A6-Regresión Poisson

Luis Maximiliano López Ramírez

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.

```
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
```

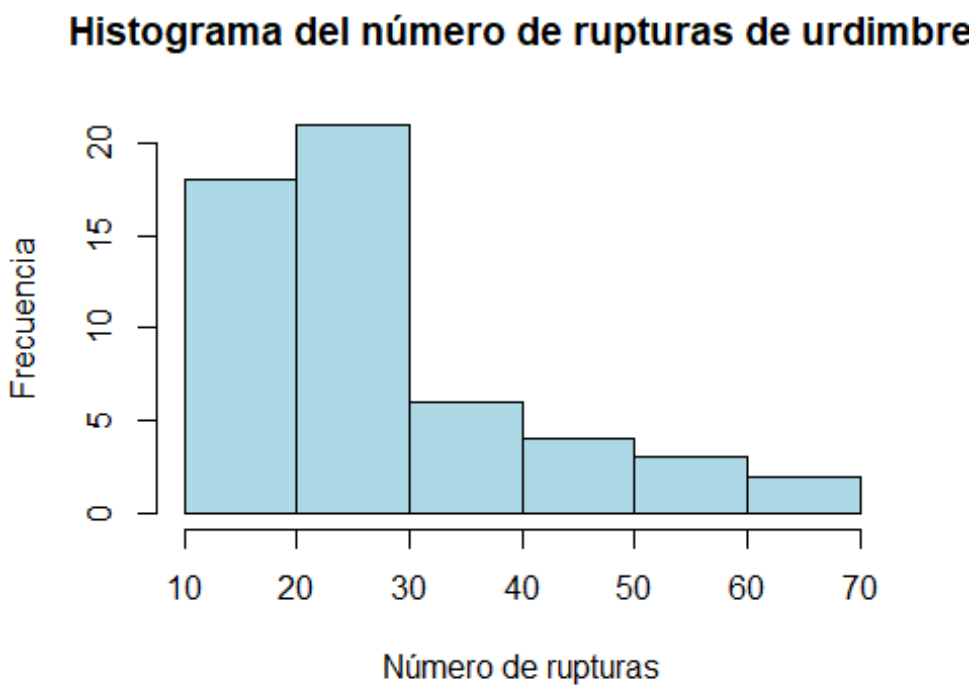
Este conjunto de datos indica cuántas roturas de urdimbre ocurrieron para diferentes tipos de telares por telar, por longitud fija de hilo:

- 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

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



Obtén la media y la varianza de la variable dependiente

```
# Calcular la media y la varianza de la variable dependiente 'breaks'
media_breaks <- mean(warpbreaks$breaks)
varianza_breaks <- var(warpbreaks$breaks)

# Mostrar resultados
cat("Media de las rupturas:", media_breaks, "\n")

## Media de las rupturas: 28.14815

cat("Varianza de las rupturas:", varianza_breaks, "\n")

## Varianza de las rupturas: 174.2041
```

Interpreta en el contexto de una Regresión Poisson

En el contexto de una regresión de Poisson, interpretamos la variable dependiente breaks, que representa el número de rupturas de urdimbre, como un conteo de eventos que ocurren dentro de una longitud fija de hilo. La regresión de Poisson es adecuada para modelar datos de conteo, ya que estos suelen seguir una distribución de Poisson, donde el valor esperado y la varianza tienden a ser iguales o similares.

Interpretación de la Media y la Varianza

Media: La media de breaks representa el número promedio de rupturas de urdimbre en el dataset. En una regresión de Poisson, esta media nos indica el valor esperado de rupturas bajo ciertas condiciones y es lo que intentamos modelar con las variables

predictoras (wool y tensión). Este valor promedio establece un punto de referencia para evaluar el efecto de las variables independientes en la frecuencia de rupturas.

Varianza: La varianza en los datos de breaks indica la dispersión o variabilidad de las rupturas en el dataset. En una distribución de Poisson, la varianza debería ser aproximadamente igual a la media. Si la varianza es notablemente mayor que la media (un fenómeno conocido como "sobredispersión"), podría ser una señal de que los datos no siguen una distribución de Poisson perfectamente, y podríamos considerar modelos alternativos, como la regresión de cuasi-Poisson o una regresión negativa binomial, que manejan mejor la sobredispersión.

