Los contajes, mejor con Poisson que con Gauss (II)

En el número anterior comenzamos una serie de artículos para utilizar la distribución de Poisson en nuestras granjas. Decíamos, a modo de recordatorio, que puede ser de gran utilidad para las variables que provienen de contajes, como puede ser el número de lechones muertos por parto, el número de abortos a la semana o el número de lechones nacidos por parto.

Alberto Morillo Alujas[1], Daniel Villalba Mata[2] y Emilio López Cano[3]

[1]Tests and Trials SLU
[2]Universidad de Lleida
[3]Universidad Rey Juan Carlos (Madrid)

Las variables que provienen de contajes cumplen determinadas características como:

- Son valores discretos. Más concretamente son contajes de números enteros no negativos (no tienen decimales), como el número de abortos en una semana, el número de repeticiones en una banda de cubrición, el número de lechones muertos por parto.
- Los sucesos que ocurren en un intervalo de tiempo determinado son independientes.
- Estos contajes pueden ser expresados también como tasas, ya que la cantidad de veces que ocurre un evento dentro de un periodo de tiempo se puede expresar bien como un conteo sin procesar (ha habido 3 abortos hoy) o como una tasa (la tasa de abortos al día es de 3 abortos/día).

También vimos que esta distribución nos puede ayudar a responder a preguntas sobre probabilidad como cuál es la probabilidad de que tenga 7 abortos a la semana si en mi granja tengo 3 abortos a la semana como media (ver Suis nº 194, enero/febrero 2023).

En la *tabla* se compara la distribución de Poisson y la de Gauss o Normal para ver claramente cuándo tenemos que usar una o la otra.

Pero quizá la forma en la que podemos observar y aprender mejor qué hace la distribución de Poisson sobre una variable es con un ejemplo. Vamos a usar los lechones nacidos muertos en el parto en una granja junto a 3 variables descriptivas y un tratamiento para estudiar esta distribución.

El conjunto de datos que vamos a usar es una modificación del conjunto de datos "cell.txt" que se usa con fines didácticos en R para estudiar la distribución de Poisson. El lector puede encontrar este archivo en nuestro blog https://www.testsandtrials.com/blog/ con el nombre de "data_poisson.txt".

Como siempre, iniciaremos R, con la instrucción *library(Rcmdr)* activaremos *RCommander* y yendo a *Datos/Importar datos desde un archivo Excel* y nombrando el conjunto de datos como "data", el sistema nos llevará a elegir dónde tenemos almacenado el archivo de datos. Aceptaremos y ya tendremos el conjunto de datos importado.

Para ver qué datos tenemos en el archivo, iremos a Estadísticos/Resúmenes/Conjunto de datos numéricos y nos aparecerá la siguiente salida:

```
> summary(data)
nac_muer    genética    edad          nave          peso
Min.: 0.000   F:376   1 a 2: 107    alta: 227     gorda: 167
1st Qu.: 0.000T:135   3 a 6: 136    baja: 284 muy gorda: 140
Median: 0.000         mas de 6: 268 normal: 204
Mean: 0.908
3rd Qu.: 1.000
Max.: 7.000
```

Nuestra hipótesis de trabajo es que las diferentes genéticas de cerdas (variable genética) que tenemos en nuestra granja influencian el número de lechones nacidos muertos (variable nac_muer). Para establecer esta hipótesis nuestro equipo técnico ha estudiado las diferentes teorías sobre los lechones nacidos muertos por parto y tras eliminar varias posibles causas, cree que en nuestro

Tabla. Diferencias entre la distribución de Poisson y la distribución de Gauss o Normal.

Distribución de Poisson	Distribución de Gauss o Normal
Variables de contajes o tasas	Variables continuas
Curva asimétrica hacia la derecha dependiendo de su media (lambda, λ)	Curva en forma de campana centrada en la media
Media = Varianza	La media y la varianza son diferentes, pero la media y la moda son iguales

caso la genética de las cerdas es la responsable de ello. Al ser un estudio observacional, se han recogido también otras variables que podrían influenciar el número de lechones nacidos muertos. En este caso, se ha elegido la paridad de las cerdas (variable edad), la nave donde han realizado la gestación (variable nave) y el estado corporal al parto de las cerdas (variable peso).

Debemos ver cómo están distribuidas las cerdas por genética en cuanto a edad, nave y peso para ver que la presencia de la dos genéticas esté equilibrada para las otras tres variables. Para ello crearemos unas tablas de doble entrada y realizaremos un test de Chi-Cuadrado que nos dirá si el número de cerdas en cada pareja de variables es homogéneo o no.

