-dispersión en los datos, esto nos indica que la regresión de Poisson tal vez no sea la mejor opción de ajuste para este set de datos.

II. Ajusta dos modelos de Regresión Poisson

Ajusta el modelo de regresión Poisson sin interacción

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

```
##
## 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
```

Este modelo de regresión Poisson sin interacción nos sugiere que el tipo de lana y el nivel de tensión tienen un efecto significativo en la frecuencia de rupturas, la lana tipo B tiene una tasa menor de rupturas conforme aumenta la tensión.

Ajusta el modelo de regresión Poisson con interacción

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

```
##
## 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
```

Este modelo de regresión Poisson con interacción nos sugiere que la interacción que hay entre wool y tension es significativa, y va a depender del nivel de tensión el efecto de la lana tipo B en la frecuencia de rupturas.

Interpreta los coeficientes de las variables Dummy. Escribe el modelo obtenido. Toma en cuenta que R genera variables Dummy para las variables categóricas. Para cada variable genera k-1 variables Dummy en k categorías.

Para poder interpretar los coeficientes de las variables Dummy en ambos modelos de regresión de Poisson, tenemos que tomar en cuenta que R toma una categoría de referencia para cada variable categórica y genera k-1 variables Dummy en una variable con k categorias.

En los modelos de regresión de Poisson, R va a usar lana tipo A y tensión baja como referencias. Cuando estamos trabajando con el modelo sin interacción y cambiamos a lana tipo B se reduce el log de rupturas en 0.206, mientras que tensiones media y alta también van a disminuir las rupturas. Y al trabajar con el modelo con interacción, los términos woolB:tensionM y woolB:tensionH van a indicar un aumento adicional en el

log de rupturas cuando la lana B se combina con tensiones media o alta. El modelo general es:

$$\log(\mathrm{E}(\mathrm{breaks})) = \beta_0 + \beta_1 \mathrm{woolB} + \beta_2 \mathrm{tensionM} + \beta_3 \mathrm{tensionH} + \beta_4 (\mathrm{woolB{:}tensionM}) + \beta_5 (\mathrm{woolB{:}tensionH})$$

III. Selección del modelo

Para seleccionar el modelo se toma en cuenta:

Desviación residual: es la suma del cuadrado de los residuos estandarizados que se obtienen bajo el modelo. Con los grados de libertad se realiza una prueba de x^2 para significancia del modelo.

```
# Para el modelo sin interacción
S_no_interaction <- summary(poisson_model_no_interaction)
gl_no_interaction <- S_no_interaction$df.null -
S_no_interaction$df.residual
dr_no_interaction <- S_no_interaction$deviance
vp_no_interaction <- 1 - pchisq(dr_no_interaction, gl_no_interaction)
cat("Estadístico de prueba (sin interacción):", dr_no_interaction, "\n")

## Estadístico de prueba (sin interacción): 210.3919

cat("Valor p (sin interacción):", vp_no_interaction, "\n")

## Valor p (sin interacción): 0

# Para el modelo con interacción
S_interaction <- summary(poisson_model_interaction)
gl_interaction <- S_interaction$df.null - S_interaction$df.residual
dr_interaction <- S_interaction$deviance
vp_interaction <- 1 - pchisq(dr_interaction, gl_interaction)
cat("Estadístico de prueba (con interacción):", dr_interaction, "\n")

## Estadístico de prueba (con interacción): 182.3051

cat("Valor p (con interacción):", vp_interaction, "\n")

## Valor p (con interacción): 0
```

AIC: Criterio de Aikaike

```
AIC_no_interaction <- AIC(poisson_model_no_interaction)
AIC_interaction <- AIC(poisson_model_interaction)
cat("AIC (sin interacción):", AIC_no_interaction, "\n")

## AIC (sin interacción): 493.056

cat("AIC (con interacción):", AIC_interaction, "\n")
```

```
## AIC (con interacción): 468.9692
```

Un AIC más bajo indican un mejor ajuste, entonces el modelo con interacción indica ser mejor en est caso.

