A4-Regresión Poisson

Diego Rodriguez

2023-10-10

Claro, procederé paso a paso con el código en R:

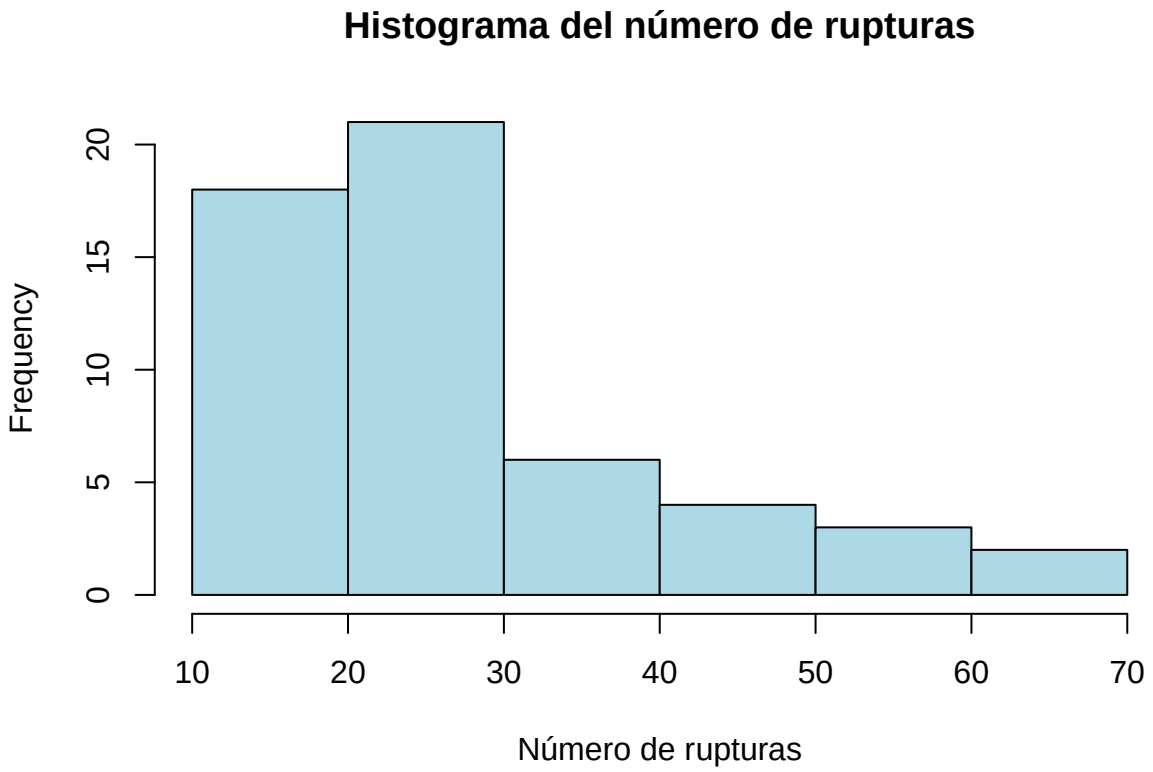
1. **Cargar el dataset y ver los primeros registros**:

```
data(warpbreaks)
head(warpbreaks, 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
```

2. **Histograma del número de rupturas**:

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



3. **Obtener la media y la varianza**:

```
mean_breaks <- mean(warpbreaks$breaks)
var_breaks <- var(warpbreaks$breaks)

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

```
## Media de rupturas: 28.14815
```

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

```
## Varianza de rupturas: 174.2041
```

4. **Ajustar el modelo de regresión Poisson y obtener el resumen**:

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

