UNIVERSIDAD D SEVILLA

FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

Trabajo Fin de Grado:

MODELO DE REGRESIÓN BINOMIAL NEGATIVA

Mario Alcaide Delgado - Grado en Matemáticas

Junio 2015

Dirigido por:
Joaquín García de las Heras

Índice general

Introducción

En este trabajo nos centraremos, principalmente, en el estudio del modelo de regresión binomial negativa. Aunque este modelo se puede derivar de un modelo de regresión Poisson en el que, tradicionalmente, se ha introducido un nuevo término o componente aleatoria, también puede ser pensado como un miembro de la familia de modelos lineales generalizados.

El planteamiento teórico como miembro de esta familia, permitirá aplicar al modelo binomial negativo los distintos test de bondad de ajuste, análisis de residuos y demás resultados desarrollados para los modelos lineales generalizados.

Si bien, el modelo de regresión de Poisson es, en general, el método más útil para modelizar datos de conteo, el requisito de igualdad de media y varianza de esta distribución dificulta su aplicabilidad, ya que en numerosas ocasiones nos encontramos con experimentos cuyos datos presentan mayor varianza que media debido a múltiples causas entre las que destaca la frecuencia alta de ceros. Estos datos son denominados *Poisson sobredispersos*, pero son más comúnmente designados como, simplemente, *sobredispersos*. Esta problemática motiva el uso de modelos alternativos que presenten una mayor flexibilidad y que se adapten mejor a este tipo de datos.

El presente trabajo se ha estructurado en cuatro capítulos. En el primer capítulo recogeremos diferentes resultados para el análisis estadístico de los modelos lineales generalizados, ya que el modelo de regresión binomial negativo puede considerarse como un caso particular del mismo. Se definen las distintas componentes del modelo lineal generalizado: componente aleatoria, componente sistemática y la función enlace. Trataremos la estimación de los parámetros del modelo mediante el método de máxima verosimilitud. Asimismo, recogemos las distintas técnicas para la evaluación del modelo, contraste de bondad de ajuste, diferentes tipos de residuos,... .

Con el fin de analizar la aplicabilidad del modelo de regresión binomial negativa, en el segundo capítulo introduciremos las variables de conteo o recuento, que cuentan con una amplia presencia en diversos ámbitos de la investigación aplicada. Nos centraremos en una breve descripción de diversas alternativas estadísticas para el estudio de este tipo de datos, algunas de ellas son casos particulares de la modelización lineal (modelo de regresión Poisson, modelo de regresión binomial negativa) que nos permite considerar y analizar el comportamiento de las variables de conteo frente a los valores del conjunto de variables explicativas. Finalizamos tratando el problema de sobredispersión que, en general, pueden presentar este tipo de datos.

En el tercer capítulo nos centraremos en el análisis y estudio del modelo de regresión

binomial negativa, tratando las diferentes derivaciones del mismo. Una vez formulado, realizaremos un estudio teórico de dicho modelo (estimación de los parámetros, medida de bondad de ajuste, interpretación, etc.) con el fin de su evaluación.

El siguiente capítulo está dedicado a dos casos prácticos, donde en ambos se ha utilizado el software R. En primer lugar se ha realizado una breve simulación de un conjunto de datos para el estudio de este modelo y en el segundo caso, se ha aplicado la técnica a un conjunto de datos dados. En ambos casos, se ha aplicado la regresión Poisson y la regresión binomial negativa. En cada una de las aplicaciones vemos cómo la regresión binomial negativa puede ser usada para modelar una variable respuesta de conteo, caracterizada por la presencia de sobredispersión en los datos.

Finalizamos con una recopilación de aquellos textos (libros, artículos, etc.) que nos han permitido realizar este trabajo.

Introduction

In this text we are going to make a study about negative binomial regression model. Although this model can be traditionally derived from a Poisson regression model in which a new term or random component is introduced, this model can be also thought of as a member of the family of generalized linear models.

The theoretical approach as a member of this family, will let us apply the different goodness-of-fit tests, analysis of residuals and other results developed for generalized linear models to the negative binomial model.

Although the Poisson regression model is the standard method used to model count response data, the equality of mean and variance in this distribution makes difficult its application, because we found too many situations in which we come up with experiments whom data presents much bigger variance tan mean due to many different causes in between we find the high frequency of zero data. This data are termed *Poisson overdispersed*, but are more commonly simply designated as *overdispersed*. This problematic motivates the use of alternative models which introduces a greater flexibility and adapt well to this type of data.

The present text is structured in four chapters. In the first chapter we collect different results for statistical analysis of generalized linear models, as the negative binomial regression model can be considered as a particular case of them. We defined the different components of generalized linear model: Random component, systematic component and the link function. We will try to estimate the parameters of the model using maximum likelihood method. We also collect the different techniques for the evaluation of the model, goodness-of-fit test, different types of residual, ...

In order to analyze the applicability of the negative binomial regression model, in the second chapter we will introduce the count variables, that have a strong presence in various fields of applied investigation. We will focus on a brief description of various statistical alternatives for the study of this type of data, some of which are special cases of the linear modeling (Poisson regression model, negative binomial regression model) that allows us to consider and analyze the behavior of count variables against the values of the set of covariables. We finished addressing the problem of overdispersion which generally can show this type of data.

In the third chapter we focus on the analysis and study of negative binomial regression model, studying its different derivations. Also, once formulated, we will make a theoretical study of the model (estimation of parameters, goodness-of-fit measure, interpretation, etc.) with the purpose of its evaluation.

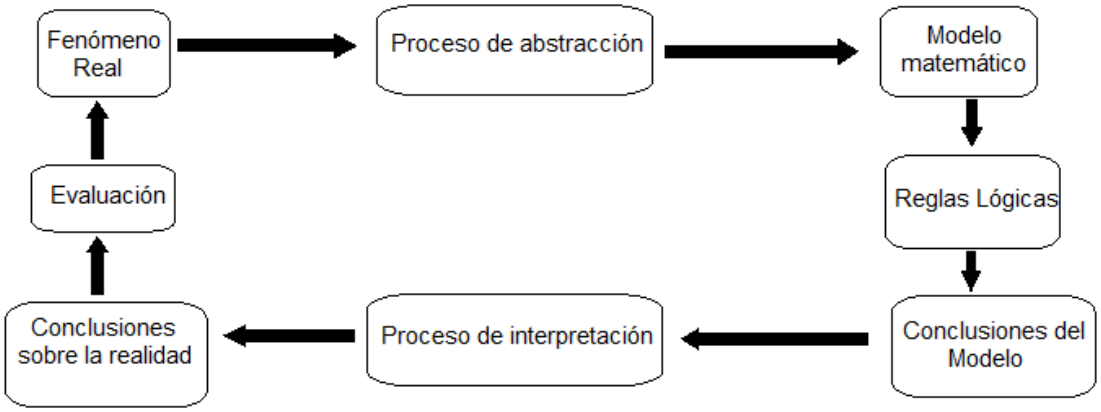
The next chapter is dedicated to two practical cases where the software R is used. Firstly we make a brief simulation of a set of data for the study of this model and in the second case, the technique has been applied to a data set. In both cases, Poisson regression and negative binomial regression have been applied. In each the applications we see how the negative binomial regression can be used to model a response count variable, characterized by the presence of overdispersion in the data.

We conclude with a compilation of those texts (books, articles, etc.) that we have used to make this text.

Capítulo 1

Modelo Lineal Generalizado.

En la mayoría de los estudios aplicados, el investigador (economista, sociólogo, psicólogo,...) a partir de los resultados del fenómeno real bajo estudio, trata de obtener un modelo simulado de la realidad que permita reconstruir el mecanismo subyacente en el fenómeno real. Un modelo es el resultado de un proceso de abstracción desde la realidad al sistema mátematico. Lo podemos reflejar mediante el siguiente diagrama:



A través de la literatura estadística muchos autores, entre los que destacan R.A.Fisher, J.Neyman y D.R.Cox, han tratado el problema de la modelización, es decir, el problema de representar la realidad y su variabilidad e incertidumbre a través de un modelo matemático que permita el estudio, el análisis y la compresión de la misma, con el objetivo de tratar de transformarla, de predecir su futuro o, simplemente, de conocerla.

Entre los diversos criterios de clasificación de modelos existentes, resulta de especial interés la clasificación en el ámbito del modelado estadístico, la distinción entre modelos deterministas y no deterministas [Lindsey, 1997]. Los modelos deterministas son aquellos con capacidad para predecir o explicar un fenómeno determinado, de forma que contienen información suficiente para representar sin error la realidad que modelan. Por el contrario, los modelos probabilísticos o estocásticos tratan con la variedad y la incertidumbre. Como ocurre en los sistemas científicos, socioeconómicos o sanitarios , en los que el fenómeno real bajo estudio presenta diversas fuentes de incertidumbre (aleatoriedad). Estos modelos

asumen la existencia de un error resultante de la desviación entre el fenómeno observado y su representación mediante el modelo, por lo que el modelo matemático resultante del proceso de abstracción incluye un error aleatorio. Son a este tipo de modelos a los que se les denomina como modelos estadísticos. Será en este último tipo de modelos en los que nos centraremos.

El objetivo de la modelización estadística consiste en, a partir de la observación o de la experimentación, explicar el comportamiento de una o más características de los individuos o elementos de una población, en base a las diferencias existentes entre las características asociadas a los individuos.

En cuanto al estudio, la variable, univariante o multivariante, que se desea explicar se conoce cómo variable respuesta o variable objetivo, mientras que las variables en las que se desea basar la explicación se denominan variables explicativas, criterios o covariables.

Inicialmente, el planteamiento de métodos explicativos, en los que se trata de explicar una o varias variables objetivos, a través de un conjunto de variables explicativas, requiere la elección de un modelo que describa la estructura de la relación entre las variables.

Para el planteamiento del modelo es importante distinguir entre el tipo de variables que intervienen (continuas, de conteo, cualitativas, etc.) y en la clase de relaciones funcionales que se admiten para analizar la relación entre la variable objetivo y las variables explicativas. Según el tipo de variables que intervienen y de la relación entre ellas, se dispondrá de un conjunto de posibles modelos más o menos adecuados, capaces de explicar la realidad.

Generalmente, el modelo más estudiado y utilizado es del tipo lineal, es decir, se modeliza la relación tratando de expresar la variable o variables explicativas, o alguna característica de ellas, a través de una combinación lineal de las variables explicativas.

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 clásica [Stigler, 1981] se puede extender a una familia más general, propuesta en [Nelder, 1972] y ampliada en [McCullagh y Nelder, 1989], conocida como modelos lineales generalizados (GLM). Esta nueva familia permite unificar tanto los modelos con variables de respuesta categórica como numérica; y considera distribuciones como binomial, Poisson, hipergeométrica, binomial negativa entre otras, y no únicamente la distribución normal. Y por otro lado supone que la esperanza μ_i está relacionada con las variables explicativas a través de una función enlace.

Como en los Modelos de Regresión Lineal, se considera el supuesto de independencia para las observaciones, sin embargo, para esta nueva familia a diferencia del modelo clásico, las distribución de la componente aleatoria no necesariamente es homocedástica, es decir no se requiere de un supuesto de homogeneidad de varianzas. Por ejemplo en el modelo de regresión de Poisson la varianza de la variable respuesta viene determinada por el valor esperado. Por tanto la varianza puede variar a medida que varíe este valor esperado, a diferencia del modelo clásico con distribución normal que tiene dos parámetros no relacionados, la media y la varianza que se suponen constante para las diferentes observaciones

[McCullagh y Nelder, 1989].

A continuación se recogen las las diferentes componentes que definen un Modelo Lineal Generalizado.

1.1. Componentes del Modelo.

Se pueden diferenciar tres componentes diferentes en la modelización lineal estadística: la componente aleatoria, la componente sistemática y la función enlace. La combinación de estas tres componentes define por completo un Modelo Lineal Generalizado.

