Regresión Poisson

Erika Martínez Meneses

2024-10-29

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. Como primer paso cargamos la base de datos y visualizamos las primeras filas.

```
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:

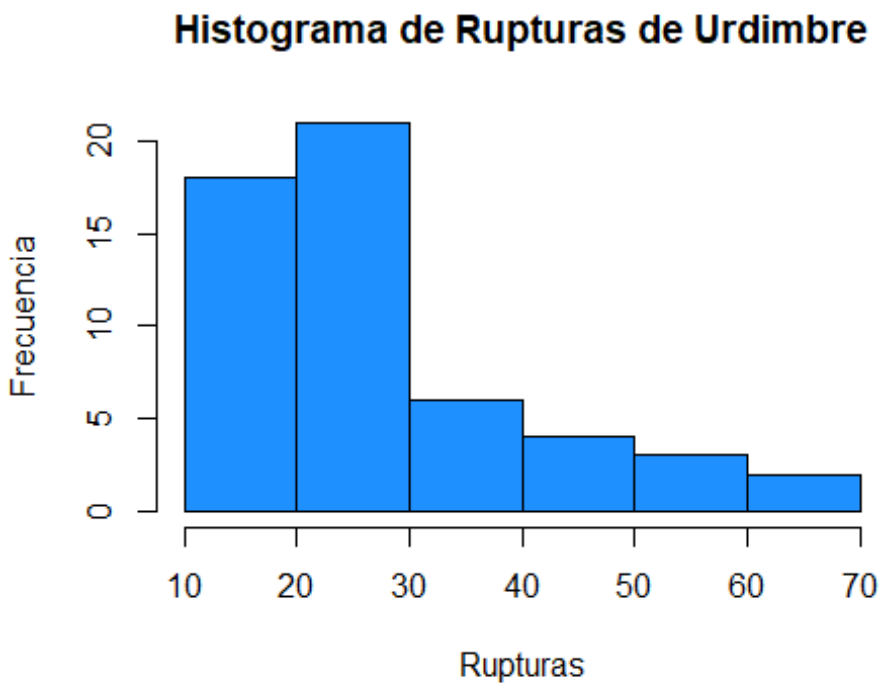
- breaks: número de rupturas
- wool: tipo de lana (A o B)
- tensión: el nivel de tensión (L, M, H)

Se sigue el siguiente procedimiento de análisis:

I. Análisis Descriptivo

Histograma del número de rupturas

```
hist(data$breaks, main = "Histograma de Rupturas de Urdimbre", xlab =
"Rupturas", ylab = "Frecuencia", col = "dodgerblue1")
```



Este histograma muestra la distribución de rupturas de urdimbre, concentrada en valores bajos (entre 10 y 30 rupturas), lo cual sugiere que las rupturas son eventos relativamente poco frecuentes. Esto hace que la variable sea adecuada para un modelo de regresión Poisson, ideal para datos de conteo.

En una regresión Poisson, podríamos analizar cómo factores como el tipo de material y la tensión afectan la frecuencia de rupturas. Sin embargo, dado el sesgo a la derecha, podría haber sobredispersión (varianza mayor que la media). Si esto ocurre, un modelo de Binomial Negativa podría ser una mejor opción para ajustar la relación entre estos factores y la frecuencia de rupturas.

Obtén la media y la varianza de la variable dependiente

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

```
## Media de rupturas: 28.14815
## Varianza de rupturas: 174.2041
```

La regresión Poisson asume que la media y la varianza de la variable dependiente son aproximadamente iguales. La alta diferencia entre la media y la varianza sugiere que el modelo Poisson podría tener problemas de ajuste debido a una posible sobredispersión.

II. Modelos de Regresión Poisson

Modelo de regresión Poisson sin interacción

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

```
##
## 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
```

- woolB: La lana tipo B reduce las rupturas en comparación con la lana tipo A (coeficiente = -0.20599).
- tensionM: Una tensión media reduce significativamente las rupturas en comparación con una tensión baja (coeficiente = -0.32132).
- tensionH: Una tensión alta reduce aún más las rupturas en comparación con una baja (coeficiente = -0.51849).

Estos coeficientes sugieren que tanto el tipo de lana como el nivel de tensión afectan la tasa de rupturas.

Modelo de regresión Poisson con interacción

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