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

# Resumen del modelo
summary(modelo_poisson)
```

```
##
## 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
```

Ajusta el modelo de regresión Poisson con interacción

```
# Ajustar el modelo de regresión Poisson sin interacción
modelo_poisson_interaccion <- glm(breaks ~ wool * tension, data =
warpbreaks, family = poisson(link = "log"))
```

```r
# Resumen del modelo
summary(modelo_poisson_interaccion)
```

```
##
## 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
```

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

woolB: El coeficiente de -0.20599 indica que, cuando la lana es de tipo B (en lugar de A), la tasa de rupturas esperada disminuye en aproximadamente un e ^ −0.20599 ≈ 0.814 o un 18.6%, manteniendo la tension constante en L.

tensionM: El coeficiente de -0.32132 sugiere que, cuando la tensión es M (media) en lugar de L (baja), la tasa de rupturas disminuye en aproximadamente un e ^ −0.32132 ≈ 0.725, o un 27.5%, manteniendo la lana en A.

tensionH: El coeficiente de -0.51849 implica que, cuando la tensión es H (alta) en lugar de L, la tasa de rupturas disminuye en aproximadamente un e ^ −0.51849 ≈ 0.596, o un 40.4%, manteniendo la lana en A.

Modelo con interacción

woolB: El coeficiente de -0.45663 indica que, en promedio, el cambio a lana de tipo B (en comparación con A) disminuye la tasa de rupturas en aproximadamente un e ^ −0.45663 ≈ 0.633, o un 36.7%, cuando la tensión es L.

tensionM: El coeficiente de -0.61868 implica que la tensión M reduce la tasa de rupturas en aproximadamente un e ^ −0.61868 ≈ 0.539, o un 46.1%, cuando se usa lana A.

tensionH: El coeficiente de -0.59580 sugiere que la tensión H reduce la tasa de rupturas en aproximadamente un e ^ −0.59580 ≈ 0.551, o un 44.9%, cuando la lana es A.

Interacción woolB:tensionM: El coeficiente positivo 0.63818 indica que el efecto combinado de wool = B y tension = M incrementa la tasa de rupturas en aproximadamente un e ^ 0.63818 ≈ 1.893, es decir, casi un 89.3% más de rupturas respecto al efecto aditivo individual de woolB y tensionM.

Interacción woolB:tensionH: El coeficiente 0.18836 sugiere que el efecto combinado de wool = B y tension = H incrementa la tasa de rupturas en un e ^ 0.18836 ≈ 1.207, es decir, un 20.7% adicional en comparación con los efectos individuales.

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

Desviación residual (Prueva 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 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))

```
# Modelo sin interacción
S_sin_interaccion <- summary(modelo_poisson)

# Calcular los grados de libertad para la desviación residual
gl_sin_interaccion <- S_sin_interaccion$df.null -
S_sin_interaccion$df.residual
# Valor frontera de la zona de rechazo al nivel de significancia del 5%
```

```r
valor_frontera_sin_interaccion <- qchisq(0.05, gl_sin_interaccion)
cat("Valor frontera sin interacción =", valor_frontera_sin_interaccion,
"\n")

## Valor frontera sin interacción = 0.3518463

# Estadístico de prueba y valor p para el modelo sin interacción
dr_sin_interaccion <- S_sin_interaccion$deviance
cat("Estadístico de prueba sin interacción =", dr_sin_interaccion, "\n")

## Estadístico de prueba sin interacción = 210.3919

vp_sin_interaccion <- 1 - pchisq(dr_sin_interaccion, gl_sin_interaccion)
cat("Valor p sin interacción =", vp_sin_interaccion, "\n")

## Valor p sin interacción = 0

# Modelo con interacción
S_con_interaccion <- summary(modelo_poisson_interaccion)

# Calcular los grados de libertad para la desviación residual
gl_con_interaccion <- S_con_interaccion$df.null -
S_con_interaccion$df.residual
# Valor frontera de la zona de rechazo al nivel de significancia del 5%
valor_frontera_con_interaccion <- qchisq(0.05, gl_con_interaccion)
cat("Valor frontera con interacción =", valor_frontera_con_interaccion,
"\n")

## Valor frontera con interacción = 1.145476

# Estadístico de prueba y valor p para el modelo con interacción
dr_con_interaccion <- S_con_interaccion$deviance
cat("Estadístico de prueba con interacción =", dr_con_interaccion, "\n")

## Estadístico de prueba con interacción = 182.3051

vp_con_interaccion <- 1 - pchisq(dr_con_interaccion, gl_con_interaccion)
cat("Valor p con interacción =", vp_con_interaccion, "\n")

## Valor p con interacción = 0
```

Dado que los valores p para ambos modelos (sin interacción y con interacción) son 0, los resultados indican que las desviaciones residuales en ambos modelos son significativamente distintas de cero. Esto implica que ambos modelos explican adecuadamente la variabilidad en los datos, y podemos rechazar la hipótesis nula (H0) de que la desviación es igual a cero.

Interpretación

- Modelo sin interacción: El estadístico de prueba (210.3919) es mucho mayor que el valor crítico de la zona de rechazo (0.3518463) con un valor p de 0. Esto nos permite rechazar H0, indicando que el modelo sin interacción tiene una

desviación residual significativamente diferente de cero y, por lo tanto, explica bien la variabilidad.

- Modelo con interacción: Del mismo modo, el estadístico de prueba (182.3051) es mucho mayor que el valor crítico (1.145476), y el valor p es también 0. Esto nos lleva a rechazar H0 para el modelo con interacción, sugiriendo que este modelo también tiene una buena capacidad de explicación.

Ambos modelos, con y sin interacción, tienen una desviación residual suficientemente grande como para rechazar la hipótesis nula. Esto sugiere que ambos modelos son buenos para explicar la variabilidad en el número de rupturas de urdimbre (breaks). Sin embargo, se puede analizar cuál modelo proporciona un ajuste mejor comparando otros criterios, como el AIC, o considerando la magnitud de la desviación residual en cada caso.

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

