A6-Regresión Poisson

Nallely Serna

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.

```
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
```

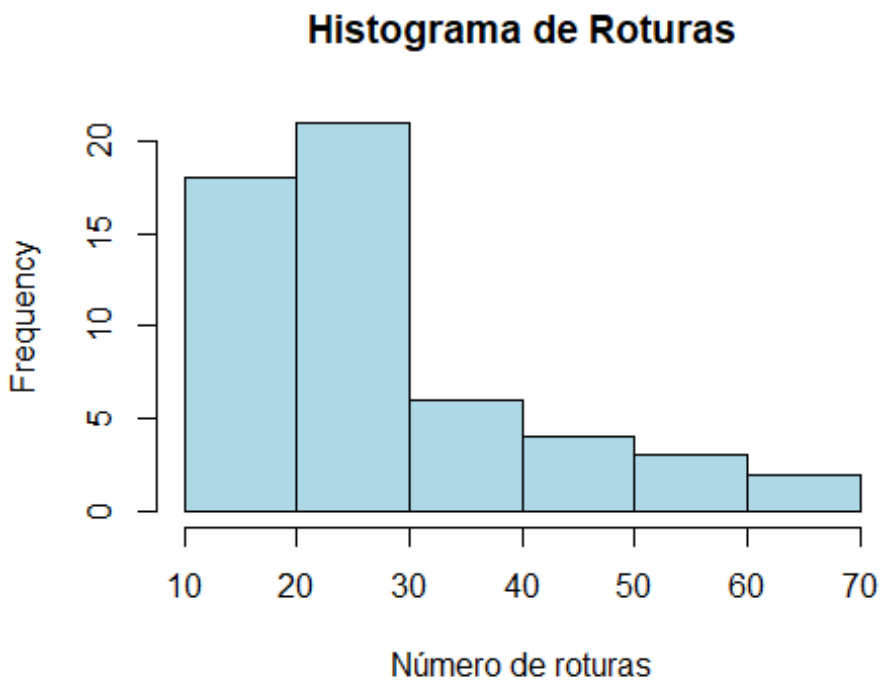
Este conjunto de datos indica cuántas roturas de urdimbre ocurrieron para diferentes tipos de telares por telar, por longitud fija de hilo:

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

I. Análisis Descriptivo

Histograma del número de rupturas

```
data <- warpbreaks
hist(data$breaks, main="Histograma de Roturas", xlab="Número de roturas",
col="lightblue", border="black")
```



Obtén la media y la varianza de la variable dependiente

```
media_breaks <- mean(data$breaks)
varianza_breaks <- var(data$breaks)
cat("Media: ", media_breaks, "\n")

## Media:  28.14815

cat("Varianza: ", varianza_breaks, "\n")

## Varianza:  174.2041
```

Interpreta en el contexto de una Regresión Poisson

En un modelo de regresión Poisson, la media y la varianza de la variable dependiente (breaks) deberían ser aproximadamente iguales. Si la varianza es significativamente mayor que la media (como lo es en este caso), podría indicar sobredispersión, lo que podría sugerir la necesidad de ajustar un modelo alternativo, como el modelo de Cuasi-Poisson o el Binomial Negativo.

II. Ajuste de los modelos de Regresión Poisson

Ajusta el modelo de regresión Poisson sin interacción

```r
poisson_model1 <- glm(breaks ~ wool + tension, data = data, family =
poisson(link = "log"))
summary(poisson_model1)
```

```
##
## Call:
## glm(formula = breaks ~ wool + tension, family = poisson(link = "log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6871  -1.6503  -0.4269   1.1902   4.2616
##
## 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
```

Ajusta el modelo de regresión Poisson con interacción

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

```
##
## Call:
## glm(formula = breaks ~ wool * tension, family = poisson(link = "log"),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3383  -1.4844  -0.1291   1.1725   3.5153
##
## 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
```

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.

R automáticamente genera variables dummy para las variables categóricas como wool y tension. Esto significa que el valor de los coeficientes representa la diferencia de cada categoría en comparación con la categoría de referencia. Para interpretar estos coeficientes, puedes observar si los coeficientes son positivos o negativos, lo que indicará si incrementan o disminuyen el número de roturas en comparación con la categoría de referencia.