```
##
## 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
```

- La interacción woolB:tensionM es significativa ya que tiene un valor p = 1.75e-07, lo que indica que la combinación de lana tipo B y tensión media afecta las rupturas de manera distinta a cada factor por separado.
- La interacción woolB:tensionH no es significativa, tiene un valor p = 0.147, siendo mayor que 0.05, sugiriendo que lana tipo B y alta tensión no tienen un efecto conjunto adicional.

Este modelo captura la interacción entre los factores, proporcionando una visión más detallada de cómo cada combinación específica de wool y tension afecta las rupturas.

Modelo obtenido.

```
cat("Modelo Poisson sin interacción:\n", "Breaks = ",
coef(poisson_model)[1], " + ", ... = coef(poisson_model)[2], "* woolB", "
+ ",coef(poisson_model)[3], "* tensionM", " + ", coef(poisson_model)[4],
"* tensionH", "\n\n")

## Modelo Poisson sin interacción:
##  Breaks =  3.691963  +  -0.2059884 * woolB  +  -0.3213204 * tensionM
+  -0.5184885 * tensionH

cat("Modelo Poisson con interacción:\n", "Breaks = ",
coef(poisson_model_inter)[1], " + ",  coef(poisson_model_inter)[2], "*
woolB", " + ",  coef(poisson_model_inter)[3], "* tensionM", " + ",
coef(poisson_model_inter)[4], "* tensionH", " + ",
coef(poisson_model_inter)[5], "* woolB:tensionM", " + ",
coef(poisson_model_inter)[6], "* woolB:tensionH", "\n\n")

## Modelo Poisson con interacción:
##  Breaks =  3.796737  +  -0.4566272 * woolB  +  -0.618683 * tensionM  +
```

```
-0.5957987 * tensionH  +  0.6381768 * woolB:tensionM  +  0.1883632 *
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.
- AIC: Criterio de Aikaike
- Comparación entre los coeficientes y los errores estándar de de ambos modelos

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:

- H_0: Devianción = 0
- H_1: Devianción > 0
- gl = gl_desviación residual (n-(p+1))

Valor frontera de la zona de rechazo

Sin interración

```
gl <- summary_poisson$null.deviance - summary_poisson$df.residual
qchisq(0.05, gl)

## [1] 211.9578
```

Con interración

```
gl2 <- summary_poisson_inter$null.deviance -
summary_poisson_inter$df.residual
qchisq(0.05, gl2)

## [1] 213.81
```

Estadístico de prueba y valor p

Sin interración

```r
dr <- summary_poisson$deviance
cat("Estadístico de Prueba =", dr, "\n")

## Estadístico de Prueba = 210.3919

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

## Valor p = 0.9575667
```

Con interacción

```r
dr2 <- summary_poisson_inter$deviance
cat("Estadístico de Prueba =", dr2, "\n")

## Estadístico de Prueba = 182.3051

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

## Valor p = 0.999506
```

Desviación Residual: La desviación residual es una medida de qué tan bien el modelo ajusta los datos. Un valor bajo de desviación residual sugiere un buen ajuste.

- En el modelo sin interacción, la desviación residual es 210.39, con 50 grados de libertad.
- En el modelo con interacción, la desviación residual es menor (182.31), con 48 grados de libertad.

Esto indica que el modelo con interacción proporciona un mejor ajuste, ya que tiene una desviación residual más baja.

La prueba de X^2 nos permite evaluar si la desviación residual del modelo es significativamente diferente de cero. En este contexto:

- Para el modelo sin interacción, el valor p de la prueba es 0.957, indicando que no podemos rechazar la hipótesis nula.
- Para el modelo con interacción, el valor p es aún mayor (0.9995), lo cual también sugiere que el modelo explica adecuadamente los datos.

Ambos modelos tienen un ajuste estadísticamente significativo, pero el modelo con interacción muestra un ajuste ligeramente superior.

Compara los AIC de cada modelo.