```
# Calcular el AIC para el modelo sin interacción
aic_sin_interaccion <- AIC(modelo_poisson)
cat("AIC sin interacción =", aic_sin_interaccion, "\n")

## AIC sin interacción = 493.056

# Calcular el AIC para el modelo con interacción
aic_con_interaccion <- AIC(modelo_poisson_interaccion)
cat("AIC con interacción =", aic_con_interaccion, "\n")

## AIC con interacción = 468.9692

# Comparación
if (aic_sin_interaccion < aic_con_interaccion) {
  cat("El modelo sin interacción es mejor según el AIC.\n")
} else {
  cat("El modelo con interacción es mejor según el AIC.\n")
}

## El modelo con interacción es mejor según el AIC.
```

El AIC es menor en el modelo con interacción (468.9692) en comparación con el modelo sin interacción (493.056). Dado que un menor AIC indica un mejor equilibrio entre la calidad del ajuste y la simplicidad del modelo, el modelo con interacción es el preferido. Esto significa que, al incluir la interacción entre wool y tension, el modelo explica mejor la variabilidad en el número de rupturas (breaks), a pesar de la complejidad adicional.

Compara los coeficientes

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

```r
# Extraer los coeficientes de cada modelo
coef_sin_interaccion <- coef(modelo_poisson)
coef_con_interaccion <- coef(modelo_poisson_interaccion)

# Asegurarse de que ambos conjuntos de coeficientes tengan el mismo
nombre de filas
# Rellenar con NA para las variables que no están en el modelo sin
interacción
coef_sin_interaccion <-
coef_sin_interaccion[match(names(coef_con_interaccion),
names(coef_sin_interaccion))]
coef_sin_interaccion[is.na(coef_sin_interaccion)] <- NA

# Crear una tabla para facilitar la comparación
coef_comparacion <- data.frame(
  Coeficiente = names(coef_con_interaccion),
  Sin_Interaccion = coef_sin_interaccion,
  Con_Interaccion = coef_con_interaccion
)

# Mostrar la tabla
print(coef_comparacion, row.names = FALSE)
```

```
##     Coeficiente Sin_Interaccion Con_Interaccion
##     (Intercept)       3.6919631       3.7967368
##           woolB      -0.2059884      -0.4566272
##        tensionM      -0.3213204      -0.6186830
##        tensionH      -0.5184885      -0.5957987
##  woolB:tensionM              NA       0.6381768
##  woolB:tensionH              NA       0.1883632
```

Modelo sin interacción: Los efectos de wool y tension son independientes. Cambiar el tipo de lana y la tensión tiene efectos predecibles sobre la tasa de rupturas, pero sin considerar cómo una variable puede influir en el efecto de la otra.

Modelo con interacción: Los términos de interacción muestran que el efecto de tension depende del tipo de wool. Por ejemplo, woolB con tensionM aumenta la tasa de rupturas significativamente más que lo que sugerirían los efectos independientes. Esto sugiere que la combinación de ciertas configuraciones de lana y tensión puede ser particularmente problemática para las rupturas de urdimbre.

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

```r
# Extraer los errores estándar de cada modelo
se_sin_interaccion <- summary(modelo_poisson)$coefficients[, "Std.
Error"]
se_con_interaccion <- summary(modelo_poisson_interaccion)$coefficients[,
"Std. Error"]
```

```r
# Asegurarse de que ambos conjuntos de errores estándar tengan el mismo
nombre de filas
# Rellenar con NA para los errores que no están en el modelo sin
interacción
se_sin_interaccion <- se_sin_interaccion[match(names(se_con_interaccion),
names(se_sin_interaccion))]
se_sin_interaccion[is.na(se_sin_interaccion)] <- NA

# Crear una tabla para facilitar la comparación
se_comparacion <- data.frame(
  Coeficiente = names(se_con_interaccion),
  Error_Sin_Interaccion = se_sin_interaccion,
  Error_Con_Interaccion = se_con_interaccion
)