1.1.1. Componente aleatoria.

Sea Y la variable aleatoria objetivo o respuesta objeto de estudio y sean las n variables aleatorias independientes e idénticamente distribuidas $Y_1, ..., Y_n$ la muestra aleatoria procedente de Y. Siendo Y denominada como componente aleatoria cuya distribución pertenece a la familia exponencial de distribuciones. Algunos miembros de la familia exponencial son las distribuciones normal, binomial, Poisson, Gamma o binomial negativa.

La distribución de una variable aleatoria Y, caracterizada por los parámetros θ y ϕ pertenece 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 $f()$ denota la función de probabilidad en el caso que Y sea una variable discreta, o la función de densidad en el caso que Y sea una variable continua, θ es el parámetro de localización o canónico, ϕ el parámetro escala y $a(\phi)$, $b(\theta)$ y $c(y,\phi)$ son funciones específicas de cada elemento de la familia. La función $a(\phi)$ es comúnmente escrita de la forma $a(\phi) = \phi/w$, donde w es una ponderación para cada observación.
Se verifica que:

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

donde $b'(\theta)$ y $b''(\theta)$ son, respectivamente, la primera y la segunda derivadas de $b(\theta)$, y donde $V(\mu)$ se denomina *función de varianza*. Esta función captura la relación entre $E(y)$ y $Var(y)$.

En la Tabla siguiente resumimos los elementos principales que caracterizan a algunas de las distribuciones más utilizadas de la familia exponencial.

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

1.1.2. Componente sistemática.

La componente sistemática recoge la variabilidad de Y expresada a través de p variables explicativas $X_1,...,X_p$, que denotaremos por $\mathbf{X}$, y de sus correspondientes parámetros $\beta=(\beta_0,\beta_1,...,\beta_p)'$. La componente sistemática, también denominado predictor lineal, se simboliza por: η

$$\eta=\mathbf{X}\beta$$

El predictor lineal del Modelo Lineal Generalizado (GLM) puede incluir términos tales como las variables explicativas originales, potencias y transformaciones de estas variables dependiendo de la función enlace que utilizemos.

1.1.3. Función enlace.

En el modelo de regresión lineal se modeliza el valor esperado como una combinación lineal de las variables explicativas, sin embargo, en muchos experimentos reales, esta relación no es adecuada, por lo que es necesario la inclusión de una función que relacione el valor esperado con las variables explicativas. Esta función se denomina función enlace y se simboliza por $g(\mu_i)$.

La función enlace que transforma el valor esperado a la escala del predictor lineal es:

$$g(\mu_i)=\eta_i=\mathbf{X}_i\beta$$

donde $\mathbf{X}_i$ representa las p variables explicativas para el i-ésimo individuo con $i=1,...,n$

La inversa de la función enlace dada por:

$$\mu_i=g^{-1}(\eta_i)=h(\eta_i)=h(\mathbf{X}_i\beta)$$

definiendo h como $h=g^{-1}$.

La elección de la función enlace no siempre resulta obvia, pueden existir diferentes funciones enlace aplicables a un problema particular, de forma que hay que decidir cual es la más apropiada en cada caso. Es importante elegir una función enlace que nos facilite la interpretación del modelo óptimo obtenido.

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

$$\theta_i = \theta(\mu_i) = \eta_i = \mathbf{X}_i\beta \qquad g(\mu_i) = \theta(\mu_i)$$

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

Binomial:

$$\theta(\mu_i) = ln\left(\frac{\mu_i}{1-\mu_i}\right) \qquad \theta^{-1}(\eta_i) = \left(\frac{exp(\eta_i)}{1+exp(\eta_i)}\right)$$

Binomial Negativa:

$$\theta(\mu_i) = ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) \qquad \theta^{-1}(\eta_i) = \frac{1}{\alpha(exp(-\eta_i)-1)}$$

siendo $\alpha = 1/r$

Gamma:

$$\theta(\mu_i) = -\frac{1}{\mu_i} \qquad \theta^{-1}(\eta_i) = -\frac{1}{\eta_i}$$

Normal:

$$\theta(\mu_i) = \mu_i \qquad \theta^{-1}(\eta_i) = \eta_i$$

Poisson:

$$\theta(\mu_i) = log(\mu_i) \qquad \theta^{-1}(\eta_i) = exp(\eta_i)$$

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

- *Modelos para respuestas binarias (Bernouilli) o binomiales.* Permiten considerar variables objetivos del tipo 0-1 (tener/no tener una característica, éxito/fracaso,...) muy útiles en medicina, análisis de riesgos, etc.
- *Modelos para datos de conteo (Poisson, Binomial Negativa).* Permiten considerar y analizar el comportamiento de variables de conteo (número de accidentes, nacimientos, individuos de una especie,...), frente a los valores del conjunto de variables explicativas.
- *Modelos para variables respuesta continuas (Normal, Exponencial, Elíptica,...).* Permiten considerar y analizar el comportamiento de variables contínuas, cuantitativas (ganancias, niveles de un compuesto químico, tiempo de vida,...), frente a los valores del conjunto de variables explicativas.

Los modelos que vamos a tratar particularmente son los denominados **modelos para datos de conteo**. En estos modelos, la distribución de Poisson es bastante utilizada, pero también encontramos, y como objeto de estudio en nuestro trabajo, el modelo de Regresión Binomial Negativa.

- **Modelo de Poisson log-linear:** Se obtiene tomando como función enlace el canónico.

$$log(\mu_i) = \eta_i = \mathbf{X}_i\beta \quad \text{ó} \quad \mu_i = exp(\mathbf{X}_i\beta)$$

- **Modelo de Regresión Binomial Negativa:** Se obtiene tomando como función enlace, la función logaritmo $g(\mu_i) = log(\mu_i)$.

$$log(\mu_i) = \eta_i = \mathbf{X}_i\beta \quad \text{ó} \quad \mu_i = exp(\mathbf{X}_i\beta)$$

En cualquiera de los dos casos la media condicional se especifica como:

$$E(Y_i|\mathbf{X}_i = x_i) = exp(x_i\beta)$$

donde x_i indica los valores observados de las p variables explicativas en el i-ésimo individuo.

1.2. Estimación de los parámetros.

Tras la especificación de uno o varios modelos, se estiman, para cada modelo especificado, los parámetros del predictor lineal y posteriormente se valora la precisión de esas estimaciones a través del cálculo de la discrepancia entre pares de modelos, con el objetivo de seleccionar el modelo óptimo.

Dos de los métodos más comunes en la estimación estadística de parámetros son el método de *Mínimos Cuadrados Ordinarios* y el método de *Máxima Verosimilitud*. Sin embargo, el más adecuado es el *Método de Máxima Verosimilitud*, que tiene las propiedades de consistencia y eficiencia asintótica [Codeiro, 2000].

Consideremos la muestra $y_1, ..., y_i, ...y_n$ junto con las covariantes $x_1, ..., x_i, ...x_n$ este método trata de maximizar la verosimilitud para obtener un estimador del vector de parámetros desconocidos β en el modelo:

$$E[Y_i|\mathbf{X}_i = x_i] = \mu_i = h(x_i\beta)$$

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

Asumiendo que cada componente de Y tiene una distribución proveniente de la familia exponencial de la forma denotada anteriormente, escribimos la función de verosimilitud como:

$$L(\theta; y) = f(y; \theta) = \prod_{i=1}^{n} f_i(y_i; \theta) \text{ con } y = (y_1, ..., y_n)'$$

dado que las observaciones son independientes la función *log-verosimilitud* viene dada por:

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

La función $c(y_i, \phi)$ que no depende de θ_i ha sido omitida. Insertando la relación $\theta_i = \theta(\mu_i)$ entre el parámetro natural y la esperanza de la i-ésima observación,

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

Dada la relación entre la esperanza y el vector de parámetros $\mu_i = h(x_i^t \beta)$, se tiene

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

Su primera derivada es la denominada **función score o función marcador:**

$$s(\beta) = \frac{\partial l}{\partial \beta} = \sum_i s_i(\beta)$$

Las contribuciones individuales a la función marcador son:

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

donde

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

En efecto, teniendo en cuenta la regla de la cadena,

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

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

Para obtener la derivada que aparece en (1.1) y (1.2), procedemos como sigue:

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

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 obtiene:

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

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

- **Matriz de información de Fisher esperada:**

  $F(\beta) = Cov\ s(\beta) = \sum_i F_i(\beta)$

  $F_i(\beta) = x_i x_i^t w_i(\beta) \qquad : \qquad w_i(\beta) = D_i^2(\beta) \sigma_i^{-2}(\beta)$

- **Matriz 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 enlace canónicas $\theta(\mu_i) = x_i^t \beta$, se simplifica la forma de las matrices de información y la función marcador:

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

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

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 positiva. Estas ecuaciones no son, generalmente, lineales por lo que ha de ser resueltas a través de métodos iterativos. En muchos casos se utiliza el método iterativo Fisher scoring o mínimos cuadrados ponderados iterados, cuyas iteraciones vienen definidas, a partir de un estimador inicial $\hat{\beta}^0$, por:

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

Sobre este método debemos realizar las siguientes consideraciones:

1. El parámetro escala ϕ se cancela en el término $F^{-1}(\hat{\beta}^{(k)}) s(\hat{\beta}^{(k)})$ por lo que se explica el comentario inicial hecho para 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 puede terminar con el criterio

$$\frac{\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\|}{\|\hat{\beta}^{(k)}\|} < \varepsilon \qquad \text{para un } \varepsilon > 0 \text{ fijado.}$$

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

Obtenemos a través de este método las estimaciones de los parámetros del modelo $\hat{\beta}$. Estas estimaciones máximo-verosímiles presentan propiedades de consistencia, eficiencia asintótica y distribución normal asintótica.

En caso de que el parámetro dispersión sea desconocido, puede considerarse el siguiente estimador consistente:

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

1.3. Adecuación de los modelos.

Una vez estimados los parámetros, se debe valorar la magnitud de discrepancia entre los datos observados y los esperados por el modelo. Una discrepancia pequeña entre los datos observados y el conjunto de valores estimados $\hat{\mu}$ puede ser tolerable, en cuanto que una discrepancia grande no.

De esta manera, si se admite una combinación satisfactoria de la distribución de la componente aleatoria y de la función enlace, el objetivo es determinar cuantos términos son necesarios en la estructura lineal para una descripción razonable de los datos. Un número grande de variables explicativas puede llevar a que un modelo explique bien los datos pero con un aumento de complejidad en su interpretación. Por otro lado, un número pequeño de variables explicativas puede llevar a un modelo de fácil interpretación pero que se ajuste pobremente a los datos. Lo que se desea en realidad es un modelo intermedio.

En el proceso del ajuste del modelo se evalúan generalmente un conjunto de modelos que constituyen aproximaciones de los datos observados. Tratamos de construir un modelo intermedio entre el modelo saturado y el modelo nulo, los cuales se caracterizan por :

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

El concepto de "*mejor modelo*" depende de la finalidad que se persiga. Cuando la finalidad del modelado es del tipo predictivo se seleccionan las variables que expliquen el mayor porcentaje de variabilidad de la respuesta, y para ello se emplean fundamentalmente los criterios estadísticos que veremos a continuación. Por otro lado, cuando la finalidad es explicativa, son los argumentos teóricos los que deben tomar un mayor protagonismo, por

lo que el proceso de selección debe ser guiado por el investigador y se basa en la especificación de un modelo máximo inicial y de un conjunto sucesivo de modelos restringidos que se comparan mediante un ajuste condicional. A partir de ahí, el proceso de adecuación seguiría, a grandes rasgos, los siguientes pasos:

- Evaluar los términos de interacción a partir de su significación estadística
- Analizar la necesidad de mantener variables de control en el modelo. Haciendo uso de criterios de relevancia práctica más que en base a criterios estadísticos, se debe evaluar si:
  1. La eliminación de variables confusas sesgará la estimación de los parámetros de interés
  2. O bien, en el caso de variables de ajuste, si su supresión del modelo implicaría una pérdida de precisión de las estimaciones.
- Valorar si las variables explicativas de interés deben permanecer en el modelo. Para ello se deben emplear tanto criterios estadísticos como de relevancia substantiva.

