A4-Regresión Poisson

AUTHOR
Alfredo García

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
```

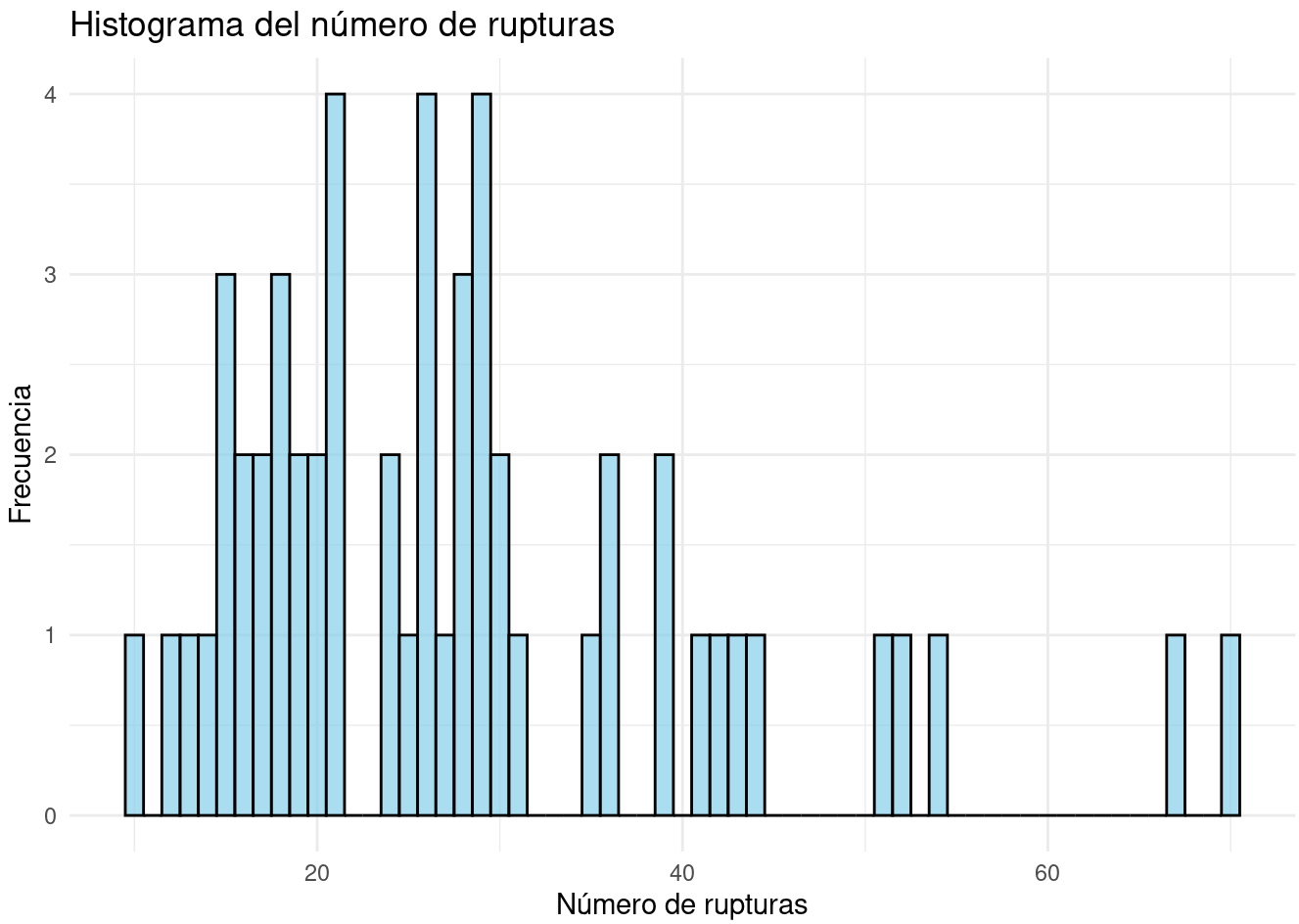
Este conjunto de datos indica cuántas roturas de urdimbre ocurrieron para diferentes tipos de telares por telar, por longitud fija de hilo:

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

Obtén: Histograma del número de rupturas

```
# Cargar la librería ggplot2 para crear gráficos
library(ggplot2)

# Histograma del número de rupturas
ggplot(data, aes(x = breaks)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black", alpha = 0.7) +
  labs(title = "Histograma del número de rupturas",
       x = "Número de rupturas",
       y = "Frecuencia") +
  theme_minimal()
```



Obtén la media y la varianza

```
# Calcular la media y la varianza del número de rupturas
media <- mean(data$breaks)
varianza <- var(data$breaks)