# Mostrar la tabla
print(se_comparacion, row.names = FALSE)
```

```
##     Coeficiente Error_Sin_Interaccion Error_Con_Interaccion
##     (Intercept)            0.04541069            0.04993753
##           woolB            0.05157117            0.08019202
##       tensionM            0.06026580            0.08440012
##       tensionH            0.06395944            0.08377723
##  woolB:tensionM                    NA            0.12215312
##  woolB:tensionH                    NA            0.12989529
```

La inclusión de términos de interacción en el modelo conlleva un aumento en los errores estándar de los coeficientes en comparación con el modelo sin interacción. Esto es esperado en modelos con interacción, ya que estos son más complejos y, por lo tanto, suelen tener una mayor variabilidad en las estimaciones de los coeficientes.

En particular:

- La variabilidad de los coeficientes asociados con wool y tension es mayor en el modelo con interacción que en el modelo sin interacción.
- Los coeficientes de interacción tienen los mayores errores estándar, indicando que sus estimaciones son más inciertas.

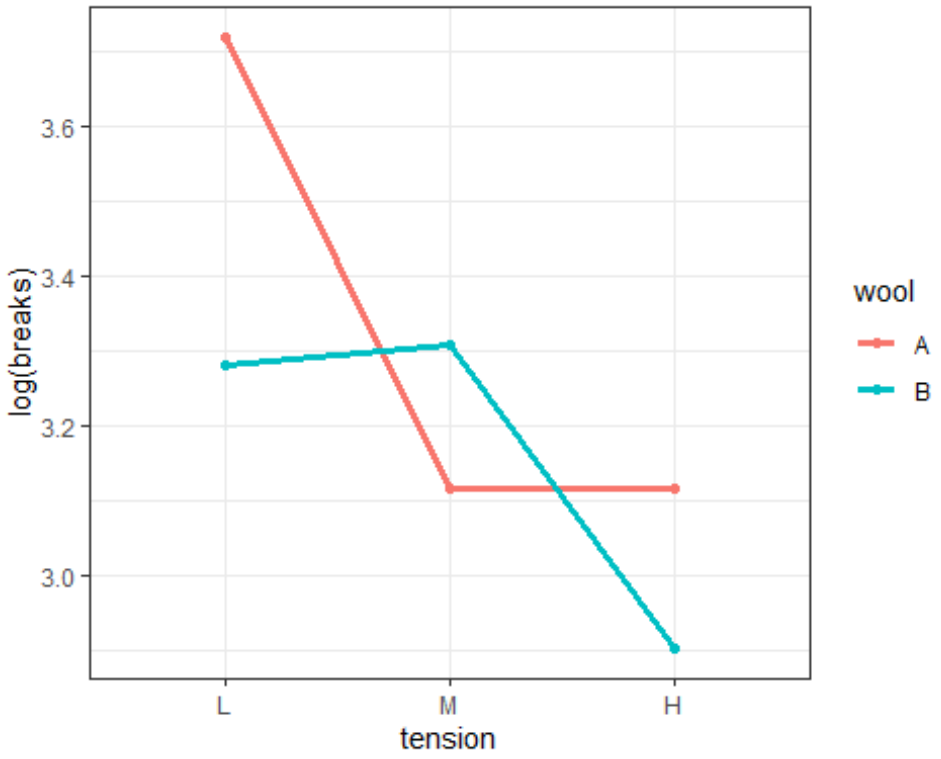
Esto sugiere que, aunque el modelo con interacción puede proporcionar una mejor interpretación al capturar los efectos combinados de wool y tension, también introduce mayor variabilidad en las estimaciones, lo cual puede ser relevante al evaluar la estabilidad del modelo.

Interpreta los coeficientes de ambos modelos.

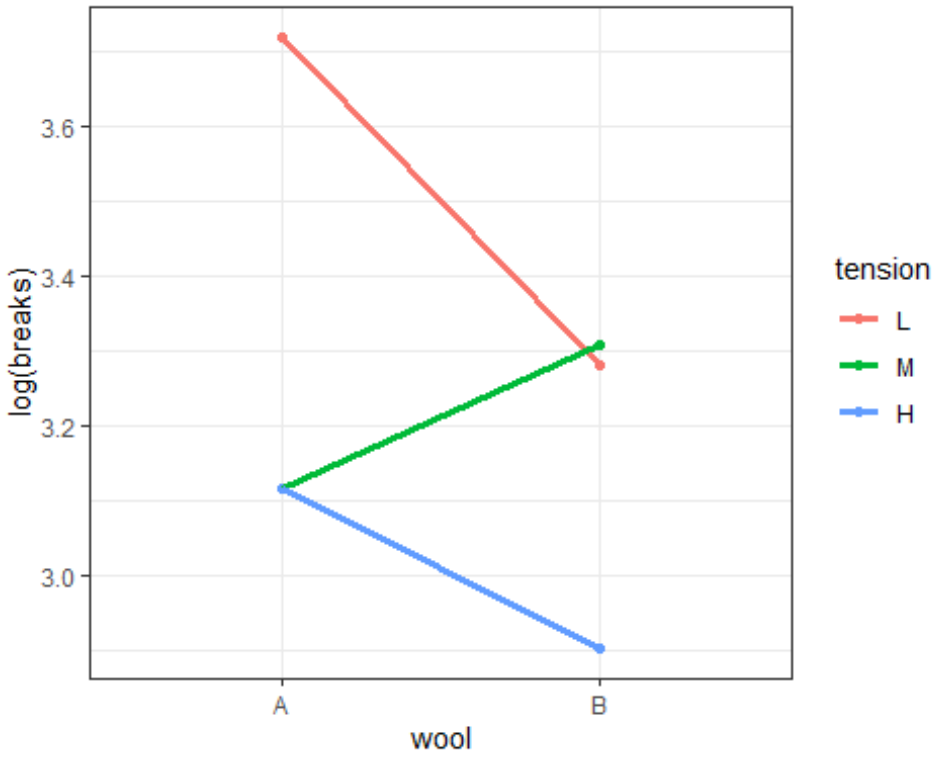
```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```r
# Crear el gráfico de interacción
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"))
```