Tal como se ha expuesto anteriormente, 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, es decir, atender a los criterios de **bondad de ajuste**.

1.3.1. Bondad de ajuste.

En un modelo lineal generalizado, la bondad del ajuste se puede evaluar de distintas formas, entre las que destacan:

- La **función ó estadístico desviación**:

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

  Que es la distancia entre el logaritmo de la función verosimilitud del modelo saturado (con n parámetros, uno por observación) y el modelo que se está investigando. 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.

  Para probar la adecuación de un Modelo Lineal Generalizado, el valor de la desviación debe ser comparado con el percentil de alguna distribución de probabilidad referente. Si el modelo es correcto el estadístico desviación se distribuye asitóticamente según una $\mathcal{X}^2_{n-p}$ con $n-p$ grados de libertad [McCullagh y Nelder, 1989].

$$D(y,\hat{\mu}) \sim \mathcal{X}^2_{n-p}$$

- El **coeficiente de determinación** R^2:

  La medida R^2 es definida como la reducción proporcional en la incertidumbre, debido a la inclusión de los regresores. Bajo ciertas condiciones también puede ser explicada como la varianza explicada por el modelo ajustado.

Este coeficiente viene dado por:

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

donde $D(y;\hat{\mu})$ y $D(y;\hat{\mu_0})$ son las funciones de desviación del modelo ajustado y nulo (modelo simple que se usa como referencia) , respectivamente y se verifica que $0 \leq R^2 \leq 1$

- El **estadístico Chi-cuadrado de Pearson**:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

donde $V(\hat{\mu})$ es la función varianza estimada para la distribución de la variable respuesta.

En cuanto a la selección del modelo, hacemos uso de los criterios de información para seleccionar entre diferentes modelos. Estos se basan en la comparación de log-verosimilitudes pero penalizando a aquellos modelos con más variables explicativas. Estos criterios son:

- El **Criterio de información Akaike (AIC)**:
En el caso general el AIC es:

$$AIC = k - 2ln(\hat{L})$$

donde k es el número de parámetros en el modelo estadístico y $\hat{L}$ es el máximo valor de la función verosimilitud para el modelo estimado. Dado un conjunto de modelos candidatos para los datos, el modelo preferido es el que tiene el valor mínimo en el AIC. Este criterio no sólo recompensa la bondad de ajuste, sino también incluye una penalidad, que es una función creciente del número de parámetros estimados.

- El **Criterio de información bayesiano (BIC)**:

$$BIC = -2ln\hat{L} + kln(n)$$

donde k es el número de parámetros libres a ser estimados, n es el tamaño de la muestra y $\hat{L}$ es el máximo valor de la función de verosimilitud. Está estrechamente relacionado con el AIC y al igual que él resuelve el problema de selección de modelos mediante la introducción de un término penalización para el número de parámetros en el modelo, el término de penalización es mayor en el BIC que el en AIC.

Se selecciona aquel modelo con menor valor en el criterio que se utilize.

Análisis de las variables explicativas

A continuación vamos a realizar inferencias sobre el vector de parámetros desconocidos β de dimensión p, dependiendo de los objetivos o de la metodología a usar para el análisis de datos.

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

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

En general, se recogen los siguientes procedimientos para este contraste:

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

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

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

- **Estadístico de Wald**, que se basa en la distibución normal asintótica del vector $\hat{\beta}$. Se define por :

  $$\xi_W = [\mathbf{C}\hat{\beta} - \xi]^T [\mathbf{C}F^{-1}(\hat{\beta})\mathbf{C}'][\mathbf{C}\hat{\beta} - \xi]$$

  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**, obtenida 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 poderada de $s(\tilde{\beta})$ a cero, es decir:

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

  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 $\mathcal{X}^2_s$.

Sea el caso especial de testar la significación de un subconjunto de covariantes:

$$H_0 : \beta_r = 0$$
$$H_1 : \beta_r \neq 0$$

donde β_r es un subvector de β. Los estadísticos anteriores se reducen a:

$$\xi_W = (\hat{\beta})^T F_r^{-1}(\hat{\beta})\hat{\beta}_r$$

$$\xi_{SR} = s_r^T(\tilde{\beta})F_r^{-1}(\tilde{\beta})s_r(\tilde{\beta})$$

Donde F_r es la submatriz de F correspondiente a los elementos de β_r y s_r el subvector de s correspondiente a dichos elementos.

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, la razón de Máxima Verosimilitud es preferida por ser un test uniformemente más potente [Codeiro, 2000].

Regiones de confianza para β

Los intervalos de confianza y/o regiones de confianza pueden contruirse usando cualquiera de estos estadísticos. Usando,por ejemplo el estadístico de Wald, una región de confianza para β con un nivel de confianza $100(1-\alpha)\,\%$ viene dada por:

$$\{ \quad \beta \in \mathbb{R}^p \quad | \quad (\hat{\beta}-\beta)^T[\hat{V}ar(\hat{\beta})]^{-1}(\hat{\beta}-\beta) < \mathcal{X}^2_{p,1-\alpha}\}$$

1.3.2. Residuos

En la práctica, puede ocurrir que aún escogiendo cuidadosamente un modelo y después al ajustarlo a un conjunto de datos, el resultado sea insatisfactorio.

Las desviaciones sistemáticas se originan por haber escogido inadecuadamente la función de enlace o las variables explicativas incluídas en el modelo. Las discrepancias aisladas pueden ocurrir debido a puntos extremos, o porque estos realmente son erróneos como resultado de una mala transcripción de los datos o por factores no controlados en el momento de la toma de datos. La verificación de la adecuación del modelo es un requisito fundamental que se realiza sobre el conjunto de datos para analizar posibles desviaciones de las suposiciones hechas para el modelo, así como la existencia de observaciones extremas con alguna interferencia desproporcionada en los resultados del ajuste.

Como en la regresión lineal, los residuos son los utilizados para verificar dicha adecuación del modelo. Expresan la discrepancia entre una observación y su valor ajustado, y también pueden indicar la presencia de valores anómalos o discordantes que puedan requerir de una investigación más detallada. Entre otros residuos los más destacados son:

- **El residuo básico**:
  Definido como la diferencia entre el valor observado, y_i, de la variable respuesta y el valor ajustado, $\hat{y}_i$, por el modelo.

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

- **EL residuo de Pearson**: Es la contribución individual al estadístico $\mathcal{X}^2$ de Pearson, se define como:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} Var(\hat{\mu}_i)}} \quad i = 1, 2, ..., n$$

siendo $\hat{\phi}$ un estimador consistente del parámetro escala ϕ.

Y su versión estudentadizada viene dada por:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\phi} Var(\hat{\mu}_i)(1 - h_i)}} \quad i = 1, 2, ..., n$$

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:

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

La ventaja de usar este residuo estudentadizado 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.

- **EL residuo desviación**, que se define como:

$$r_i^D = sign(y_i - \hat{\mu}_i) \sqrt{d_i} \quad i = 1, 2, ..., n$$

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

Y su versión estudentadizada:

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

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.

Es importante destacar que los errores de especificación mencionados pueden provocar una violación del supuesto distribucional de relación media-varianza de diversos modelos lineales generalizados, como es el caso del modelo de regresión de Poisson.

1.4. Interpretación

Una vez se ha obtenido el modelo adecuado haciendo uso de los criterios antes mencionados de bondad de ajuste, estudio de residuos,... el proceso de modelado se cierra con la interpretación del modelo.

Recordando que la transformación provocada por la aplicación de una función enlace (excepto en la función de enlace identidad) da lugar en la mayoría de los casos a una ecuación del modelo expresada en términos multiplicativos, en la que la interpretación de los parámetros se realiza en términos del factor de cambio en el valor esperado para un incremento unitario de las variables explicativas.

Capítulo 2

Modelos para Datos de Recuento

Las variables de recuento cuentan con una amplia presencia en diversos ámbitos de la investigación aplicada tanto en las Ciencias Sociales como en las Ciencias de la Salud. En este sentido, encontramos ejemplos de investigaciones aplicadas con variables de recuento en disciplinas como Demografía [Merklesson y Roth, 2000], Farmacología [Lindsey, 2001], Relaciones Laborales [Sturman, 1999]ó Criminología [Osgood, 2000], por citar algunos.

Mucho más numerosas resultan las investigaciones que contienen variables de recuento en Medicina [Böhning, 1994], Ciencias Políticas [King, G., 1989] y Ciencias Económicas [Meliciani, 2000]. De hecho, estas tres disciplinas merecen ser consideradas aparte, puesto que no sólo cuentan con una extensa aplicación de investigaciones con variables de recuento, sino que han hecho valiosas aportaciones en el tratamiento estadístico de este tipo de variables [Cameron y Trivedi, 1986], [Winkelmann, 2000].

En este capítulo nos centraremos en los modelos de datos de recuento o conteo, un caso particular de modelización lineal que nos permite considerar y analizar el comportamiento de las variables de conteo frente a los valores del conjunto de variables explicativas. Introduciremos los distintos tipos de modelos específicos para este tipo de variables y las distintas características que estos presentan.

2.1. Datos de recuento

Se denominan variables de recuento (count data) a aquellas variables que determinan el número de sucesos o eventos que ocurren en una misma unidad de observación en un intervalo espacial o temporal definido.

A partir de esta definición propuesta por [Lindsey, 1995], se derivan dos características principales de una variable recuento, que la diferencia de una variable cuantitativa continua, estas son, su naturaleza discreta y no negativa.

La variable a tratar Y, que toma los valores $0, 1, 2, ...$, se caracteriza por tomar infinitos números de valores que podemos ordenar en orden creciente, y cuya probabilidad va en descenso a medida que sea mayor el valor de la variable.

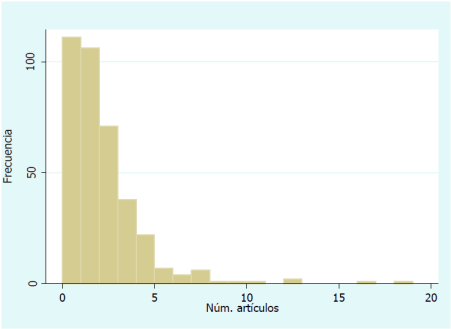
Ejemplos de tales eventos o conteos pueden ser:

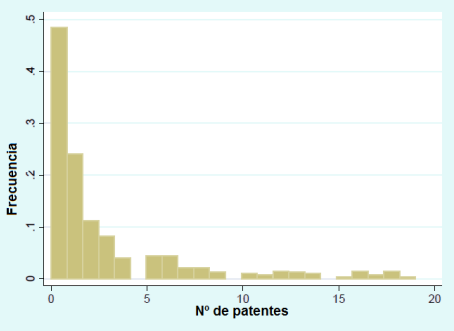
Conteos en el tiempo

- Número de accidentes de tráfico en un tramo de cierta carretera en un mes.
- Número del registro de partículas de una desintegración radioactiva por segundo.
- Número de patentes solicitadas por una empresa en 1 año.

Conteos en el espacio

- Número de células sanguíneas en una muestra de sangre.
- Número de árboles infectados por hectárea en un bosque.
- Número de artículos publicados por una revista.





Los modelos de datos de conteo se caracterizan porque no tienen un límite superior natural, toman el valor cero (en un porcentaje no despreciable) para algunos miembros de la población y suelen tomar pocos valores distintos.

El objetivo,como hemos señalado, es analizar Y como función de las variables explicativas $X_1, ..., X_p$, a través de:

$$E(Y/\mathbf{X} = x)$$

podiendo ser cada una de las variables explicativas de cualquier tipo.

2.2. Modelos específicos para variables de recuento

No todos los modelos de predicción son aplicables a este tipo de variables, pues pueden originarse problemas como pérdida de información o de inconsistencia. La problemática que surgiría al construir un modelo cualquiera para este tipo de variables sería:

- **Modelo de regresión lineal**

  1. Las predicciones de Y pueden salirse del rango de valores en el que está definido.
  2. Las estimaciones pueden ser inconsistentes.
  3. Puede tener validez para hacer una exploración previa de las relaciones.

