FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Trabajo Fin de Grado:

MODELOS DE REGRESIÓN CON DATOS DE CONTEO. APLICACIÓN A COMPETICIONES DEPORTIVAS

Pablo Atoche Calzada

Grado en Matemáticas

Junio 2017

Dirigido por:
Joaquín García de las Heras
José Luis Pino Mejías

Índice general

Resumen

Este trabajo se centrará en el estudio de los modelos de regresión para datos de conteo, incluidos dentro de la familia de modelos lineales generalizados. Por esta razón, primero se estudiará el modelo lineal generalizado (componentes del modelo, inferencia estadística sobre el MLG, adecuación e interpretación). A continuación, se explicarán y estudiarán diferentes modelos de regresión para estos tipos de datos, como el modelo de regresión de Poisson o el modelo de regresión Binomial Negativa, además de otros modelos menos usuales como el modelo de regresión Poisson Inversa Gaussiana o el modelo de regresión Poisson Generalizado. También se estudiarán los dos problemas fundamentales que se dan en estos modelos: la sobredispersión e infradispersión de los datos y las modificaciones producidas por los ceros. Finalmente, se verá la aplicabilidad de estos modelos en competiciones deportivas, a partir datos reales usando el software estadístico R.

Abstract

This work will focus on the study of regression models for counting data, included within the family of generalized linear models. For this reason, we will first study the generalized linear model (components of the model, statistical inference about MLG, adequacy and interpretation). We will then explain and study different regression models for these types of data, such as the Poisson regression model or the Negative Binomial regression model in addition to other less usual models such as the Poisson Inverse Gaussian regression model or the regression model Generalized Poisson. We will also study the two fundamental problems that happen in these models: overdispersion and underdispersion of data and modifications produced by zeros. Finally, we will see the applicability of these models in sports competitions, from real data using the statistical software R.

Introducción

El presente trabajo se ha estructurado en siete capítulos. En el primero, se realiza una introducción al concepto de modelo matemático y en concreto del modelo estadístico. A continuación, se recogen diferentes resultados sobre el modelo lineal generalizado (MLG) ya que los modelos de conteo se pueden abordar como un caso particular del mismo. Se definen las tres componentes del modelo lineal generalizado: componente aleatoria, componente sistemática y la función enlace. Se obtienen los estimadores de los parámetros desconocidos y se evalúa la adecuación del modelo mediante diferentes medidas de bondad de ajuste y el estudio de los residuos. A continuación, se estudia la importancia de la interpretación del modelo y se finaliza con una breve explicación de los modelos lineales mixtos generalizados.

En el siguiente capítulo, se define que son los datos de conteo o recuento y las variables de conteo. Se dan ejemplos de diversos tipos con orígenes muy diferentes demostrando así su gran utilidad y aplicabilidad. Seguidamente, se expone que son los modelos para datos de conteo y se citan y comparan los modelos que posteriormente se estudiarán.

En el capítulo tres se estudia el modelo de regresión de Poisson, modelo base para modelizar datos de conteo, aún así, el requisito de igualdad de media y varianza (equidispersión) dificulta su aplicabilidad, ya que en numerosas ocasiones los datos presentan mayor varianza que media (datos sobredispersos). Esto motiva el uso de modelos que se adapten mejor a este tipo de datos.

El modelo de regresión Binomial Negativa es el mejor candidato para estudiar datos de conteo con sobredispersión. En el cuarto capítulo se estudia este modelo y se ve que se puede derivar de un modelo de regresión de Poisson compuesto con una distribución Gamma, aunque también se puede pensar como un miembro de la familia de modelos lineales generalizados y por lo tanto aplicar los test de bondad de ajuste, análisis de residuos y cualquier otro estudio de los desarrollados en el primer capítulo.

En el capítulo cinco se estudian otros modelos para datos de conteo que son menos usuales, pero se adaptan mejor a algunos tipos de datos de conteo que los anteriormente estudiados. Entre ellos está el modelo Poisson Inversa Gaussiana, que se trata de un modelo mixto y se utiliza para datos de recuento con un pico inicial muy alto y que

pueden estar sesgados extremadamente a la derecha. Otro modelo que se estudia es el modelo Poisson Generalizado que se suele usar para datos con infradispersión (varianza menor que la media).

Al trabajar con datos de conteo suelen darse problemas con el valor cero. En el sexto capítulo se estudian los modelos truncados por ceros (aquellos en los que la variable no puede tomar el valor 0) y los modelos con excesos de ceros. En esta segunda parte se estudia la diferencia entre los falsos ceros y los auténticos. Se exponen los *modelos inflados* y los *modelos en dos partes* como modelos que se adaptan al exceso de ceros pero siguiendo diferentes estrategias.

En el último capítulo, para ver la aplicabilidad real de la estadística a las competiciones deportivas se menciona el sistema AMISCO usado en la actualidad. Y, a continuación, se realiza una aplicación sobre datos de conteo en competiciones deportivas usando el software R.

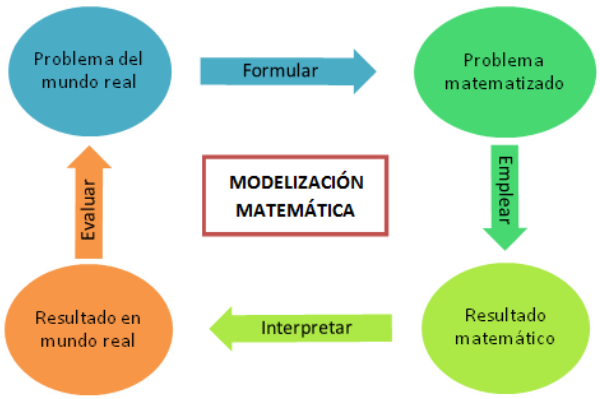
Finalmente, se recoge una revisión bibliográfica con aquellos libros, trabajos, artículos, etc., usados para la realización de este trabajo.

Capítulo 1

Modelo Lineal Generalizado

1.1. Conceptos previos

¿Qué se entiende por un modelo? En las diferentes áreas del conocimiento al realizarse un estudio, el investigador (químico, economista, sociólogo,...) parte de la observación del fenómeno real bajo estudio y desarrolla un sistema o teoría matemática. A través de éste obtiene un modelo abstracto de la realidad que le permite reconstruir el mecanismo interno que se da en el fenómeno real. Por lo tanto, un **modelo** es un proceso de abstracción desde la realidad al sistema matemático con el objetivo de facilitar la comprensión del suceso que se está estudiando. Esta idea se puede reflejar en el siguiente diagrama:



En estos sistemas (científicos, socioeconómicos, biosanitarios,...) el fenómeno real bajo estudio presenta incertidumbre (aleatoriedad), por lo que el modelo resultante del proceso de abstracción debe ser un **modelo estadístico**. Muchos autores entre los que destacan R.A.Fisher, J.Neyman y D.R.Cox han tratado el problema de la modelización estadística, es decir, representar la realidad y su variabilidad a través de un modelo matemático que permita el estudio, el análisis y la comprensión de la misma con el objetivo de transformarla, predecir su futuro, o simplemente conocerla.

Figura 1.1: **Ronald Aylmer Fisher** (1890-1962), matemático-bioestadístico cuyas aportaciones son fundamentales en el desarrollo de los modelos estadísticos.

El objetivo de la modelización estadística es por tanto, a través de la observación o experimentación, explicar el comportamiento de una o más variables en los individuos de una población en base a la diferencia entre las características asociadas a estos individuos (otras variables).

La variable, univariante o multivariante que se desea explicar se conoce como variable objetivo o variable respuesta, mientras que las variables en las que se basan la explicación se denominan variables explicativas.

Por lo tanto se busca un modelo que describa la estructura de la relación entre las variables o variable objetivo y las variables explicativas. Para el planteamiento de este modelo es muy importante distinguir entre el tipo de variable que intervienen (continuas, cualitativas, de conteo,...) y en la clase de relaciones funcionales que se admiten para analizar la relación entre la variable objetivo y las variables explicativas. Según la naturaleza de las variables que intervengan y la relación entre ellas se dispondrá de un conjunto de posibles modelos más o menos eficaces, capaces de explicar la realidad.

La construcción del modelo se realiza atendiendo a las siguientes etapas:

1. **Especificación del modelo teórico**, determminando que variables son de interés, así como cuáles son las relaciones entre ellas. Que el modelo describa de la forma más simple posible, o bien que la concordancia entre el modelo y los datos sea lo más completa posible.

2. **Estimación de parámetros**, calculando el valor de los coeficientes del modelo examinado a partir del conjunto de datos observados, con el objetivo de ver si

el modelo teórico propuesto es aceptable como representación aproximada de los datos.

3. **Selección del modelo**, valorando si el nivel de discrepancia entre entre los datos observados y los datos ajustados es suficientemente bajo como para optar por el modelo, o en caso contrario rechazarlo y buscar otro.

4. **Evaluación del modelo**, examinando las observaciones individuales, los datos influyentes y los datos anómalos, así como comprobando todos los supuestos que caractericen el modelo

5. **Interpretación del modelo**, comprendiendo sus implicaciones con respecto a la variable respuesta.

Generalmente el modelo más utilizado es el tipo **lineal**, es decir, se modeliza la relación tratando de expresar las variables o variable objetivo a través de una combinación lineal de las variables explicativas.

Los modelos que tienen especial interés y que pueden formalizarse a través de la modelización lineal son los siguientes:

- **Modelos para variables de respuesta continuas**, permiten estudiar y analizar el comportamiento de variables contínuas, cuantitativas (ganancias, tiempo de vida, desintegración radiactiva,...) frente a los valores del conjunto de variables explicativas.

- **Modelos para respuestas binarias o binomiales**, permiten considerar variables objetivos del tipo 0-1 (éxito/fracaso), muy útiles en medicina, análisis de riesgos, etc.

- **Modelos para datos de conteo**, los cuales se tratarán posteriormente, permiten considerar y analizar el comportamiento de variables de conteo (número de accidentes, individuos de una especie,...) frente a los valores del conjunto de variables explicativas.

La primera forma de regresión lineal documentada fue el método de los mínimos cuadrados que fue publicada por Legendre en 1805, Gauss publicó un trabajo en donde desarrollaba de manera más profunda este método, en donde además se incluía una versión del teorema de Gauss-Márkov. El término regresión se utilizó por primera vez en el estudio de variables antropométricas al comparar la estatura de padres e hijos, donde resultó que los hijos cuyos padres tenían una estatura muy superior al valor medio, tendían a igualarse a éste, mientras que aquellos cuyos padres eran muy bajos tendían a reducir su diferencia respecto a la estatura media; es decir, regresaban al promedio.

El **modelo lineal clásico** consiste en expresar la esperanza condicionada de la variable objetivo Y como combinación lineal de las variables explicativas X:

$$E[Y|X = x_i] = \mu_i = z_i^t \beta$$

$$y_i = z_i^t \beta + \varepsilon_i$$

donde z_i es un vector de diseño, es decir, una función apropiada de las variables explicativas, β es un vector de parámetros desconocidos y $\varepsilon_i \approx N(0, \sigma^2)$ independientes.

Por lo tanto el modelo lineal clásico consiste en expresar la esperanza condicionada de la variable objetivo como combinación lineal de las variables explicativas bajo la suposición de normalidad y homocedasticidad. Esta modelización lineal clásica se puede extender a una familia mas general [Nelder y Wedderburn, 1972] y ampliada por [McCullagh y Nelder, 1989] conocida como **modelos lineales generalizados (MLG)**. Esta familia permite unificar los modelos con variables de respuestas categóricas como numéricas, y considera otras distribuciones, no únicamente la distribución normal. Además supone que la esperanza μ_i está relacionada con las variables explicativas a través de una función enlace.

Se considera el supuesto de independencia para las observaciones, sin embargo, para esta nueva familia a diferencia del modelo clásico, la distribución puede ser heterocedástica, es decir no se requiere un supuesto de homogeneidad de varianzas.

En un MLG, la media μ de la distribución de la variable objetivo depende de las variables independientes X, a través de la fórmula:

$$E(Y) = \mu = g^{-1}(X^t \beta)$$

donde $E(Y)$ es el valor esperado de Y, $(X^t \beta)$ es el predictor lineal (combinación lineal de parámetros desconocidos), g es la función enlace (función link o vínculo).

1.2. Componentes del MLG

Un modelo lineal generalizado tiene tres componentes básicas: **componente aleatoria** (identifica la variable respuesta cuya distribución pertenece a la familia exponencial), **componente sistemática** (especifica las variables explicativas utilizadas en la función predictor lineal), **función enlace** (es una función del valor esperado de la variable objetivo como una combinación lineal de las variables explicativas)

1.2.1. Componente aleatoria

En un MLG, se asume que la función de distribución de la variable dependiente (variable objetivo) Y pertenece a la familia exponencial.

La distribución de una variable objetivo, caracterizada por los parámetros θ y ϕ pertene a la **familia exponencial** si presenta la forma:

$$f(y;\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

donde f es la función de probabilidad en el caso de que Y sea discreta, o la función de densidad en el caso de que sea continua, θ es el **parámetro canónico o natural**, ϕ es el **parámetro de escala o dispersión** y $a(\phi)$, $b(\theta)$ y $c(y,\phi)$ son funciones específicas de cada elemento de la familia. La función $a(\phi)$ se suele escribir como ϕ/w, donde w es una ponderación para cada observación.

Esta formulación confiere a las distribuciones de esta familia una serie de propiedades algebraicas y estadísticas muy ricas. Algunos miembros de la familia exponencial son las distribuciones normal, exponencial, gamma, beta, Bernoulli, binomial, Poisson, binomial negativa o geométrica. Sin embargo, las distribuciones uniforme y de Cauchy no forman parte de esta familia.

Para cualquier distribución de esta familia se verifica que:

$$E(Y) = \mu = b'(\theta) = \frac{\partial b(\theta)}{\partial \theta}$$

$$Var(Y) = \sigma^2 = a(\phi)\frac{\partial^2 b(\theta)}{\partial \theta^2} = a(\phi)V(\mu)$$

Donde $V(\mu)$ se denomina **función de varianza**. Esta función caracteriza la relación entre $E(y)$ y $Var(y)$.

En la siguiente tabla se resumen los elementos principales que caracterizan a algunas de las distribuciones más utilizadas de la familia exponencial:

Distribuciones	Rango	θ	$a(\phi)$	$b(\theta)$	$V(\mu)$
Bernoulli:$B(p)$	$\{0,1\}$	$\ln(p/(1-p))$	1	$\ln(1+e^{\theta})$	$p(1-p)$
Binomial:$B(n,p)$	$[0,n]$	$\ln(p/(1-p))$	1	$n\ln(1+e^{\theta})$	$np(1-p)$
Normal:$N(\mu,\sigma^2)$	$(-\infty,\infty)$	μ	σ^2	$\theta^2/2$	1
Gamma:$G(\mu,v)$	$(0,\infty)$	$-1/\mu$	$1/v$	-ln(-θ)	μ^2
Poisson:$P(\mu)$	Ent$[0,\infty)$	$\ln(\mu)$	1	e^{θ}	μ
Binom.Negativa:$BN(p,r)$	Ent$[0,\infty)$	ln(1-p)	1	-r(ln(1-e^{θ}))	$\frac{r(1-p)}{p^2}$

1.2.2. Componente sistemática

La componente sistemática también se denomina predictor lineal y se denota η, recoge la variabilidad de Y expresada a través de las p variables explicativas $X_1, ..., X_p$ junto con sus correspondientes parámetros $\beta = (\beta_0\, \beta_1 \, ... \, \beta_p)'$

$$\eta = X'\beta$$

donde $X' = (1\, X_1 \, ... \, X_p)$

1.2.3. Función enlace

Como se ha visto, en el modelo de regresión lineal se expresa el valor esperado de la variable objetivo como una combinación lineal de las variables explicativas, pero en la mayoría de los casos, en experimentos reales, esta relación no es adecuada, por lo que es necesario incluir una función que relacione el valor esperado con las variables explicativas. Esta función se denomina **función enlace o vínculo** y se denota por $g(\mu_i)$.