Si vamos a Estadísticos/Tablas de contingencia/Tabla de doble entrada… nos aparecerá un panel donde elegiremos por ejemplo en primer lugar la genética y la edad. En la pestaña Estadísticos, activaremos la opción Porcentajes por filas, aceptaremos y obtendremos la siguiente salida:

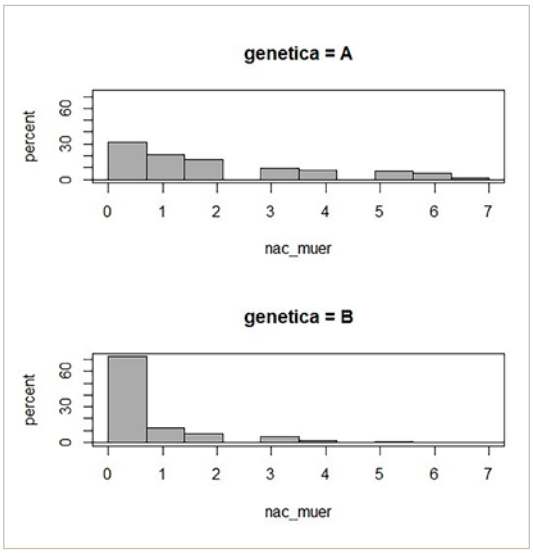

Figura 1. Histograma de los lechones nacidos muertos por genética en porcentaje.

```
Frequency table:
    edad
genética      1 a 2    3 a 6    mas de 6
    A           62       40          33
    B           45       96         235
    Pearson's Chi-squared test
data:  .Table
X-squared = 82.76, df = 2, p-value < 2.2e-16
```

Podemos ver que tenemos 62 cerdas de la genética A de 1 a 2 partos, 45 de la genética B del mismo rango de partos y así sucesivamente. El estadístico Chi-Cuadrado (X-squared) es igual a 82,76, con 2 grados de libertad y el valor p es muy inferior a 0,05, lo que nos informa de que el número de partos no está igualmente distribuido en las diferentes paridades. El lector puede llevar a cabo el mismo procedimiento con las variables nave y peso. Si evaluáramos el efecto de la genética sin tener en cuenta el resto de variables, podríamos estar cometiendo un error, y por tanto es necesario hacer un análisis conjunto.

Podríamos realizar también un análisis gráfico mediante histogramas. Yendo a Gráficas/Histograma… eligiendo en Gráficas por grupos... genética y en las opciones elegimos porcentajes y obtendremos la *figura 1*.

Una vez estudiadas la tabla y la figura anteriores, pasaremos a realizar un análisis de regresión, pero con un modelo lineal generalizado (GLM, en su abreviación en inglés) cuya variable respuesta, en nuestro caso el número de lechones nacidos muertos sigue una distribución diferente a la normal o gaussiana del tal forma que la relación entre la variable respuesta y la predictora no es lineal. Así, los modelos de regresión de Poisson se usan para modelizar variables de contajes y tablas de contingencia. Para ello iremos a Estadísticos/Ajustes de modelos/Modelo lineal generalizado... y nos aparecerá una ventana como la de la *figura 2*, donde construiremos el modelo. Allí, elegiremos la variable *nac_muer* como variable respuesta y el resto de las variables como predictoras. En la parte inferior derecha, elegiremos la *Familia poisson* y en la Función de enlace elegiremos *log* (*figura 3*).

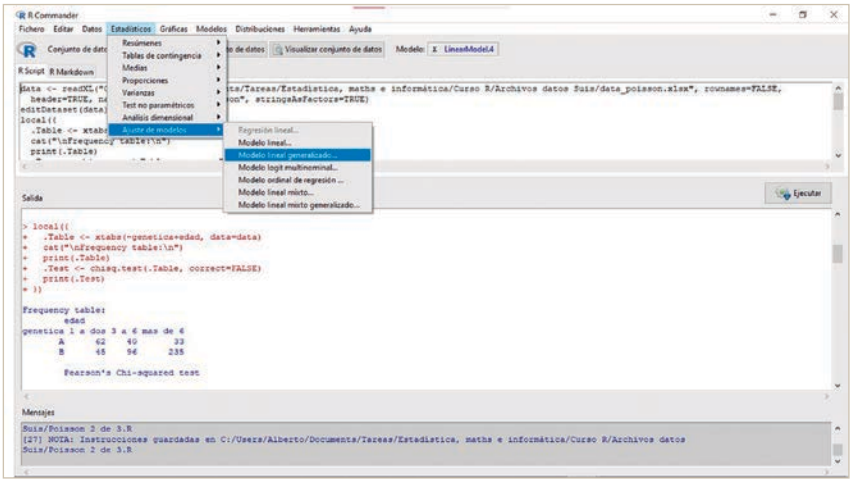

Figura 2. Elección del modelo lineal generalizado en *RCommander*.