El AIC nos permite comparar modelos: un menor valor de AIC indica un mejor balance entre la calidad del ajuste y la complejidad del modelo.

```r
AIC(poisson_model)
```

```
## [1] 493.056
```

```
AIC(poisson_model_inter)
```

```
## [1] 468.9692
```

- El AIC del modelo sin interacción es 493.06.
- El AIC del modelo con interacción es 468.97.

Dado que el AIC es significativamente menor en el modelo con interacción, este modelo es preferible en términos de parsimonia y ajuste.

Compara los coeficientes

Los coeficientes indican el efecto de cada variable en el número de rupturas (breaks), en términos de log-ratio.

Parámetro	Modelo sin Interacción	Error Estándar	Modelo con Interacción	Error Estándar
Intercepto	3.69196	0.04541	3.79674	0.04994
woolB	-0.20599	0.05157	-0.45663	0.08019
tensionM	-0.32132	0.06027	-0.61868	0.08440
tensionH	-0.51849	0.06396	-0.59580	0.08378
woolBtensionM	-	-	0.63818	0.12215
woolBtensionH	-	-	0.18836	0.12990

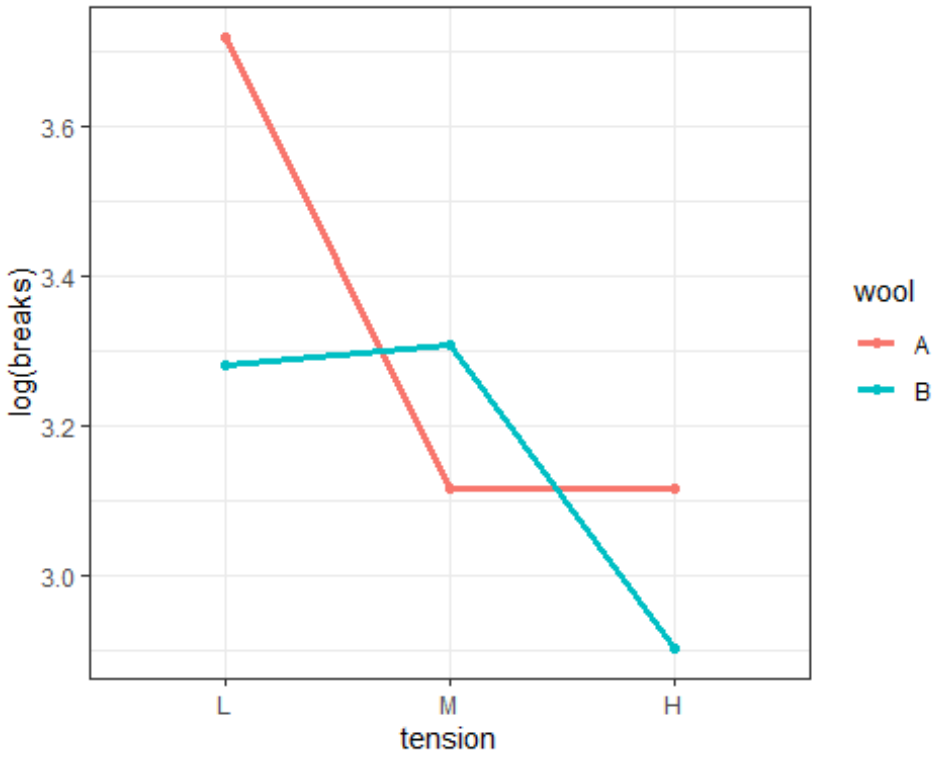
- En el modelo sin interacción, los coeficientes de woolB y tension son negativos, lo que sugiere que el tipo de lana B y los niveles de tensión M y H reducen el número de rupturas en comparación con las categorías de referencia (woolA y tensionL).
- En el modelo con interacción, los coeficientes muestran además el efecto combinado de wool y tension. En particular, el término woolB:tensionM es positivo, lo que sugiere que cuando se usa lana B con tensión M, el efecto en el número de rupturas es menos negativo (o incluso positivo) en comparación con los efectos individuales de woolB y tensionM.

Los errores estándar son menores en el modelo con interacción para algunos términos clave, lo cual indica una estimación más precisa de los efectos individuales y combinados. Esto refuerza la elección del modelo con interacción, ya que proporciona una mayor precisión en los coeficientes estimados.