La función enlace que transforma el valor esperado en el predictor lineal es:

$$g(\mu_i) = \eta_i = X_i^t\beta$$

donde $X_i^t = (1\, X_{i1} \, ... \, X_{ip})$ representa las p variables explicativas para el i-ésimo individuo con $i = 1, ..., n$, donde n es el tamaño de la muestra (número de individuos).

La función inversa de la función enlace se denota como h, de modo que se verifica:

$$\mu_i = g^{-1}(\eta_i) = h(\eta_i) = h(X_i^t\beta)$$

La elección de la función enlace no es simpre evidente, pueden existir varias funciones enlaces aplicables a un problema particular de regresión, luego hay que decidir cuál es la más apropiada en cada caso. Por lo tanto es muy importante la elección de una buena función enlace que facilite la interpretación del modelo óptimo obtenido.

Para cada elemento de la familia exponencial existe una función enlace denominada **función enlace canónica o natural**, que consisten en relacionar el parámetro natural directamente con el predictor lineal:

$$\theta_i = \theta(\mu_i) = \eta_i = X_i'\beta \quad g(\mu_i) = \theta(\mu_i)$$

Así, para las siguientes distribuciones se tiene la función enlace canónica correspondiente:

Bernoulli:

$$g(\mu_i) = \theta(\mu_i) = ln(\frac{\mu_i}{1-\mu_i}) = X_i^t\beta = \eta_i$$

Binomial:

$$g(\mu_i) = \theta(\mu_i) = ln(\frac{\mu_i}{1-\mu_i}) = X_i^t\beta = \eta_i$$

Normal:

$$g(\mu_i) = \theta(\mu_i) = \mu_i = X_i^t\beta = \eta_i$$

Gamma:

$$g(\mu_i) = \theta(\mu_i) = -\frac{1}{\mu_i} = X_i^t\beta = \eta_i$$

Poisson:

$$g(\mu_i) = \theta(\mu_i) = ln(\mu_i) = X_i^t\beta = \eta_i$$

Binomial Negativa:

$$g(\mu_i) = \theta(\mu_i) = ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) = X_i^t\beta = \eta_i$$

donde $\alpha = 1/r$

Podemos por lo tanto definir también $h = g^{-1}$ en función del predictor lineal η_i para cada distribución:

Bernoulli:

$$h(X_i^t\beta) = g^{-1}(\eta_i) = \theta^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

Binomial:

$$h(X_i^t\beta) = g^{-1}(\eta_i) = \theta^{-1}(\eta_i) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

Normal:

$$h(X_i^t\beta) = g^{-1}(\eta_i) = \theta^{-1}(\eta_i) = \eta_i$$

Gamma:

$$h(X_i^t\beta) = g^{-1}(\eta_i) = \theta^{-1}(\eta_i) = -\frac{1}{\eta_i}$$

Poisson:

$$h(X_i^t\beta) = g^{-1}(\eta_i) = \theta^{-1}(\eta_i) = e^{\eta_i}$$

Binomial Negativa:

$$h(X_i^t\beta) = g^{-1}(\eta_i) = \theta^{-1}(\eta_i) = \frac{1}{\alpha(e^{-\eta_i}-1)}$$

donde $\alpha = 1/r$

1.3. Inferencia en el modelo lineal generalizado

Una vez elegido el modelo o modelos, se estima para cada uno de ellos el valor de los parámetros del predictor lineal y a continuación se valora la precisión de esas estimaciones a través del cálculo de la discrepancia entre pares de modelos, con el objetivo de seleccionar el mejor.

Dos de los métodos más estudiados y comunes de la estimación paramétrica son el método de **Mínimos Cuadrados** y el método de **Máxima Verosimilitud**, el cuál es el más adecuado pues tiene las propiedades de consistencia y eficiencia asintótica.

Para poder realizar un estudio se necesita disponer de una **muestra aleatoria** $Y = y_1, ..., y_n$ junto con sus correspondientes variables explicativas $x_1, ..., x_i, ..., x_n$ (x_i es un vector de dimensión p, es decir, del número de variables explicativas que se estén estudiando para los n individuos). Este método trata de maximizar la verosimilitud para obtener un estimador del vector de parámetros desconocidos $\beta = (\beta_0\, \beta_1 \dots \beta_p)'$ en el modelo muestral:

$$E[Y/X_i = x_i] = \mu_i = h(x_i'\beta) \qquad i = 1, ..., n$$

Supuesto que el parámetro de escala ϕ es conocido, y dado que aparece como factor en la función de verosimilitud, se puede considerar $\phi = 1$, sin pérdida de generalidad. Posteriormente se puede obtener un estimador de ϕ a través del método de los momentos.

Asumiendo que Y pertenece a la familia exponencial, la **función de verosimilitud** viene dada por:

$$L(\theta; y) = f(y; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$$

donde $y = (y_1 \dots y_n)'$

Por lo tanto la función **log-verosimilitud** viene dada por:

$$l(\theta, \phi, y) = \sum_{i=1}^{n} l_i(\theta, \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi) \right\}$$

Se omite la función $c(y_i, \phi)$ que no depende de θ_i y se inserta la relación $\theta_i = \theta(\mu_i)$ entre el parámetro natural y la esperanza de la i-ésima observación, obteniéndose la función:

$$l(\mu_i, \phi, y) = \sum_{i=1}^{n} l_i(\mu_i, \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i\theta_i(\mu_i) - b(\theta_i(\mu_i))}{a(\phi)} \right\}$$

Debido a la relación entre la esperanza y la inversa de la función enlace (h) aplicada al predictor lineal, se tiene que $\mu_i = h(x_i'\beta)$, por lo tanto:

$$l(\beta, \phi, y) = \sum_{i=1}^{n} l_i(\beta, \phi, y_i) = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i(h(x_i'\beta)) - b(\theta_i(h(x_i'\beta)))}{a(\phi)} \right\}$$

La primera derivada respecto a β de la función log-verosimilitud es la llamada **función marcador o función score**:

$$s(\beta) = \frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} s_i(\beta)$$

Las contribuciones individuales a la función marcador, es decir, la expresión de cada uno de los sumandos es:

$$s_i(\beta) = x_i D_i(\beta) \sigma_i^{(-2)}(\beta)[y_i - \mu_i(\beta)]$$

donde

$$\begin{cases} \mu_i(\beta) = h(x_i'\beta) \\ \sigma_i^{-2}(\beta) = a(\phi) v(h(x_i'\beta)) \\ V(\mu) = \partial^2 b(\theta)/\partial\theta^2 \\ D_i(\beta) = \partial h(x_i'\beta)/\partial\eta \, \text{(primera derivada de la función } h \text{ evaluada en } \eta_i = x_i'\beta) \end{cases}$$

Aplicando la regla de la cadena se obtiene lo siguiente:
(NOTA: Se cambia $'$ por t para denotar el vector traspuesto, y se utiliza $'$ para denotar la derivada en el siguiente desarrollo)

$$\frac{\partial}{\partial \beta} \theta(h(x_i^t\beta)) = \theta'(h(x_i^t\beta)) h'(x_i^t\beta) x_i = x_i D_i(\beta) \theta'(h(x_i^t\beta)) \tag{1.1}$$

$$\begin{aligned} \frac{\partial}{\partial \beta} b(\theta(h(x_i^t\beta))) &= \frac{\partial}{\partial \theta} b(\theta(h(x_i^t\beta))) \frac{\partial}{\partial h} \theta(h(x_i^t\beta)) \frac{\partial}{\partial \eta_i} h(x_i^t\beta) x_i = \\ &= \mu_i(\beta) \frac{\partial}{\partial h} \theta(h(x_i^t\beta)) D_i(\beta) x_i = \mu_i(\beta) \left(\frac{\partial}{\partial \mu_i} \theta(\mu_i) \right) D_i(\beta) x_i \end{aligned} \tag{1.2}$$

Para obtener la derivada que aparece en (1.1) y (1.2):

$$\mu(\theta) = b'(\theta) = \frac{\partial b(\theta}{\partial \theta} \Longrightarrow \frac{\partial \mu(\theta)}{\partial \theta} = b''(\theta)$$

por la derivada de la función inversa:

$$\frac{\partial}{\partial \mu_i}\theta(\mu_i) = \frac{1}{b''(\theta(\mu_i))} = \frac{1}{V(\mu_i)} = a(\phi)\sigma_i^{-2}(\beta) \tag{1.3}$$

Sustituyendo (1.3) en (1.1) y (1.2) se obtiene:

$$\frac{\partial}{\partial \beta}\theta(h(x_i^t\beta)) = a(\phi)x_i D_i(\beta)\sigma_i^{-2}(\beta) \tag{1.4}$$

$$\frac{\partial}{\partial \beta}b(\theta(h(x_i^t\beta))) = a(\phi)\mu_i(\beta)\sigma_i^{-2}(\beta)D_i(\beta)x_i \tag{1.5}$$

Y teniendo en cuenta (1.4) y (1.5) se llega a:

$$s_i(\beta) = \frac{\partial}{\partial \beta}l_i(\beta, \phi, y_i) = y_i x_i D_i(\beta)\sigma_i^{-2}(\beta) - \mu_i(\beta)x_i D_i(\beta)\sigma_i^{-2}(\beta) =$$
$$= x_i D_i(\beta)\sigma_i^{-2}(\beta)(\beta)\left[y_i - \mu_i(\beta)\right]$$

Otros conceptos importantes que aparecen en la estimación máximo-verosímil del vector de parámetros son:

- Matriz de información de Fisher esperada:

$$F(\beta) = Cov\, s(\beta) = \sum_i F_i(\beta)$$
$$F_i(\beta) = x_i x_i^t w_i(\beta), \text{ donde } w_i(\beta) = D_i^2(\beta)\sigma_i^{-2}(\beta)$$

- Matriz de información de Fisher observada:

$$F_{obs}(\beta) = -\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^t}$$

  Se puede comprobar que $F(\beta) = E(F_{obs}(\beta))$

Para las funciones enlaces naturales o canónicas $\theta(\mu_i) = x_i^t\beta$, se simplifica la forma de las matrices de información así como la función marcador:

$$s(\beta) = \frac{1}{a(\phi)}\sum_i x_i[y_i - \mu_i(\beta)]$$

$$F(\beta) = \frac{1}{a(\phi)}\sum_i V(\mu_i(\beta))x_i x_i^t, \text{ donde} F(\beta) = F_{obs}(\beta)$$

La obtención de la estimación máxima-verosímil no se plantea, generalmente, como el cálculo de un máximo global, sino como las soluciones de las ecuaciones de verosimilitud $s(\hat{\beta}) = 0$, lo que corresponde a un máximo local, es decir, con la matriz de segundas derivadas $F_{obs}(\hat{\beta})$ definida negativa. Estas ecuaciones no son normalmente lineales por lo que han de ser resueltas con métodos iterativos. En muchos casos se utiliza el método iterativo *marcador de Fisher* o *mínimos cuadrados ponderados iterados*, cuyas iteraciones vienen definidas, a partir de un estimador inicial $\hat{\beta}^{(0)}$, por:

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + F^{-1}(\hat{\beta}^{(k)})s(\hat{\beta}^{(k)}) \quad k = 0, 1, 2, ...$$

Algunas consideraciones sobre este método:

1. El parámetro escala ϕ se cancela en el término $F^{-1}(\hat{\beta}^{(k)})s(\hat{\beta}^{(k)})$ por lo que toma sentido el comentario inicial realizado sobre dicho parámetro.
2. Como punto inicial del proceso iterativo $\hat{\beta}^{(0)}$ se puede utilizar el estimador mínimo cuadrático de los puntos $(g(y_i), x_i)$.
3. El proceso iterativo suele terminar con el criterio:
   $$\frac{\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\|}{\|\hat{\beta}^{(k)}\|} < \varepsilon \quad \text{para un } \varepsilon > 0 \text{ fijado}$$

Otros métodos alternativos son el método de *Newton-Raphson* y métodos Quasi-Newton.

Por lo tanto se obtienen a través de estos métodos las estimaciones de los parámetros β del modelo. Estas estimaciones máximo verosímiles presentan las siguientes propiedades:

- **Existencia y unicidad asintótica**: La probabilidad de que exista una solución y sea (localmente) única tiende a 1 cuando $n \longrightarrow \infty$.
- **Consistencia**: Si se denota por β el "verdadero" valor del parámetro, la solución converge a β cuando $n \longrightarrow \infty$ en probabilidad (consistencia débil) o con probabilidad 1 (consistencia fuerte).
- **Normalidad asintótica**: $\hat{\beta}$ es asintóticamente normal con matriz de varianzas y covarianzas $F^{-1}(\hat{\beta})$, es decir:
  $$\hat{\beta} \overset{a}{\sim} N_p(\beta, F^{-1}(\hat{\beta}))$$
  En el caso de que el parámetro de dispersión ϕ sea desconocido, el resultado sigue siendo válido si se sustituye dicho parámetro por un estimador consistente de él.

En este último caso para ϕ desconocido, puede considerarse el siguiente estimador consistente:

$$\hat{\phi} = \frac{1}{n-p} \sum_i \frac{[y_i - \mu_i(\hat{\beta})]^2}{v(\mu_i(\hat{\beta}))}$$

1.4. Adecuación e interpretación de los modelos

Una vez estimados los parámetros, se debe valorar la discrepancia que hay entre los datos observados y los esperados por el modelo.

Si se admite una combinación satisfactoria de la distribución de la componente aleatoria y la función enlace, el objetivo es determinar cuantos términos son necesarios en la estructura lineal para una descripción razonable de los datos. Por ejemplo un gran número de variables explicativas puede hacer que un modelo explique bien los datos pero aumenta la complejidad de su interpretación. Y al contario, un modelo con pocas variables explicativas puede ser de fácil interpretación pero se ajustará poco a los datos. Lo que se busca es un modelo intermedio.

En el proceso del ajuste del modelo se evalúan un conjunto de modelos que constituyen unas aproximaciones de los datos observados. Y se trata de construir un modelo intermedio entre los modelos siguientes:

- **Modelo saturado**: El número de parámetros estimados es igual al número de observaciones. En estos datos individuales, este modelo no constituye parámetros a estimar, sólo reproduce lo que está ocurriendo.
- **Modelo nulo**: Es muy simple, se utiliza como modelo de referencia. Contiene como único parámetro el valor esperado μ para todas las observaciones. Es incapaz de representar adecuadamente la estructura de los datos, asume un efecto nulo de las variables explicativas.

Obviamente el concepto o idea de *mejor* o *peor* modelo depende de la finalidad del modelo. Cuando la finalidad es de tipo **predictivo** se seleccionan las variables que expliquen el mayor porcentaje de variabilidad de la variable objetivo. Cuando la finalidad es **explicativa** tienen mayor protagonismo los argumentos teóricos, por lo que el proceso de selección de variables debe ser guiado por el investigador y el proceso de adecuación seguirá estos tres pasos aproximadamente:

1. Evaluar los términos de interacción a través de su significación estadística.
2. Analizar la necesidad de mantener variables en el modelo. Utilizando criterios de relevancia práctica más que estadística, se debe evaluar si:

a) La eliminación de variables confusas sesgará la estimación de los parámetros de interés.

b) Si la eliminación de una variable implicará una pérdida de precisión de las estimaciones.