# Mostrar la media y la varianza
print(paste("Media:", media))
```

```
[1] "Media: 28.1481481481481"
```

```
print(paste("Varianza:", varianza))
```

```
[1] "Varianza: 174.204053109713"
```

Ajusta el modelo de regresión Poisson. Usa el mando: poisson.model<-glm(breaks ~ wool + tension, data, family = poisson(link = “log”)) summary(poisson.model)

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

```
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
```

Interpreta la información obtenida. Toma en cuenta que R genera variables Dummy para las variables categóricas. Para cada variable genera k-1 variables Dummy en k categorías.

1. **Intercepto (Intercept)**: El intercepto en este modelo es 3.69196. Esto indica el logaritmo de la tasa media de rupturas cuando todas las otras variables predictoras son cero. Dado que las variables predictoras son categóricas, esto se interpreta como el logaritmo de la tasa media de rupturas cuando el tipo de lana es “A” y la tensión es “L”.

2. **Coeficientes para las Variables Dummy de wool y tension**:

    - **woolB**: La variable dummy para el tipo de lana “B” tiene un coeficiente de -0.20599. Esto significa que, en comparación con el tipo de lana “A” (que es el grupo de referencia ya que no tiene una variable dummy), el logaritmo de la tasa media de rupturas es menor en 0.20599 unidades para el tipo de lana “B”.

    - **tensionM**: La variable dummy para la tensión “M” tiene un coeficiente de -0.32132. Esto significa que, en comparación con la tensión “L” (grupo de referencia), el logaritmo de la tasa media de rupturas es menor en 0.32132 unidades para la tensión “M”.

    - **tensionH**: La variable dummy para la tensión “H” tiene un coeficiente de -0.51849. Esto significa que, en comparación con la tensión “L” (grupo de referencia), el logaritmo de la tasa media de rupturas es menor en 0.51849 unidades para la tensión “H”.

3. **Devianza**: La devianza es una medida de ajuste del modelo. En este caso, la devianza nula (Null deviance) representa qué tan bien se ajusta un modelo con intercepto solamente (sin variables predictoras) a los datos. La devianza residual (Residual deviance) representa qué tan bien se ajusta el modelo completo (con las variables predictoras) a los datos. En este modelo, la devianza residual es 210.39, lo que indica un buen ajuste del modelo a los datos.

4. **Significancia de los Coeficientes**: Los coeficientes para las variables dummy de wool y tension son significativos a un nivel de significancia muy alto (p-values muy cercanos a cero), lo que sugiere que tanto el tipo de lana como la tensión tienen un efecto significativo en el número de rupturas.

5. **Variables Dummy**: En R, para variables categóricas con k categorías, se crean k-1 variables dummy. Esto se hace para evitar la multicolinealidad en el modelo. En este caso, hay dos categorías para la variable "wool" (A y B) y tres categorías para la variable "tension" (L, M, H), por lo que se generan tres variables dummy en total.

La desviación residual debe ser mayor que los grados de libertad para asegurarse que no exista una dispersión excesiva. Una diferencia menor, significará que aunque las estimaciones son correctas, los errores estándar son incorrectos y el modelo no los toma en cuenta.

En el resultado obtenido, la devianza residual es 210.39 y hay 50 grados de libertad (grados de libertad residual), lo que implica que hay más datos que parámetros en el modelo. En este caso, la devianza residual es mayor que los grados de libertad, lo que indica que el modelo está proporcionando un buen ajuste a los datos.

La desviación excesiva nula muestra que tan bien se predice la variable de respuesta mediante un modelo que incluye solo el intercepto (gran media) mientras que el residual con la inclusión de variables. Una diferencia en los valores significa un mal ajuste.

Podemos ver que la diferencia entre los vaores es 'grande' por lo que podemos tener sospechas de que el modelo no es del todo bueno y por ello vamos a también hacer la parte de generar el modelo quasipoisson y comparar los resultados yver si hay una mejora

Si hay un mal modelo, recurre a usar un modelo cuasi Poisson, si los coeficientes son los mismos, el modelo es bueno:

poisson.model2<-glm(breaks ~ wool + tension, data = data, family = quasipoisson(link = "log"))
summary(poisson.model2)

```
poisson.model2<-glm(breaks ~ wool + tension, data = data, family = quasipoisson(link = "l
summary(poisson.model2)
```

```
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
```

En resumen, este modelo cuasi Poisson parece ser una mejora sobre el modelo de Poisson estándar, ya que aborda la sobredispersión en los datos. Las variables tensionM y tensionH siguen siendo significativas en la predicción del número de rupturas, mientras que woolB no es significativa a un nivel de 0.05. La devianza residual menor que la devianza nula sugiere que el modelo está capturando bien la variabilidad en los datos además de cuplir con la condición de que los cieficientes de la regresión poisson sean similares a la de la regresión quasipoisson.