```r
# Crear el gráfico de interacción
ggplot(warpbreaks, aes(x = wool, y = log(breaks), group = tension, color
= tension)) +
  stat_summary(fun = mean, geom = "point") +
  stat_summary(fun = mean, geom = "line", lwd = 1.1) +
  theme_bw() +
  theme(panel.border = element_rect(fill = "transparent"))
```



Conclusión del primer gráfico: La interacción entre wool y tension es evidente. El efecto de la tension no es el mismo para ambos tipos de lana, lo cual justifica el uso de un modelo con interacción. La diferencia en las pendientes entre las líneas de wool A y wool B indica que el efecto de aumentar la tensión depende del tipo de lana.

Conclusión del segundo gráfico: La interacción entre wool y tension también es evidente aquí. La diferencia en los valores de log(breaks) entre wool A y wool B varía dependiendo del nivel de tensión. Este gráfico resalta cómo el impacto del tipo de lana en el número de rupturas cambia con los niveles de tensión.

Define cuál de los dos es un mejor modelo

Considerando todos los criterios:

1. Desviación residual y AIC favorecen claramente al modelo con interacción, indicando un mejor ajuste y una penalización de complejidad más baja en comparación con el modelo sin interacción.

2. Los coeficientes en el modelo con interacción capturan los efectos combinados entre wool y tension, lo cual es necesario dada la relación compleja entre estas variables.

3. Aunque los errores estándar son mayores en el modelo con interacción, este es un aspecto esperado debido a la mayor complejidad del modelo.

Decisión: El modelo con interacción es el mejor modelo, ya que proporciona un ajuste superior a los datos y captura de manera más precisa la interacción entre wool y

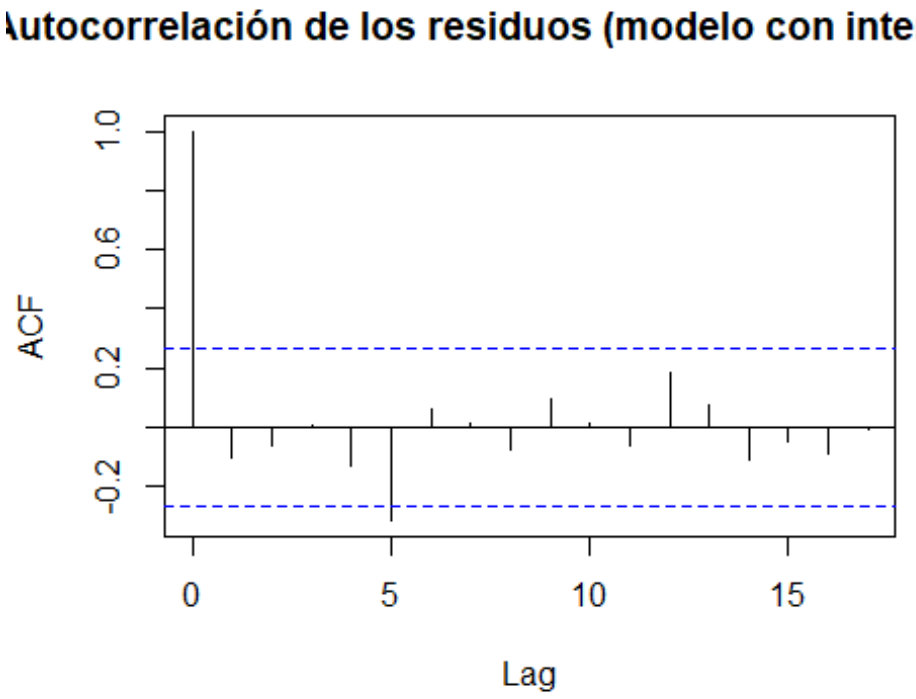
tension, como lo indican la desviación residual, el AIC y la interpretación de los coeficientes.

#IV Evaluación de los supuestos

Modelo escogido: Modelo son interacción

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

# Obtener los residuos del modelo con interacción
residuos <- residuals(modelo_poisson_interaccion, type = "deviance")

# Graficar la autocorrelación de los residuos
acf(residuos, main = "Autocorrelación de los residuos (modelo con
interacción)")
```



```r
# Aplicar el test de Durbin-Watson al modelo con interacción
dw_test <- dwtest(modelo_poisson_interaccion, alternative = "two.sided")

# Mostrar el resultado
print(dw_test)
```