- **Modelo de elección binaria**

  Si la variable Y toma muchos valores, plantear un modelo de elección binaria nos conduce a una pérdida de eficiencia (porque perdemos información) ya que agregamos todos los valores mayores que uno en un solo valor.

- **Modelos ordenados**

  1. Si la variable Y toma muchos valores o si tiene pocas observaciones en algunas de las variables, es necesario agrupar si queremos estimar un modelo ordenado. Esto en determinados contextos, puede suponer pérdida de información.
  2. Un aspecto positivo de los modelos ordenados es que podemos utilizarlos cuando queremos analizar variables que toman valores enteros negativos

A continuación presentamos un breve resumen de algunos de los modelos propuestos que ofrecen una mejor aproximación para este tipo de datos, son los **Modelos para variables de recuento** y estos son :

→ Modelo de regresión de Poisson

→ Modelo de regresión Binomial Negativa

→ Modelo en dos partes

→ Modelo con exceso de ceros.

Donde los modelos de regresión de Poisson y de regresión binomial negativa, a diferencia del resto de los mencionados, pertenecen a la familia de modelos lineales generalizados.

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

2.2.1. Modelo de Regresión de Poisson

La distribución habitual para procesos de conteo es la **Distribución de Poisson**

Sea

$$Y \sim P(\mu)$$

con función de probabilidad:

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

de parámetro $\mu > 0$

la media y la varianza viene dada por:

$$E(Y) = \mu$$

$$Var(Y) = \mu$$

Esta igualdad media-varianza se conoce como la equidispersión de la distribución de Poisson.

Es a través de la construcción de un modelo, en la que la variable independiente sigue una distribución Poisson, donde podemos especificar el parámetro μ_i como una forma funcional de las variables explicativas $X_1, ..., X_p$. La especificación más habitual es una exponencial lineal con el fin de garantizar que $\mu_i > 0$, es decir, se hace uso de la función enlace canónica para la formación del modelo.

$$\mu_i = exp(x_i\beta)$$

Se trata del **Modelo exponencial**

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

Por tanto la distribución de Poisson condicionada a las variables explicativas $\mathbf{X}$ viene dada por:

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

donde

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

Esta formulación se conoce como **Modelo de Regresión de Poisson**

La principal bondad del modelo de regresión de Poisson es que es capaz de capturar la naturaleza discreta y no negativa de los datos de recuento, en especial cuando tales datos proceden de eventos raros.

Notamos que:

$Var(Y_i/x_i) = E(Y_i/x_i)$ (propiedad de equidispersión).

Dado que el modelo de regresión de Poisson pertenece a la familia de modelos lineales generalizados, podremos hacer uso de los resultados estudiados en el primer capítulo, obteniéndose para este modelo las siguientes expresiones:

- La función desviación:

$$\begin{aligned} D(y; \hat{\mu}) &= 2\{l(y; y) - l(\hat{\mu}; y)\} \\ &= 2\sum_{i=1}^{n}\{y_i log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\} \end{aligned}$$

- El coeficiente de determinación R^2:

$$\begin{aligned} R^2 &= 1 - \frac{D(y; \hat{\mu})}{D(y; \hat{\mu_0})} \\ &= \frac{\sum_{i=1}^{n}\{y_i log(\hat{\mu}_i/y_i) - (\hat{\mu}_i - y_i)\}}{\sum_{i=1}^{n}\{y_i log(y_i/\bar{y}_i)\}} \end{aligned}$$

- El estadístico Chi-cuadrado de Pearson:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

- El residuo de Pearson:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \quad \text{con } i = 1, 2, ..., n$$

- El residuo de Pearson estudentadizado:

$$r_i^p = \frac{y_i - \hat{\mu}_i}{(\sqrt{\hat{\mu}_i})(1 - h_i)} \quad \text{con } i = 1, 2, ..., n$$

  donde h_i es el i-ésimo elemento de la matriz de proyección

El modelo de regresión Poisson se caracteriza por:

- Es un modelo heterocedástico, aquel en el que las varianzas de las perturbaciones no son constantes, por lo tanto, la variabilidad es diferente para cada observación.
- Tiene la propiedad de equidispersión.

Si bien es cierto que el modelo de regresión Poisson presenta indudables mejoras con respecto al modelo lineal general, sin embargo, puede resultar inapropiado en otros aspectos. En este sentido, tal y como expone [Winkelmann, 2000], «es común encontrar en el trabajo aplicado con datos de recuento (...) que ciertas asunciones del modelo de regresión de Poisson son sistemáticamente rechazadas por los datos».

Luego, a pesar de que éste es el modelo de referencia en estudios de variables de recuento, y que resulta especialmente adecuado para modelar valores enteros no negativos, especialmente cuando la frecuencia de ocurrencia es baja, presenta varios problemas a la hora de tratar con datos en los que la media y la varianza condicionadas no coinciden.

2.2.2. Modelo de Regresión Binomial Negativa

Este modelo, más general para este tipo de datos, puede ser usado en presencia de sobredispersión, es decir, en casos en los que $Var(Y_i/x_i) \geq E(Y_i/x_i)$.

Su representación tradicional es obtenida mediante la incorporación de un término de perturbación en la medida del modelo de Poisson, una aleatoriedad en el parámetro μ_i.

$$\mu_i^* = exp(x_i\beta + \varepsilon_i) = \mu_i exp(\varepsilon_i)$$

donde el término de perturbación ε_i sigue una distribución Gamma.

Siendo su función de probabilidad:

$$P(Y = y_i/x_i) = \frac{\Gamma(y_i + v_i)}{\Gamma(y_i + 1)\Gamma(v_i)} \left(\frac{v_i}{v_i + \mu_i}\right)^{v_i} \left(\frac{\mu_i}{v_i + \mu_i}\right)^{y_i}$$

con $\mu_i = E[Y_i/x_i] = exp(x_i\beta)$ y definimos $v_i = (1/\alpha)\mu_i^t$ donde $t = 0, 1$

La especificación final depende de cómo definamos v

- Si $v = (1/\alpha)$ aparece el **Modelo de Regresión Binomial Negativo 1 (NB1)** donde:

$$E(Y_i|x_i) = exp(x_i\beta)$$
$$Var(Y_i|x_i) = (1+\alpha)exp(x_i\beta)$$

- Si $v = (1/\alpha)\mu$ aparece la **Modelo de Regresión Binomial Negativo 2 (NB2)** donde:

$$E(Y_i|x_i) = exp(x_i\beta)$$
$$Var(Y_i|x_i) = exp(x_i\beta)(1+\alpha exp(x_i\beta))$$

En estos modelos si $\alpha > 0$ entonces $Var(Y_i|x_i) \geq E(Y_i|x_i)$, lo que sugiere que los datos presentan **sobredispersión**.

Este modelo de regresión lo estudiaremos con más profundidad en el próximo capítulo tratando sus parametrizaciones más conocidas, la estimación de sus parámetros entre otros estudios.

2.2.3. Exceso de ceros

Algunas variables de recuento muestran un porcentaje de ceros muy grande. Esa cantidad de ceros no es consistente con las distribuciones Poisson o Binomial Negativa (generalmente es mayor)

Dos de los modelos más utilizados para abordar este tipo de situaciones son:

- Modelo en dos partes
- Modelo con exceso de ceros

Ambos modelos no pertenecen a la familia de modelos lineales generalizados. La idea básica de estos modelos es que los ceros (todos o parte de ellos) no proceden del mismo proceso generador de datos que el resto de valores.

Un ejemplo

Nº de veces que va una persona a pescar en el último mes.

Aquellos que han ido 0 veces puede ser por dos motivos:

- No son pescadores

- Si son pescadores, pero ese mes no han ido a pescar debido a restricciones de tiempo, dinero...

Modelo con exceso de ceros

Otro modelo que permite manejar la presencia de sobredispersión y el exceso de ceros es el denominado modelo en dos partes. En éste se supone que el proceso que siguen las observaciones cero es diferente del que sigue las observaciones positivas. Para ello se combina un modelo binario y otro truncado.

De forma general se tiene:

$$P(Y_i = 0) = f_1(0)$$

$$P(Y_i = y_i) = (1 - f_1(0))\frac{f_2(y_i)}{1 - f_2(0)} \qquad y_i = 1, 2, ...$$

donde $f_1()$ y $f_2()$ representan distribuciones de probabilidad, que pueden ser o no del mismo tipo, $1 - f_1(0)$ recoge la probabilidad de cruzar el primer estado de decisión y $1 - f_2(0)$ se aplica para truncar en el cero la segunda distribución. Los diferentes supuestos que se realizan sobre estas distribuciones conducen a las versiones truncadas Poisson y Binomial Negativa.

Modelo en dos partes

Los modelos con exceso de ceros suponen otra alternativa para recoger de forma adecuada la abundancia de valores nulos, en los que se considera que estos valores pueden haber sido generados por dos situaciones diferentes, lo que lleva a una probabilidad suplementaria para estos valores a la obtenida por una distribución estándar. De forma similar al modelo en dos partes, en éste se combina un proceso binario y una distribución para los recuentos de ceros y positivos. Formalmente,

$$P(Y_i = 0) = f_1(0) + (1 - f_1(0)f_2(0)$$

$$P(Y_i = y_i) = (1 - f_1(0))f_2(y_i)\frac{f_2(y_i)}{1 - f_2(0)} \qquad y_i = 1, 2..$$

donde $f_1()$ representa la distribución de probabilidad del proceso binario habitualmente recogida mediante el modelo logit y $f_2()$ la distribución de los recuentos, que será Poisson o Binomial Negativa.

Este modelo puede considerarse también como un modelo mixto de dos componentes, donde uno de ellos sigue una distribución degenerada en cero y el otro, es un modelo de regresión estándar para datos de recuento.

2.3. Problema de sobredispersión

Como hemos señalado anteriormente, a pesar de que el Modelo de Regresión de Poisson se presenta como un modelo de indudable utilidad para representar datos de conteo, éste puede resultar inapropiado debido al incumplimiento de ciertos supuestos, el más común es la ausencia de equidispersión. En la práctica, puede ocurrir que se presente infradispersión o sobredispersión en el conjunto de datos, esta última aparece con mayor frecuencia.

Recordemos que la equidispersión constituye un supuesto básico, es decir, se asume que $V(Y) = \sigma^2 E(Y)$, donde el parámetro de dispersión $\sigma^2 = 1$. La sobredispersión ocurre cuando $V(Y) > E(Y)$, es decir $\sigma^2 > 1$. Cuando existe exceso de variación en los datos,

las estimaciones de los errores estándar pueden resultar sesgadas, pudiendo presentarse errores en las inferencias a partir de los parámetros del modelo de regresión.

La incidencia y el grado de sobredispersión encontrado dependen mucho del campo de aplicación. Son muchos los autores que desde diferentes disciplinas han tratado este tema: [Hausman, Hall y Griliches, 1984] y [Osgood, 2000]
Entre las diversas causas de la sobredispersión destacan:

- Alta variabilidad en los datos.
- Los datos no provienen de una distribución Poisson.
- Los eventos no ocurren independientemente a través del tiempo.
- Falta de estabilidad, es decir, la probabilidad de ocurrencia de un evento puede ser independiente de la ocurrencia de un evento previo pero no es constante.
- Errores de especificación de la media μ [Winkelmann, 2000] 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 fue apropiado el escoger el enlace log-lineal.

Existen diversas propuestas para detectar sobredispersión, por ejemplo [Lindsey, 1995] propone aplicar el coeficiente de variación CV:

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

Este coeficiente teóricamente debería tomar el valor 1, si se cumpliera la equidispersión.

Generalmente la sobredispersión se evalúa a través de la relación entre el estadístico de Pearson $\mathcal{X}^2$ o la función desviación D y sus respectivos grados de libertad (gl), es decir evaluar:

$$\frac{\mathcal{X}^2}{gl} \qquad \text{ó} \qquad \frac{D}{gl}$$

Si estos valores son mayores que 1, indican sobredispersión.

