Actividad 6: Regresión Poisson

Daniela Jiménez Téllez

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.

Importación de datos

```
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:

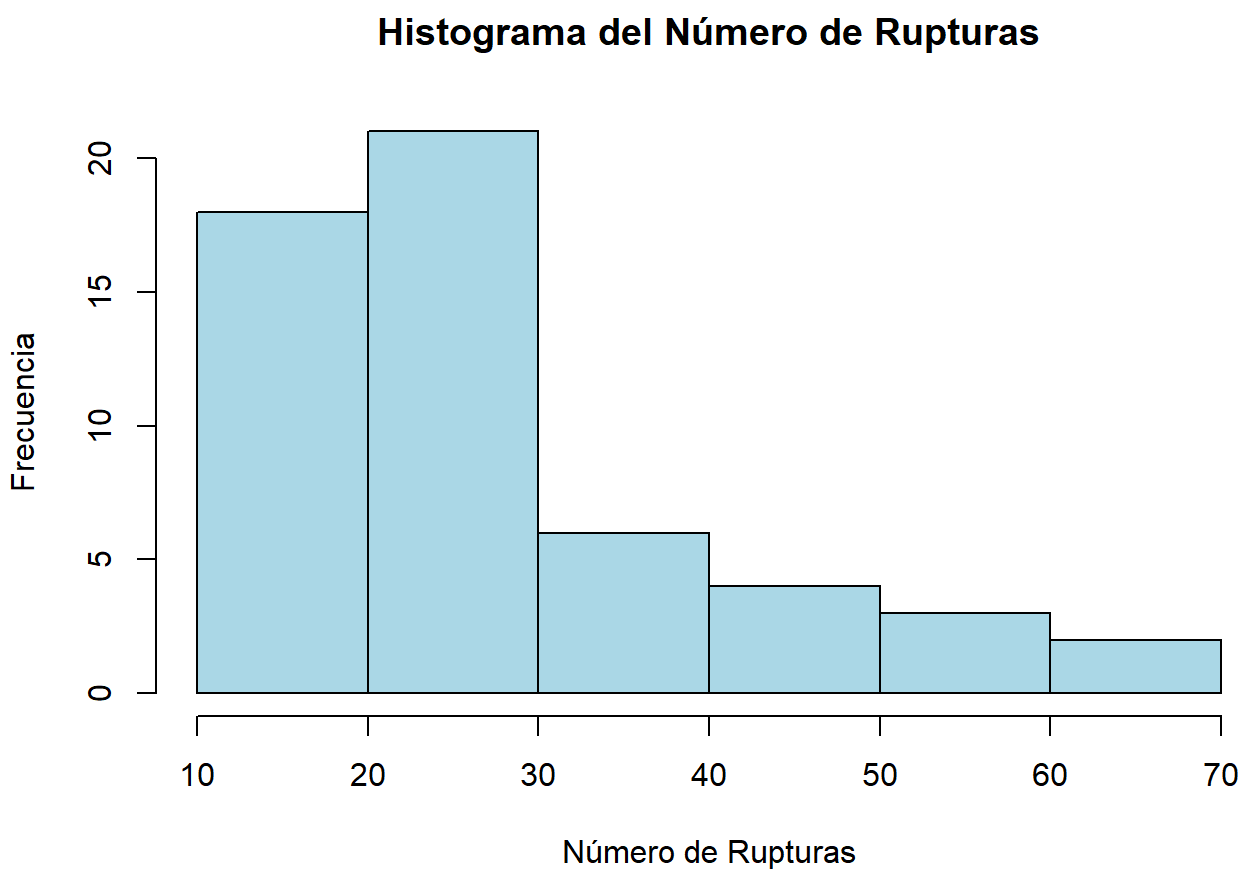
- breaks: número de rupturas
- wool: tipo de lana (A o B)
- tensión: el nivel de tensión (L, M, H)

Sigue el siguiente procedimiento de análisis:

Análisis descriptivo

- Histograma del número de rupturas
- Obtén la media y la varianza de la variable dependiente
- Interpreta en el contexto de una Regresión Poisson