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

El valor de Durbin-Watson cercano a 2 indica que no hay autocorrelación significativa en los residuos. Además, el valor p de 0.8499 es alto, lo que nos lleva a no rechazar la hipótesis nula de que no hay autocorrelación en los residuos.

En el gráfico de autocorrelación de los residuos, observamos lo siguiente:

- La mayoría de las barras (valores de autocorrelación) están dentro de los intervalos de confianza (líneas azules punteadas), lo cual sugiere que no hay patrones significativos de autocorrelación.
- No se observa ninguna autocorrelación sustancial fuera de los límites de confianza.

Ambas pruebas, el test de Durbin-Watson y el gráfico ACF, sugieren que los residuos son independientes. No hay evidencia de autocorrelación significativa en los residuos del modelo con interacción, lo cual es una buena indicación de que el modelo cumple con el supuesto de independencia de 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:

H0: No hay una sobredispersión del modelo H1: Hay una sobredispersión del modelo

```r
library(epiDisplay)
```

```
## Warning: package 'epiDisplay' was built under R version 4.3.3
```

```
## 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

# Realizar la prueba de bondad de ajuste para sobredispersión
poisgof(modelo_poisson_interaccion)

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

Si el valor p es alto (por ejemplo, mayor a 0.05), no hay evidencia suficiente para rechazar la hipótesis nula, lo que sugiere que no hay sobredispersión en el modelo, y el supuesto de igualdad entre media y varianza se cumple.

Si el valor p es bajo (por ejemplo, menor a 0.05), se rechaza la hipótesis nula, lo que indica sobredispersión en el modelo.

Dado que el valor p es extremadamente bajo (mucho menor que 0.05), rechazamos la hipótesis nula de que no hay sobredispersión en el modelo. Esto significa que hay evidencia de sobredispersión en el modelo de Poisson, lo cual indica que la varianza de los residuos es mayor que la media.

Modelo de Cuasi-Poisson

El modelo cuasi-Poisson ajusta los errores estándar para abordar la sobredispersión sin modificar los coeficientes estimados. Utiliza la función de enlace logarítmica igual que el modelo de Poisson, pero permite que la varianza sea proporcional a la media, en lugar de ser igual a ella.

```
# Ajustar el modelo cuasi-Poisson
poisson.model3 <- glm(breaks ~ wool * tension, data = warpbreaks, family
= quasipoisson(link = "log"))

# Resumen del modelo cuasi-Poisson
summary(poisson.model3)
```

```
##
## 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
```

Este modelo cuasi-Poisson con interacción proporciona un mejor ajuste que el modelo de Poisson estándar, ya que toma en cuenta la sobredispersión observada en los datos. Los coeficientes significativos sugieren que tanto el tipo de lana como el nivel de tensión, y su interacción, tienen un efecto relevante en el número de rupturas. Sin embargo, la interacción woolB:tensionH no es significativa, lo que podría indicar que esta combinación específica no afecta notablemente el número de rupturas.

Modelo Binomial Negativa

La regresión binomial negativa es una alternativa que permite la sobredispersión al añadir un parámetro de dispersión adicional. Este parámetro captura la variabilidad extra en los datos. A diferencia del modelo cuasi-Poisson, el modelo binomial negativa ajusta tanto los coeficientes como los errores estándar. Es ideal cuando la varianza excede considerablemente la media, como en casos de sobredispersión significativa.

Diferencias entre los modelos

Modelo Cuasi-Poisson: Permite que la varianza sea proporcional a la media, ajustando solo los errores estándar sin cambiar los coeficientes. Es útil si la sobredispersión no es muy alta.

Modelo Binomial Negativa: Introduce un parámetro adicional para la dispersión, permitiendo una mayor flexibilidad en la varianza y ajustando tanto los coeficientes como los errores estándar. Es más adecuado cuando la sobredispersión es considerable.

