A6 - Regresión Poisson

Juan Bernal

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)
```

breaks	wool	tension
26	A	L
30	A	L
54	A	L
25	A	L
70	A	L
52	A	L

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

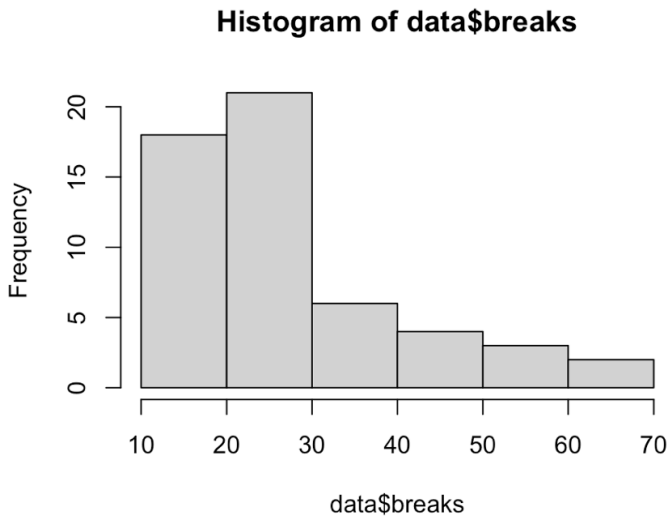
- breaks: número de rupturas
- wool: tipo de lana (A o B)
- tensión: el nivel de tensión (L, M, H)

Sigue el siguiente procedimiento de análisis:

I. Análisis Descriptivo

a. Histograma del número de rupturas

```
hist(data$breaks)
```



El gráfico anterior demuestra que la distribución es asimétrica hacia la derecha, lo que indica que la mayoría de las observaciones se encuentran en los intervalos más bajos (cercanos a 10-30). A medida que se avanza hacia valores más altos (50-70), la frecuencia de observaciones disminuye. Esto sugiere que los datos tienen una tendencia hacia valores más pequeños, con pocas observaciones en el extremo superior.

b. Obtén la media y la varianza de la variable dependiente

```
mu = mean(data$breaks)
ds = sd(data$breaks)
var = ds^2

cat('La media de las rupturas es ', mu, ' y la varianza es ',var)
```

```
## La media de las rupturas es  28.14815  y la varianza es  174.2041
```

c. Interpreta en el contexto de una Regresión Poisson

La regresión de Poisson se utiliza cuando modelamos datos de conteo, es decir, cuando la variable de respuesta es un número de ocurrencias. Esta distribución se caracteriza por el hecho de que la media y la varianza teóricamente deberían ser iguales. La media de los datos es 28.14815, lo que indica que, en promedio, hay aproximadamente 28 rupturas (breaks) en la muestra considerada. La varianza es 174.2041, que es considerablemente mayor que la media. Por lo tanto, no estamos tratando con una regresión de Poisson.

II. Ajusta dos modelos de Regresión Poisson

a. Ajusta el modelo de regresión Poisson sin interacción

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

```
##
## 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 < 0.0000000000000002 ***
## woolB       -0.20599    0.05157  -3.994 0.000064897752817398 ***
## tensionM    -0.32132    0.06027  -5.332 0.000000097286419487 ***
## tensionH    -0.51849    0.06396  -8.107 0.000000000000000521 ***
## ---
## 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
```

Suponiendo un nivel de significancia de $\alpha = 0.05$, dados los p-values de todos los coeficientes, encontramos que todos los coeficientes son significantes para el modelo.

b. Ajusta el modelo de regresión Poisson con interacción

```
poisson_model2<-glm(breaks ~ wool*tension, data, family = poisson(link = "log"))
S2=summary(poisson_model2)
S2
```