```
##
## 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
```

De acuerdo, vamos a interpretar los resultados del modelo dado:

Coeficientes: - **(Intercept): 3.69196**: Este coeficiente es el logaritmo del número esperado de rupturas cuando se utiliza `woolA` (porque `woolB` es la variable dummy y `woolA` es la categoría de referencia) y la tensión es `L` (porque `tensionM` y `tensionH` son las variables dummy y `tensionL` es la categoría de referencia). El número esperado de rupturas para estas condiciones es $e^{3.69196}$ o aproximadamente 40 rupturas.

- **woolB: -0.20599**: Cuando cambiamos de `woolA` a `woolB` (manteniendo constante la tensión), el logaritmo del número esperado de rupturas disminuye en 0.20599. Esto indica que, todo lo demás constante, `woolB` tiende a tener menos rupturas que `woolA`.
- **tensionM: -0.32132**: Con respecto a una tensión baja `L`, una tensión media `M` tiene un logaritmo del número esperado de rupturas que es 0.32132 unidades menor. Por lo tanto, la tensión media tiende a tener menos rupturas que una tensión baja, todo lo demás constante.
- **tensionH: -0.51849**: Similarmente, con respecto a una tensión baja `L`, una tensión alta `H` tiene un logaritmo del número esperado de rupturas que es 0.51849 unidades menor. La tensión alta tiende a tener aún menos rupturas que una tensión baja y que una tensión media, manteniendo todo lo demás constante.

Significancia: Todos los coeficientes son altamente significativos (como lo indica `***`), lo que implica que cada uno de estos factores (`wool` y `tension`) tienen un efecto significativo en el número de rupturas.

Desviación: - **Null deviance 297.37**: Esta es la desviación del modelo que solo tiene un intercepto. Es una medida de la variabilidad de los datos.

- **Residual deviance 210.39**: Esta es la desviación del modelo que incluye las predictoras. Si este valor es mucho más pequeño que la desviación nula, indica que las predictoras están explicando una porción significativa de la variabilidad en los datos. En este caso, se ha reducido de 297.37 a 210.39, lo que sugiere que las variables `wool` y `tension` proporcionan información valiosa sobre el número de rupturas.

El hecho de que la desviación residual (210.39) sea considerablemente menor que la nula (297.37) y que haya una diferencia notable entre los grados de libertad (53 para la nula y 50 para la residual) es indicativo de un buen ajuste del modelo.

AIC 493.06: El Criterio de Información de Akaike (AIC) es una métrica que se utiliza para comparar modelos. Cuanto menor sea el AIC, mejor es el modelo en términos de ajuste y simplicidad.

Número de iteraciones de Fisher Scoring: 4: Esto simplemente nos dice cuántas iteraciones tomó para que el algoritmo convergiera. Normalmente no nos preocupamos por este número a menos que sea inusualmente alto, lo que podría indicar problemas de convergencia.

En resumen, el tipo de lana (`wool`) y el nivel de tensión (`tension`) tienen efectos significativos en el número de rupturas de hilo. Específicamente, la lana tipo B tiende a tener menos roturas que la tipo A, y a medida que aumenta la tensión (de baja a media o alta), el número de roturas tiende a disminuir.

5. **Verificar la desviación residual**:

Para revisar si hay una dispersión excesiva, puedes ver la desviación residual y compararla con los grados de libertad:

```r
cat("Desviación Residual:", deviance(poisson.model), "\n")
```

```
## Desviación Residual: 210.3919
```

```r
cat("Grados de Libertad:", df.residual(poisson.model), "\n")
```

```
## Grados de Libertad: 50
```

En este caso, tenemos una desviación residual de 210.3919 y 50 grados de libertad. La desviación residual es bastante más alta que los grados de libertad, lo que podría ser una señal de sobre-dispersión en los datos.

La sobre-dispersión sugiere que hay más variabilidad en los datos de la que puede ser explicada por el modelo de Poisson. Esto puede ser debido a que el modelo de Poisson no captura completamente la estructura de correlación en los datos, o que hay otras variables no consideradas en el modelo que están influyendo en el número de rupturas.

En resumen, dado que la desviación residual es bastante mayor que los grados de libertad, es posible que el modelo de Poisson no sea el ajuste más adecuado y se podría considerar la exploración de otros modelos, como el modelo cuasipoisson, que es más flexible y puede manejar la sobre-dispersión.

6. **Si hay sobre-dispersión, ajustar el modelo quasipoisson y obtener el resumen**:

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

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