3. Valorar si las variables explicativas de interés deben permanecer en el modelo. Para comprobarlo se emplean tanto criterios estadísticos como criterios de relevancia.

En conclusión, el objetivo del proceso de modelado es la obtención de un modelo que sea capaz de representar los datos y al mismo tiempo reducir la complejidad.

1.4.1. Bondad de ajuste

La medida de la bondad de ajuste de un modelo describe lo bien que se ajusta el mismo a un conjunto de observaciones. Las medidas de bondad en general resumen la discrepancia entre los valores observados y los valores esperados en el modelo de estudio. En el modelo lineal generalizado la bondad de ajuste se puede evaluar de diferentes maneras entre las que destacan:

Función o estadístico desviación D:

$$D(y;\hat{\mu}) = 2\phi\left\{l(y;y) - l(\hat{\mu};y)\right\}$$

Se trata de la distancia entre el logaritmo de la función de verosimilitud del modelo saturado y el modelo que se está estudiando. Un valor pequeño de la desviación indica que para un número menor de parámetros se obtiene un ajuste tan bueno como cuando se ajusta el modelo saturado. La notación $l(y;y)$ es un reflejo de que el modelo saturado selecciona perfectamente la variable respuesta. Así los valores que predice el modelo son $\hat{\mu} = y$.

Para probar la adecuación de un MLG, el valor de la desviación debe ser comparado con el percentil de alguna distribución de probabilidad referente. Si el modelo es adecuado el estadítico desviación se distribuye asintóticamente según una χ^2_{n-p} con $n-p$ grado de libertad [McCullagh y Nelder, 1989].

$$D(y;\hat{\mu}) \approx \chi^2_{n-p}$$

Coeficiente de determinación R^2: determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo. Viene dado por:

$$R^2 = 1 - \frac{D(y;\mu)}{D(y;\mu_0)}$$

donde $D(y;\mu)$ y $D(y;\mu_0)$ son las funciones de desviación del modelo ajustado y nulo respectivamente, y se verifica $0 \leq R^2 \leq 1$. El valor del coeficiente de determinación aumenta cuando se incluyen nuevas variables en el modelo, incluso cuando éstas son poco significativas o tienen poca correlación con la variable dependiente. El coeficiente de determinación corregido mide el porcentaje de variación de la variable dependiente (al igual que el coeficiente de determinación) pero tiene en cuenta además el número de variables incluidas en el modelo.

Estadístico Chi-cuadrado de Pearson:

$$\chi^2 = \sum_{1=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

donde $V(\mu_i)$ es la función varianza estimada para la distribución de la variable respuesta.

Análisis de las variable explicativas

A continuación se realiza inferencia estadística sobre el vector de parámetros desconocidos $\beta = (\beta_0\, \beta_1 \ldots \beta_p)^t$ de dimensión $p+1$.

La mayoría de las cuestiones que se plantean en la realidad pueden ser formuladas a través de la hipótesis lineal de la forma $\mathbf{C}\beta$, siendo $\mathbf{C}$ una matriz de rango total $s \leq p+1$, y ξ un vector de constantes conocido de dimensión s se tiene:

$$H_0 : \mathbf{C}\beta = \xi$$
$$H_1 : \mathbf{C}\beta \neq \xi$$

En general se recogen los siguientes procedimientos para este contraste:

- **Estadístico de razón de Verosimilitud.** Definido por:

  $$\Lambda_{RV} = -2l(\overline{\beta}; y) - l(\hat{\beta}; y)$$

  que compara el máximo del logaritmo de verosimilitud con el máximo obtenido bajo la restricción definida por H_0.

- **Estadístico de Wald.** Se basa en la distribución normal asintótica del vector β. Se define por:

  $$\xi_W = [\mathbf{C}\hat{\beta} - \xi]^t [\mathbf{C}F^{-1}(\hat{\beta})\mathbf{C}^t][\mathbf{C}\hat{\beta} - \xi]$$

  Determina la distancia ponderada entre el estimador $\mathbf{C}\hat{\beta}$ y $\mathbf{C}\beta$ y su valor determinado por la hipótesis nula, donde $F^{-1}(\hat{\beta})$ denota la estimación de la matriz de información de Fisher de $\hat{\beta}$.

- **Estadístico score**. Se obtiene a partir de la función score. Se basa en el hecho de que la función score se anula en el estimador de máxima verosimilitud, por lo que la evaluación de ésta en el estimador obtenido bajo restricción lineal, el resultado será significativamente diferente de cero si la hipótesis nula no es cierta. Así se utiliza la distancia ponderada de $s(\beta)$ a cero, es decir:

$$\xi_{SR} = [s(\overline{\beta})]^t F^{-1}(\overline{\beta}) s(\overline{\beta})$$

donde $F(\hat{\beta})$ es la matriz de covarianza asintótica estimada bajo $H_0 : \mathbf{C}\beta = \xi$.

Asintóticamente y bajo hipótesis nula, los tres estadísticos definidos anteriormente se distribuyen según una ley Chi-cuadrado con s grados de libertad χ^2_s.

Para las hipótesis relativas a un único coeficiente β_i, el estadístico de Wald es el más utilizado. Éste coincide con el cuadrado del estadístico t^2:

$$t_j = \frac{\hat{\beta}}{\hat{a_{jj}}}$$

donde $\hat{a_{jj}}$ es el elemento j-ésimo diagonal de la matriz de covarianzas asintóticas $F(\hat{\beta})$ de $\hat{\beta}$.

Para hipótesis relativas a varios coeficientes, el test de razón de verosimilitud es preferible por ser un test uniformemente más potente [Cordeiro, 2000].

Regiones de confianza para β

Las regiones de confianza pueden construirse usando cualquiera de los estadísticos propuestos anteriormente. Usando por ejemplo el estadístico de Wald, una región de confianza para β con un nivel de confianza $1 - \alpha$ viene dada por:

$$\{\beta \in \mathbb{R}^{p+1} \mid (\hat{\beta} - \beta)^t [Var(\hat{\beta})]^{-1} (\hat{\beta} - \beta) < \chi^2_{p+1,n-1}\}$$

1.4.2. Residuos

A veces puede ocurrir que aún escogiendo cuidadosamente un modelo al ajustarlo posteriormente a un conjunto de datos el resultado sea insatisfactorio. Estos errores se originan por no haber elegido bien la función enlace o las variables explicativas incluidas en el modelo. Las discrepancias aisladas se han podido producir debido a algún dato erróneo. La verificación de la adecuación del modelo es fundamental para analizar posibles desviaciones o la existencia de observaciones anómalas (outliers).

Como en la regresión lineal, los residuos son los utilizados para verificar la adecuación del modelo. Expresan la diferencia entre una observación y su valor ajustado, y también

la presencia de valores anómalos que requieran una atención más detallada. Los residuos mas destacados son:

- **Residuos básicos**: Se trata de la diferencia entre el valor observado y_i, de la variable respuesta y el valor ajustado $\hat{y}_i$ por el modelo:

$$r_i^b = y_i - \hat{y}_i \quad \text{con } i = 1, ..., n$$

- **Residuos de Pearson**: Se definen como:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} Var(\hat{\mu}_i)}} \text{ con} \quad i = 1, ..., n$$

  siendo $\hat{\phi}$ un estimador consistente del parámetro escala ϕ. Y su versión estundentadizada viene dada por:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} Var(\hat{\mu}_i(1 - h_i)}} \quad i = 1, ..., n$$

  siendo h_i el elemento diagonal de la matriz de proyección, H, donde:

$$H = W^{1/2} X (X^T W X)^{-1} X^T W^{1/2}$$

  con W una matriz diagonal, cuyos elementos de la diagonal principal vienen dados por:

$$w_i = \frac{1}{Var(\mu_i)} \left(\frac{\partial \mu_i}{\partial \eta} \right)^2$$

  La ventaja de usar este residuo estundentadizado frente al anterior es que capta mejor la variabilidad de los datos, debido a que usa el valor de h_i, el cual es útil para medir la influencia de la i-ésima observación.

- **Residuos de desviación**: se definen como:

$$r_i^D = sign(y_i - \hat{\mu}_i) \sqrt{d_i} \quad i = 1, ..., n$$

  donde d_i es llamada la componente desviación, $d_i = 2(l(y_i, y_i) - l(\hat{\mu}_i, y_i))$

  Y su versión estundentadizada:

$$r_i^{D'} = \frac{r_i^D}{\sqrt{\hat{\phi}(1 - h_i)}}$$

  donde h_i es el i-ésimo elemento de la diagonal de la matriz H y $\hat{\phi}$ es la estimación del parámetro escala ϕ.

- Según otros autores otros residuos importantes pero menos utilizados son los siguientes: Residuos de Anscombe, Residuos de puntuación (score residuals), Residuos parciales o Residuos de probabilidad.

1.4.3. Interpretación del modelo

Una vez obtenido el modelo adecuado, usando los criterios de bondad de ajuste y estudio de los residuos, se finaliza el proceso de la modelización con la interpretación del modelo elegido.

Esta parte es sin duda esencial, pues una mala interpretación del modelo hace que todo lo anterior pierda su utilidad, pues nunca hay que olvidar que el analista o matemático todo el tiempo tiene como objetivo modelizar y estudiar un sistema a partir de unos datos para poder llegar a conclusiones que expliquen la relación entre variables e incluso puedan servir para preveer situaciones futuras.

1.5. Modelos lineales mixtos generalizados

En general, los modelos paramétricos se basan en el análisis de relaciones lineales entre cierta variable de interés (variable respuesta) y ciertas variables que contribuyen a explicar su comportamiento (variables explicativas). Estas características suelen medirse en individuos o unidades experimentales que son independientes entre sí. Además se asume que el error que se produce tiene la propiedad de seguir una distribución Normal homogénea.

Puede suceder que la variable objetivo sólo tome valores en un intervalo o bien, que no sea continua o ni siquiera cuantitativa. Por lo tanto los errores no pueden ser Normales. Esta generalización da lugar como se ha visto anteriormente a los modelos lineales generalizados (MLG).

Otra posible generalización consiste en reducir restricciones sobre los errores manteniendo sin embargo la propiedad de Normalidad, es decir, contemplando errores no independientes o heterogéneos. Esta generalización que permite dotar de estructura a la variabilidad de los errores del modelo, da lugar a los modelos mixtos.

Los modelos lineales generalizados y los modelos mixtos pueden ser fusionados dando lugar a los modelos lineales mixtos generalizados (MLMG) adaptando las propiedades de ambas propuestas de modelización. Su aplicación es enorme en multitud de estudios de distintos ámbitos.

Luego, en resumen, los **modelos lineales mixtos generalizados (MLMG)** tienen las siguientes características:

- La variable objetivo está linealmente relacionada con la variable explicativa mediante una función enlace especificada.
- La variable objetivo puede tener una distribución no normal.

- Puede existir correlación entre las observaciones.

Los MLMG cubren una amplia variedad de modelos, desde modelos de regresión lineal simplie hasta modelos altamente complejos.

Ejemplo: Los responsables educativos pueden utilizar un modelo lineal mixto generalizado para determinar si un método educativo es eficaz para mejorar las notas en una asignatura. Los estudiantes de la misma clase deben correlacionarse ya que les enseña el mismo profesor, a su vez, las clases del mismo colegio también deben correlacionarse. Por lo tanto, se pueden incluir efectos aleatorios a nivel de colegio y de clase para explicar las diferentes fuentes de variabilidad.

Capítulo 2

Datos y variables de conteo

2.1. Datos de conteo

Al tratar el modelado de datos de conteo es importante aclarar que se entiende por un **recuento** o **conteo**, además de que son los **datos de conteo** y las **variables de conteo**.

Contar o hacer un recuento no es más que enumerar unidades, artículos o eventos en un intervalo espacial o temporal. Podemos contar elementos tan dispares como por ejemplo el número de accidentes observados en un tramo de carretera o el número de erupciones solares observadas por año. Por otra parte, entendemos por datos de conteo a las observaciones sobre eventos o sucesos que contamos.

Se denominan **variables de conteo o recuento** (count data) a aquellas que determinan el número de sucesos o eventos que ocurren en una misma unidad de observación en un intervalo espacial o temporal definido [Lindsey, 1995]. Luego, a partir de esta definición, se pueden observar dos características fundamentales en una variable de conteo: su naturaleza discreta y no negativa. Esta variable objetivo, que toma los valores 0, 1, 2,... se caracteriza por tomar infinitos números de valores que podemos ordenar en orden creciente y cuya variabilidad va en descenso a medida que sea mayor el valor de la variable.

Ejemplos de variables de conteo son:

Conteo de artículos o sucesos que ocurren en un área geográfica o espacial dada:

- Número de colores primarios en un cuadro.
- Número de hijos de cada mujer en un municipio.
- Número de medallas conseguidas en deportistas olímpicos ya retirados.
- Número de preguntas acertadas en un examen tipo test.

Conteo de eventos o sucesos que ocurren dentro de un periodo de tiempo:

- Número de visitas a un museo en un mes.
- Número de artículos vendidos en una tienda deportiva durante un año.
- Número de accidentes en una carretera en un año.
- Número de lesiones deportivas por equipo durante una competición.
- Número de goles por jugador de fútbol en una temporada.

A continuación, se muestra un ejemplo real de datos de conteo que provienen de competiciones deportivas realizado mediante el software R:

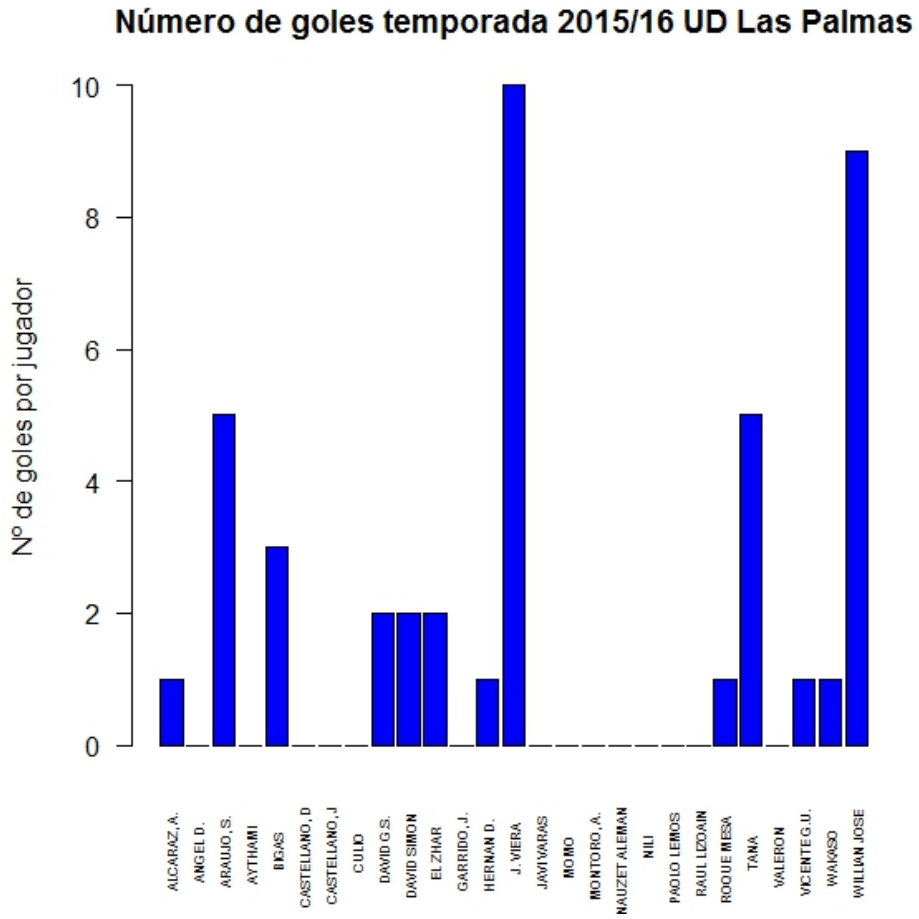


Figura 2.1: Goles por jugador UD Las Palmas

2.2. Modelos para datos de conteo

Los modelos de datos de conteo se caracterizan teóricamente porque no tienen un límite superior natural, aunque desde el punto de vista práctico siempre se limitan a un valor, generalmente el valor máximo del conteo que se está estudiando. También se caracterizan por tomar el valor cero (en un porcentaje no despreciable) para algunos miembros de la población estudiada y no tomar valores muy altos.