```
##
## 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 < 0.0000000000000002 ***
## woolB          -0.45663    0.08019  -5.694      0.00000001239721 ***
## tensionM       -0.61868    0.08440  -7.330      0.00000000000023 ***
## tensionH       -0.59580    0.08378  -7.112      0.00000000000115 ***
## woolB:tensionM  0.63818    0.12215   5.224      0.00000017472034 ***
## 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
```

Suponiendo un nivel de significancia de $\alpha = 0.05$, dados los p-values de todos los coeficientes, encontramos que todos los coeficientes a excepción de woolB:tensionH son significantes para el modelo.

c. Usa los comandos:

poisson_model<-glm(breaks ~ wool + tension, data, family = poisson(link = “log”)) S=summary(poisson_model)

d. 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.

Para el modelo sin interacción, la variable dependiente se estima como:

- $break = 3.6919 - 0.2059 \times woolB - 0.3213 \times tensionM - 0.5185 \times tensionH$

El intercepto de 3.6919 representa el valor esperado de break cuando las variables woolB, tensionM, y tensionH son cero. woolB: tener el tipo de lana “B” en lugar de “A” disminuye break en 0.2059 unidades. tensionM: tener tensión media disminuye break en 0.3213 unidades en comparación con la tensión baja. tensionH: tener tensión alta disminuye break en 0.5185 unidades en comparación con la tensión baja.

Y para el modelo con interacción, la variable dependiente se estima como:

- $break = 3.7967 - 0.4563 \times woolB - 0.6186 \times tensionM - 0.5958 \times tensionH + 0.6381 \times BM + 0.1883 \times BH$

El intercepto de 3.7967 representa el valor esperado de break cuando las variables woolB, tensionM, tensionH, BM y BH son cero. woolB: tener el tipo de lana “B” en lugar de “A” disminuye break en 0.4563 unidades. tensionM: tener tensión media disminuye break en 0.6186 unidades en comparación con la tensión baja. tensionH: tener tensión alta disminuye break en 0.5958 unidades en comparación con la tensión baja. BM: tener lana “B” con tensión media aumenta break en 0.6381 unidades. BH: tener lana “B” con tensión alta aumenta break en 0.1883 unidades.

III. Selección del modelo

a. 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.

La desviación residual del modelo sin interacción es de 210.39 con 50 grados de libertad. Mientras que la desviación residual del modelo con interacción es de 182.31 con 48 grados de libertad. De tal manera que el modelo con interacción se ajusta mejor a los datos.

• AIC: Criterio de Aikaike

El criterio de AIC en el modelo sin interacción tiene un puntaje de 493.06, mientras que en el modelo con interacción tiene un puntaje de 468.97, lo que sugiere que el modelo con interacción tiene un mejor balance entre ajuste y complejidad del modelo.

• Comparación entre los coeficientes y los errores estándar de ambos modelos

Coeficiente	Modelo sin interacción	Modelo con interacción
Intercepto	3.69196	3.79674
woolB	-0.20599	-0.45663
tensionM	-0.32132	-0.61868
tensionH	-0.51849	-0.59580
woolB:tensionM	-	0.63818
woolB:tensionH	-	0.18836

En el modelo sin interacción, el efecto de cada variable es independiente. Por ejemplo, woolB reduce el valor de break en 0.20599, mientras que tensionM y tensionH tienen efectos negativos adicionales. En el modelo con interacción, el coeficiente de woolB se vuelve más negativo (-0.45663), lo que sugiere que su impacto es más fuerte en combinación con la tensión. Además, los coeficientes de interacción (woolB:tensionM y woolB:tensionH) indican que la combinación de lana tipo B con tensiones M y H tiene un impacto positivo adicional en break.

Coeficiente	Error Estándar (sin interacción)	Error Estándar (con interacción)
Intercepto	0.04541	0.04994
woolB	0.05157	0.08019
tensionM	0.06027	0.08440
tensionH	0.06396	0.08378
woolB:tensionM	-	0.12215
woolB:tensionH	-	0.12990

Los errores estándar son más altos en el modelo con interacción, especialmente para las variables woolB, tensionM, y los términos de interacción. Esto es común cuando se incluyen términos de interacción, ya que se introduce complejidad adicional al modelo y potencial colinealidad.

b. Desviación residual (Prueba de Chi cuadrada)

• 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.