Comparación entre los coeficientes y los errores estándar de de ambos modelos

```
coef_no_interaction <- coef(summary(poisson_model_no_interaction))
coef_interaction <- coef(summary(poisson_model_interaction))

comparison_table <- data.frame(
  Variable = c(rownames(coef_no_interaction),
rownames(coef_interaction)),
  Modelo = c(rep("Sin Interacción", nrow(coef_no_interaction)), rep("Con
Interacción", nrow(coef_interaction))),
  Estimate = c(coef_no_interaction[, "Estimate"], coef_interaction[,
"Estimate"]),
  Std.Error = c(coef_no_interaction[, "Std. Error"], coef_interaction[,
"Std. Error"])
)

comparison_table
```

```
##          Variable          Modelo   Estimate  Std.Error
## 1     (Intercept) Sin Interacción  3.6919631 0.04541069
## 2           woolB Sin Interacción -0.2059884 0.05157117
## 3        tensionM Sin Interacción -0.3213204 0.06026580
## 4        tensionH Sin Interacción -0.5184885 0.06395944
## 5     (Intercept) Con Interacción  3.7967368 0.04993753
## 6           woolB Con Interacción -0.4566272 0.08019202
## 7        tensionM Con Interacción -0.6186830 0.08440012
## 8        tensionH Con Interacción -0.5957987 0.08377723
## 9  woolB:tensionM Con Interacción  0.6381768 0.12215312
## 10 woolB:tensionH Con Interacción  0.1883632 0.12989529
```

En esta tabla comparamos los coeficientes y errores estándar de los dos modelos:con y sin interacción. En los dos, woolB, tensionM, y tensionH afectan significativamente las rupturas, con reducción en lana B y tensiones media y alta. En el modelo que tiene interacción, las combinaciones woolB:tensionM y woolB:tensionH muestran que el efecto de la lana B depende de la tensión, englobando mejor la relación entre lana y tensión en las rupturas.

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))

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

Los resultados obtenidos fueron: AIC (sin interacción): 493.056 AIC (con interacción): 468.9692 como un AIC más bajo indican un mejor ajuste, entonces el modelo con interacción indica ser mejor en est caso.

Compara los coeficientes

Con tabla que se creó anterorirmente para la comparación de los coeficientes y errores estándar de ambos modelos podemos observar cómo los coeficientes de woolB, tensionM, y tensionH varían entre los modelos, además de la adición de términos de interacción (woolB:tensionM y woolB:tensionH) en el modelo con interacción.

Compara los coeficientes de ambos modelos (haz una tabla para que se facilite la comparación)

La interpretación nos sugiere que en el modelo con interacción, las combinaciones de lana tipo B con tensiones media y alta (woolB:tensionM y woolB:tensionH) modifican el efecto de la lana en el número de rupturas, lo cual no está capturado en el modelo sin interacción. Esto nos sugiere que el modelo con interacción ofrece una mejor comprensión de cómo la lana y la tensión interactúan para afectar las rupturas.

Compara el error estándar de cada estimador de B_i de ambos modelos (haz una tabla para que se facilite la comparación)

Si comparamos los errores estándar entre los modelos, vemos que en el modelo con interacción, los errores estándar de woolB, tensionM y tensionH aumentan en comparación con el modelo sin interacción, indicando una ligera reducción en la precisión de estos estimadores. Los términos de interacción (woolB:tensionM y woolB:tensionH) también presentan errores estándar un poco altos, lo cual no es bueno porque eso es incertidumbre adicional al incluir interacciones. Apesar de que el

modelo con interacción es menos preciso en algunos estimadores, ofrece una interpretación más completa de los efectos combinados entre wool y tension.

Define cuál de los dos es un mejor modelo

Con el análisis que se realizó en esta sección podemos decir que el modelo con interacción es el mejor modelo:

*El modelo con interacción tiene un AIC menor (468.97), lo que indica un mejor ajuste al balancear precisión y complejidad

- La inclusión de los términos de interacción (woolB:tensionM y woolB:tensionH) en el modelo con interacción permite capturar la relación conjunta que hay entre wool y tension, lo cual sirve para facilitar una interpretación más detallada de cómo la combinación de estas variables afecta el número de rupturas.

*Apesar de que los errores estándar son un poco mayores en el modelo con interacción, indicando una leve pérdida de precisión en algunos estimadores, este incremento es moderado y podemos no tomarlo como algo muy negativo ya que al final obtenemos el valor añadido en la interpretación de los efectos combinados.