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 para significancia del modelo. - AIC: Criterio de Aikaike

```
# Modelo sin interacción
S1 <- summary(poisson_model1)
gl1 <- S1$df.null - S1$df.residual
qchisq(0.05, gl1)
```

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

```
dr1 <- S1$deviance
cat("Estadístico de prueba para el modelo 1:", dr1, "\n")
```

```
## Estadístico de prueba para el modelo 1: 210.3919
```

```
vp1 <- 1 - pchisq(dr1, gl1)
cat("Valor p para el modelo 1:", vp1, "\n")
```

```
## Valor p para el modelo 1: 0
```

```r
# Modelo con interacción
S2 <- summary(poisson_model2)
gl2 <- S2$df.null - S2$df.residual
qchisq(0.05, gl2)
```

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

```r
dr2 <- S2$deviance
cat("Estadístico de prueba para el modelo 2:", dr2, "\n")
```

```
## Estadístico de prueba para el modelo 2: 182.3051
```

```r
vp2 <- 1 - pchisq(dr2, gl2)
cat("Valor p para el modelo 2:", vp2, "\n")
```

```
## Valor p para el modelo 2: 0
```

```r
# Comparación de AIC
cat("AIC Modelo 1: ", AIC(poisson_model1), "\n")
```

```
## AIC Modelo 1:  493.056
```

```r
cat("AIC Modelo 2: ", AIC(poisson_model2), "\n")
```

```
## AIC Modelo 2:  468.9692
```

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

Desviación residual (Prueba de x2)

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 x2 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))

Usa los siguientes comandos: Valor frontera de la zona de rechazo (S es la variable que denota el summary del modelo): gl = S$null.deviance - S$df.residual qchisq(0.05,gl)

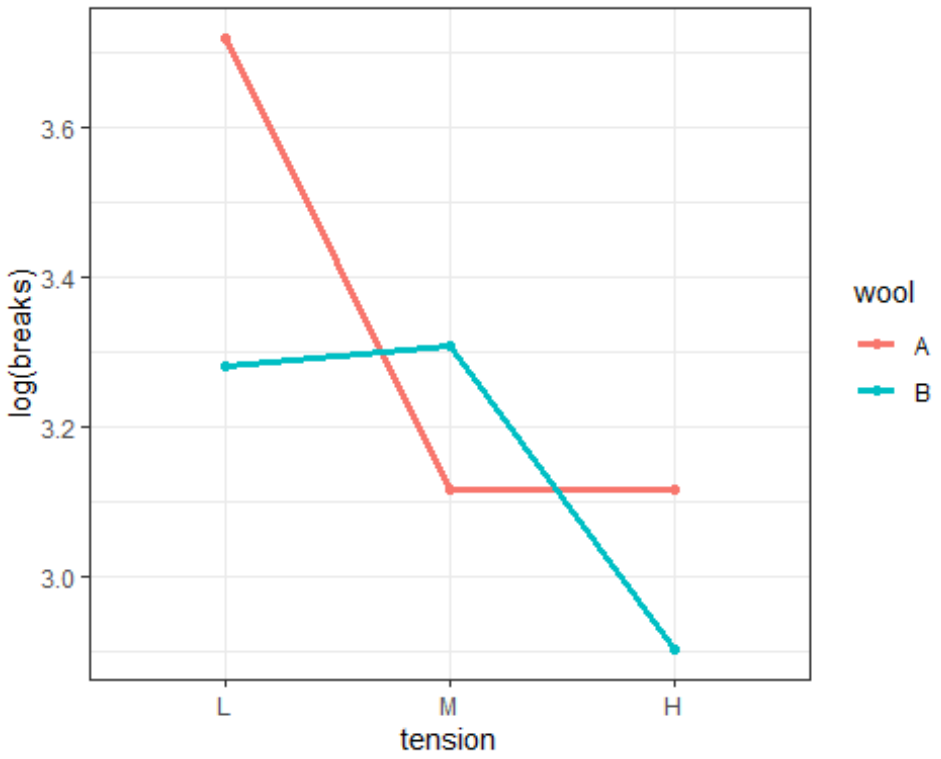
Estadístico de prueba y valor p: dr = S$deviance cat("Estadístico de prueba =",dr, "") vp = 1-pchisq(dr,gl) cat("Valor p =",vp)

Compara los AIC de cada modelo. Recuerda que un menor AIC indica un mejor modelo. Compara los coeficientes Compara los coeficientes de ambos modelos (haz una tabla para que se facilite la comparación) Compara el error estándar de cada

estimador de Bi de ambos modelos (haz una tabla para que se facilite la comparación) Interpreta los coeficientes de ambos modelos. Define cuál de los dos es un mejor modelo.

```
# Extraer coeficientes del modelo 1 (sin interacción)
coef1 <- summary(poisson_model1)$coefficients

# Extraer coeficientes del modelo 2 (con interacción)
coef2 <- summary(poisson_model2)$coefficients

# Crear un data frame con los coeficientes comunes (las primeras 4 filas)
coef_comparison_common <- data.frame(
  Coef_Model1 = coef1[, 1],  # Coeficientes del modelo 1
  Std_Error_Model1 = coef1[, 2],  # Error estándar del modelo 1
  Coef_Model2 = coef2[1:4, 1],  # Coeficientes del modelo 2 sin
interacción
  Std_Error_Model2 = coef2[1:4, 2]  # Error estándar del modelo 2 sin
interacción
)

# Imprimir la tabla de comparación para los coeficientes comunes
cat("Comparación de coeficientes comunes (sin interacción):\n")

## Comparación de coeficientes comunes (sin interacción):

print(coef_comparison_common)

##             Coef_Model1 Std_Error_Model1 Coef_Model2 Std_Error_Model2
## (Intercept)   3.6919631       0.04541069   3.7967368       0.04993753
## woolB        -0.2059884       0.05157117  -0.4566272       0.08019202
## tensionM     -0.3213204       0.06026580  -0.6186830       0.08440012
## tensionH     -0.5184885       0.06395944  -0.5957987       0.08377723

# Comparar también los coeficientes de interacción del modelo 2
cat("\nCoeficientes de interacción en el modelo 2:\n")

##
## Coeficientes de interacción en el modelo 2:

coef_interaction <- coef2[5:6, ]
print(coef_interaction)

##                 Estimate Std. Error  z value     Pr(>|z|)
## woolB:tensionM 0.6381768  0.1221531 5.224400 1.747203e-07
## woolB:tensionH 0.1883632  0.1298953 1.450115 1.470263e-01

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(fill="transparent"))
```