No todos los modelos de predicción son aplicables a este tipo de variable, pues pueden surgir problemas como:

1. Las predicciones de la variable respuesta pueden salirse del rango de valores en el que está definido.
2. Las estimaciones pueden ser inconsistentes.
3. Si la variable objetivo toma más de dos valores, plantear un modelo de elección binaria para su estudio conduce a una pérdida de eficacia (se pierde información) ya que se tratarán varios valores como un solo valor.
4. No tienen porqué cumplirse la hipótesis de normalidad u homocedasticidad.

Modelos que tienen especial interés para datos de conteo son **Poisson (P)** y **Binomial Negativa (BN)**, estos modelos de regresión permiten considerar y analizar el comportamiento de variables de conteo frente a los valores del conjunto de variables explicativas. Otros modelos para datos de conteo son los modelos **Poisson inversa Gaussiana (PIG)**, **Binomial Negativa de tres parámetros de Greene (BN-P)** y **Poisson generalizado (PG)**.

La distribución de Poisson tiene un único parámetro a estimar, μ, la media, el cual a veces es llamado parámetro de localización. La principal característica de esta distribución es que la media y la varianza son iguales. Luego cuanto mayor sea la media mayor será la varianza o variabilidad en los datos. Esta propiedad en la distribución de Poisson es llamada **propiedad de equidispersión**. La sobredispersión es un gran problema para la regresión de Poisson y ocurre cuando la varianza es mayor que la media. Probablemente el método más popular de tratar con la sobredispersión es modelar los datos usando un modelo de regresión Binomial Negativa, el cuál incorpora un parámetro de dispersión α. En cambio para la infradispersión o subdispersión (varianza del modelo menor que la media) no se puede usar el modelo de regresión Binomial Negativa. Sobre todo esto se profundizará en los siguientes capítulos.

La razón por la que el modelo **Poisson inversa Gaussiana (PIG)** no ha tenido amplio uso es que hasta hace poco no había software para su estimación, a menos que los analistas de datos crearan el software ellos mismos. La programación del modelo

PIG no es fácil, si no se programa bien, un algoritmo para PIG puede tomar mucho tiempo para converger y obtener estimaciones de parámetros apropiadas.

El modelo **Poisson Generalizado (PG)** incorpora respecto al modelo de Poisson un segundo parámetro α, también denominado dispersión o parámetro escala. El modelo de Poisson generalizado al igual que todos los anteriormente introducidos se reduce al modelo de Poisson cuando dicho parámetro es 0. La ventaja es que el parámetro de dispersión puede tomar valores negativos, lo cual puede proporcionar un buen ajuste para datos con subdispersión.

Respecto al modelo **Binomial Negativa de tres parámetros de Greene (BN-P)** diseñado por William Greene en la universidad de Nueva York tiene como tercer parámetro ρ, que permite que la dispersión varíe a través de la observación, proporcionando una mejor oportunidad para ajustar datos binomiales negativos. La función de distribución de probabilidad subyacente al modelo es una variación de la distribución binomial negativa y es usualmente empleada por los analistas de datos para decidir si emplear un modelo BN1 o modelo BN2 para modelizar los datos, el significado de BN1 y BN2 se estudiará posteriormente en el capítulo sobre el modelo de regresión Binomial Negativa.

En la siguiente tabla se muestra la media y la varianza de estos modelos de conteo muy relacionados entre sí:

Modelo para datos de conteo	Media	Varianza
Poisson	μ	μ
Binomial Negativa (BN1)	μ	$\mu(1+\alpha)$
Binomial Negativa (BN2)	μ	$\mu(1+\alpha\mu)$
Poisson inversa Gaussiana	μ	$\mu(1+\alpha\mu^2)$
Binomial Negativa (BN-P)	μ	$\mu(1+\alpha\mu^\rho)$
Poisson Generalizado	μ	$\mu(1+\alpha\mu)^2$

También se utilizan modelos para variables de conteo que no pertenecen a la familia de modelos lineales generalizados como el **Modelo en dos partes** y **Modelo con exceso de ceros**.

Los modelos para variables de conteo se encuentran en la intersección entre el modelo lineal generalizado y el estudio de las variables de conteo o recuento. De entre todos estos modelos destaca, por su papel como modelo de referencia en el estudio de variables de conteo, el modelo de regresión de Poisson.

Capítulo 3

Modelo de Regresión de Poisson

El modelo de Regresión de Poisson es fundamental para modelizar datos de conteo. Este fue el primer modelo usado específicamente para datos de conteo, y es todavía la base de muchos tipos de modelos de conteo usados por analistas de datos. Sin embargo el asumir la propiedad de equidispersión (media = varianza) hace que su uso en estudios con datos reales sea normalmente insatisfactorio. A veces, es posible hacer ajustes para remediar el problema de la sobredispersión o el problema de la infradispersión o subdispersión, pero desafortunadamente en la mayoría de los casos no es posible.

En este modelo, la variable respuesta (variable objetivo) $Y \approx P(\mu)$ con función de probabilidad:

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!} \quad y = 0, 1, 2, ...$$

O en la forma de la familia exponencial:

$$f(y; \mu) = exp\left\{yln(\mu) - \mu - ln\Gamma(y + 1)\right\} \quad y = 0, 1, 2, ...$$

donde el parámetro $\mu > 0$

Por lo tanto el parámetro natural o canónico θ y la función $b(\theta)$ vienen dados por:

$$\theta = ln(\mu)$$
$$b(\theta) = \mu$$

El **enlace canónico** es la función logaritmo. Por lo tanto la inversa de la función enlace es $exp(\eta)$ donde η es el predictor lineal. En consecuencia se trata del **modelo exponencial**:

$$E(Y_i/x_i) = exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_n x_{in})$$

Utilizando esta función enlace las variables explicativas tienen un efecto multiplicativo en vez de aditivo sobre la media. Se pueden usar otras funciones enlaces alternativas cuando falla el enlace canónico como la identidad ($g(\mu) = \mu$) o la raíz cuadrada ($g(\mu) = \sqrt{\mu}$), sin embargo estas funciones enlaces podrían ser problematicas, pues las predicciones podrian ser negativas.

Dado que el modelo de Poisson pertenece a la familia de modelos lineales generalizados, se calculan las funciones media y varianza como en el capítulo uno, sustituyendo y derivando respecto θ:

$$E(Y) = \mu = \frac{\partial b(\theta)}{\partial \theta} = \frac{\partial b}{\partial \mu}\frac{\partial \mu}{\partial \theta} = (1)(\mu) = \mu$$

$$Var(Y) = a(\phi)\frac{\partial^2 b(\theta)}{\partial \theta^2} = (1)\left(\frac{\partial^2 b}{\partial \mu^2}\left(\frac{\partial \mu}{\partial \theta}\right)^2 + \frac{\partial b}{\partial \mu}\frac{\partial^2 \mu}{\partial \theta^2}\right) = (1)((0)(1)^2 + (\mu)(1)) = \mu$$

Por tanto la media y la varianza vienen dadas en este modelo por:

$$E(Y) = Var(Y) = \mu$$

Esta es la igualdad de media-varianza de la que ya se ha comentado que caracteriza a la distribución de Poisson (propiedad de equidispersión).

Esta distribución se debe a **Siméon-Denis Poisson**, que la dio a conocer en 1838 en su trabajo Recherches sur la probabilité des jugements en matières criminelles et matière civile (Investigación sobre la probabilidad de los juicios en materias criminales y civiles).

Figura 3.1: Siméon Denis Poisson (1781-1840)

Además, la distribución de Poisson es el límite de la distribución Binomial cuando $n \longrightarrow \infty$ y $p \longrightarrow 0$ mientras que np se mantiene constante.

La suma de variables aleatorias de Poisson independientes es otra variable de Poisson cuyo parámetro es la suma de los parámetros de las originales.

A medida que μ crece, la distribución de Poisson se aproxima a una distribución normal por el **Teorema Central del Límite**, es decir, sea $Y \approx P(\mu)$, si $\mu \longrightarrow \infty$ entonces:

$$\frac{Y-\mu}{\sqrt{\mu}} \approx N(0,1)$$

La distribución de Poisson condicionadas a las n variables explicativas $X = (X_1, ..., X_n)'$ viene dada por:

$$P(Y_i = y_i/x_i) = \frac{e^{-\mu_i(x_i)}\mu_i(x_i)^{y_i}}{y_i!} \quad y_i = 0, 1, 2, ...$$

donde

$$E(Y_i/x_i) = \mu_i(x_i) = \mu(x_{i1}, ..., x_{in}) = exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_n x_{in})$$

Esta formulación es la que se conoce como **Modelo de Regresión de Poisson**.

El modelo de Regresión de Poisson tiene las siguientes características:

1. La distribución es discreta con un único parámetro, la media, usualmente denotada por μ. La media también se entiende como un parámetro de velocidad, es decir, el número esperado de veces que un suceso o evento ocurre por unidad de tiempo, área o volúmen.

2. Los valores de la variable objetivo Y son enteros no negativos.

3. Las observaciones son independientes entre sí.

4. No hay conjuntos grandes de datos que estén muy por encima o muy por debajo de la media de la distribución muestral.

5. La media y la varianza son idénticas, es decir, distribución de Poisson con una media alta tiene una gran variabilidad.

Los datos de conteo de un modelo de Poisson deben ser generados por un proceso que satisfaga las siguentes cuatro propiedades:

- La probabilidad de que se dé un único suceso o evento sobre un pequeño intervalo es aproximadamente proporcional al tamaño del intervalo.

- La probabilidad de que dos sucesos ocurran en un mismo intervalo que tiende a cero es despreciable.
- La probabilidad de que se dé un suceso dentro de un intervalo de una cierta longitud no cambia en otro intervalo diferente de la misma longitud.
- La probabilidad de un suceso en un intervalo es independiente de la probabilidad del suceso en otro intervalo no superpuesto al anterior.

Cuando las dos últimas propiedades no se cumplen en los datos, el modelo puede tener una varianza que exceda la media.

Luego, a pesar de ser un modelo de referencia en estudios de variables de conteo y resultar muy adecuado para modelar valores enteros no negativos, en concreto cuando la frecuencia de ocurrencia es baja, presenta muchos problemas al tratar con datos en que la media y variancia no son iguales.

3.1. Problema de la infradispersión y sobredispersión

Como ya se ha visto, el Modelo de Regresión de Poisson a pesar de ser un buen modelo para datos de conteo puede resultar inapropiado al no cumplir ciertos supuestos, el más común es la ausencia de equidispersión. Cuando se trabaja con datos reales frecuentemente presentan **infradispersión** ($Var(Y) < E(Y)$) o **sobredispersión** ($Var(Y) > E(Y)$), esta última aparece con mas asiduidad, por lo que se estudia con más detalle. De hecho por abuso del lenguaje, las pruebas para evaluar la equidispersión son denominadas habitualmente pruebas de sobredispersión.

Se pueden destacar las siguientes causas de sobredispersión:

- Alta variabilidad en los datos.
- Los datos no provienen de una distribución de Poisson.
- Los sucesos no ocurren independientemente.
- Falta de estabilidad, la probabilidad de ocurrencia de un suceso puede ser independiente de la ocurrencia de un suceso previo pero no es constante.
- Errores de la especificación de la media, como omitir variables explicativas, o que entran al modelo a través de alguna transformación en lugar de linealmente.
- Errores al elegir la función enlace, es decir tal vez no es apropiado el enlace loglineal.

Existen varias formas de detectar la sobredispersión, por ejemplo [Lindsey, 1995] propone aplicar el coeficiente de variación CV:

$$CV = \frac{Var(\mu_i)}{\mu_i}$$

Para que se cumpla la propiedad de equidispersión este coeficiente debe valer 1, este sencillo índice constituye una muy simple aproximación para la detección de sobredispersión, aunque existen más criterios de detección.

Generalmente la sobredispersión se evalúa mediante la relación entre el estadístico de Pearson χ^2 o la función desviación D y sus respectivos grados de libertad.

Si estos valores son mayores que 1 indican sobredispersión.

Otra forma es una prueba de Razón de Verosimilitud basada en las distribuciones de Poisson y Binomial Negativa (BN2). Para la distribución de Poisson $V(Y) = \mu$ y para la distrubución Binomial Negativa (BN2) $V(Y) = \mu(1 + \alpha\mu) = \mu + \alpha\mu^2$.

Por tanto si $\alpha = 0$ la distribución Binomial Negativa se reducirá a una Poisson. Por tanto las hipótesis que se plantean son las siguientes:

$$H_0 : \alpha = 0$$
$$H_1 : \alpha > 0$$

Para llevar a cabo esta prueba, se deberán ajustar los datos a los dos modelos: Poisson y Binomial Negativa (BN2). Para cada modelo se obtendrá su respectiva función de log-verosimilitud(l). El estadístico propuesto para esta prueba es:

$$RV = -(2(l(Poisson) - l(BN)))$$

Según [Cameron y Trivedi, 1998] este estadístico tiene una distribución asintótica χ^2_1. Por tanto rechazamos la hipóteis nula H_0 si el estadístico es mayor que $\chi^2_{1,1-\alpha}$, donde α es el nivel de significación.

Otra alternativa sería usar métodos de estimación de Quasi Verosimilitud, los cuales nos permiten estimar el parámetro de dispersión e incluirlo en el modelo, estos métodos comprenden una teoría bastante más compleja.

Capítulo 4

Modelo de Regresión Binomial Negativa

Como se ha comentado en el capítulo anterior, a pesar de ser el modelo de regresión de Poisson un modelo de referencia para datos de conteo, este presenta ciertos problemas a la hora de tratar con datos en los que la media y varianza no coinciden, por ejemplo cuando se da la sobredispersión. El modelo paramétrico estandar para datos de conteo con sobredispersión es el Modelo de Regresión Binomial Negativa.

El modelo de regresión binomial negativa es un modelo estadístico atípico, aunque los investigadores se refieren a éste como un único modelo, en realidad existen distintos modelos binomiales negativas, que dependerán del tipo de problema de fondo que se está abordando.[Boxwell y Patil, 1970] identificaron 13 tipos distintos que derivan de la distribución binomial negativa mientras que otros autores exponen que existen más.

La **distribución binomial negativa** estudia la probabilidad de que en una variable aleatoria Y se observen y fracasos antes del r-ésimo éxito en una serie de experimentos Bernoulli independientes. Bajo esta descripción, r debe ser por lo tanto un entero positivo. La *distribución geométrica* es el caso de la binomial negativa cuando $r = 1$ (la distribución geométrica que cuenta el número de fallos antes del primer éxito $P(Y = y) = (1-p)^y p$, para $y = 0, 1, 2...$).