Otro diagnóstico está basado en una prueba de Razón de Verosimilitud (RV) basada en las distribuciones Poisson y tradicional Binomial Negativa (BN2) .

- Para la distribución Poisson $V(Y) = \mu$.
- Para la distribución Binomial Negativa (BN2) $V(Y) = \mu + k\mu^2$.

Si $k = 0$, entonces la distribución Binomial Negativa se reducirá a una Poisson. Por tanto las hipótesis que se plantean son:

$$H_0 : k = 0$$
$$H_1 : k > 0$$

Para llevar a cabo esta prueba, se deberán ajustar los 2 modelos: Poisson y Binomial Negativa. Para cada modelo se obtendrá su respectiva función de log-verosimilitud (l) . El estadístico propuesto es :

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

Según [Cameron y Trivedi, 1998] este estadístico tiene una distribución asintótica $\mathcal{X}_1^2$. Por tanto, rechazaremos H_0 si el estadístico es mayor que $\mathcal{X}_{1,1-\alpha}^2$.

En tal caso, sería más conveniente modelar el número de ocurrencias a través de una Binomial Negativa. La interpretación de los resultados sería la misma que en el caso de la Regresión Poisson.

Es pues necesario, la introducción de un nuevo modelo que efectúe una mejor aproximación de las predicciones. Se trata del **Modelo de Regresión Binomial Negativo**, el cual analizaremos con más detenimiento en el siguiente capítulo.

Capítulo 3

Modelos de Regresión Binomial Negativa

Como hemos visto en el capítulo anterior, a pesar de que el modelo de referencia en estudios de variables de recuento es el modelo de regresión de Poisson (MRP), éste presenta varios problemas a la hora de tratar con datos en los que la media y la varianza condicionadas no coinciden, en concreto, con datos que presentan sobredispersión.

Una forma de relajar esta restricción de igualdad media-varianza del Modelo de Regresión de Poisson es especificar una distribución que permita un modelado más flexible. En este sentido, el modelo paramétrico estandar para datos de recuento con presencia de sobredispersión es el Modelo de Regresión Binomial Negativa (MRBN).

A pesar de las ventajas potenciales de este modelo para datos de conteo, son pocos los estudios publicados que han usado esta distribución. Esos estudios se han centrado mayormente en los recuentos de aves raras, accidentes de tráfico, la utilización de servicios de salud, lesiones neurológicas y leucocitos [Welsh, Cunningham y Chambers, 2000], [Abdel y Radwan, 2000], [Sormani, Bruzzi y Miller, 1999] y [Finch y Chen, 1999]. En el artículo sobre un ensayo clínico realizado por [Byers, Allore, Gill y Peduzzi 2003] podemos ver un estudio practico más detallado del modelo de regresión binomial negativa.

El modelo de regresión binomial negativa es un modelo estadístico atípico, en el sentido de que los investigadores se refieren a éste como un único modelo, como, por ejemplo, el modelo de regresión Poisson, la regresión logística,... cuando en realidad existen distintos modelos binomiales negativos, que dependerán del tipo de problema de fondo que se esté abordando. [Boswell y Patil, 1970] identificaron 13 tipos distintos que derivan de la distribución binomial negativa mientras que otros autores argumentan que existen más.

Podemos describir la distribucíon binomial negativa como aquella variable que estudia la probabilidad de observar un número determinado de fracasos, y, antes del r-ésimo éxito en una serie de experimentos Bernoulli independientes. Bajo tal descripción de r, este sería un entero positivo, sin embargo, no hay razón matemática de peso para limitar este parámetro para números enteros; solo limitar r como positivo.

Diremos que la variable aleatoria de conteo Y_i con $i = 1, ..., n$ sigue una distribución Binomial Negativa de parámetros r y p, $Y_i \sim BN(r, p)$, con función de probabilidad dada

por:

$$P(Y_i = y_i) = \binom{y_i + r - 1}{r - 1} p^r (1-p)^{y_i} \tag{3.1}$$

donde $0 < p < 1$, $r > 0$, $y_i = 0, 1, 2....$

El valor esperado y la varianza vienen dados por:

$$E(Y_i) = \frac{r(1-p)}{p} \tag{3.2}$$

$$V(Y_i) = \frac{r(1-p)}{p^2} \tag{3.3}$$

estableciéndose entre ambos la siguiente relación:

$$V(Y_i) = \frac{1}{p} E(Y_i)$$

Como $0 < p < 1$, se verifica que $V(Y_i) > E(Y_i)$ lo que justifica la aptitud natural de esta distribución para modelar datos que se caracterizan por la existencia de sobredispersión.

En este capítulo nos centraremos principalmente en la derivación del modelo de regresión binomial negativo, trataremos los distintos métodos para la estimación de los parámetros del mismo y estudiaremos la bondad de ajuste, los residuos asociados,... del modelo con el fin de evaluarlo.

3.1. Derivación del Modelo

Podemos recoger las distintas formas de derivar el modelo de regresión binomial negativo (MRBN) en dos orígenes diferenciados.

Por una parte el MRBN se puede derivar como una distribución Poisson compuesta con una Gamma, en el cual la distribución Gamma es usada para ajustar los datos Poisson que presentan sobredispersión. De esta forma se deriva el modelo tradicional binomial negativo, el cual es comunmente simbolizado como modelo de regresión BN2, [Cameron y Trivedi, 1998].

Por otra parte el MRBN puede ser pensado como miembro de la familia exponencial de distribuciones y, por tanto, ser considerado como un modelo lineal generalizado. Este enfoque sólo se tiene si el parámetro de dispersión o heterogeneidad es introducido en la distribución como una constante [Hilbe 1993]. Tal interpretación permite a los investigadores aplicar al MRBN los test de bondad de ajuste, análisis de residuos y cualquier otro estudio que hemos desarrollado para los Modelos Lineales Generalizados.

3.1.1. Derivación del modelo a partir de una distribución Poisson

Partiendo de un modelo de regresión Poisson (MRP), podemos derivar el modelo de regresión binomial negativo (MRBN) siguiendo dos enfoques distintos.

Un primer enfoque, el más común, introducido por [GTM, 1984], en el que la variable respuesta sigue una distribución Poisson, cuya media está especificada de forma incompleta debido a una situación de heterogeneidad no observada, para suplir dicha situación se introduce un nuevo término error.

Un segundo enfoque, [Hausman, Hall y Griliches, 1984], que supone que la variable respuesta sigue una distribución Poisson en la que su media no se considera como un parámetro fijo, sino que se interpreta como un parámetro estocástico que varía aleatoriamente como una distribución Gamma.

En cualquiera de los dos casos, el modelo Poisson es insuficiente para la modelización de los datos, debido a la heterogeneidad que estos presentan. Ambas motivaciones conducen a una distribución binomial negativa.

A continuación recogemos la función de densidad de la variable Gamma:

$$T \sim G(\tau, \omega) \Rightarrow f(t) = \frac{1}{w^{\tau}\Gamma(\tau)} t^{\tau-1} e^{\frac{t}{\omega}}$$

con $t, \omega, \tau > 0$

y siendo la función Gamma:

$$\Gamma(v) = \int_0^{\infty} t^{v-1} e^{-t} dt$$

1^{er} Caso

Mientras que en el MRP la media condicional de Y_i es:

$$\mu_i = exp(x_i\beta)$$

en el modelo que se va a realizar esta media es reemplazada por la variable aleatoria $\tilde{\mu}_i$, verificándose la siguiente ecuación estocástica:

$$\tilde{\mu}_i = exp(x_i\beta + \varepsilon_i)$$

bajo la hipótesis de que el término error considerado, ε_i, está incorrelacionado con las variables explicativas. El término de error ε_i puede ser el resultado del efecto conjunto de variables no incluidas en el modelo o bien una fuente de aleatoriedad intrínseca. Sea cual sea su origen, ε_i representa la heterogeneidad no observada de los datos.

En el MRP, la variación en μ_i es introducida a través de la heterogeneidad observada, de forma que diferentes valores de x_i dan diferentes valores de μ_i. Así, todos los individuos con los valores x_i tienen la misma μ_i. En este modelo propuesto, la variación en $\tilde{\mu}$ es debida tanto a la variación en x entre los individuos, como a la heterogeneidad no observada introducida a través de ε. Para una combinación de valores en las variables independientes, existe una distribución de diversas $\tilde{\mu}_i$ en lugar de una μ_i única.

La relación entre ambas medias viene dada por:

$$\tilde{\mu}_i = exp(x_i\beta)exp(\varepsilon_i) = \mu_i \cdot exp(\varepsilon_i) = \mu_i\delta_i$$

siendo δ_i equivalente a $exp(\varepsilon_i)$. La aproximación del MRBN dependerá de la asignación ó hipótesis acerca de la media del término de error. La asignación más conveniente y usual es que:

$$E(\delta_i) = 1$$

$$V(\delta_i) = \omega_i$$

que implicará que el recuento esperado después de añadir la nueva fuente de variación sea la misma que para el MRP:

$$E(\tilde{\mu}_i) = E(\mu_i\delta_i) = \mu_i \cdot E(\delta_i) = \mu_i$$

Bajo estas suposiciones se verifica que:

$$P(Y_i = y_i|x_i, \delta) = \frac{exp(-\tilde{\mu}_i)\tilde{\mu}_i^{y_i}}{y_i!} = \frac{exp(-\mu_i\delta_i)(\mu_i\delta_i)^{y_i}}{y_i!}$$

por lo que se vuelve a obtener una distribución Poisson.

Sin embargo, puesto que el error δ_i es desconocido no podemos calcular $P(y_i|x_i, \delta_i)$.

Para calcular $P(y_i|x_i)$ sin tener en cuenta δ_i, promediamos $P(y_i|x_i\delta_i)$ por la probabilidad de cada valor de δ_i. Si g es la función de densidad de probabilidad de δ_i, entonces la densidad marginal de Y_i de i-ésimo individuo puede ser obtenida integrando con respecto a δ_i:

$$P(Y_i = y_i|x_i) = \int_0^\infty [P(Y_i = y_i|x_i\delta_i)]g(\delta_i)d\delta_i = \int_0^\infty \frac{e^{-exp(x_i\beta_i)\delta_i}exp(x_i\beta_i)^{y_i}\delta^{y_i}}{y_i!}g(\delta_i)d\delta_i \tag{3.4}$$

Esta expresión define la distribución de Poisson compuesta [Cameron y Trivedi, 1986]. Tal como indican estos mismos autores, las distribuciones de Poisson compuestas proporcionan una generalización natural de los modelos de Poisson básicos y su aplicación obedece generalmente a una necesidad de mayor flexibilidad, especialmente en situaciones de sobredispersión.

La ecuación de la distribución de Poisson compuesta (3.4) calcula la probabilidad de la variable respuesta Y_i como una mixtura de dos distribuciones de probabilidad. Asimismo, la forma de (3.4) depende de la selección de $g(\delta_i)$, es decir, de la función de densidad de probabilidad que se asuma para δ_i.

En este sentido, suponiendo que la variable error δ_i sigue una distribución $G(\tau, \omega)$, con función de densidad:

$$g(\delta_i) = \frac{1}{\omega^\tau\Gamma(\tau)}\delta^{\tau-1}exp(-\frac{\delta_i}{\omega}) \quad para \quad \tau, \omega > 0$$

El resultado de la ecuación de la regresión de Poisson compuesta (3.4) conduce a una distribución binomial negativa, es decir, tal como indica [Poortema, 1999], la distribución binomial negativa es la distribución compuesta resultante si la conjugada de la distribución Poisson, es decir, la distribución Gamma, es utilizada para la composición.