```
# Histograma

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




```
# Media y varianza de la variable dependiente

mean_breaks <- mean(data$breaks)
var_breaks <- var(data$breaks)

cat("Media del número de rupturas:", mean_breaks, "\n")
```

```
## Media del número de rupturas: 28.14815
```

```
cat("Varianza del número de rupturas:", var_breaks, "\n")
```

```
## Varianza del número de rupturas: 174.2041
```

En una Regresión Poisson, se modela una variable de conteo como función de una o más variables independientes. La distribución Poisson asume que la media y la varianza de la variable dependiente deberían ser muy parecidas. Si la varianza es significativamente mayor que la media, podría decirnos que hay sobredispersión. En este caso, la media y la varianza están muy alejadas entre sí, siendo la varianza más grande que la media.

Ajuste de modelos de Regresión Poisson

- Ajusta el modelo de regresión Poisson sin interacción
- Ajusta el modelo de regresión Poisson con interacción

- 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.

```
# Modelo sin interacción

modelo_sin_interaccion <- glm(breaks ~ wool + tension, data = data, family = poisson(link = "lo
g"))

cat("----- MODELO SIN INTERACIÓN ----- \n")
```

```
## ----- MODELO SIN INTERACIÓN -----
```

```
summary(modelo_sin_interaccion)
```

```
##
## 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
```

```
# Modelo con interacción

modelo_con_interaccion <- glm(breaks ~ wool * tension, data = data, family = poisson(link = "lo
g"))

cat("----- MODELO CON INTERACIÓN ----- \n")
```

```
## ----- MODELO CON INTERACIÓN -----
```

```
summary(modelo_con_interaccion)
```

```
##
## 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
```

En esta sección se hicieron 2 modelos: uno con interacción y uno sin. En el modelo sin interacción, cada variable independiente (wool y tension) fue representada con variables Dummy, donde wool incluye dos categorías (“A” y “B”) y tension tres niveles (“L”, “M”, “H”). En este modelo, cada coeficiente muestra el cambio en el logaritmo del valor esperado de breaks asociado con un cambio de categoría en una sola variable, manteniendo la otra constante. En este caso, los coeficientes $\beta_1, \beta_2, \mathrm{y} \beta_3$ representan los efectos de wool y tension en comparación con sus referencias.

Por otro lado, el modelo con interacción permite analizar cómo el efecto de wool cambia según el nivel de tension. En este caso, se muestra como $\beta_4 \mathrm{y} \beta_5$ son la combinación de los efectos de las dos variables independientes. La interacción permite observar que el impacto de wool en el número de breaks depende de los niveles de tension, lo que nos dice que hay una relación no aditiva entre ambas variables.

Selección de 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

1. 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:

$$H_0 : Deviance = 0$$

$$H_1 : Deviance > 0$$

$$gl = \text{gl desviación residual}(n - (p + 1))$$

```
# Modelo sin interacción

S_sin_interaccion <- summary(modelo_sin_interaccion)

gl_sin_interaccion <- S_sin_interaccion$df.null - S_sin_interaccion$df.residual
cat("Modelo sin interacción:\n")
```

```
## Modelo sin interacción:
```

```
cat("Grados de libertad (gl):", gl_sin_interaccion, "\n")
```

```
## Grados de libertad (gl): 3
```

```
valor_frontera_sin_interaccion <- qchisq(0.05, gl_sin_interaccion, lower.tail = FALSE)
cat("Valor frontera para la zona de rechazo (alfa = 0.05):", valor_frontera_sin_interaccion,
"\n")
```

```
## Valor frontera para la zona de rechazo (alfa = 0.05): 7.814728
```

```
dr_sin_interaccion <- S_sin_interaccion$deviance
cat("Estadístico de prueba (desviación residual):", dr_sin_interaccion, "\n")
```

```
## Estadístico de prueba (desviación residual): 210.3919
```

```
vp_sin_interaccion <- 1 - pchisq(dr_sin_interaccion, gl_sin_interaccion)
cat("Valor p:", vp_sin_interaccion, "\n\n")
```

```
## Valor p: 0
```

```
# Modelo con interacción

S_con_interaccion <- summary(modelo_con_interaccion)

gl_con_interaccion <- S_con_interaccion$df.null - S_con_interaccion$df.residual
cat("Modelo con interacción:\n")
```

```
## Modelo con interacción:
```

```
cat("Grados de libertad (gl):", gl_con_interaccion, "\n")
```

```
## Grados de libertad (gl): 5
```

```
valor_frontera_con_interaccion <- qchisq(0.05, gl_con_interaccion, lower.tail = FALSE)
cat("Valor frontera para la zona de rechazo (alfa = 0.05):", valor_frontera_con_interaccion,
"\n")
```

```
## Valor frontera para la zona de rechazo (alfa = 0.05): 11.0705
```

```
dr_con_interaccion <- S_con_interaccion$deviance
cat("Estadístico de prueba (desviación residual):", dr_con_interaccion, "\n")
```

```
## Estadístico de prueba (desviación residual): 182.3051
```

```
vp_con_interaccion <- 1 - pchisq(dr_con_interaccion, gl_con_interaccion)
cat("Valor p:", vp_con_interaccion, "\n")
```

```
## Valor p: 0
```