```
gl = S$df.null-S$df.residual
qchisq(0.05,gl)
```

```
## [1] 0.3518463
```

```
gl2 = S2$df.null-S2$df.residual
qchisq(0.05,gl2)
```

```
## [1] 1.145476
```

Este código se utiliza para comparar la deviancia o una estadística chi-cuadrado de un modelo contra un valor crítico para decidir si el modelo con variables predictoras tiene un mejor ajuste que el modelo nulo. Si el valor de la estadística chi-cuadrado es mayor que este valor crítico (qchisq(0.05, gl)), entonces rechazamos la hipótesis nula al nivel de significancia del 5%, lo que indica que el modelo ajustado es significativamente mejor que el modelo nulo.

• La prueba de chi cuadrada 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:

- $H_0 : Deviance = 0$ El modelo no explica la variabilidad de los datos
- $H_1 : Deviance > 0$ El modelo explica cierta variabilidad de los datos
- gl = gl_desviación residual (n-(p+1))

```
dr = S$deviance
cat("Estadístico de prueba =",dr, "\n")
```

```
## Estadístico de prueba = 210.3919
```

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

```
## Valor p = 0
```

Suponiendo un valor de significancia estándar del 0.05, y dado que el p_value obtenido fue 0, entonces se cuenta con la evidencia suficiente para rechazar la hipótesis inicial, por lo que el modelo sin interacción explica cierta variabilidad de los datos. Además, notemos también que el estadístico de prueba Deviance es mayor a 0, reafirmando el análisis anterior.

```
dr2 = S2$deviance
cat("Estadístico de prueba =",dr2, "\n")
```

```
## Estadístico de prueba = 182.3051
```

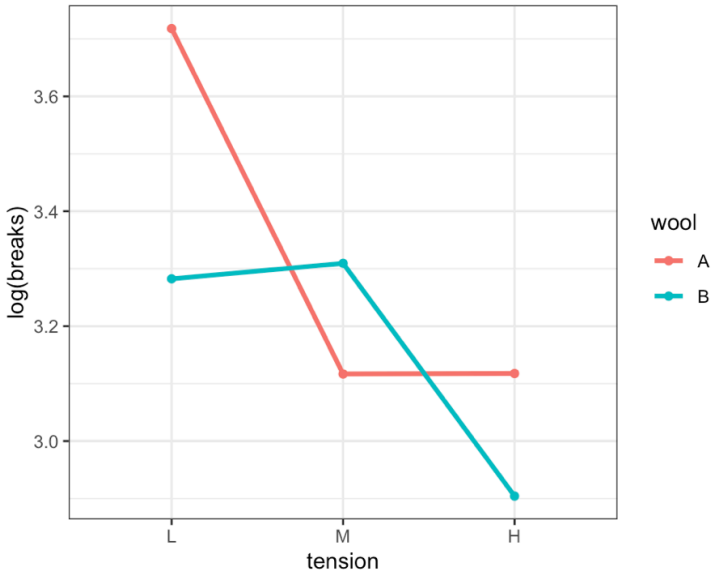
```
vp2 = 1-pchisq(dr2,gl2)
cat("Valor p =",vp2)
```

```
## Valor p = 0
```

Suponiendo un valor de significancia estándar del 0.05, y dado que el p_value obtenido fue 0, entonces se cuenta con la evidencia suficiente para rechazar la hipótesis inicial, por lo que el modelo con interacción explica cierta variabilidad de los datos. Además, notemos también que el estadístico de prueba Deviance es mayor a 0, reafirmando el análisis anterior.

• Interpreta los coeficientes de ambos modelos.