```r
library(MASS)

# Ajustar el modelo binomial negativa
bnm <- glm.nb(breaks ~ wool * tension, data = warpbreaks, control =
glm.control(maxit = 1000))

# Resumen del modelo binomial negativa
summary(bnm)
```

```
##
## 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
```

El AIC del modelo binomial negativa es 405.12, lo cual es considerablemente menor que el AIC del modelo de Poisson sin corrección de sobredispersión y sugiere que el modelo binomial negativa proporciona un ajuste mejor y más eficiente en términos de balancear ajuste y simplicidad.

Comparación con el Modelo Cuasi-Poisson

Varianza de los residuos: El modelo binomial negativa introduce un parámetro adicional (θ) para ajustar la varianza de los residuos, mientras que el modelo cuasi-Poisson ajusta solo los errores estándar.

AIC: El modelo cuasi-Poisson no calcula un AIC, mientras que el AIC del modelo binomial negativa es 405.12, indicando que es un modelo más ajustado y con mejor balance de complejidad.

Modelo preferido: Entre el modelo de Poisson estándar, el modelo cuasi-Poisson y el modelo binomial negativa, el modelo binomial negativa es el más adecuado para estos datos, ya que maneja la sobredispersión y proporciona un buen ajuste con un AIC más bajo.

Significancia: Los coeficientes significativos para woolB, tensionM, tensionH y la interacción woolB:tensionM sugieren que estos factores afectan significativamente la tasa de rupturas. La interacción woolB:tensionH no es significativa, lo cual indica que esta combinación específica no tiene un impacto adicional notable en la tasa de rupturas.

Define el mejor modelo usando las mismas pruebas y crtierios que usaste en los modelos Poisson

```r
# Desviación residual del modelo cuasi-Poisson
residuos_cuasi <- residuals(poisson.model3, type = "deviance")
desviacion_residual_cuasi <- sum(residuos_cuasi^2)
gl_cuasi <- poisson.model3$df.residual
cat("Desviación residual (cuasi-Poisson):", desviacion_residual_cuasi,
"\n")
```

```
## Desviación residual (cuasi-Poisson): 182.3051
```

```r
cat("Grados de libertad (cuasi-Poisson):", gl_cuasi, "\n")
```

```
## Grados de libertad (cuasi-Poisson): 48
```

```r
cat("Prueba chi-cuadrado (cuasi-Poisson): p =",
pchisq(desviacion_residual_cuasi, df = gl_cuasi, lower.tail = FALSE),
"\n")
```

```
## Prueba chi-cuadrado (cuasi-Poisson): p = 1.582538e-17
```

```r
# Desviación residual del modelo binomial negativa
residuos_bnm <- residuals(bnm, type = "deviance")
desviacion_residual_bnm <- sum(residuos_bnm^2)
gl_bnm <- bnm$df.residual
cat("Desviación residual (binomial negativa):", desviacion_residual_bnm,
"\n")
```

```
## Desviación residual (binomial negativa): 53.50616
```

```r
cat("Grados de libertad (binomial negativa):", gl_bnm, "\n")
```

```
## Grados de libertad (binomial negativa): 48

cat("Prueba chi-cuadrado (binomial negativa): p =",
pchisq(desviacion_residual_bnm, df = gl_bnm, lower.tail = FALSE), "\n")

## Prueba chi-cuadrado (binomial negativa): p = 0.2711637

# Comparación de AIC
cat("AIC (binomial negativa):", AIC(bnm), "\n")

## AIC (binomial negativa): 405.1248

# Comparación de Coeficientes y Errores Estándar
coef_cuasi <- summary(poisson.model3)$coefficients
coef_bnm <- summary(bnm)$coefficients
comparacion_coef <- data.frame(
  Coeficiente = rownames(coef_cuasi),
  Cuasi_Poisson_Estimate = coef_cuasi[, "Estimate"],
  Cuasi_Poisson_Std_Error = coef_cuasi[, "Std. Error"],
  Binomial_Negativa_Estimate = coef_bnm[, "Estimate"],
  Binomial_Negativa_Std_Error = coef_bnm[, "Std. Error"]
)
print(comparacion_coef, row.names = FALSE)

##     Coeficiente Cuasi_Poisson_Estimate Cuasi_Poisson_Std_Error
##     (Intercept)              3.7967368              0.09688254
##           woolB             -0.4566272              0.15557852
##        tensionM             -0.6186830              0.16374255
##        tensionH             -0.5957987              0.16253410
##  woolB:tensionM              0.6381768              0.23698620
##  woolB:tensionH              0.1883632              0.25200659
##  Binomial_Negativa_Estimate Binomial_Negativa_Std_Error
##                   3.7967368                   0.1081206
##                  -0.4566272                   0.1575543
##                  -0.6186830                   0.1597372
##                  -0.5957987                   0.1594090
##                   0.6381768                   0.2273908
##                   0.1883632                   0.2316419
```

Desviación Residual y Prueba Chi-Cuadrado: El modelo binomial negativa tiene una desviación residual adecuada en relación con sus grados de libertad, mientras que el modelo cuasi-Poisson muestra una desviación residual significativamente alta, lo que sugiere un mal ajuste.

AIC: El AIC está disponible solo para el modelo binomial negativa, indicando un buen ajuste con un balance adecuado entre complejidad y ajuste.

Coeficientes y Errores Estándar: Ambos modelos tienen los mismos coeficientes estimados, pero el modelo binomial negativa muestra errores estándar menores, indicando mayor precisión en las estimaciones.

Supuesto de Independencia