IV. Evaluación de los supuestos

Los supuestos principales que se deben cumplir son:

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

```
residual_deviance <- deviance(poisson_model_interaction)
residual_df <- df.residual(poisson_model_interaction)
dispersion_ratio <- residual_deviance / residual_df

cat("Razón de desviación:", dispersion_ratio, "\n")

## Razón de desviación: 3.798024

if(dispersion_ratio > 1.5) {
  cat("Hay evidencia de sobredispersión.\n")
} else {
  cat("No hay evidencia de sobredispersión.\n")
}
```

```
## Hay evidencia de sobredispersión.
```

Modelo Cuasi-Poisson con interacción

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

##
## 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 de Binomial Negativa con interacción

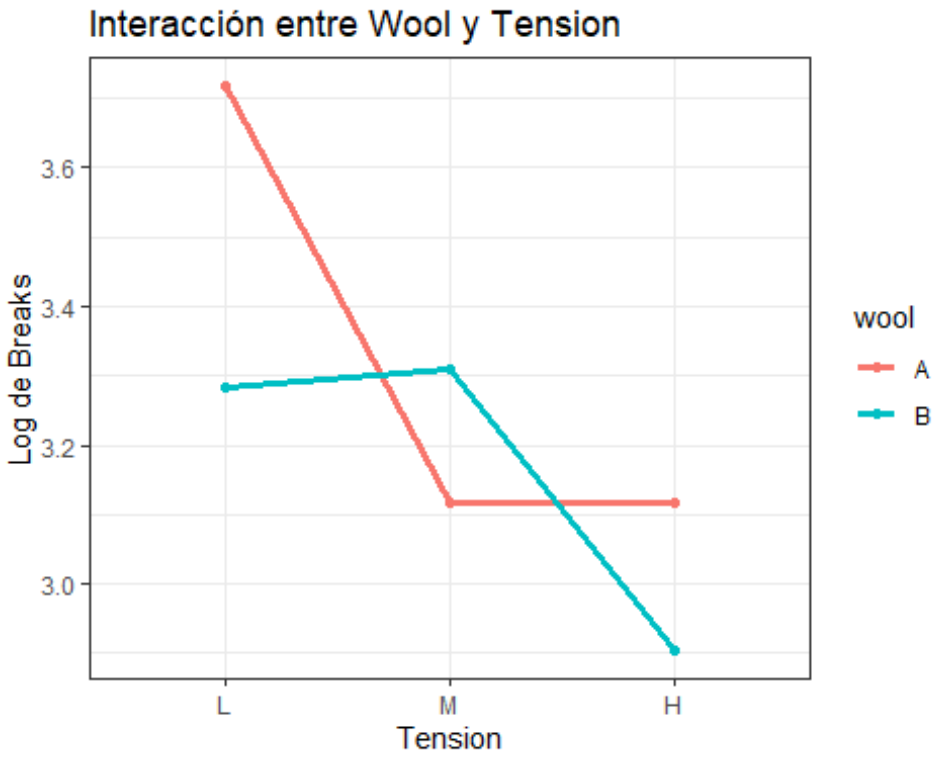
```
library(MASS)
bnm_model <- glm.nb(breaks ~ wool * tension, data = warpbreaks, control =
glm.control(maxit = 1000))
summary(bnm_model)

##
## 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

library(ggplot2)

ggplot(warpbreaks, 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")) +
  labs(title = "Interacción entre Wool y Tension", x = "Tension", y =
"Log de Breaks")
```



El grafico nos indica que la lana A reduce las rupturas significativamente con el aumento de tensión, y por otro lado la lana B se mantiene estable en tensión baja y media, solamente disminuye en alta. Esto evidencia una interacción entre wool y tension.

Define cuál es tu mejor modelo

En este caso para el problema que estamos trabajando el mejor modelo es el de Binomial Negativa con interacción por las siguientes razones:

AIC: El modelo de Binomial Negativa cuenta con un AIC menor en comparación con los modelos de Poisson, lo cual nos indica un mejor ajuste.

Desviación Residual: La desviación residual en el modelo de Binomial Negativa es significativamente menor que en el otro modelo, el Cuasi-Poisson, esto nos sugiere que el modelo de Binomial Negativa se ajusta mejor a la variabilidad de los datos.

Sobredispersión: El modelo de Binomial Negativa maneja adecuadamente la sobredispersión, que fue un problema que se observó en el modelo de Poisson y el parámetro de dispersión de este modelo ajusta la variabilidad adicional en los datos.

En conclusión el modelo de Binomial Negativacon interacción es el selccionado ya que se considera el más adecuado para estos datos debido a su mejor ajuste general y capacidad para manejar la sobredispersión.