En resumen, bajo estas consideraciones y partiendo de que $\delta_i \sim G(\tau, \omega)$ podemos probar que

$$Y_i|x_i \sim BN\left(\tau, \frac{1}{1+\omega\mu_i}\right)$$

con función de probabilidad:

$$P(Y_i = y_i/x_i) = \frac{\Gamma(y_i+\tau)}{\Gamma(y_i+1)\Gamma(\tau)}\left(\frac{1/\omega}{1/\omega+\mu_i}\right)^{\tau}\left(\frac{\mu_i}{1/\omega+\mu_i}\right)^{y_i} \quad para \quad y_i = 0, 1, 2... \tag{3.5}$$

y cuyo valor esperado es:

$$E(Y_i|x_i) = \tau\omega\mu_i$$

y la varianza condicional:

$$Var(Y_i|x_i) = \tau\omega\mu_i(1+\omega\mu_i) \tag{3.6}$$

2º Caso

En este enfoque se supone que el parámetro μ_i no es un parámetro fijo, sino que este se distribuye de forma aleatoria según una ley Gamma.

Se supone que:

$$\begin{aligned} &Y_i|x_i, \mu_i \sim P(\mu_i) \\ &\mu_i|x_i \sim G(\tau_i, \omega), \qquad \tau_i, \omega > 0 \end{aligned}$$

y sabiendo que

$$P(Y_i = y_i/x_i) = \int_0^{\infty} P(Y_i = y_i|x_i, \mu_i) f(\mu_i|x_i) d\mu_i$$

demostramos que $Y_i|x_i$ sigue una distribución binomial negativa:

$$\begin{aligned} P(Y_i = y_i|x_i) &= \int_0^{\infty} \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\frac{1}{w^{\tau_i}\Gamma(\tau_i)}\mu_i^{\tau_i-1}e^{\frac{\mu_i}{\omega}}d\mu_i \\ &= \frac{1}{\omega^{\tau_i}\Gamma(\tau_i)y_i!}\int_0^{\infty} e^{-\left(1+\frac{1}{\omega}\right)\mu_i}\mu_i^{y_i+\tau_i-1}d\mu_i \end{aligned}$$

Esta última integral corresponde a $\frac{\Gamma(y_i+\tau_i)}{\left(1+\frac{1}{\omega}\right)^{y_i+\tau_i}}$ y $y_i! = \Gamma(y_i+1)$, por lo que:

$$\begin{aligned} P(Y_i = y_i|x_i) &= \frac{\Gamma(y_i+\tau_i)}{\Gamma(y_i+1)\Gamma(\tau_i)}\left[\frac{1}{\omega\left(1+\frac{1}{\omega}\right)}\right]^{\tau_i}\left(\frac{1}{1+\frac{1}{\omega}}\right)^{y_i} \\ &= \frac{\Gamma(y_i+\tau_i)}{\Gamma(y_i+1)\Gamma(\tau_i)}\left(\frac{1}{1+\omega}\right)^{\tau_i}\left(\frac{\omega}{1+\omega}\right)^{y_i} \end{aligned}$$

Comparándola con (3.1), se verifica que $Y_i|x_i \sim BN\left(\tau_i, \frac{1}{1+\omega}\right)$, y haciendolo con (3.2) y (3.3) concluimos que:

$$E(Y_i) = \omega\tau_i \tag{3.7}$$

$$V(Y_i) = \omega(1+\omega)\tau_i \tag{3.8}$$

El modelo de regresión se establece como:

$$E(Y_i) = exp(x_i\beta)$$

La relación entre la varianza y el valor esperado condicional es lineal y viene dada por la expresión:

$$V(Y_i|x_i) = (1+\omega)E(Y_i|x_i)$$

Parametrizaciones más comunes

A continuación, presentaremos las parametrizaciones más comunes del modelo de regresión binomial negativa, dependiendo del enfoque que sigamos.

Siguiendo el primer enfoque señalado, [Long, 1997] afirma que la asignación más común es que δ_i sigue una distribución $G(\frac{1}{\omega}, \omega)$, con función de densidad:

$$g(\delta_i) = \frac{1}{\omega^{1/\omega}\Gamma(1/\omega)}\delta_i^{\frac{1}{\omega}-1}exp(-\frac{\delta_i}{\omega}) \quad para \quad \omega > 0$$

[Johnson,Kotz y Balakrishnan 1994] demuestran que si δ_i sigue una distribución Gamma, entonces se verifica también que $E(\delta_i) = 1$ y $Var(\delta_i) = \omega$, como habiamos considerado previamente.

En resumen, bajo estas consideraciones y partiendo de que $\delta_i \sim G(\frac{1}{\omega}, \omega)$ podemos probar que

$$Y_i|x_i \sim BN\left(\frac{1}{\omega}, \frac{1}{1+\omega\mu_i}\right)$$

con función de probabilidad:

$$P(Y_i = y_i/x_i) = \frac{\Gamma(y_i + 1/\omega)}{\Gamma(y_i+1)\Gamma(1/\omega)}\left(\frac{1/\omega}{1/\omega+\mu_i}\right)^{1/\omega}\left(\frac{\mu_i}{1/\omega+\mu_i}\right)^{y_i} \quad para \quad y_i = 0, 1, 2... \tag{3.9}$$

y cuyo valor esperado es el mismo que para la distribución de Poisson:

$$E(Y_i|x_i) = exp(x_i\beta) = \mu_i$$

sin embargo, la varianza condicional si difiere en relación a la de la distribución de Poisson:

$$Var(Y_i|x_i) = \mu_i(1+\omega\mu_i) = exp(x_i\beta) + \omega exp(x_i\beta)^2 \tag{3.10}$$

Puesto que $\mu_i > 0$ y $\omega > 0$, la varianza condicional de Y_i del individuo i-ésimo en el MRBN será mayor que la media condicional $exp(x_i\beta)$. Se observa que a medida que ω tiende a

cero la varianza tiende a μ_i $Var(Y_i|x_i) \to \mu_i$, obteniéndose la hipótesis de equidispersión.

Por otro lado, el incremento de la frecuencia relativa de valores de recuento altos y bajos ocasiona una varianza condicional elevada en Y_i. De esta forma, en una situación de sobredispersión, la distribución binomial negativa corrige, especialmente, la probabilidad asociada a valores bajos de recuento que habitualmente presentan un ajuste deficiente a través del MRP.

Siguiendo el segundo enfoque señalado, [Cameron y Trivedi, 1986] construyó un modelo más general que contempla varias relaciones posibles entre la media y varianza condicionada.

Suponiendo que:

$$\begin{aligned} &Y_i|x_i, \mu_i \sim P(\mu_i) \\ &\mu_i|x_i \sim G(v_i, \frac{\mu_i}{v_i}), \qquad v_i, \mu_i > 0 \end{aligned}$$

Y usando un razonamiento similar al anterior se obtiene que:

$$Y_i|x_i \sim BN\left(v_i, \frac{v_i}{\mu_i + v_i}\right)$$

con media y varianza condicional dada por:

$$E(Y_i|x_i) = \mu_i \tag{3.11}$$

$$V(Y_i|x_i) = \mu_i + \frac{1}{v_i}\mu_i^2 \tag{3.12}$$

Los modelos de regresión se obtienen definiendo:

$$\mu_i = exp(x_i\beta) \tag{3.13}$$

$$v_i = \frac{1}{\alpha}[exp(x_i\beta)]^a, \qquad \alpha > 0, \qquad a = 0, 1, 2, ... \tag{3.14}$$

Sustituyendo en el valor de v_i en (1.10) se obtiene:

$$\begin{aligned} Var(Y_i|x_i) &= exp(x_i\beta) + \alpha[exp(x_i\beta)]^{-a}[exp(x_i\beta)]^2 \\ &= exp(x_i\beta) + \alpha[exp(x_i\beta)]^{2-a} \\ &= \mu_i + \alpha\mu_i^{2-a} \end{aligned}$$

Esta nueva expresión origina una amplia variedad de formas para relacionar la varianza y la media. Bajo esta expresión ya se pueden empezar a considerar dos modelos principales del tipo Binomial Negativa que recogemos a continuación.

- El **modelo Binomial Negativa I (BN1):** (relación lineal entre la varianza y la media)

  Es el modelo resultante si consideramos $a = 1$, entonces $v_i = \frac{\mu_i}{\alpha}$. Luego:

  $$Y_i|x_i \sim BN\left(\frac{\mu_i}{\alpha}, \frac{1}{1+\alpha}\right)$$

  con una función de probabilidad:

  $$P(y_i/x_i) = \frac{\Gamma(y_i + \alpha^{-1}\mu_i)}{\Gamma(y_i+1)\Gamma(\alpha^{-1}\mu_i)}\left(\frac{1}{1+\alpha}\right)^{\frac{\mu_i}{\alpha}}\left(\frac{\alpha}{1+\alpha}\right)^{y_i} \quad para \quad y_i = 0, 1, 2...$$

  y su varianza condicional es:

  $$Var(Y_i|x_i) = \mu_i + \alpha\mu_i = (1+\alpha)\mu_i$$

- El **modelo Binomial Negativa II (BN2):** (relación cuadrática entre varianza y media)

  Es el modelo resultante si consideramos $a = 0$, entonces $v_i = \frac{1}{\alpha}$. Luego:

  $$Y_i|x_i \sim BN\left(\frac{1}{\alpha}, \frac{1}{1+\alpha\mu_i}\right)$$

  con función de probabilidad:

  $$P(y_i/x_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i} \quad para \quad y_i = 0, 1, 2... \tag{3.15}$$

  Al hacer efectiva la igualdad $v_i = \alpha^{-1}$, se hace evidente que al incrementar α, que es conocido como el parámetro de dispersión, aumenta la varianza condicional de Y:

  $$Var(Y_i|x_i) = \mu_i\left(1 + \frac{\mu_i}{\alpha^{-1}}\right) = \mu_i(1+\alpha\mu_i) = \mu_i + \alpha\mu_i^2 \tag{3.16}$$

  Se observa que si el parámetro de dispersión $\alpha = 0$ habría equidispersión, pues $Var(Y_i|x_i) = \mu_i + \alpha\mu_i^2 = \mu_i$

  La función de probabilidad (3.15) y la varianza (3.16) caracterizan al modelo de regresión binomial negativo de tipo 2 (NB2) al que se conoce como el modelo tradicional binomial negativo.

De la forma en que hemos deducido el modelo binomial negativo, se es bastante exigente en términos de supuestos, en particular al exigir que el error de especificación siga una distribución Gamma. Es por ello que, por lo general, se comienza estimando el modelo más sencillo el de Poisson, sujeto a una serie de pruebas y, sólo si se considera conveniente,

analizar el resto de modelos, que también deberán ser evaluados.

Además de los modelos BN1 y BN2 considerados anteriormente, algunos autores como [Cameron y Trivedi, 1986], o [Winkelmann y Zimmermann, 1991] proponen un modelo binomial negativo más general, el denominado hipermodelo Binomial Negativo K(BN k), en el cual $Var(Y_i|x_i) = \mu_i + \alpha\mu_i^{2-k}$. Consideramos que esta extensión queda fuera de nuestro estudio.

3.1.2. Derivación del modelo como modelo lineal generalizado.

Él método que abordamos a continuación, es el método más usual para derivar cualquier miembro de la familia de modelos lineales generalizados.

La distribución binomial negativa es un miembro de la familia exponencial siempre que el parámetro de dispersión o heterogeneidad sea introducido en la distribución como una constante, por tanto la podemos usar como distribución de la componente aleatoria del modelo lineal generalizado. Como se ha comentado anteriormente, otra decisión importante es la selección del enlace y por lo tanto, según la elección de éste saldrán a su vez diferentes modelos. En la literatura se suelen elegir dos enlaces diferentes el enlace canónico y el enlace logarítmico. La ventaja de utilizar éste último se debe a que nos permite hacer una comparación con el modelo de regresión Poisson.