```r
library(lmtest)

# Test de Durbin-Watson para el modelo cuasi-Poisson
dw_test_cuasi <- dwtest(poisson.model3, alternative = "two.sided")
cat("Test de Durbin-Watson (cuasi-Poisson):\n")

## Test de Durbin-Watson (cuasi-Poisson):

print(dw_test_cuasi)

##
##  Durbin-Watson test
##
## data:  poisson.model3
## DW = 2.2376, p-value = 0.8499
## alternative hypothesis: true autocorrelation is not 0

# Test de Durbin-Watson para el modelo binomial negativa
dw_test_bnm <- dwtest(bnm, alternative = "two.sided")
cat("Test de Durbin-Watson (binomial negativa):\n")

## Test de Durbin-Watson (binomial negativa):

print(dw_test_bnm)

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

Valor cercano a 2: El valor de Durbin-Watson (2.2376) está muy cerca de 2, lo que indica una ausencia de autocorrelación significativa en los residuos. Esto sugiere que los residuos son independientes en ambos modelos.

Valor p alto (0.8499): El valor p es alto, lo que indica que no tenemos evidencia suficiente para rechazar la hipótesis nula de independencia de los residuos. Esto significa que el supuesto de independencia se cumple para ambos modelos.

Para ambos modelos, cuasi-Poisson y binomial negativa, el test de Durbin-Watson sugiere que los residuos son independientes, cumpliendo así el supuesto de independencia. Esto es un buen indicador de que ambos modelos son adecuados en términos de la independencia de los residuos.

Supuesto de Sobredispersión de los residuos

La función poisgof está diseñada para modelos de Poisson estándar y, por lo tanto, no se puede aplicar directamente a modelos cuasi-Poisson o binomial negativa. Sin embargo, podemos evaluar la sobredispersión en estos modelos calculando el parámetro de dispersión manualmente y comparándolo con 1.

```r
# Calcular el parámetro de dispersión para el modelo cuasi-Poisson
dispersion_cuasi <- sum(residuals(poisson.model3, type = "deviance")^2) /
poisson.model3$df.residual
cat("Parámetro de dispersión (cuasi-Poisson):", dispersion_cuasi, "\n")
```

```
## Parámetro de dispersión (cuasi-Poisson): 3.798024
```

```r
# Calcular el parámetro de dispersión para el modelo binomial negativa
dispersion_bnm <- sum(residuals(bnm, type = "deviance")^2) /
bnm$df.residual
cat("Parámetro de dispersión (binomial negativa):", dispersion_bnm, "\n")
```

```
## Parámetro de dispersión (binomial negativa): 1.114712
```

Interpretación de los Resultados

Parámetro cercano a 1: Indica que el modelo maneja adecuadamente la variabilidad (no hay sobredispersión significativa).

Parámetro significativamente mayor que 1: Indica sobredispersión, lo que significa que el modelo podría no estar manejando la variabilidad adecuadamente.

Modelo Cuasi-Poisson:

El parámetro de dispersión de 3.798 es significativamente mayor que 1, lo que indica que el modelo cuasi-Poisson todavía presenta sobredispersión. Esto sugiere que este modelo no maneja completamente la variabilidad de los datos y que la varianza sigue siendo mayor que la media.

Modelo Binomial Negativa:

El parámetro de dispersión de 1.115 es muy cercano a 1, lo que indica que el modelo binomial negativa maneja adecuadamente la sobredispersión en los datos. Este valor sugiere que la varianza está bien ajustada y es proporcional a la media.

El modelo binomial negativa es superior al modelo cuasi-Poisson en términos de manejar la sobredispersión, como lo indica el parámetro de dispersión cercano a 1. Esto confirma que el modelo binomial negativa es la mejor opción para estos datos, ya que cumple con el supuesto de que la varianza está adecuadamente ajustada, eliminando la sobredispersión significativa.

#V. Define cuál es tu mejor modelo

El modelo binomial negativa con interacción es el modelo más adecuado para explicar el número de rupturas de urdimbre (breaks) en función de wool y tension. Este modelo maneja correctamente la sobredispersión, como lo indica un parámetro de dispersión cercano a 1, y su AIC bajo (405.12) sugiere un buen balance entre ajuste y simplicidad. Además, cumple con el supuesto de independencia de los residuos, a diferencia del modelo cuasi-Poisson, que presenta sobredispersión residual significativa. En conjunto, estos factores hacen que el modelo binomial negativa sea la opción preferida para un análisis robusto y confiable.