La variable aleatoria de conteo Y sigue una distribución Binomial Negativa de parámetros $r > 0$ y $0 < p < 1$, $Y \approx BN(r, p)$, si tiene como función de probabilidad:

$$P(Y = y) = \left(\begin{array}{c} y + r - 1 \\ r - 1 \end{array} \right) p^r (1-p)^y = \frac{\Gamma(y+r)}{y!\Gamma(r)} p^r (1-p)^y$$

donde y=0,1,2,...

Ejemplo: Supuesto que la probabilidad en un cierto equipo de baloncesto de que un jugador enceste es 0.4. Se realiza el siguiente experimento:

Los jugadores lanzan individualmente a canasta uno tras otro ¿Cuál es la probabilidad de que el décimo jugador que lance a canasta sea el tercero en encestar?

En este caso, Y es el número de lanzamientos encestados, $r = 10$ número de ensayos (de experiemntos de Bernoulli independientes), $p = 0{,}4$ (se considera éxito encestar):

$$P(Y = 7) = \begin{pmatrix} 7+10-1 \\ 10-1 \end{pmatrix} 0{,}4^{10}(1-0{,}4)^7 = \begin{pmatrix} 16 \\ 9 \end{pmatrix} 0{,}4^{10}(0{,}6)^7 = 0{,}0335$$

En la distribución Binomial Negativa el valor esperado y la varianza vienen dados por:

$$E(Y) = \frac{r(1-p)}{p} \qquad Var(Y) = \frac{r(1-p)}{p^2}$$

Se establece por tanto la relación entre ambos : $Var(Y) = \frac{1}{p}E(Y)$, como $0 < p < 1$, se verifica que $\mathbf{Var}(\mathbf{Y}) > \mathbf{E}(\mathbf{Y})$, lo que justifica la predisposición natural de esta distribución para modelar datos que se caracterizan por la existencia de sobredispersión.

El modelo de regresión binomial negativa fue estudiado por primera vez en 1949 por Anscombe. Muchos autores lo han citado, señalando su utilidad para datos de conteo con sobredispersión. [Lawless, 1987] detalló la parametrización del modelo mixto, obteniendo fórmulas para la log-verosimilitud, media, varianza y momentos. Más tarde [Breslow,1990] citó el trabajo de Lawless mientras manipulaba el modelo de Poisson para ajustarse a parámetros binomiales negativos. Desde sus inicios hasta finales de la década de los 80 el modelo de regresión binomial negativa era construido como un modelo mixto usado para datos de Poisson con sobredispersión.

[McCullagh y Nelder, 1989] mencionan que el modelo de regresión binomial negativa puede considerarse un MLG, los autores mencionan la existencia de un enlace canónico, rompiendo con el concepto de modelo mixto, pero no llegan a desarrollarlo. No es hasta mitad de los 90 cuando la binomial negativa es construida formalmente como un miembro de la familia de modelos lineales generalizados[Hilbe,1993].

4.1. Derivación del modelo

Se pueden resumir las diferentes formas de derivar el modelo de regresión binomial negativo en dos orígenes:

1. A partir de una **distribución de Poisson compuesta con una Gamma**, en la cual la distribución Gamma es usada para ajustar los datos de Poisson que presentan sobredispersión. De esta forma se deriva el modelo tradicional binomial negativo, que se denota BN2.

2. Como miembro de la familia exponencial de distribuciones y, por tanto ser **considerado un MLG**. Tal interpretación permite a los investigadores aplicar al modelo los test de bondad de ajuste, análisis de residuos y cualquier otro estudio desarrollado para los Modelos Lineales Generalizados.

4.1.1. Derivación del modelo a partir del modelo de Poisson compuesto con Gamma

Partiendo de un modelo de regresión de Poisson, se deriva el modelo de regresión binomial negativa siguiendo dos enfoques distintos:

- El enfoque más común, en el que la variable respuesta sigue una distribución de Poisson, cuya media está especificada de forma incompleta debido a una situación de heterogeneidad no observada, para solucionar dicha situación se introduce un nuevo término de error. El término de error puede ser el resultado del efecto conjunto de las variables no incluidas en el modelo o bien una fuente de aleatoriedad intrínsica. Sea cual sea su origen, representa la heterogeneidad no observada de los datos.
- Otro enfoque supone que la variable respuesta sigue una distribución Poisson en la que su media no se considera un parámetro fijo, sino que se interpreta como un parámetro que varía aleatoriamente como una distribución Gamma.

En los dos casos, ambos enfoques conducen a una distribución binomial negativa.

A continuación se recoge la función de densidad de la distribución Gamma:

$$Y \approx G(\tau, \omega) \quad \text{si} \quad f(y; \tau, \omega) = \frac{1}{\omega^{\tau}\Gamma(\tau)} y^{\tau-1} e^{y/\omega}$$

com $x, \omega, \tau > 0$ y siendo la función Gamma:

$$\Gamma(v) = \int_0^{\infty} x^{v-1} e^{-x} dx$$

4.1.2. Derivación del modelo como modelo lineal generalizado

La distribución binomial negativa pertenece a la familia exponencial siempre que el parámetro de dispersión sea introducido en la distribución como una constante, por tanto se puede usar esta distribución como componente aleatoria del modelo lineal generalizado. También es importante la selección de la función enlace, según la que se utilice se obtienen diferentes modelos. Normalmente, se suele usar el enlace canónico y el enlace logarítmico. La ventaja de este segundo enlace es que permite una comparación con el modelo de regresión de Poisson [Hardin y Hilbe, 2012].

Componente aleatoria: La función de probabilidad se puede expresar como miembro de la familia exponencial paramétrica con la siguiente estructura:

$$P(Y=y) = f(y;r,p) = exp\left\{yln(1-p) + rln(p) + ln\begin{pmatrix} y+r-1 \\ r-1 \end{pmatrix}\right\}$$

donde:

$$\theta = ln(1-p) \Longrightarrow p = 1 - e^{\theta}$$
$$b(\theta) = -rln(p) = -rln(1-e^{\theta})$$
$$a(\phi) = 1$$

Por lo tanto para hallar la media y la varianza se calcula la primera y segunda derivada de $b(\theta)$ respecto de θ:

$$b'(\theta) = \frac{\partial b}{\partial p}\frac{\partial p}{\partial \theta} = -\frac{r}{p}\{-(1-p)\} = \frac{r(1-p)}{p} = \mu$$
$$b''(\theta) = \frac{\partial^2 b}{\partial p^2}\left(\frac{\partial p}{\partial \theta}\right)^2 + \frac{\partial b}{\partial p}\frac{\partial^2 p}{\partial \theta^2} = \frac{r}{p^2}(1-p)^2 + \frac{r}{p}(1-p) = \frac{r(1-p)}{p^2} = \sigma^2$$

Reparametrizando $\alpha = 1/r$, así se obtiene p y r en función de μ y α:

$$\frac{r(1-p)}{p} = \frac{(1-p)}{\alpha p} = \mu \Longrightarrow$$
$$\frac{(1-p)}{p} = \alpha\mu \Longrightarrow p = \frac{1}{1+\alpha\mu}$$

Por lo tanto se pueden obtener los siguientes términos en función de μ y α:

$$\theta = ln(1-p) = ln\left(\frac{\alpha\mu}{1+\alpha\mu}\right)$$
$$b(\theta) = -\frac{1}{\alpha}ln(p) = \frac{1}{\alpha}ln(1+\alpha\mu)$$
$$b'(\theta) = \frac{1-p}{\alpha p} = \mu = \frac{1}{\alpha(e^{-\theta}-1)}$$
$$b''(\theta) = \frac{1-p}{\alpha p^2} = \mu + \alpha\mu^2$$
$$V(\mu) = b''(\theta) = \mu + \frac{\mu^2}{r}$$

Por lo tanto, dados los valores definidos de μ y α se puede volver a expresar la función de probabilidad de la distribución binomial negativa como:

$$P(Y = y) = \begin{pmatrix} y + 1/\alpha - 1 \\ 1/\alpha - 1 \end{pmatrix} \left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y}$$

Función enlace: La función de enlace canónica de esta distribución, donde se parametriza la relación entre la media μ y las variables predictoras, viene dada por:

$$g(\mu) = \eta = \theta = ln\left(\frac{\alpha\mu}{1+\alpha\mu}\right) = -ln\left(\frac{1}{\alpha\mu} + 1\right)$$

Y su inversa:

$$h(\eta) = g^{-1}(\eta) = \mu = \frac{1}{\alpha(e^{-\theta} - 1)}$$

Si bien desde el punto de vista teórico, el enlace canónico representa una simplificación del estudio del modelo, desde el punto de vista aplicado la mayoría de los investigadores proponen el enlace logarítmico $g(\mu) = ln(\mu) = \eta$.

Bajo este enlace el modelo que se obtiene es el modelo BN2 o modelo tradicional de regresión binomial negativa.

Independientemente de la manera en que se obtenga el modelo de regresión binomial negativa, éste es casi siempre usado para modelar datos con sobredispersión. Las ventajas del enfoque como modelo lineal generalizado está en poder usar las técnicas estadísticas específicas para el MLG que vienen en la mayoría de los softwares.

El inconveniente con la versión MLG es que se debe estimar el parámetro de dispersión α e introducirse en el modelo como una constante.

Capítulo 5

Otros Modelos de Regresión para datos de conteo

En este capítulo, se estudian otros modelos de regresión para datos de conteo menos conocidos, por ser mas complejos matemáticamente y más dificiles de tratar computacionalmente. Aún así, en determinados estudios, estos modelos pueden adaptarse mejor a los datos que los modelos anteriormente estudiados.

5.1. PIG, PG y BN-P

- Modelo **POISSON INVERSA GAUSSIANA (PIG)**:

  Al igual que el modelo de regresión binomial negativa se trata de un modelo mixto. El modelo PIG es una mezcla de distribución de Poisson y **distribución inversa Gaussiana** (también conocida como distribución de Wald), cuya función de densidad de probabilidad viene dada por:

$$f(y;\mu,\alpha) = \left(\frac{\alpha}{2\pi y^3}\right)^{1/2} exp\left\{\frac{-\alpha(y-\mu)^2}{2\mu^2 y}\right\}$$

  para $y > 0$, donde $\mu > 0$ y $\alpha > 0$ son dos parámetros (μ es la media).

  La distribución inversa Gaussiana tiene varias propiedades análogas a la distribución Gaussiana. Esta distribución parece haber sido derivada por primera vez por Schrödinger en 1915. El nombre inverso Gaussiano fue propuesto por Tweedie en 1945. Wald redefinió esta distribución en 1947 como la forma límite de una muestra en una prueba de razón de probabilidad secuencial. El nombre puede ser engañoso, pues no se trata de la inversa de la distribución Gaussiana.

La distribución inversa gaussiana pertenece a la familia exponencial biparamétrica. El modelo PIG también es reconocido como una versión biparamétrica de la distribución de Sichel.

La regresión PIG se utiliza para modelar datos de recuento que tienen un pico inicial alto y que pueden estar sesgados extremadamente a la derecha.

Con un parámetro de dispersión mayor que 1, mayores valores de la media de la variable objetivo en un modelo de regresión PIG proporciona un ajuste de mayores cantidades de sobredispersión que el modelo binomial negativo. Luego el modelo PIG puede tratar mejor con datos altamente dispersos que la regresión binomial negativa, particularmente cuando los datos se agrupan fuertemente entre el valor 1 y 2.

Este modelo se puede aplicar al siguiente ejemplo: Se quiere estudiar el número de días ingresados en un hospital de pacientes que llegan a urgencias, la mayoría de los pacientes son dados de alta en los primeros días y luego disminuye rápidamente el número de pacientes ingresados, con pocos pacientes persistentes durante mucho tiempo.[Hilbe, 2014]

- Modelo **POISSON GENERALIZADO (PG)**:

Raramente se encuentra **infradispersión** (o **subdispersión**) cuando se trata con datos reales. Pero a veces puede ocurrir y que no se le preste atención a la infradispersión y modelar los datos por medios normales, entonces los errores estándar del modelo resultante se sobreestiman. Esto puede llevar a pensar que los predictores no son significativos cuando de hecho si lo son.

En este modelo se asume que la variable respuesta sigue una distribución Poisson generalizada con función de probabilidad [Harrris, Yang and Hardin, 2012]:

$$f(y;\theta_i,\delta) = \frac{\theta_i(\theta_i+\delta y_i)^{y_i-1}e^{-\theta_i-\delta y_i}}{y_i!} \qquad y_i = 0,1,2,...$$

donde $\theta_i > 0 \quad 0 \leq \delta < 1$. [Joe and Zhu, 2005]

Se ve que:

$$E(Y_i) = \mu_i = \frac{\theta_i}{1-\delta}$$
$$V(Y_i) = \frac{\theta_i}{(1-\delta)^3} = \frac{1}{(1-\delta)^2}E(Y_i) = \phi E(Y_i)$$

El término $\phi = \dfrac{1}{(1-\delta)^2}$ caracteriza la dispersión en la distribución PG. Claramente, cuando $\delta = 0$ la distribución Poisson Generalizada se reduce a la usual distribución de Poisson con parametro θ_i. Además cuando $\delta < 0$ el modelo presenta subdispersión y cuando $\delta > 0$ presenta sobredispersión.

- Algunos autores recogen otros modelos para datos de conteo, combiando modelos ya existentes o creando variaciones de los anteriormente citados. Luego, con el objetivo de encontrar un modelo que se ajuste lo mejor posible a los datos de un estudio concreto se desarrollan estos modelos más específicos y elaborados. Entre estos otros modelos se encuentra el modelo **BINOMIAL NEGATIVA DE TRES PARÁMETROS DE GREEN (BN-P)**.

Capítulo 6

Problemas con el valor 0

En este capítulo se analizan dos casos especiales que pueden aparecer cuando se trabaja con datos de conteo. Ambos casos son completamente diferentes pero tienen el mismo elemento en común: el valor **0**.

Primero se tratan los modelos que no permiten que la variable objetivo tome el valor 0, **Modelos truncados por ceros**. Y, posteriormente, los **Modelos con exceso de ceros**, en los que por algún motivo hay más observaciones con valor 0 de las que cabría esperar de acuerdo a una distribución Poisson o Binomial Negativa.

6.1. Modelos truncados por ceros

Se comienza mostrando algunos ejemplos de estos tipos de datos, se tratan de datos que por su propia naturaleza no pueden tomar nunca el valor 0:

- Número de delitos por los que ha sido condenado un preso.
- Número de veces que un conductor tuvo que hacer el examen práctico.
- Número de individuos que forman un grupo.
- Número de idiomas que habla una persona.

Por trabajar con un modelo de regresión para datos de conteo concreto, se elige por ejemplo el modelo de regresión de Poisson. Por lo tanto se asume que la variable respuesta se distribuye según una función de Poisson con un parámetro μ (que es tanto la media como la varianza). El problema que se plantea es que la distribución de Poisson no excluye los ceros, es decir, predice valores de 0 para la variable objetivo, **especialmente cuando los valores de μ son bajos**.