Aceptaremos y obtendremos la siguiente salida:

```
Call:
glm(formula = nac_muer ~ genetica+edad+nave+peso,
family = poisson(log), data = data)
Deviance Residuals:
Min        1Q      Median    3Q      Max
-2.4787    -1.2493 -0.7866   0.5244  3.6826

Coefficients:
                    Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)         1.122342   0.105018    10.687   < 2e-16 ***
genetica[T.B]       -1.305209  0.111522    -11.704  < 2e-16 ***
edad[T.3 a 6]       0.003507   0.130616    0.027    0.978580
edad[T.mas de 6]    0.058954   0.123849    0.476    0.634063
nave[T.baja]        -0.124044  0.109401    -1.134   0.256860
peso[T.muy gorda]   -0.393412  0.112328    -3.502   0.000461***
peso[T.normal]      -0.925295  0.117704    -7.861   3.8e-15 ***

---
Signif. codes:  0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 1052.95  on 510  degrees of freedom
Residual deviance:  803.72  on 504  degrees of freedom
AIC: 1331.4
```

Vamos a fijarnos de momento en la tabla de datos que hemos marcado en cursiva y en concreto en la columna Pr(>|z|) donde se nos informa del valor *p*. Podemos ver que la variable edad y nave no son significativas y no están influyendo en el número de nacidos muertos por lo que las eliminaremos del modelo y construiremos uno nuevo (*figura 4*) siguiendo la sistemática anterior, pero esta vez sin las variables edad y nave. Y obtendremos la siguiente salida:

```
Call:
glm(formula = nac_muer ~ genetica+peso, family =
poisson(log), data = data)
Deviance Residuals:
Min        1Q      Median    3Q      Max
-2.4779    -1.2706 -0.7908   0.5627  3.6575

Coefficients:
                    Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)         1.12167    0.08284     13.540   < 2e-16 ***
genetica[T.B]       -1.33584   0.09398     -14.214  < 2e-16 ***
peso[T.muy gorda]   -0.40664   0.11116     -3.658   0.000254***
peso[T.normal]      -0.94845   0.11498     -8.249   < 2e-16 ***

---
Signif. codes:  0 “***” 0.001 “**” 0.01 “*” 0.05 “.” 0.1 “ ” 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 1052.95  on 510  degrees of freedom
Residual deviance:  805.14  on 507  degrees of freedom
AIC: 1326.8
```

En el próximo número interpretaremos todos estos datos y veremos todas las posibilidades que nos ofrece la distribución de Poisson.

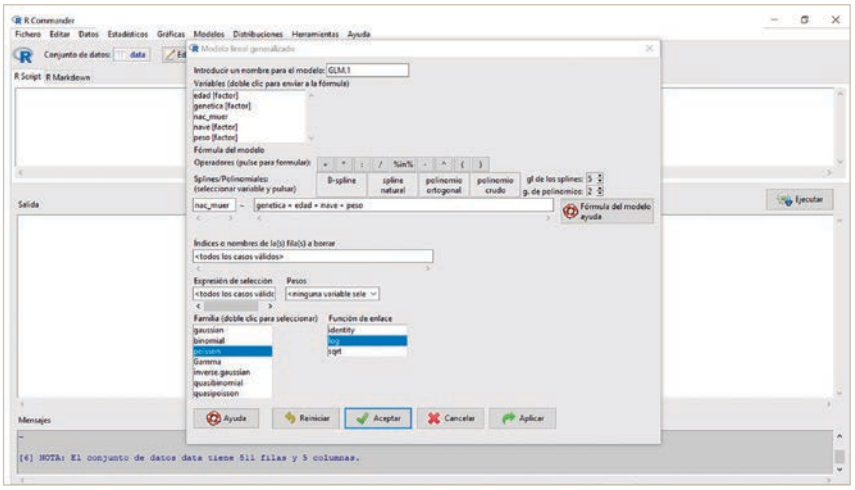

Figura 3. Construcción del modelo Poisson de nuestro ejemplo en *RCommander*.

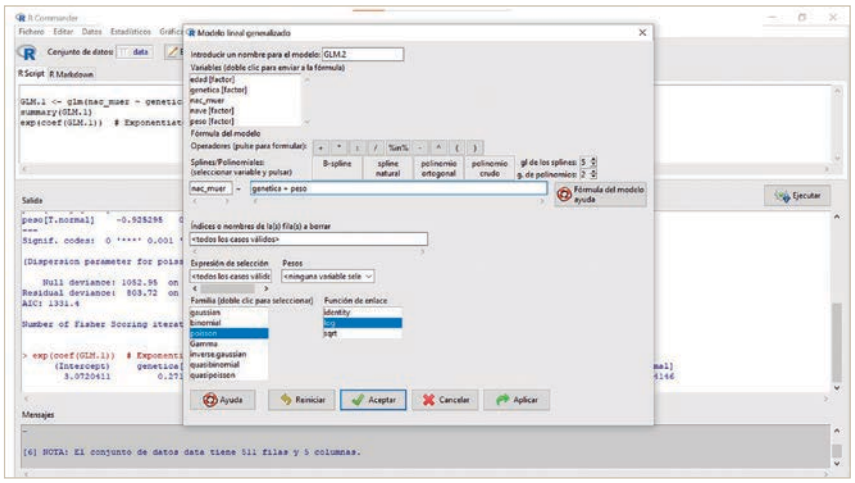

Figura 4. GLM sin las variables edad y nave.