Interpretación de la Interacción

Se visualiza la interacción entre wool y tension para entender cómo afecta la tasa de rupturas:

```
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"))
```



En la gráfica se observa el efecto de la interacción entre wool (tipo de lana) y tension (nivel de tensión) en la variable log(breaks).

- Para tensión baja (L): La lana A presenta un valor de log(breaks) más alto que la lana B, lo que sugiere que hay más rupturas con la lana A bajo baja tensión.
- Para tensión media (M): La lana B tiene un valor de log(breaks) ligeramente superior al de la lana A. Esto indica que, bajo tensión media, la lana B experimenta más rupturas que la lana A.

- Para tensión alta (H): La lana B muestra una disminución significativa en el número de rupturas, alcanzando un valor de log(breaks) más bajo que la lana A, lo que indica que, con tensión alta, la lana B es más resistente a las rupturas en comparación con la lana A.

En concluisón la interacción entre wool y tension indica que el efecto de cada tipo de lana en el número de rupturas depende del nivel de tensión. En tensión baja, la lana A genera más rupturas, pero en tensión media, la lana B tiene una tasa de rupturas ligeramente mayor. Finalmente, en tensión alta, la lana B tiene menos rupturas que la lana A, lo cual sugiere que la lana B se comporta mejor en condiciones de alta tensión.

Define cuál de los dos es un mejor modelo.

IV. Evaluación de los supuestos

Los supuestos principales que se deben cumplir son:

Independencia

Prueba de hipótesis (prueba de Durbin-Watson)

H_0: No existe autocorrelación en los residuos. H_1: Existe autocorrelación en los residuos.

```
library(lmtest)
```

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

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

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

```
dw_test_no_inter <- dwtest(poisson_model, alternative = "two.sided")
cat("Modelo Poisson sin interacción:\n")
```

```
## Modelo Poisson sin interacción:
```

```
print(dw_test_no_inter)
```

```
##
##  Durbin-Watson test
##
## data:  poisson_model
## DW = 2.0332, p-value = 0.7791
## alternative hypothesis: true autocorrelation is not 0
```

```r
dw_test_inter <- dwtest(poisson_model_inter, alternative = "two.sided")
cat("\nModelo Poisson con interacción:\n")
```

```
##
## Modelo Poisson con interacción:
```

```r
print(dw_test_inter)
```

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

Un valor de Durbin-Watson cercano a 2 indica independencia (no hay autocorrelación significativa). Valores significativamente distintos de 2 (cerca de 0 o 4) pueden indicar problemas de autocorrelación, sugiriendo que los residuos no son independientes. Ambos valores de Durbin-Watson cercanos a 2 sugieren independencia en los residuos.

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 X^2 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:

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

## 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_model_inter)
```

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

Ambos modelos presentan evidencia de sobredispersión, lo cual indica que el modelo Poisson podría no ser el más adecuado, y podría ser necesario considerar un modelo de Poisson con sobre-dispersión o un modelo binomial negativa.

Recurrimos a los siguientes modelos para observar si encontramos un mejor modelo

Otros Modelos

Modelo cuasi Poisson

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

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

Este modelo estima el número de rupturas (breaks) con una distribución cuasi Poisson, permitiendo manejar la sobredispersión en los datos (como lo indica el parámetro de dispersión de 4.26, que sugiere variabilidad mayor a la esperada bajo una distribución Poisson estándar).

- El término Intercept tiene un estimador de 3.692, lo que implica que el número de rupturas promedio en el logaritmo cuando wool = A y tension = L es aproximadamente 3.69.
- El coeficiente asociado a woolB es -0.206. Esto indica que, cuando se usa lana tipo B (comparada con lana tipo A) y manteniendo constante la tensión, se espera una ligera disminución en el número de rupturas (aunque no es estadísticamente significativo al 5%, con un p-valor de 0.059).
- El coeficiente para tensionM es -0.321, indicando que al pasar de tensión baja (L) a media (M) se espera una disminución significativa en el número de rupturas. Este efecto es estadísticamente significativo ($p < 0.05$).
- El coeficiente para tensionH es -0.518, lo que sugiere una reducción aún mayor en el número de rupturas cuando se incrementa la tensión a alta (H), comparado con baja (L). Esta relación es significativa ($p < 0.001$).

En el modelo cuasi Poisson, la lana B parece reducir el número de rupturas en comparación con la lana A, aunque esta reducción no es estadísticamente significativa. En cambio, los incrementos en el nivel de tensión sí reducen significativamente el número de rupturas.

Modelo Binomial Negativa (intenta imaginar qué es lo que cambia en este modelo con respecto al Poisson):

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

##
## 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
```

Este modelo ajusta el número de rupturas (breaks) utilizando una distribución binomial negativa, lo cual es adecuado para manejar la sobredispersión en los datos, con el parámetro theta de 12.08 indicando la sobreabundancia de varianza.