A continuación, se muestra esto en el siguiente ejemplo realizado con el software R a partir de 4 muestras aleatorias de Poisson de tamaño 1000 con los valores de μ indicados:

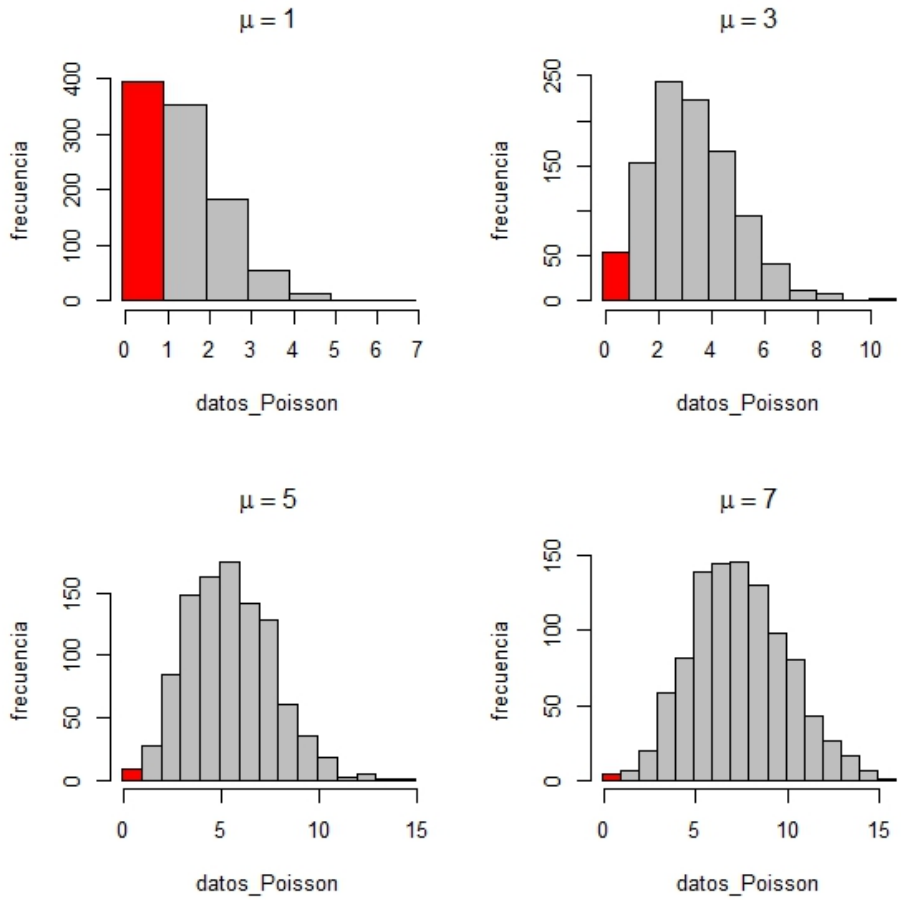


Figura 6.1: Ejemplos de la frecuencia con la que la distribución de Poisson toma el valor cero, incrementandose para valores de μ bajos

Para solucionar este problema los modelos truncados por ceros reparten la probabilidad de que la variable respuesta sea 0 entre todos los demás valores posibles. A continuación, se muestra como se hace esto a partir de una distribución de Poisson, se desarrolla así la construcción del modelo de Poisson truncado por ceros:

Se recuerda que la función de distribución de probabilidad de Poisson es:

$$f(y_i; \mu | y_i \geq 0) = \frac{e^{-\mu}\mu^{y_i}}{y_i!}$$

donde $i = 1, ..., n$

La probabilidad de que la variable respuesta tome el valor 0:

$$f(0; \mu) = \frac{e^{-\mu}\mu^0}{0!} = e^{-\mu}$$

En los modelos sin ceros, se excluye la posibilidad de que la variable objetivo tome el valor 0 al dividir las probabilidades del resto de valores posibles por $(1 - f(0))$. Por lo tanto ya se pueden recalcular las probabilidades del resto de valores:

$$f(y_i; \mu | y_i > 0) = \frac{e^{-\mu}\mu^{y_i}}{y_i!(1 - e^{-\mu})}$$

Consecuentemente al haber excluido el cero, ahora la probabilidad del resto de valores es más alta.

Así, los modelos Poisson truncados por ceros son iguales a los modelos generalizados lineales con errores de Poisson, con la diferencia de que utilizan distribuciones de probabilidad en las que se eliminan los ceros y se reajustan las probabilidades del resto de valores.

Cuando μ es grande, la porción a repartir (probabilidad de tomar el valor 0) es muy pequeña y no vale la pena sustituir un modelo de Poisson por uno truncado por ceros.

En caso de que, en lugar de una distribución de Poisson, se estuviese usando una Binomial Negativa, el procedimiento seria análogo y se obtendría el modelo de Regresión Binomial Negativa truncado por ceros. Y así para cualquier otro modelo sobre datos de conteo, como los estudiados en el capítulo anterior (PIG,PG,BN-P).

6.2. Modelos con excesos de ceros

Es usual encontrar datos de conteo en los que hay muchas observaciones que toman el valor cero. Esta cantidad de ceros no es consistente con los modelos que se están usando (Poisson, BinomialNegativa, etc).

De entre los datos cuyo valor es cero podemos distinguir dos tipos de cero: los falsos ceros y los ceros auténticos. La presencia de estos **falsos ceros** puede llevar a que haya

una sobreabundancia de ceros en la base de datos. Para aclarar este concepto se muestra un ejemplo relacionado con las competiciones deportivas:

Ejemplo: En el estudio del número de goles de un delantero de fútbol en la útima temporada, aquellos que han marcado 0 goles puede ser por dos motivos:

- No han jugado en toda la temporada (lesionados la temporada completa, nunca han sido convocados...) *Falsos ceros*
- Sí han jugado, pero han fallado todos los tiros realizados o nunca han tirado a puerta. *Ceros auntén*ticos

Hay básicamente dos estrategias para lidiar con el problema del exceso de ceros:

1. Asumir que los ceros proceden de dos procesos distintos: el proceso binomial y el proceso de Poisson. Igual que en los Modelos en dos partes, se hace un modelo lineal generalizado binomial para modelizar la probabilidad de medir un 0 (los falsos ceros). Posteriormente se modeliza la probabilidad de obtener el resto de valores, incluyendo ceros (los ceros auténticos). Estos modelos se incluyen en los llamados Modelos lineales generalizados mezclados o mixtos (mixture models) y se denominan **Modelos inflados por ceros**.
2. Asumir que todos los ceros son iguales (sin distinguir entre falsos y auténticos). Este tipo de modelos constan de dos partes: en una primera parte se consideran todos los datos como ceros o no-ceros y se modela la probabilidad de que una observación sea cero (en función de las variables explicativas seleccionadas) usando un modelo binomial. Posteriormente, las observaciones que no son cero se modelizan usando modelos truncados por ceros, como los que se acaban de ver. Estos modelos se llaman **Modelos en dos partes (Hurdle models)**.

Ambos tipos de modelos permiten abordar el exceso de ceros. La elección de uno u otro modelo debería basarse en conocimiento a priori de las fuentes de ceros en el problema. En principio, si se considera que por algún motivo hay una gran cantidad de falsos ceros en los datos, se deben usar los Modelos inflados por cero, que van a permitir descubrir los motivos que llevan a medir falsos ceros, y tal vez tomar medidas para que se eviten en futuros trabajos. Si por el contrario se piensa que la mayoría de los ceros son auténticos, aunque haya muchos, se debe optar por los Modelos en dos partes.

Como ya se ha expuesto, la principal diferencia entre los Modelos en dos partes y los Modelos inflados por ceros es que en estos últimos se está interesado en distinguir los distintos orígenes de los ceros observados, es decir, considerar que hay ceros auténticos y ceros falsos.

6.2.1. Modelos inflados con ceros

Estos modelos suponen que los ceros se generan de dos formas, por una parte se tienen los ceros *falsos* y por otra los ceros que proceden de la distribución que se ha supuesto (Poisson, Binomial Negativa,...) y que también es la que genera los valores mayores que cero. Por lo tanto:

$$P(Y_i = y_i|y_i = 0) = g + (1 - g)f(0)$$
$$P(Y_i = y_i|y_i > 0) = (1 - g)f(y_i)$$

donde g representa la probabilidad de los ceros *falsos*, viene definida por un proceso de decisión binario. Y $f(0)$ es la probabilidad de observar cero en aquellos individuos que no pertenecen a los ceros falsos, f se trata de una distribución de recuento (Poisson, Binomial Negativa,...)

6.2.2. Modelos en dos partes

La idea básica de este método es que hay una decisión binaria que determina si el resultado es cero o no, y una segunda parte de decisión que determina la probabilidad de los valores mayores que cero cuando se pasa la primera decisión.

El proceso se divide en dos: Modelo de decisión binaria (generado por una distribución g) y modelo truncado en cero (generado por una distribución f)

$$P(Y_i = y_i|y_i = 0) = g(0)$$
$$P(Y_i = y_i|y_i > 0) = (1 - g(0))\frac{f(y_i)}{1 - f(0)}$$

donde $1 - g(0)$ es la probabilidad de pasar la primera decisión, es decir, de no ser el valor 0, y $(f(y_i)/(1 - f(0)))$ la probabilidad de tomar un valor y_i, sabiendo que $y_i > 0$.

Capítulo 7

Aplicación a competiciones deportivas

En este capítulo se lleva a cabo la aplicación práctica de las técnicas estadísticas en los modelos para datos de conteo y se muestra su aplicabilidad en competiciones deportivas.

7.1. Estadísticas en el Deporte

Actualmente son muchas las fuentes que suministran información estadística para los deportes más populares. Para ilustrar las posibilidades de estas fuentes se muestra como ejemplo el sistema AMISCO en el fútbol.

Amisco es un programa informático de la empresa francesa Sport Universal Process. Se trata de un sistema de captación, procesamiento y análisis de los datos que se obtienen a través de las cámaras que se instalan en los estadios de fútbol. Su principal función es la de medir, almacenar y descifrar los datos estadísticos que se desprenden en un partido de fútbol, desde parámetros tanto tácticos, técnicos como físicos.

El programa AmiscoPro tiene 8 cámaras situadas alrededor de todo el estadio que graban todo lo que pasa en el terreno de juego, desde las acciones técnico-tácticas de cada jugador (pase, chut, remate, desmarque, cobertura, etc.), a trayectorias del balón o recorridos realizados por los árbitros. Estas imágenes se envían a una central, donde se utilizan para reproducir virtualmente todas las acciones que se dan durante el partido.

Este programa lo utilizan en la actualidad varios equipos de la Liga BBVA, entre los que podemos incluir al Real Madrid, Valencia o Villarreal, y otros grandes clubes de Europa como Liverpool, Manchester, Ajax o Borussia Dortmund. Solo con estos datos, es evidente que el uso de este sistema aporta grandes beneficios a aquellos equipos que lo utilizan, entre los que se pueden destacar:

- Para los entrenadores: Comprobar si se realizan correctamente las acciones técnico-tácticas entrenadas. Medir distancias entre jugadores y respecto a sus marcas. Ocupación de zonas.

- Para los preparadores físicos: Velocidad y distancia recorrida por los jugadores en un partido, que sirve para demostrar que el estado de la condición física permite por ejemplo realizar un sprint en el minuto 88.

- Para los jugadores: Reciben un mensaje que les permite comprobar la participación y la eficacia de sus acciones con y sin balón.

- Para el propio deporte: Se ha aplicado también en el ámbito de la investigación, con el desarrollo de diferentes estudios para analizar la actividad de los jugadores en función a su posición, del tipo de competición, etc.

En conclusión, se puede afirmar que este sistema actual de análisis ofrece a los clubes que lo utilizan, una amplia gama de información única e innovadora sobre los partidos, así como la posibilidad de una preparación táctica ventajosa de sus jugadores sobre futuros rivales.

Esto demuestra que la tecnología junto con la estadística puede aportar enormes beneficios a la cada vez más demandante competitividad del futbol, ya que son datos reales y fiables. Sin embargo, hay que tener claro que el triunfo lo obtendrá aquel equipo que mejor uso le haya dado a esa información, y no el que más tenga.

7.2. Aplicación a una base de datos deportiva

En esta sección, a partir de datos obtenidos mediante la página web de la liga de fútbol profesional española (http://www.laliga.es) se estudia el número de goles anotados por los jugadores pertenecientes a los cinco primeros equipos de la temporada 2015-2016.

Se han programado distintos scripts en R para poder realizar tal estudio, interpretando posteriormente los resultados obtenidos.

7.2.1. Descripción de los datos

La base de datos está formada por 125 jugadores de los siguientes cinco equipos (Ath.Bilbao, Atl. de Madrid, F.C. Barcelona, Real Madrid y Villareal FC). El objetivo será modelizar el número de goles por jugador (variable objetivo) en función de una

serie de variables explicativas relacionadas con las caracterísiticas del jugador (peso, altura, edad, equipo,...), persiguiendo un objetivo fundamentalmente explicativo.

Instrucciones iniciales necesarias para el estudio a realizar y **resumen descriptivo** de la base de datos "GolesJugadoresLiga_20152016.sav" que se encuentra en un archivo SPSS.

```
> library(foreign) #Para leer ficheros de SPSS
> X<-data.frame(read.spss("GolesJugadoresLiga_20152016.sav"))
> jugador<-X$Jugador
> X<-X[,2:16]      #Seleción de las variables
> rownames(X)<-jugador

> str(X)
```

```
'data.frame':   125 obs. of  15 variables:
 $ Altura       : num  182 174 177 175 180 185 184 181 180 191 ...
 $ Peso         : num  78 69 76 72 76 79 78 77 71 87 ...
 $ Edad         : num  36 23 28 29 27 29 29 28 26 35 ...
 $ Nacionalidad : Factor w/ 2 levels "Esp",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Pos          : Factor w/ 4 levels "centrocampista",..: 3 1 2 1 1 ...
 $ Equipo       : Factor w/ 5 levels "Ath.Bilbao",..: 1 1 1 1 1 1 1 ...
 $ PJ           : num  34 3 34 36 34 30 11 23 16 37 ...
 $ PJCompl      : num  25 1 31 22 30 26 3 13 4 36 ...
 $ Partidos_Tit : num  30 2 33 32 33 28 4 15 8 37 ...
 $ Partidos_Sust: num  4 1 1 4 1 2 7 8 8 0 ...
 $ Min_Jugados  : num  2776 186 2905 2894 2886 ...
 $ Tarjetas     : num  9 0 6 8 6 5 1 1 2 1 ...
 $ Expulsiones  : num  0 0 0 0 0 0 0 0 0 1 ...
 $ Penaltis     : num  2 0 0 0 0 0 0 0 0 0 ...
 $ GOLES        : num  20 0 0 1 1 0 1 0 2 0 ...
```