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

```
# Desviación residual

cat("Desviación Residual (sin interacción):", deviance(modelo_sin_interaccion), "\n")
```

```
## Desviación Residual (sin interacción): 210.3919
```

```
cat("Desviación Residual (con interacción):", deviance(modelo_con_interaccion), "\n")
```

```
## Desviación Residual (con interacción): 182.3051
```

```
# AIC

cat("AIC (sin interacción):", AIC(modelo_sin_interaccion), "\n")
```

```
## AIC (sin interacción): 493.056
```

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

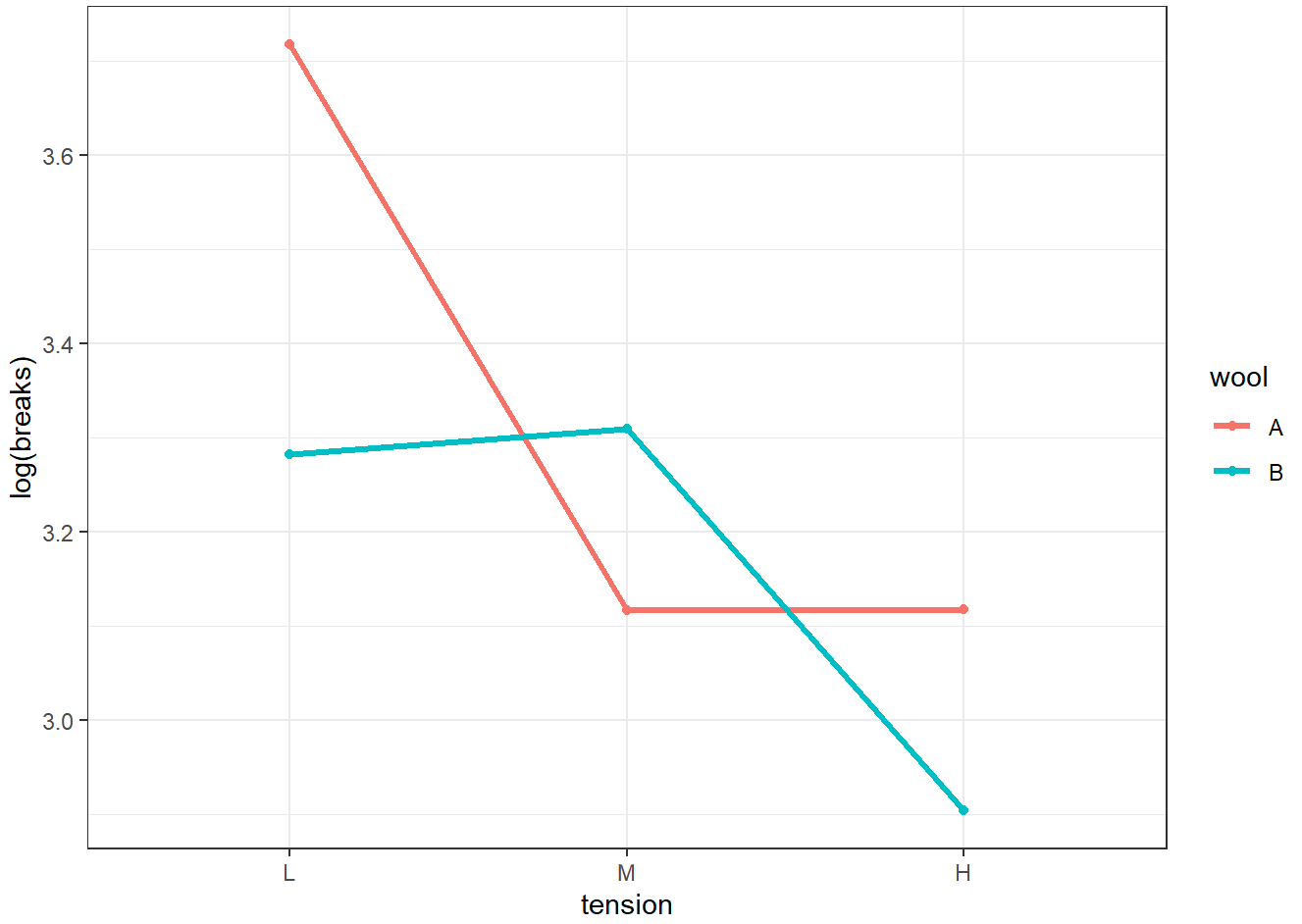
```
## AIC (con interacción): 468.9692
```