- Similar al modelo cuasi Poisson, el valor estimado del intercepto (3.797) representa el número esperado de rupturas en el logaritmo cuando wool = A y tension = L.
- El coeficiente para woolB es -0.457, indicando una reducción significativa en el número de rupturas cuando se usa lana B en lugar de lana A ($p < 0.01$).
- El coeficiente de tensionM es -0.619, lo que indica que el cambio de tensión baja (L) a media (M) produce una disminución significativa en el número de rupturas ($p < 0.001$).
- El coeficiente de tensionH es -0.596, indicando una reducción significativa en el número de rupturas cuando la tensión es alta (H), en comparación con baja (L) ($p < 0.001$).
- woolB: El coeficiente de 0.638 indica que la combinación de lana B y tensión media (M) aumenta el número de rupturas en comparación con la lana A y tensión media. Este efecto es estadísticamente significativo ($p < 0.01$).

- woolB: El coeficiente de 0.188 sugiere un leve aumento en las rupturas con lana B y tensión alta (H) en comparación con lana A, aunque este efecto no es estadísticamente significativo (p = 0.416).

En el modelo de binomial negativa, tanto el tipo de lana como el nivel de tensión influyen significativamente en el número de rupturas. En general, la lana B reduce el número de rupturas, excepto cuando se combina con tensión media, donde se observa un aumento significativo de rupturas en comparación con la lana A.

Ambos modelos sugieren que el nivel de tensión tiene un efecto significativo en la cantidad de rupturas, mientras que el tipo de lana también afecta, pero de forma menos consistente.

Valor frontera de la zona de rechazo

Cuasi

```
gl3 <- summary(poisson_model_quasi)$null.deviance -
summary(poisson_model_quasi)$df.residual
qchisq(0.05, gl3)

## [1] 211.9578
```

Binomial Negativa

```
gl4 <- summary(poisson_model_nb)$null.deviance -
summary(poisson_model_nb)$df.residual
qchisq(0.05, gl4)

## [1] 25.49922
```

Estadístico de prueba y valor p

Cuasi

```
dr3 <- summary(poisson_model_quasi)$deviance
cat("Estadístico de Prueba =", dr3, "\n")

## Estadístico de Prueba = 210.3919

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

## Valor p = 0.9575667
```

Binomial negativa

```
dr4 <- summary(poisson_model_nb)$deviance
cat("Estadístico de Prueba =", dr4, "\n")

## Estadístico de Prueba = 53.50616
```

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

## Valor p = 0.05777862
```

Independencia

```r
dw_test_quasi <- dwtest(poisson_model_quasi, alternative = "two.sided")
cat("Modelo cuasi Poisson:\n")

## Modelo cuasi Poisson:

print(dw_test_quasi)

##
##  Durbin-Watson test
##
## data:  poisson_model_quasi
## DW = 2.0332, p-value = 0.7791
## alternative hypothesis: true autocorrelation is not 0
```

En este caso, el DW de 2.0332, junto con el p-valor alto (0.7791), indica que no hay evidencia de autocorrelación significativa en los residuos del modelo cuasi Poisson. Esto es positivo, ya que la ausencia de autocorrelación en los residuos es un supuesto clave para la validez del modelo.

```r
dw_test_nb <- dwtest(poisson_model_nb, alternative = "two.sided")
cat("\nModelo Binomial Negativa:\n")

##
## Modelo Binomial Negativa:

print(dw_test_nb)

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

Al igual que en el modelo cuasi Poisson, el estadístico DW está cerca de 2, y el p-valor alto (0.8499) indica que no hay evidencia de autocorrelación significativa en los residuos del modelo binomial negativa. Esto sugiere que los residuos del modelo están independientemente distribuidos, cumpliendo con el supuesto de no autocorrelación.