Las variables que requieren una descripción adicional a su nombre son: Pos (Posición), PJ (Partidos jugados), PJCompl (Partidos jugados completos), Partidos_Tit (Partidos jugados como titular) y Partidos_Sust (Partidos jugados como sustituto). A continuación se muestra un resumen de los valores que toman cada variable.

```
> summary(X)
```

```
     Altura          Peso            Edad       Nacionalidad
 Min.   :169.0   Min.   :62.00   Min.   :19.00   Esp     :71
 1st Qu.:177.0   1st Qu.:72.00   1st Qu.:24.00   NoEsp   :54
 Median :182.0   Median :74.00   Median :27.00
 Mean   :181.1   Mean   :75.19   Mean   :27.04
 3rd Qu.:185.0   3rd Qu.:78.00   3rd Qu.:30.00
 Max.   :195.0   Max.   :94.00   Max.   :36.00


    Pos                    Equipo                PJ
 centrocampista :45   Ath.Bilbao          :25   Min.   :1.00
 defensa        :39   Atl. de Madrid      :25   1st Qu.:11.00
 delantero      :31   FC Barcelona        :25   Median :23.00
 portero        :10   Real Madrid         :24   Mean   :20.94
                      Villarreal CF       :26   3rd Qu.:31.00
                                                Max.   :38.00

    PJCompl       Partidos_Tit   Partidos_Sust     Min_Jugados
 Min.   : 0.00   Min.   : 0.00   Min.   : 0.000   Min.   :  10
 1st Qu.: 2.00   1st Qu.: 7.00   1st Qu.: 1.000   1st Qu.: 657
 Median : 9.00   Median :15.00   Median : 3.000   Median :1467
 Mean   :12.37   Mean   :16.72   Mean   : 4.216   Mean   :1501
 3rd Qu.:22.00   3rd Qu.:28.00   3rd Qu.: 7.000   3rd Qu.:2475
 Max.   :38.00   Max.   :38.00   Max.   :21.000   Max.   :3420

    Tarjetas       Expulsiones       Penaltis         GOLES
 Min.   : 0.000   Min.   :0.000   Min.   :0.000   Min.   : 0.000
 1st Qu.: 1.000   1st Qu.:0.000   1st Qu.:0.000   1st Qu.: 0.000
 Median : 3.000   Median :0.000   Median :0.000   Median : 1.000
 Mean   : 3.368   Mean   :0.152   Mean   :0.112   Mean   : 3.024
 3rd Qu.: 5.000   3rd Qu.:0.000   3rd Qu.:0.000   3rd Qu.: 2.000
 Max.   :13.000   Max.   :2.000   Max.   :4.000   Max.   :40.000
```

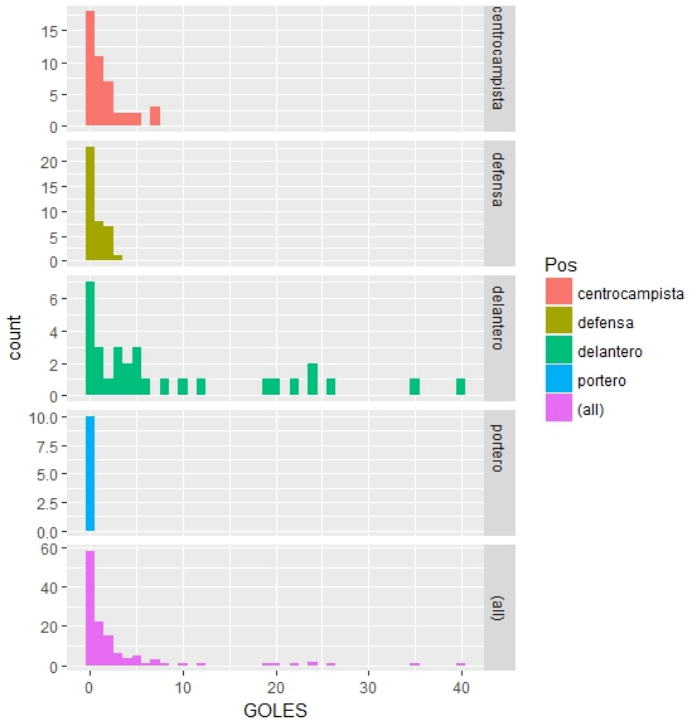
```
> mean(X$GOLES)

[1] 3.024

> var(X$GOLES)
```

```
[1] 44.92684
```

Se observa que la varianza muestral del número de goles (44.93) es bastante mayor a la media (3.024), lo que nos hace empezar a cuestionarnos el problema de la sobredispersión en los datos.

Con un objetivo descriptivo se realizan los siguientes histogramas que representan el número de jugadores que han anotado una determinada cantidad de goles, dependiendo de la posición en la que juegan. Como es de esperar los delanteros tienen más jugadores que anotan muchos goles que el resto de posiciones. También observamos como los todos los porteros se acumulan, obviamente, en el valor 0:

```
> ggplot(X, aes(GOLES, fill = Pos)) +
+   geom_histogram(binwidth=2) +
+   facet_grid(Pos ~ ., margins=TRUE, scales="free")
```



Análogamente se pueden obtener usando las mismas instrucciones, los histogramas que representan el número de jugadores que han anotado una determinada cantidad de goles por equipos o por nacionalidad (Españoles y no Españoles), e interpretar los resultados gráficos.

7.2.2. Aplicación del modelo de regresión de Poisson

A continuación se analiza el conjunto de datos a través de la regresión de Poisson mediante el método paso a paso. Se han seleccionado las variables explicativas: Altura, Nacionalidad, Pos, Equipo, PJ, PJCompl, Min_Jugados y Penaltis.

A continuación se muestra la instrucción del paso a paso y el modelo de regresión de Poisson con su respectiva salida:

```
> modpoisson <- step(glm(GOLES ~ . , family="poisson", data=X),
direction="both")
> summary(modpoisson <-glm(GOLES~Altura + Nacionalidad + Pos + Equipo +
PJ + PJCompl + Min_Jugados + Penaltis,family="poisson", data=X))


Call:
glm(formula = GOLES ~ Altura + Nacionalidad + Pos + Equipo +
    PJ + PJCompl + Min_Jugados + Penaltis, family = "poisson",
    data = X)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.1787  -1.0845  -0.3958   0.3845   2.9828

Coefficients:
                            Estimate   Std. Error  z value  Pr(>|z|)
(Intercept)                -6.802e+00  2.239e+00  -3.038  0.00238 **
Altura                      2.729e-02  1.233e-02   2.213  0.02691 *
NacionalidadNoEsp           3.930e-01  1.568e-01   2.506  0.01220 *
Posdefensa                 -7.200e-01  2.462e-01  -2.925  0.00345 **
Posdelantero                1.446e+00  1.662e-01   8.700  < 2e-16 ***
Posportero                 -1.755e+01  8.538e+02  -0.021  0.98360
EquipoAtl. de Madrid       -5.851e-01  2.254e-01  -2.596  0.00943 **
EquipoFC Barcelona          1.330e-01  2.263e-01   0.588  0.55677
EquipoReal Madrid           1.514e-01  2.127e-01   0.711  0.47678
EquipoVillarreal CF        -3.869e-01  2.426e-01  -1.595  0.11070
PJ                          6.287e-02  2.286e-02   2.750  0.00596 **
PJCompl                    -3.778e-02  1.912e-02  -1.976  0.04818 *
Min_Jugados                 6.333e-04  3.656e-04   1.732  0.08324 .
Penaltis                    1.823e-01  8.050e-02   2.265  0.02354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 997.81  on 124  degrees of freedom
Residual deviance: 176.35  on 111  degrees of freedom
AIC: 404.13

Number of Fisher Scoring iterations: 15
```

Se observa que puede haber un problema de colinealidad entre las variables PJComp y Min_Jugados, luego se estudia la colinealidad de las variables mediante el Factor de inflación de varianza:

```
> library(car)
> vif(modpoisson)
```

```
                 GVIF Df GVIF^(1/(2*Df))
Altura       1.563021  1        1.250208
Nacionalidad 2.082131  1        1.442959
Pos          1.889924  3        1.111921
Equipo       4.257626  4        1.198522
PJ           9.855946  1        3.139418
PJCompl     19.500017  1        4.415882
Min_Jugados 34.684988  1        5.889396
Penaltis     4.523385  1        2.126825
```

Efectivamente se observa colinealidad, en la variable Min_Jugados GVIF=34.68. Luego se realiza el modelo de regresión de Poisson sin esta variable:

```
> summary(m1poisson <-glm(GOLES~Altura + Nacionalidad + Pos + Equipo + PJ +
+  PJCompl + Penaltis,family="poisson", data=X))
```

```
Call:
glm(formula = GOLES ~ Altura + Nacionalidad + Pos + Equipo +
    PJ + PJCompl + Penaltis, family = "poisson", data = X)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0041  -1.0631  -0.3650   0.3803   2.7638

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -7.232e+00  2.250e+00  -3.215  0.00131 **
Altura                       2.926e-02  1.239e-02   2.362  0.01819 *
NacionalidadNoEsp            4.664e-01  1.502e-01   3.106  0.00190 **
```

```
Posdefensa                  -7.572e-01  2.448e-01  -3.094  0.00198 **
Posdelantero                 1.461e+00  1.651e-01   8.848  < 2e-16 ***
Posportero                  -1.860e+01  1.354e+03  -0.014  0.98904
EquipoAtl. de Madrid        -6.335e-01  2.226e-01  -2.846  0.00443 **
EquipoFC Barcelona           2.825e-02  2.188e-01   0.129  0.89725
EquipoReal Madrid            6.384e-02  2.040e-01   0.313  0.75431
EquipoVillarreal CF         -5.203e-01  2.339e-01  -2.225  0.02609 *
PJ                           9.667e-02  1.149e-02   8.416  < 2e-16 ***
PJCompl                     -8.446e-03  9.151e-03  -0.923  0.35605
Penaltis                     1.470e-01  7.790e-02   1.887  0.05911 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 997.81  on 124  degrees of freedom
Residual deviance: 179.37  on 112  degrees of freedom
AIC: 405.15

Number of Fisher Scoring iterations: 16
```

Para la interpretación del modelo se debe tener en cuenta las categorías de referencia para las variables cualitativas: **Nacionalidad**: Española, **Posición**: centrocampista y **Equipo**: Ath.Bilbao.

En la salida, entre los resultados obtenidos se encuentran las estimaciones de los coeficientes de regresión para cada una de las variables en el modelo de Regresión de Poisson. A través de los p-valores observamos que no todas las variables explicativas influyen en la variable objetivo para un nivel de significación de 0.05.

A nivel interpretativo tenemos que un aumento de la variable **Altura**, **PJ** y **Penaltis** produce un aumento del valor esperado del número de goles. Respecto a las variables explicativas dummy definidas, **NacionalidadNoEsp** aumenta el valor esperado de goles. **Posdefensa** disminuye frente a **Poscentrocampista**, y **Posdelantero** aumenta el valor esperado de goles frente a **Poscentrocampista**. Se observa que **EquipoAtl. de Madrid** disminuye el valor esperado de goles respecto **EquipoAth.Bilbao** y **EquipoVillarreal CF** también disminuye el valor esperado de goles respecto **EquipoAth.Bilbao** .

Además en la salida también se muestra la desviación calculada a partir del modelo con solo el coeficiente β_0, el residuo desviación y se muestra el criterio AIC. También se

obtiene el estadístico desviación, que como se vió en el capítulo 3 detecta sobredispersión en los datos:

$$\frac{D}{gl} = \frac{176{,}35}{111} = 1{,}59 > 1$$

Luego se da la **sobredispersión** en los datos. Los modelos de Poisson sobredispersos algunas veces conducen a la confusión de que algunas variables explicativas contribuyen significativamente al modelo, cuando en realidad no es así.

El coeficiente de variación **CV** constituye también otra forma de detección de sobredispersión, para que se cumpla la propiedad de equidispersión este coeficiente debe ser 1, vemos a continuación que en ninguno de los 5 equipos se cumple, es más, en algunos equipos CV toma valores muy grandes:

```
> cv<-function(x) {sd(x)/abs(mean(x))}
> with(X, tapply(GOLES,Equipo, function(x) {
+  sprintf("(Media, Var, CV) = (%1.2f, %1.2f, %1.2f)",
+  mean(x), var(x), cv(x))}))


              Ath.Bilbao
"(Media, Var, CV) = (2.28, 18.46, 1.88)"
            Atl. de Madrid
"(Media, Var, CV) = (2.48, 22.43, 1.91)"
            FC Barcelona
"(Media, Var, CV) = (4.36, 102.07, 2.32)"
             Real Madrid
"(Media, Var, CV) = (4.50, 78.00, 1.96)"
            Villarreal CF
"(Media, Var, CV) = (1.62, 7.05, 1.64)"
```

Por lo tanto, se prueba aplicar un modelo de regresión Binomial Negativa.

7.2.3. Aplicación del modelo de regresión Binomial Negativa

El lenguaje R permite un estudio del Modelo de Regresión Binomial Negativa desde los dos enfoques planteados en el capítulo 4, como una Poisson compuesta con Gamma y como miembro de la familia de MLG. Se analizará el conjunto de datos enfocando el modelo de regresión binomial negativa como un miembro de la familia de modelos lineales generalizados. Para ello se usa **glm.nb** del paquete MASS. Al igual que con el modelo de Poisson, se aplica previamente el método de paso a paso para seleccionar las variables con la que se realiza la regresión Binomial Negativa. Se recogen las siguientes instrucciones con sus salidas correspondientes:

```
> library("MASS")
> modBN <- step(glm.nb(GOLES ~ . , data=X),direction="both")
> summary(modBN <-glm.nb(GOLES~Altura + Nacionalidad + Pos + PJ + PJCompl +
+     Penaltis, data=X))


Call:
glm.nb(formula = GOLES ~ Altura + Nacionalidad + Pos + PJ + PJCompl +
    Penaltis, data = X, init.theta = 3.169403907, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3316  -0.9428  -0.4586   0.2772   2.5200

Coefficients:
                            Estimate   Std. Error  z value Pr(>|z|)
(Intercept)                -1.017e+01  3.536e+00  -2.875  0.00404 **
Altura                      4.539e-02  1.952e-02   2.325  0.02007 *
NacionalidadNoEsp           3.220e-01  1.964e-01   1.639  0.10116
Posdefensa                 -6.806e-01  3.047e-01  -2.234  0.02551 *
Posdelantero                1.043e+00  2.352e-01   4.433 9.28e-06 ***
Posportero                 -3.715e+01  1.980e+07   0.000  1.00000
PJ                          1.076e-01  1.650e-02   6.524 6.86e-11 ***
PJCompl                    -3.001e-02  1.529e-02  -1.962  0.04974 *
Penaltis                    4.949e-01  1.536e-01   3.222  0.00127 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(3.1694) family taken to be 1)

    Null deviance: 470.51  on 124  degrees of freedom
Residual deviance: 118.30  on 116  degrees of freedom
AIC: 391.52

Number of Fisher Scoring iterations: 1


              Theta:  3.17
          Std. Err.:  1.17

 2 x log-likelihood:  -371.524
```

init.theta es el valor inicial para el parámetro dispersión, si se omite se estima mediante el MLG de Poisson. Hay que aclarar que en esta salida 3.16940 no es la estimación de θ sino de su invera, siendo su valor 0.315.

Como se hizo con el modelo de regresión de Poisson se realiza un estudio de la colinealidad de las variables que han sido seleccionadas. Se observa que no hay colinealidad, luego mantenemos todas las variables explivativas seleccionadas por el paso a paso:

```
> library(car)
> vif(modBN)
```

```
                GVIF Df GVIF^(1/(2*Df))
Altura      1.183049  1        1.087681
Nacionalidad 1.131378  1        1.063662
Pos         1.836403  3        1.106610
PJ          2.317967  1        1.522487
PJCompl     3.415688  1        1.848158
Penaltis    2.254691  1        1.501563
```

El modelo de Poisson **con las mismas variables seleccionadas** puede relacionarse con el modelo binomial negativo. Se usa la siguiente prueba de razón de verosimilitud para comparar estos dos modelos.

```
> m2poisson <- glm(GOLES~Altura + Nacionalidad + Pos + PJ + PJCompl +
+     Penaltis , family="poisson", data=X)
```

```
Call:
glm(formula = GOLES ~ Altura + Nacionalidad + Pos + PJ + PJCompl +
    Penaltis, family = "poisson", data = X)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8715  -1.2718  -0.4643   0.5114   3.8785

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -9.377273   1.981888  -4.731 2.23e-06 ***
Altura                          0.042204   0.010901   3.872 0.000108 ***
NacionalidadNoEsp               0.452887   0.121606   3.724 0.000196 ***
Posdefensa                     -0.893733   0.245174  -3.645 0.000267 ***
```

```
Posdelantero                  1.234151   0.154462   7.990 1.35e-15 ***
Posportero                  -17.903312 868.907578  -0.021 0.983561
PJ                            0.083456   0.010590   7.880 3.26e-15 ***
PJCompl                      -0.004321   0.008595  -0.503 0.615168
Penaltis                      0.291521   0.069503   4.194 2.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 997.81  on 124  degrees of freedom
Residual deviance: 199.88  on 116  degrees of freedom
AIC: 417.66

Number of Fisher Scoring iterations: 15
```

```
> X2 <- 2 * (logLik(modBN) - logLik(m2poisson))
> pchisq(X2, df = 1, lower.tail=FALSE)
```

```
'log Lik.' 1.129599e-07 (df=10)
```

El valor del chi-cuadrado asociado es de $1{,}129 \cdot 10^{-7}$, con 10 grados de libertad. Esto indica que el modelo de regresión Binomial Negativo con la estimación del parámetro de dispersión es más apropiado que el modelo de Poisson utilizando las mismas variables explicativas.