- 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 β_i de ambos modelos (haz una tabla para que se facilite la comparación)
    - Interpreta los coeficientes de ambos modelos.

```
coef_sin_interaccion <- summary(modelo_sin_interaccion)$coefficients
coef_con_interaccion <- summary(modelo_con_interaccion)$coefficients

tabla_sin_interaccion <- data.frame(
  Modelo = "Sin Interacción",
  Coeficiente = rownames(coef_sin_interaccion),
  Estimación = coef_sin_interaccion[, "Estimate"],
  Error_Estandar = coef_sin_interaccion[, "Std. Error"])

tabla_con_interaccion <- data.frame(
  Modelo = "Con Interacción",
  Coeficiente = rownames(coef_con_interaccion),
  Estimación = coef_con_interaccion[, "Estimate"],
  Error_Estandar = coef_con_interaccion[, "Std. Error"])

tabla_comparacion <- rbind(tabla_sin_interaccion, tabla_con_interaccion)

print(tabla_comparacion)
```

```
##                         Modelo    Coeficiente Estimación Error_Estandar
## (Intercept)    Sin Interacción    (Intercept)  3.6919631     0.04541069
## woolB          Sin Interacción          woolB -0.2059884     0.05157117
## tensionM       Sin Interacción       tensionM -0.3213204     0.06026580
## tensionH       Sin Interacción       tensionH -0.5184885     0.06395944
## (Intercept)1   Con Interacción    (Intercept)  3.7967368     0.04993753
## woolB1         Con Interacción          woolB -0.4566272     0.08019202
## tensionM1      Con Interacción       tensionM -0.6186830     0.08440012
## tensionH1      Con Interacción       tensionH -0.5957987     0.08377723
## woolB:tensionM Con Interacción woolB:tensionM  0.6381768     0.12215312
## woolB:tensionH Con Interacción woolB:tensionH  0.1883632     0.12989529
```

```
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"))
```



- Define cuál de los dos es un mejor modelo.

Con base en los análisis anteriores se puede decir que:

- *Desviación Residual:* La desviación residual del modelo sin interacción es 210.39, mientras que la del modelo con interacción es 182.31. Dado que una desviación residual más baja quiere decir que hay un mejor ajuste a los datos, el modelo con interacción es mejor tomando en cuenta este criterio.
- *AIC:* El AIC del modelo sin interacción es 493.06 y el del modelo con interacción es de 468.97. Al igual que la desviación residual, un AIC menor es un mejor ajuste. Por lo tanto, el mejor modelo de acuerdo a este análisis es el modelo con interacción.
- *Significancia de los coeficientes y errores estándar:* En este caso, podemos ver que en el modelo con interacción se incluyen más términos para las combinaciones de wool y tension. La tabla de coeficientes muestra que los coeficientes de interacción son significativos y presentan errores estándar normales, por lo que al igual que los análisis anteriores, el mejor modelo es el que tiene interacción.
- *Gráfica:* Para finalizar se tiene que la gráfica muestra que los efectos de tension sobre el logaritmo del número de breaks varían dependiendo de la cantidad de wool, lo que muestra que la interacción refleja bien el comportamiento.