Componente aleatoria

La forma de la función de probabilidad binomial negativa con la que comenzamos se expresa como:

$$Y \sim BN(r, p)$$

$$P(Y = y) = \binom{y + r - 1}{r - 1} p^r (1 - p)^y \tag{3.17}$$

con $y = 0, 1, ...$
que se puede expresar como miembro de la familia exponencial negativa con la siguiente estructura:

$$P(Y = y) = exp\left\{ yln(1 - p) + r(ln(p)) + ln\left(\binom{y + r - 1}{r - 1}\right)\right\} \tag{3.18}$$

Aplicando los resultados obtenidos en el capítulo 1, se obtiene que:

$$\theta = ln(1 - p) \Rightarrow p = 1 - exp(\theta) \tag{3.19}$$

$$b(\theta) = -rln(p) \Rightarrow -r(1 - exp(\theta)) \tag{3.20}$$

$$a(\phi), \text{ es el escalar que tomamos como 1} \tag{3.21}$$

La primera y la segunda derivada de $b(\theta)$, con respecto de θ, respectivamente dan lugar a la función media y varianza:

$$b'(\theta) = \frac{\partial b}{\partial p}\frac{\partial p}{\partial \theta} = -\frac{r}{p}\{-(1-p)\} = \frac{r(1-p)}{p} = \mu \tag{3.22}$$

$$b''(\theta) = \frac{\partial^2 b}{\partial p^2}\left(\frac{\partial p}{\partial \theta}\right)^2 + \frac{\partial b}{\partial p}\frac{\partial^2 p}{\partial \theta^2} = \frac{r}{p^2}(1-p)^2 - \frac{r}{p}(1-p) = \frac{r(1-p)}{p^2} = \sigma^2 \tag{3.23}$$

Y por lo tanto $V(\mu) = r(1-p)/p^2$.

Expresando p y r en términos de μ y α se obtiene que:

$$(1-p)/(\alpha p) = \mu$$

$$(1-p)/p = \alpha\mu$$

$$p = 1/(1+\alpha\mu)$$

donde $\alpha = 1/r$

Dados los valores definidos de μ y α, podemos volver a expresar la función de probabilidad de la distribución binomial negativa tal que:

$$P(Y=y) = \binom{y+1/\alpha-1}{1/\alpha-1}\left(\frac{1}{1+\alpha\mu}\right)^{1/\alpha}\left(\frac{\alpha\mu}{1+\alpha\mu}\right)^{y} \tag{3.24}$$

A partir de aquí obtenemos los siguientes términos en función de μ y α:

$$b'(\theta) = \frac{\partial b}{\partial \mu}\frac{\partial \mu}{\partial \theta} = \frac{1}{1+\alpha\mu}\mu(1+\alpha\mu) = \mu \tag{3.25}$$

$$b''(\theta) = V(\mu) = \frac{\partial^2 b}{\partial \mu^2}\left(\frac{\partial \mu}{\partial \theta}\right)^2 + \frac{\partial b \partial^2 \mu}{\partial \mu \partial \theta^2} = \mu + \alpha\mu^2 \tag{3.26}$$

Función enlace

El modelo de Regresión Binomial Negativo se deriva a partir de la función enlace, donde se parametriza la relación entre la media, μ y las variables predictoras. Ésta función enlace viene dada por:

$$g(\mu) = \theta = ln((\alpha\mu)/(1+\alpha\mu)) = -ln(1/\alpha\mu + 1)$$

y su inversa:

$$g^{-1}(\mu) = \mu = 1/\{\alpha(e^{-\theta}-1)\}$$

Usando la función enlace canónica el modelo tiene la siguiente forma:

$$g(\mu) = \eta = x\beta$$

$$ln(1/\alpha\mu + 1) = \eta = x\beta$$

Desde el punto de vista teórico, el enlace canónico presenta una simplificación del estudio del modelo, pero, desde el punto de vista aplicado la mayoría de los investigadores proponen el enlace logarítmico, $g(\mu) = ln(\mu)$, obtenemos el siguiente modelo:

$$g(\mu) = \eta = x\beta$$

$$ln(\mu) = \eta = x\beta$$

Bajo este enlace el modelo que obtenemos es el modelo BN2 o modelo tradicional de regresión binomial negativa. Este modelo de binomial negativa fue llamado la *log-binomial negativa* en[Hilbe 1993] por el enlace elegido. El hecho de que el modelo BN2 coincida con el modelo de regresión binomial negativa con enlace logarítmica es lo que nos permita aplicar todos los resultados obtenidos en el primer capítulo para este modelo.

Se produce un modelo de regresión binomial negativo interpretado como modelo lineal generalizado (GLM), que produce estimaciones de los parámetros idénticos a los calculados para el modelo de tipo compuesto Poisson-Gamma. Aunque los errores estándar calculados como GLM difieren ligeramente de modelo de regresión binomial negativo de tipo compuesto, que se estima normalmente mediante un procedimiento de máxima verosimilitud mediante el algoritmo Newton-Raphson.

Independientemente de la manera en que se estima el modelo de regresión binomial negativa, éste es casi siempre usado para modelar la sobredispersión de los datos. Las ventajas del enfoque GLM está en su capacidad para utilizar las estadísticas especializadas y residuos para modelos lineales generalizados que vienen con la mayoría de softwares GLM.

El único inconveniente con la versión de GLM es que el parámetro de dispersión,α, no se estima directamente,sino que se tiene que introducir en el modelo GLM como una constante.

3.2. Estimación

El método más usual para la estimación de los parámetros en los modelos de recuento es el método de **Máxima verosimilitud**. Dada la complejidad de los ecuaciones de verosimilitud obtenidas, se debe recurrir a procedimientos numéricos para la resolución de las mismas. La mayoría de los investigadores proponen el método de Newton-Raphson o el método Fisher scoring, éste último método está intrínsecamente relacionado a la estimación de modelos lineales generalizados. Ambos métodos son usados comunmente para el análisis de datos tanto Poisson como binomiales negativos.

Cuando el modelo es concebido como un modelo compuesto Poisson-Gamma, se estima mediante el método de máxima verosimilitud usando el algoritmo de Newton-Raphson. Tal método permite la estimación tanto del parámetro media μ, así como del parámetro binomial negativo de dispersión, α.

Por otro lado , si el modelo de regresión binomial negativa es considerado como miembro de la familia de modelos lineales generalizados, habitualmente la estimación en este caso toma la forma del método Fisher scoring, que permite la estimación de sólo el parámetro media, μ o $exp(\beta x)$, por lo que el parámetro de dispersión α debe de ser incluído en el algoritmo como una constante conocida. Aunque esto es un inconveniente para su utilidad, la capacidad de evaluar el ajuste compensa este problema. [Breslow 1984] y [Hilbe 1993] dan diferentes procedimientos para la estimación de α, cuyo resultado difiere, aunque no demasiado, del valor estimado usando máxima verosimilitud.

La mayoría de los autores optan por usar ambos métodos al emplear una tarea de modelado. Se realiza una estimación inicial utilizando un procedimiento de máxima verosimilitud Newton-Raphson, con el resultante valor estimado de α que se inserta a continuación en el algoritmo conocido para modelos lineales generalizados. Ambos métodos son pues utilizados para el modelado global.

En este apartado recogemos una breve descripción de ambos métodos mencionados para estimar modelos de recuento, el Newton-Raphson y el Fisher scoring.

Como se procede en relación a la estimación de modelos lineales generalizados visto en el primer capítulo, se presentan en esta sección las expresiones relativas, como la función log-verosimilitud, la función score y los elementos de la matriz hessiana de cada uno de los modelos binomiales negativos principales.

La función de log-verosimilitud de las dos variedades principales del modelo BN son las siguientes:

$$BN1 : \mathcal{L}(\beta, \alpha) = \sum_{i=1}^{n} \left[-ln(y_i!) + \sum_{j=1}^{y_i} ln(\alpha y_i + \mu_i - \alpha_j) - \left(\frac{\mu_i}{\alpha} + y_i \right) ln(1+\alpha) \right] \quad (3.27)$$

$$BN2 : \mathcal{L}(\beta, \alpha) = \sum_{i=1}^{n} \left[-ln(y_i!) + \sum_{j=1}^{y_i} ln(\alpha y_i + 1 - \alpha_j) + y_i ln(\mu_i) - \left(\frac{1}{\alpha} + y_i \right) ln(1+\alpha\mu_i) \right] \quad (3.28)$$

Hemos visto que el modelo binomial negativo BN2 se puede derivar como una composición Poisson-Gamma y como un modelo lineal generalizado, lo que no ocurre en el modelo BN1, ya que, para ello, una condición necesaria es que el parámetro v_i sea fijo, no dependiendo de μ_i (3.14).

Por lo tanto, en el caso del modelo BN2, tanto el vector de parámetros del modelo, β, como el parámetro de dispersión α, pueden ser estimados mediante el método de máxima verosimilitud conservando las propiedades teóricas de los estimadores obtenidos por éste método, como vimos en el primer capítulo.

Sin embargo, la estimación simultánea de α y β puede conducir a resultados inconsistentes en el caso en que la distribución real de la variable respuesta no sea BN2. Quizás es por esta razón que el modelo BN2 es más "popular"que el BN1, pues en este caso, la aplicación del método de máxima verosimilitud sólo produce estimadores consistentes de β, y no de α.

Las **funciones score** vienen dadas por:

$$s(\beta,\alpha)=\left\{\frac{\partial\mathcal{L}(\beta,\alpha)}{\partial\beta},\quad\frac{\partial\mathcal{L}(\beta,\alpha)}{\partial\alpha}\right\}$$

Los estimadores de máxima verosimilitud $\hat{\beta}$ y $\hat{\alpha}$ son los valores de β y α que maximizan $\mathcal{L}(\beta,\alpha)$ sobre el rango válido donde se definen los parámetros. Consideramos que la verosimilitud tiene un único máximo global. Las ecuaciones de verosimilitud para encontrar los estimadores de máxima verosimilitud vienen dadas por:

BN1:

$$\frac{\partial\mathcal{L}(\beta,\alpha)}{\partial\beta}=\sum_{i=1}^{n}x_i'\mu_i\left[\sum_{j=1}^{y_i}\frac{1}{\alpha y_i+\mu_i-\alpha_j}-\frac{ln(1+\alpha)}{\alpha}\right]=0 \tag{3.29}$$

$$\frac{\partial\mathcal{L}(\beta,\alpha)}{\partial\alpha}=\sum_{i=1}^{n}x_i'\mu_i\left\{\sum_{j=1}^{y_i}\frac{y_i-j}{\alpha y_i+\mu_i-\alpha_j}+\frac{1}{\alpha}\left[\frac{\mu_i ln(1+\alpha)}{\alpha}-\frac{\mu_i+\alpha y_i}{1+\alpha}\right]\right\}=0 \tag{3.30}$$

BN2:

$$\frac{\partial\mathcal{L}(\beta,\alpha)}{\partial\beta}=\sum_{i=1}^{n}\left(x_i'\frac{y_i-\mu_i}{1+\alpha\mu_i}\right)=0 \tag{3.31}$$

$$\frac{\partial\mathcal{L}(\beta,\alpha)}{\partial\alpha}=\sum_{i=1}^{n}x_i'\mu_i\left\{\sum_{j=1}^{y_i}\frac{y_i-j}{\alpha y_i+1-\alpha_j}+\frac{1}{\alpha}\left[\frac{ln(1+\alpha\mu_i)}{\alpha}-\frac{\mu_i(1+\alpha y_i)}{1+\alpha\mu_i}\right]\right\}=0 \tag{3.32}$$

La **matriz Hessiana** viene dada por:

$$H(\beta,\alpha)=\begin{bmatrix}\dfrac{\partial^2\mathcal{L}(\beta,\alpha)}{\partial\beta\beta'} & \dfrac{\partial^2\mathcal{L}(\beta,\alpha)}{\partial\beta\partial\alpha}\\ \dfrac{\partial^2\mathcal{L}(\beta,\alpha)}{\partial\alpha\partial\beta} & \dfrac{\partial^2\mathcal{L}(\beta,\alpha)}{\partial\alpha^2}\end{bmatrix}$$