IV. Evaluación de los supuestos

1. Independencia

```
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 el modelo sin interacción
dw_test1 <- dwtest(poisson_model1, alternative = "two.sided")
cat("Durbin-Watson test for Model 1 (sin interacción):\n")

## Durbin-Watson test for Model 1 (sin interacción):

print(dw_test1)

##
##  Durbin-Watson test
```

```
##
## data:  poisson_model1
## DW = 2.0332, p-value = 0.7791
## alternative hypothesis: true autocorrelation is not 0
```

```
# Prueba de Durbin-Watson para el modelo con interacción
dw_test2 <- dwtest(poisson_model2, alternative = "two.sided")
cat("\nDurbin-Watson test for Model 2 (con interacción):\n")
```

```
##
## Durbin-Watson test for Model 2 (con interacción):
```

```
print(dw_test2)
```

```
##
##  Durbin-Watson test
##
## data:  poisson_model2
## DW = 2.2376, p-value = 0.8499
## alternative hypothesis: true autocorrelation is not 0
```

2. Sobredispersión de los residuos

```
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_model2)
```

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

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

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)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6871  -1.6503  -0.4269   1.1902   4.2616
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.69196    0.09374  39.384  < 2e-16 ***
## woolB       -0.20599    0.10646  -1.935 0.058673 .
## tensionM    -0.32132    0.12441  -2.583 0.012775 *
## tensionH    -0.51849    0.13203  -3.927 0.000264 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 4.261537)
##
##     Null deviance: 297.37  on 53  degrees of freedom
## Residual deviance: 210.39  on 50  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

V. Define cuál es tu mejor modelo

1. Ajuste del modelo (Deviance Residuals y Estadístico de bondad de ajuste): Modelo 1 (sin interacción): Residual deviance = 210.39, con 50 grados de libertad. Modelo 2 (con interacción): Residual deviance = 182.31, con 48 grados de libertad. Interpretación: Un valor más bajo de la deviance residual sugiere un mejor ajuste. El Modelo 2 tiene un valor más bajo de deviance residual, lo que indica que se ajusta mejor a los datos que el Modelo 1.

2. AIC (Criterio de Información de Akaike): AIC Modelo 1: 493.06 AIC Modelo 2: 468.97 Interpretación: Un valor más bajo del AIC indica un mejor modelo en términos de equilibrio entre bondad de ajuste y parsimonia (menos

sobreajuste). El Modelo 2 tiene un AIC más bajo, lo que sugiere que es un modelo más eficiente.

3. Significancia de los coeficientes: Modelo 1: Todos los coeficientes (woolB, tensionM, tensionH) son significativos con p-valores < 0.001. Modelo 2: Además de los coeficientes de las variables principales, los coeficientes de interacción (woolB) son significativos, aunque woolB:tensionH no lo es. Interpretación: El Modelo 2 captura interacciones significativas entre las variables, lo que podría proporcionar una representación más realista de los datos.

4. Prueba de Durbin-Watson (Independencia de residuos): Modelo 1: DW = 2.0332 (p-value = 0.7791), lo que sugiere que no hay autocorrelación significativa en los residuos. Modelo 2: DW = 2.2376 (p-value = 0.8499), también sugiere que no hay autocorrelación significativa. Interpretación: Ambos modelos cumplen con el supuesto de independencia de los residuos.

5. Goodness-of-fit para la distribución de Poisson: Modelo 1: Estadístico de bondad de ajuste = 210.39, con 50 grados de libertad. Modelo 2: Estadístico de bondad de ajuste = 182.31, con 48 grados de libertad. Interpretación: El Modelo 2 muestra un mejor ajuste según esta métrica.

Conclusión: El Modelo 2 (con interacción) es superior al Modelo 1 en términos de ajuste a los datos (menor deviance residual), menor AIC y captura de interacciones significativas. Aunque uno de los coeficientes de interacción no es significativo, el mejor ajuste global del Modelo 2 lo hace una mejor opción en este caso.