Sobredispersión de los residuos

```r
poisgof(poisson_model_nb)

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

```
## [1] 53.50616
##
## $df
## [1] 48
##
## $p.value
## [1] 0.2711637
```

Esta prueba evalúa si el modelo binomial negativa se ajusta bien a los datos. El p-valor de 0.2712 es mayor al nivel de significancia típico de 0.05, lo que sugiere que no hay evidencia suficiente para rechazar la hipótesis nula de que el modelo se ajusta bien a los datos. En otras palabras, el modelo binomial negativa proporciona un ajuste adecuado a los datos.

Ambos modelos, cuasi Poisson y binomial negativa, cumplen con el supuesto de independencia de los residuos, como indica la prueba de Durbin-Watson en cada caso. Además, el modelo binomial negativa pasa la prueba de bondad de ajuste (poisgof), lo que sugiere que este modelo se ajusta bien a los datos, probablemente capturando mejor la sobredispersión en comparación con el modelo cuasi Poisson.

V. Define cuál es tu mejor modelo

Se evaluaron cuatro modelos de regresión para modelar la variable de respuesta breaks en función de las variables wool y tension. Los modelos considerados fueron:

1. Modelo de Regresión Poisson sin interacción
2. Modelo de Regresión Poisson con interacción
3. Modelo Cuasi Poisson
4. Modelo de Binomial Negativa

- Modelo Poisson sin Interacción: Este modelo mostró un buen ajuste inicial, pero presentó problemas de sobredispersión, como lo indica el valor de la deviance residual comparado con los grados de libertad. Adicionalmente, el test de Durbin-Watson arrojó un valor de $p = 0.7791$, lo cual indica que no hay autocorrelación significativa. Sin embargo, la prueba de bondad de ajuste (poisgof) tuvo un valor p extremadamente bajo ($p < 0.05$), lo que sugiere que el modelo Poisson no es adecuado para los datos.
- Modelo Poisson con Interacción: La inclusión de la interacción entre wool y tension mejoró el ajuste del modelo, reduciendo la deviance residual. Sin embargo, al igual que el modelo sin interacción, este también mostró problemas de ajuste según la prueba de bondad de ajuste (poisgof), con un valor p menor a 0.05. A pesar de la mejora en el AIC respecto al modelo sin interacción (468.97 vs 493.06), el problema de sobredispersión persiste.

- Modelo Cuasi Poisson: Este modelo aborda la sobredispersión al ajustar el parámetro de dispersión. La deviance residual y el test de bondad de ajuste indican un mejor ajuste en comparación con los modelos Poisson, y el valor p de la prueba de bondad de ajuste es alto ($p = 0.957$), lo cual sugiere un ajuste adecuado.

- Modelo Binomial Negativa: Este modelo resultó ser el más adecuado para los datos. Al igual que el modelo Poisson con interacción, incluye la interacción entre wool y tension, pero aborda de manera más efectiva la sobredispersión. La deviance residual es la más baja entre todos los modelos (53.506), y la prueba de bondad de ajuste (poisgof) muestra un valor p alto ($p = 0.271$), indicando un buen ajuste. Además, el AIC es el más bajo (405.12), lo cual es un fuerte indicador de que este modelo proporciona el mejor balance entre ajuste y parsimonia.

Selección del Mejor Modelo

En base a los resultados de los distintos indicadores de ajuste (deviance, AIC, y pruebas de bondad de ajuste), se concluye que el Modelo de Binomial Negativa es el mejor modelo para este conjunto de datos. Este modelo maneja efectivamente la sobredispersión y proporciona un ajuste adecuado sin problemas de autocorrelación, además de incluir la interacción entre wool y tension, que ha demostrado ser significativa.

En conclusión el modelo de Binomial Negativa con interacción entre wool y tension es el más adecuado para explicar la variabilidad en breaks. Este modelo permite una interpretación robusta y precisa de los efectos de las variables wool y tension sobre breaks, haciendo frente a la sobredispersión y optimizando el balance entre ajuste y simplicidad del modelo.