Por lo tanto, el mejor modelo es el que tiene interacción.

Evaluación de los supuestos

Los supuestos principales que se deben cumplir son:

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

- 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:

$$H_0 : \text{No hay una sobredispersión del modelo}$$

$$H_1 : \text{Hay una sobredispersión del modelo}$$

```
# Prueba de independencia

cat("----- MODELO SIN INTERACIÓN ----- \n")
```

```
## ----- MODELO SIN INTERACIÓN -----
```

```
dwtest(modelo_sin_interaccion, alternative = "two.sided")
```

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

```
cat("----- MODELO CON INTERACIÓN ----- \n")
```

```
## ----- MODELO CON INTERACIÓN -----
```

```
dwtest(modelo_con_interaccion, alternative = "two.sided")
```

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

```
# Prueba de sobredispersión

cat("----- MODELO SIN INTERACIÓN ----- \n")
```

```
## ----- MODELO SIN INTERACIÓN -----
```

```
poisgof(modelo_sin_interaccion)
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 210.3919
##
## $df
## [1] 50
##
## $p.value
## [1] 1.44606e-21
```

```
cat("----- MODELO CON INTERACIÓN ----- \n")
```

```
## ----- MODELO CON INTERACIÓN -----
```

```
poisgof(modelo_con_interaccion)
```

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

Con base en los resultados anteriores se puede observar que en ambos modelos existe una sobredispersión en los datos ya que tienen valores p muy bajos en la prueba de bondad de ajuste, lo que nos dice que la varianza de los residuos es mayor que la media, lo que contradice el supuesto de igualdad entre media y varianza del modelo Poisson.

- Si hay un mal modelo, recurre a usar:
    - Modelo cuasi Poisson
    - Modelo Binomial Negativa (intenta imaginar qué es lo que cambia en este modelo con respecto al Poisson)

```
# Quasi Poisson

poisson.model3 <- glm(breaks ~ wool * tension, data = data, family = quasipoisson(link = "log"))

cat("----- QUASI POISSON ----- \n")
```

```
## ----- QUASI POISSON -----
```

```
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
```

```
cat("\n\n")
```

```
cat("Desviación Residual (Quasi-Poisson con interacción):", deviance(poisson.model3), "\n")
```

```
## Desviación Residual (Quasi-Poisson con interacción): 182.3051
```

```
cat("Grados de libertad residual (Quasi-Poisson con interacción):", df.residual(poisson.model3),
"\n")
```

```
## Grados de libertad residual (Quasi-Poisson con interacción): 48
```

```
cat("AIC (Quasi-Poisson con interacción):", AIC(poisson.model3), "\n")
```

```
## AIC (Quasi-Poisson con interacción): NA
```

```
cat("\n\n")
```

```
# Binomial negativa

bnm <- glm.nb(breaks ~ wool * tension, data = data, control = glm.control(maxit = 1000))

cat("----- BINOMIAL NEGATIVA ----- \n")
```

```
## ----- BINOMIAL NEGATIVA -----
```

```
summary(bnm)
```

```
##
## 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
```

```
cat("\n\n")
```

```
cat("Desviación Residual (Binomial Negativa con interacción):", deviance(bnm), "\n")
```

```
## Desviación Residual (Binomial Negativa con interacción): 53.50616
```

```
cat("Grados de libertad residual (Binomial Negativa con interacción):", df.residual(bnm), "\n")
```

```
## Grados de libertad residual (Binomial Negativa con interacción): 48
```

```
cat("AIC (Binomial Negativa con interacción):", AIC(bnm), "\n")
```

```
## AIC (Binomial Negativa con interacción): 405.1248
```

Definición de cuál es el mejor modelo

Finalmente, habiendo hecho todos estos análisis se puede concluir que el mejor modelo es Binomial Negativa, ya que en general tiene mejores resultados con respecto a los análisis. Este tiene una desviación residual mucho menor que el modelo Quasi-Poisson y tiene un AIC (el Quasi Poisson no). Igualmente, este modelo maneja adecuadamente la sobredispersión en los datos y tiene un ajuste superior en términos de desviación residual.