A continuación se compara el primer modelo de Poisson que se realizó (modelo con las variables seleccionadas paso a paso y eliminando la variable que daba problemas de colinealidad) con el modelo de regresión BN precedente. Para comparar estos dos modelos no se puede usar el método anterior, pues los modelos tienen distintas variables explicativas, luego se usa el test de Vuong, cuya hipótesis nula es que los modelos son equivalentes. Se obtiene lo siguiente:

```
> install.packages("pscl")
> library(pscl)
> vuong(m1poisson,modBN)
```

```
Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A   p-value
Raw                  -0.6748969 model2 > model1 0.2498706
AIC-corrected        -1.3825215 model2 > model1 0.0834058
BIC-corrected        -2.3832137 model2 > model1 0.0085811
```

La elección entre el test y sus correciones AIC y BIC se basa en la existencia o no de ceros inflados que se estudiará en el siguiente apartado. En general cuando no hay ceros inflados se utiliza el test sin correción y se aceptaría que son indistinguibles.[https://stats.stackexchange.com/questions/182020/ zero-inflated-poisson-regression-vuong-test-raw-aic-or-bic-corrected-results/217869]

Por lo tanto, en caso de que no hubiera exceso de ceros, ambos modelos son equivalentes.

Dado que el modelo de regresión BN se trata de un modelo cuya función enlace (link) es logarítmica, a la hora de estimar los coeficientes se tienen que exponenciar los coeficientes del modelo. Los coeficientes tienen un efecto multiplicativo.

```
> est <- cbind(Estimate = coef(modBN))
> est
```

```
                                Estimate
(Intercept)                -10.16574143
Altura                       0.04538793
NacionalidadNoEsp            0.32196426
Posdefensa                  -0.68059355
Posdelantero                 1.04267610
Posportero                 -37.14561482
PJ                           0.10762305
PJCompl                     -0.03000689
Penaltis                     0.49491672
```

7.2.4. Modelo inflado con ceros y modelo en dos partes BN

Para esta sección en la que se aborda el exceso de ceros en el modelo, se crea una nueva variable en la base de datos que indique si un jugador es o no portero con la siguiente instrucción:

```
> for(i in 1:nrow(X))
+ if (X$Pos[i]=="portero") X$Portero[i]=1 else X$Portero[i]=0
> X <- within(X, {Portero <- factor(Portero, levels=0:1,
+ labels=c("noportero", "portero"))})
> str(X)


'data.frame':   125 obs. of  16 variables:
 $ Altura       : num  182 174 177 175 180 185 184 181 180 191 ...
 $ Peso         : num  78 69 76 72 76 79 78 77 71 87 ...
 $ Edad         : num  36 23 28 29 27 29 29 28 26 35 ...
 $ Nacionalidad : Factor w/ 2 levels "Esp",..: 1 1 1 1 1 1 1 1 1 ...
 $ Pos          : Factor w/ 4 levels "centrocampista",..: 3 1 2 1 1 ...
 $ Equipo       : Factor w/ 5 levels "Ath.Bilbao",..: 1 1 1 1 1 1 ...
 $ PJ           : num  34 3 34 36 34 30 11 23 16 37 ...
 $ PJCompl      : num  25 1 31 22 30 26 3 13 4 36 ...
 $ Partidos_Tit : num  30 2 33 32 33 28 4 15 8 37 ...
 $ Partidos_Sust: num  4 1 1 4 1 2 7 8 8 0 ...
 $ Min_Jugados  : num  2776 186 2905 2894 2886 ...
 $ Tarjetas     : num  9 0 6 8 6 5 1 1 2 1 ...
 $ Expulsiones  : num  0 0 0 0 0 0 0 0 0 1 ...
 $ Penaltis     : num  2 0 0 0 0 0 0 0 0 0 ...
 $ GOLES        : num  20 0 0 1 1 0 1 0 2 0 ...
 $ Portero      : Factor w/ 2 levels "noportero","portero": 1 1 1 1 ...
```

Para tratar el exceso de ceros se utiliza inicialmente un modelo en dos partes (hurdle) en el que se supone que todos los ceros son iguales. Se tienen que introducir dos fórmulas, la primera corresponde al modelo de Binomial Negativa y la segunda al Binomial, hurdle estima la probabilidad de que el conteo no sea cero. En el siguiente modelo hurdle se ha eliminado la variable explicativa **Pos** y se ha añadido la variable **Portero** para estimar la probabilidad de no ser cero. Se observa lo siguiente:

```
> library(pscl)
> mhurdlebn <- hurdle(GOLES ~ Altura + Nacionalidad + PJ + PJCompl +
+     Penaltis|Portero , dist="negbin", data = X)
> summary(mhurdlebn)
```

```
Call:
hurdle(formula = GOLES ~ Altura + Nacionalidad + PJ + PJCompl + Penaltis |
    Portero, data = X, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.0670 -0.8281 -0.1930  0.3401  6.7757

Count model coefficients (truncated negbin with log link):
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -15.77148    5.34067  -2.953 0.003146 **
Altura                         0.07696    0.02874   2.678 0.007411 **
NacionalidadNoEsp              0.42970    0.27706   1.551 0.120919
PJ                             0.13753    0.02987   4.605 4.13e-06 ***
PJCompl                       -0.07933    0.02098  -3.781 0.000156 ***
Penaltis                       1.09556    0.23461   4.670 3.02e-06 ***
Log(theta)                     0.46393    0.42679   1.087 0.277024
Zero hurdle model coefficients (binomial with logit link):
               Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.3335     0.1891   1.764   0.0778 .
Porteroportero -17.8996  2062.6394  -0.009   0.9931
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta: count = 1.5903
Number of iterations in BFGS optimization: 17
Log-likelihood: -216.1 on 9 Df
```

Se observa como ser portero tiene un coeficiente negativo en la probabilidad de que el conteo de goles no sea cero. Es coherente, aunque no es significativo, pues el p-valor es 0.9931.

Ahora con las mismas variables se aplica el modelo inflado por cero en el cuál se supone que los ceros son de dos tipos (ceros verdaderos y ceros falsos). La función **zeroinfl** estima la probabilidad de que la variable de conteo valga 0.

```
> mzeroinflbn <- zeroinfl(GOLES ~ Altura + Nacionalidad + PJ +
+ PJCompl + Penaltis|Portero , data = X, dist = "negbin")
> summary(mzeroinflbn)
```

```
Call:
zeroinfl(formula = GOLES ~ Altura + Nacionalidad + PJ + PJCompl + Penaltis |
    Portero, data = X, dist = "negbin")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.0361 -0.5812 -0.3593  0.3296  7.0723

Count model coefficients (negbin with log link):
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -13.23741    4.24218  -3.120  0.00181 **
Altura                         0.06160    0.02321   2.654  0.00795 **
NacionalidadNoEsp              0.28519    0.23783   1.199  0.23049
PJ                             0.14809    0.02028   7.303 2.82e-13 ***
PJCompl                       -0.07943    0.01739  -4.568 4.93e-06 ***
Penaltis                       1.08889    0.21323   5.107 3.28e-07 ***
Log(theta)                     0.29471    0.26047   1.131  0.25786

Zero-inflation model coefficients (binomial with logit link):
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.901     55.852  -0.141    0.887
Porteroportero  17.905     86.715   0.206    0.836
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 1.3427
Number of iterations in BFGS optimization: 39
Log-likelihood:  -199 on 9 Df
```

Se observa como ser portero tiene un coeficiente positivo en el modelo que estima la odd-ratio de que el conteo de goles sea cero. Es coherente, aunque no es significativo, pues el p-valor es 0.836.

Se comparan los dos modelos que se han usado para abordar el exceso de ceros con el objetivo de ver cual se adapta mejor a los datos:

```
> vuong(mhurdlebn,mzeroinflbn)


Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
```

```
-------------------------------------------------------------
              Vuong z-statistic             H_A    p-value
Raw                  -3.434199 model2 > model1 0.00029715
AIC-corrected        -3.434199 model2 > model1 0.00029715
BIC-corrected        -3.434199 model2 > model1 0.00029715
```

Se rechaza que los modelos sean indistinguibles. El modelo inflado BN se adapta mejor a los datos que el modelo hurdle BN.

Finalmente, se compara el modelo de regresión BN con el modelo inflado BN y se ve si se adapta mejor a los datos:

```
> vuong(modBN,mzeroinflbn)

Vuong Non-Nested Hypothesis Test-Statistic:
(test-statistic is asymptotically distributed N(0,1) under the
 null that the models are indistinguishible)
-------------------------------------------------------------
              Vuong z-statistic             H_A   p-value
Raw                   2.705442 model1 > model2 0.0034107
AIC-corrected         2.501679 model1 > model2 0.0061803
BIC-corrected         2.213526 model1 > model2 0.0134307
```

Según la salida obtenida, el modelo BN se adapta mejor a nuestros datos que el modelo inflado BN que se ha diseñado.

Una razón que puede justificar esta elección es que la diferencia entre los goles marcados por los defensas y los porteros no es estadísticamente significativa.

7.2.5. Valores pronosticados

Según se ha estudiado, los modelos que mejor se adaptan a los datos son: el primer modelo de Poisson con la variable Min_Jugados eliminada (m1poisson) y el modelo BN (modBN), ambos equivalentes.

Se observa que para predecir el número de goles de un nuevo jugador, en el modelo de Poisson influye el equipo al que este pertenezca, en cambio en modelo BN esta variable no influye en la predicción (el equipo al que pertenezca el jugador es indiferente).

Para predecir el número esperado de goles de un nuevo futbolista perteneciente al Villarreal CF utilizando el modelo de regresión de Poisson, se realizan las siguientes instrucciones y se observa en la salida el número de goles esperado para este futbolista:

```
> nuevofutbolista=data.frame(Altura=178,Nacionalidad="Esp",
Pos="delantero",Equipo="Villarreal CF",PJ=34,PJCompl=30,Penaltis=1)
> (GolesPredichos<- predict(m1poisson, nuevofutbolista, type = "response"))

       1
8.143466
```

A continuación se usa el modelo de regresión Binomial Negativa para predecir el número esperado de goles del mismo fútbolista. Hay que tener en cuenta que en este modelo el equipo al que pertenece no aporta nada, si se cambia el equipo por otro cualquiera se obtendría la misma predicción.

```
> (GolesPredichos<- predict(modBN, nuevofutbolista, type = "response"))

       1
9.114624
```

Se pueden observar por ejemplo los valores predichos para los delanteros españoles y no españoles usando el modelo de regresión BN, con las siguientes instrucciones:

```
delanteros=data.frame(Altura=mean(X$Altura),
Nacionalidad=factor(1:2,levels=1:2,labels=levels(X$Nacionalidad)),
Pos="delantero",PJ=mean(X$PJ),
PJCompl=mean(X$PJCompl),Penaltis=mean(X$Penaltis))
GolesPredichos <- predict(modBN, delanteros, type = "response")
cbind(delanteros,GolesPredichos)[,c(2,7)]

  Nacionalidad  GolesPredichos
1 Esp           2.814361
2 NoEsp         3.883355
```

Según el modelo BN los goles esperados en los delanteros de nacionalidad extranjera son mayores que en los españoles.

Para realizar estas predicciones utilizando el modelo de Poisson habría que tener en cuenta el equipo al que pertenecen los jugadores. A continuación se realiza el mismo estudio anterior pero solamente con los jugadores del Villarreal CF.

```
> delanterosVillarreal=data.frame(Altura=mean(X$Altura),
+ Nacionalidad=factor(1:2,levels=1:2,labels=levels(X$Nacionalidad)),
+ Pos="delantero",Equipo="Villarreal CF",PJ=mean(X$PJ),
+ PJCompl=mean(X$PJCompl),Penaltis=mean(X$Penaltis))
> GolesPredichosVillarreal <- predict(m1poisson, delanterosVillarreal,
+ type = "response")
> cbind(delanterosVillarreal,GolesPredichosVillarreal)[,c(2,8)]
```

```
  Nacionalidad GolesPredichosVillarreal
1 Esp                          2.569421
2 NoEsp                        4.096366
```

Según el modelo de Poisson los goles esperados en los delanteros de nacionalidad extranjera son mayores que en los españoles en el Villarreal CF.

Conclusión:

La modelización de cualquier situación real no es un proceso unívoco sino que depende de muchos factores tales como calidad de datos, modelo teórico subyacente, etc. Con esta aplicación se han mostrado las posibilidades y limitaciones de la aplicación de técnicas estadísticas para datos de conteo en el campo del deporte de competición.

Bibliografía

[1] Agresti, A. *Categorical Data Analysis*
Wiley Series in Probability and Statistics, 1996

[2] Alcaide, M. *Modelo de Regresión Binomial Negativa*
Trabajo fin de grado, 2015

[3] Cameron, A.C. ; Trivedi, P.K. *Regression Analysis of Count Data*
New York: Cambridge University Press, 1998

[4] Cascales, B. ; Lucas, P. ; Mira, J.M. ; Pallarés, A. ; Sánchez-Pedreño S. *El libro de Latex*
Prentice Hall, 2003

[5] Cordeiro, G.M. ; Vasconcellos, K.L.P. ; Barroso, L.P. *Improved estimation for robust econometric regression models*
Brazilian Journal of Probability and Statistics, 2000

[6] Hardin, J.W. ; Hilbe, J.M. *Generalized Linear Models and Extensions*
A State Press Publication StataCorp LP, College Station, Texas, third edition, 2012

[7] Hilbe, J.M. *Negative Binomial Regression*
Cambridge University Press,second Edition, 2011

[8] Hilbe, J.M. *Modeling count data*
Cambridge University Press, 2014

[9] Kopka, H. ; Daly, P.W. *A guide to LATEX2*
Addison-Wesley, 1995

[10] Lindsey, J.K. *Modelling Frecuency and Count Data*
Oxford Univerity Press, 1995

[11] Lindsey, J.K. *Applying Generalized Linear Models*
New York: Springer-Verlag, 1997

[12] Madsen, H. ; Thyregod, P. *Introduction to General and Generalized Linear Models*
CRC Press Taylor & Francis Group, 2011

[13] McCullagh, P. ; Nelder, J.A. *Generalized Linear Models*
London:Chapman and Hall, 1989

[14] Muñoz, J.M. *Apuntes Asignatura Modelos lineales y diseño de experimentos*
Universidad de Sevilla, 2015

[15] Nelder, J.A. ; Wedderburn, R.W.M. *Generalized Linear Models*
Journal of the Royal Statistical Society, 1972

[16] Pino, J.L. *Apuntes Master Big Data y Data Science*
Universidad de Sevilla, 2016