```
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(fil="transparent"))
```



La gráfica anterior sugiere una interacción entre el tipo de lana y la tensión aplicada, donde la lana A es más susceptible a romperse con baja tensión, pero ambas lanas muestran comportamientos más similares a medida que aumenta la tensión, con la lana B mostrando mayor resistencia en tensión alta.

e. Define cuál de los dos es un mejor modelo.

Aunque el modelo con interacción tiene errores estándar más altos, sus métricas de ajuste (deviancia residual y AIC) indican que captura mejor la relación entre las variables. Además, los coeficientes de interacción significativos sugieren que el efecto combinado de lana y tensión es relevante para explicar la variable de respuesta. Por lo tanto, el modelo con interacción es mejor en términos de ajuste y capacidad para capturar las relaciones subyacentes entre las variables. Este modelo proporciona una representación más precisa y detallada de los datos, ya que considera que el efecto de la lana depende de la tensión, lo cual es respaldado por la reducción en deviance y AIC.

IV. 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.

- H_0 : Hay independencia entre las variables
- H_1 : No hay independencia entre las variables

```
library(lmtest)
```

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

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

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

```
dwtest(poisson_model)
```

```
##
##  Durbin-Watson test
##
## data:  poisson_model
## DW = 2.0332, p-value = 0.3896
## alternative hypothesis: true autocorrelation is greater than 0
```

Suponiendo un valor de significancia estándar del 0.05, y dado que el p_value obtenido fue 0.3896, entonces no se cuenta con la evidencia suficiente para rechazar la hipótesis inicial, por lo que en el modelo sin interacción hay independencia entre las variables.

```
dwtest(poisson_model2)
```

```
##
##  Durbin-Watson test
##
## data:  poisson_model2
## DW = 2.2376, p-value = 0.575
## alternative hypothesis: true autocorrelation is greater than 0
```

Suponiendo un valor de significancia estándar del 0.05, y dado que el p_value obtenido fue 0.575, entonces no se cuenta con la evidencia suficiente para rechazar la hipótesis inicial, por lo que en el modelo con interacción hay independencia entre las variables.

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 χ^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

```
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] 0.00000000000000000000144606
```

Suponiendo un valor de significancia estándar del 0.05, y dado que el p_value obtenido fue aproximadamente 0, entonces se cuenta con la evidencia suficiente para rechazar la hipótesis inicial, por lo que en el modelo sin interacción hay una sobredispersión.

```
poisgof(poisson_model2)
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 182.3051
##
## $df
## [1] 48
##
## $p.value
## [1] 0.00000000000000001582538
```

Suponiendo un valor de significancia estándar del 0.05, y dado que el p_value obtenido fue aproximadamente 0, entonces se cuenta con la evidencia suficiente para rechazar la hipótesis inicial, por lo que en el modelo con interacción hay una sobredispersión.

Si hay un mal modelo, recurre a usar:

De acuerdo con los resultados obtenidos en análisis anteriores, el mejor modelo fue el que tenía interacción, pero este no cumplió con todos los supuestos, pues demuestra una sobredispersión.

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

A diferencia de los modelos Poisson que suponen que la media y la varianza son iguales, el modelo binomial negativo es adecuado para conteos con varianza mayor que la media. Y de acuerdo con el análisis descriptivo, la varianza es mucho mayor que la media, por lo que este modelo será el siguiente en analizar. Además, tomaremos en cuenta la interacción entre Wool y Tension, dado que el mejor modelo hasta al momento fue el poisson con interacción.

```
bnm = model.nb = glm.nb(breaks ~ wool * tension, data, control = glm.control(maxit=1000))
S3 = summary(bnm)
S3
```

```
##
## 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 < 0.0000000000000002 ***
## 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
```

Para el modelo binomial negativo con interacción, la variable dependiente se estima como:

- $break = 3.7967 - 0.4566 \times woolB - 0.6187 \times tensionM - 0.5958 \times tensionH + 0.6382 \times BM + 0.1884 \times BH$

El intercepto de 3.7967 representa el valor esperado de break cuando las variables woolB, tensionM, tensionH, BM y BH son cero. woolB: tener el tipo de lana “B” en lugar de “A” disminuye break en 0.4566 unidades. tensionM: tener tensión media disminuye break en 0.6187 unidades en comparación con la tensión baja. tensionH: tener tensión alta disminuye break en 0.5958 unidades en comparación con la tensión baja. BM: tener lana “B” con tensión media aumenta break en 0.6382 unidades. BH: tener lana “B” con tensión alta aumenta break en 0.1884 unidades.

Además, la desviación residual es muchísimo menor a la de los modelos análizados anteriormente, por lo que el modelo negativo binomial resulta ajustarse mejor a los datos. También presenta menor puntaje AIC, por tanto tiene un mejor balance entre ajuste y complejidad que los modelos Poisson anteriores.

Prueba de Chi Cuadrada

- $H_0 : Deviance = 0$ El modelo no explica la variabilidad de los datos
- $H_1 : Deviance > 0$ El modelo explica cierta variabilidad de los datos

```
gl3 = S3$df.null-S3$df.residual
dr3 = S3$deviance
cat("Estadístico de prueba =",dr3, "\n")
```

```
## Estadístico de prueba = 53.50616
```

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

```
## Valor p = 0.0000000002647427
```

Suponiendo un valor de significancia estándar del 0.05, y dado que el p_value obtenido fue aproximadamente 0, entonces se cuenta con la evidencia suficiente para rechazar la hipótesis inicial, por lo que el modelo binomial negativo con interacción explica cierta variabilidad de los datos. Además, notemos también que el estadístico de prueba Deviance es mayor a 0, reafirmando el análisis anterior.

Prueba de independencia

- H_0 : Hay independencia entre las variables
- H_1 : No hay independencia entre las variables

```
dwtest(bnm)
```

```
##
##  Durbin-Watson test
##
## data:  bnm
## DW = 2.2376, p-value = 0.575
## alternative hypothesis: true autocorrelation is greater than 0
```

Suponiendo un valor de significancia estándar del 0.05, y dado que el p_value obtenido fue 0.575, entonces no se cuenta con la evidencia suficiente para rechazar la hipótesis inicial, por lo que en el modelo binomial negativo con interacción hay independencia entre las variables.

Prueba de sobredispersión

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

```
poisgof(bnm)
```

```
## $results
## [1] "Goodness-of-fit test for Poisson assumption"
##
## $chisq
## [1] 53.50616
##
## $df
## [1] 48
##
## $p.value
## [1] 0.2711637
```

Suponiendo un nivel de significancia de α = 0.05, y dado que el p valor obtenido fue 0.2711, entonces no se cuenta con evidencia suficiente para rechazar la hipótesis inicial. Por lo que en el modelo binomial negativo con interacción no hay sobredispersión.

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

El modelo binomial negativo con interacción es el mejor modelo, pues mostró mejor adaptabilidad a los datos, así como también mejor balance entre ajuste y complejidad que los modelos Poisson anteriores, de acuerdo con el AIC y la desviación residual. Además, cumplió con los supuestos de explicación de variabilidad, independencia entre variables y la no sobredispersión del modelo.

V. Define cuál es tu mejor modelo

Se presentó un análisis del número de rupturas de urdimbre (breaks) en el conjunto de datos warpbreaks, utilizando dos variables categóricas: el tipo de lana (wool) y el nivel de tensión (tension). Inicialmente, se ajustaron modelos de regresión Poisson, tanto sin como con interacciones entre las variables. Sin embargo, debido a que la varianza en los datos es considerablemente mayor que la media, se detectó sobredispersión, lo cual indica que el modelo Poisson podría no ser el más adecuado.

Posteriormente, se introduce un modelo de Binomial Negativa, el cual es más apropiado para datos con sobredispersión. Al comparar métricas de ajuste como la desviación residual y el AIC, se observó que el modelo de Binomial Negativa ofrece un mejor desempeño en comparación con los modelos Poisson. Además, este modelo proporcionó errores estándar más estables y coeficientes significativos, lo que respalda su capacidad para capturar la variabilidad en los datos de forma más precisa.

En conclusión, el modelo de Binomial Negativa resulta ser el más adecuado para el análisis debido a que gestiona de manera efectiva la sobredispersión al permitir que la varianza sea mayor que la media. Además de proporcionar menor desviación residual y AIC. Este modelo permite representar los datos de forma más precisa y proporciona inferencias más robustas.