La **matriz de información de Fisher esperada**:

$$I_e(\beta,\alpha)=E[-H(\beta,\alpha)]$$

y la **matriz de información observada en la muestra**

$$I_{Obs}(\beta,\alpha)=-H(\beta,\alpha)$$

De donde, para el modelo BN1 se obtiene:

$$\frac{\partial^2\mathcal{L}(\beta,\alpha)}{\partial\beta\beta'}=\sum_{i=1}^{n}x_i'x_i\mu_i\left\{\sum_{j=1}^{y_i}\frac{\alpha(y_i-j)}{[\alpha y_i+\mu_i-\alpha_j]^2}-\frac{ln(1+\alpha)}{\alpha}\right\}$$

$$\frac{\partial^2\mathcal{L}(\beta,\alpha)}{\partial\beta\partial\alpha}=\left[\frac{\partial^2\mathcal{L}(\beta,\alpha)}{\partial\alpha\partial\beta}\right]'=\sum_{i=1}^{n}x_i'\mu_i\left[\frac{ln(1+\alpha}{\alpha^2}-\sum_{j=1}^{y_i}\frac{y_i-j}{[\alpha y_i+\mu_i-\alpha_j]^2}-\frac{1}{\alpha(1+\alpha}\right]$$

$$\frac{\partial^2\mathcal{L}(\beta,\alpha)}{\partial\alpha^2}=\sum_{i=1}^{n}\left\{-\sum_{j=1}^{y_i}\left(\frac{y_i-j}{\alpha y_i+\mu_i-\alpha_j}\right)^2+\frac{1}{\alpha^2}\left[\frac{2\mu_i+3\alpha\mu_i+\alpha^2y_i}{(1+\alpha)^2}-\frac{2\mu_i ln(1+\alpha)}{\alpha}\right]\right\}$$

Y para el modelo BN2:

$$\frac{\partial^2 \mathcal{L}(\beta,\alpha)}{\partial\beta\beta'} = \sum_{i=1}^{n} x_i' x_i \mu_i \frac{1+\alpha y_i}{(1+\alpha\mu_i)^2}$$

$$\frac{\partial^2 \mathcal{L}(\beta,\alpha)}{\partial\beta\partial\alpha} = \left[\frac{\partial^2 \mathcal{L}(\beta,\alpha)}{\partial\alpha\partial\beta}\right]' = \sum_{i=1}^{n} x_i' \mu_i \frac{y_i-\mu_i}{(1+\alpha\mu_i)^2}$$

$$\frac{\partial^2 \mathcal{L}(\beta,\alpha)}{\partial\alpha^2} = \sum_{i=1}^{n}\left\{-\sum_{j=1}^{y_i}\left(\frac{y_i-j}{\alpha y_i+1-\alpha_j}\right)^2 + \frac{1}{\alpha^2}\left[\frac{2\mu_i+3\alpha\mu_i^2+\alpha^2\mu_i^2 y_i}{(1+\alpha)^2} - \frac{2ln(1+\alpha\mu_i)}{\alpha}\right]\right\}$$

En el caso del modelo de regresión binomial negativo BN2, si designamos por θ el vector que contiene los parámetros $\begin{pmatrix}\beta\\ \alpha\end{pmatrix}$ se tiene que su distribución es asintóticamente normal.

$$\sqrt{n}(\hat{\theta}-\theta) \sim N[0, nI_e^{-1}(\theta)]$$

Para la obtención de los estimadores de máxima verosimilitud siguiendo el método de máxima verosimilitud, haremos uso de dos algoritmos iterativos:

→ **El Algoritmo Newton- Raphson**:

$$\theta_{k+1} = \theta_k + \frac{s(\theta)}{I_{Obs}(\theta_k)}$$

donde $s(\theta) = \dfrac{\partial \mathcal{L}(\theta)}{\partial\theta}$

A través de este algoritmo iterativo podremos calcular de forma simultánea un estimador consistente de máxima verosimilitud de β y α, determinado por $\hat{\theta} = \begin{pmatrix}\hat{\beta}\\ \hat{\alpha}\end{pmatrix}$.

Este algoritmo comienza con un valor inicial θ_0, generando una secuencia $\{\theta_k\}$, $k = 1, 2, ...$ que converge a $\hat{\theta}$ bajo unas condiciones apropiadas.

→ **El Algoritmo Fisher Scoring**:

$$\beta_{k+1} = \beta_k + \frac{s(\beta_k)}{I_e(\beta_k)}$$

El algoritmo Fisher scoring es una simplificación del método de máxima verosimilitud Newton-Raphson. Es el algoritmo tradicionalmente usado para la estimación de modelos lineales generalizados, y permite una estimación consistente unícamente del parámetro β. El parámetro de dispersión se incluye como una constante conocida.

Este algoritmo requiere más iteraciones que el Newton-Raphson, pero los cálculos de I_e son más simples I_{Obs}

Los errores estándar producidos por el método Fisher scoring están basados generalmente en la matriz de información esperada. En el caso de un modelo lineal generalizado con enlace canónico, la matriz de información observada se reduce a la esperada, resultando en un error estándar con el mismo valor al producido por el método Newton-Rapshon.

Por ejemplo el modelo de regresión Poisson con enlace logarítmico, el cual es su enlace canónico, puede ser estimada usando el algoritmo Newton-Raphson empleando la matriz de información observada, o considerada como miembro de la familia de modelos lineales generalizados, usando la matriz de información esperada. En cada uno de los casos el error estándar calculado será identico, excepto para quizás errores muy pequeños. El enlace no canónico para modelos lineales generalizados producirá errores estándar que son diferentes a los producidos usando el algoritmo Newton-Raphson.

3.3. Adecuación del modelo

En este apartado nos centraremos en la adecuación del modelo, concretamente, nos centraremos en la adecuación del modelo tradicional de regresión binomial negativa (BN2). Este modelo destaca porque podremos considerarlo como un modelo lineal generalizado y como un modelo compuesto Poisson-Gamma. El valor del enfoque del modelo binomial negativa como modelo lineal generalizado está en la capacidad de evaluar el modelo usando los diferentes estadísticos de bondad de ajuste y los resíduos asociados a los modelos lineales generalizados señalados en el primer capítulo. Podemos pues adecuar los resultados del primer capítulo para el modelo de regresión binomial negativo BN2, entre los que destacamos:

- El **estadístico desviación**:

$$\begin{aligned} D &= 2\{l(y;y) - l(\hat{\mu};y)\} \\ &= 2\sum_{i=1}^{n}\{y_i ln(y_i/\hat{\mu}_i) - (y_i + 1/\hat{\alpha})/\hat{\alpha} \cdot ln((1+\hat{\alpha}y_i)/(1+\hat{\alpha}\hat{\mu}_i))\} \end{aligned}$$

- El **estadístico Chi-cuadrado de Pearson**:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{(\hat{\mu}_i + \hat{\alpha}\hat{\mu}_i)^2}$$

En cuanto al análisis de residuos del modelo de regresión binomial negativa, se siguen generalmente las variedades como miembros de la familia de modelos lineales generalizados. Entre estas variedades se incluyen el residuo básico, los residuos de Pearson y desviación estandarizados y no estandarizados, así como los estudentadizados de ambos.

- El **residuo básico**:

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

- El **residuo de Pearson**:

$$r_i^p = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\mu}_i + \hat{\alpha}\hat{\mu}_i^2}}$$

siendo su versión estudentadizado:

$$r_i^{pt} = \frac{(y_i - \hat{\mu}_i)}{\hat{\phi}\sqrt{\hat{\mu}_i + \hat{\alpha}\hat{\mu}_i^2}}$$

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

$$r_i^{pst} = \frac{(y_i - \hat{\mu}_i)}{\sqrt{(1-h_i)(\hat{\mu}_i + \hat{\alpha}\hat{\mu}_i^2)}}$$

siendo h_i el i-ésimo elemento de la matriz de proyección.

- El **residuo desviación**:

$$r_i^d = sgn(y_i - \hat{\mu}_i)\sqrt{D}$$

siendo D el estadístico desviación para el MRBN anteriormente señalado. Su versión estudentadizado viene dada por:

$$r_i^{dt} = \frac{sgn(y_i - \hat{\mu}_i)\sqrt{D}}{\hat{\phi}}$$

Y su versión estandarizada:

$$r_i^{dst} = \frac{sgn(y_i - \hat{\mu}_i)\sqrt{D}}{\sqrt{(1-h_i)}}$$

Además, consideraremos el residuo Anscombe, desarrollado para la binomial negativa NB2 [Hilbe, 1994], [Hardin y Hilbe, 2001]. El residuo Anscombe [Ascombe, 1972] toma valores cercanos a los de la desviación estandarizada. Aunque hay veces en la que este no es el caso, y el residuo Ascombe realiza una mejor interpretación que el residuo desviación. El residuo Ascombe trata de normalizar la diferencia entre valores observados y ajustados de manera que la heterogeneidad en los datos llega a ser fácilmente identificable.

Los residuos Anscombe usan la función de varianza del modelo

$$\begin{aligned} Poisson : V(\mu) &= \mu \\ B2 : V(\mu) &= \mu(1 + \alpha\mu) \end{aligned}$$

Este residuo se define de la siguiente forma:

$$r_i^A = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}}$$

donde

$$A(.) = \int d\mu_i / V^{1/3}(\mu_i)$$

El residual Anscombe calculado para un modelo de regresión binomial negativa viene dado por:

$$r^A = \frac{\{3/\hat{\alpha}\{(1+\hat{\alpha}y_i)^{2/3} - (1+\hat{\alpha}\hat{\mu}_i)^{2/3}\} + 3(y_i^{2/3} - \hat{\mu}_i^{2/3})\}}{2(\hat{\alpha}\hat{\mu}_i^2 + \hat{\mu}_i)^{1/6}}$$

3.4. Interpretación del modelo

Una vez se ha obtenido el modelo adecuado de regresión binomial negativa haciendo uso de los criterios antes mencionados de bondad de ajuste, estudio de residuos,... el proceso de modelado se cierra con la interpretación del modelo.

Es importante recordar que la transformación provocada por la aplicación de una función enlace da lugar, en la mayoría de los casos, a una ecuación del modelo expresada en términos multiplicativos, en la que, como señalamos en el primer capítulo, la interpretación de los parámetros se realiza en términos del factor de cambio en el valor esperado para un incremento unitario de las variables explicativas.

Particularmente el modelo tradicional de regresión binomial negativa viene dado por:

$$ln(E[Y]) = \hat{\beta}_0 + X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + ... + X_p\hat{\beta}_p$$

cuya función enlace es la logarítmica. Dependiendo del valor del β estimado, cada variable explicativa puede influir de forma positiva o negativa a la variable de interés.

Al igual que en la regresión logística si una variable explicativa es cualitativa, se deben introducir las variables "*dummys*". En este caso, si la variable explicativa X_k es una variable dummy, entonces se tiene la siguiente relación:

$$\frac{E(Y_i/X_{ik}=1)}{E(Y_i/X_{ik}=0)} = exp(\hat{\beta}_k)$$

la media condicional es $exp(\hat{\beta}_k)$ veces mayor si X_k toma valor 1 en vez de 0.

En el caso de que sea una variable explicativa X_k con $k = 1...p$ una variable continua cada unidad de aumento de esta variable produce un aumento de $\hat{\beta}_j$ del log valor esperado de la variable respuesta. Esta interpretación proporciona una restricción a valores positivos de la variable respuesta, adecuándose al tipo de datos de recuento.

Capítulo 4

Aplicaciones

En este capítulo trataremos dos ejemplos prácticos trabajados en R, teniendo como fin el correcto uso del modelo de regresión binomial negativa (MRBN). En cada uno de ellos veremos como la regresión binomial negativa puede ser usada para modelar una variable respuesta de conteo, caracterizada por la presencia de sobredispersión en los datos.

En el primer caso se ha realizado una simulación de datos que se ajustan a una distribución binomial negativa en la que estudiaremos las distintas mejoras que presenta el modelo MRBN frente al modelo de regresión Poisson (MRP).

Y en el segundo caso aplicamos la regresión binomial negativa a una base de datos dada con el posterior análisis de los resultados obtenidos.

En ambos ejemplos haremos uso de los distintos comandos implementados en R para poder realizar tales estudios, interpretando posteriormente los resultados obtenidos.

4.1. Simulación

En este apartado vamos a realizar una breve simulación a través de R, comenzaremos construyendo un conjunto simulado de datos que satisfaga ciertas especificaciones para demostrar el efecto del ajuste del modelo de regresión binomial negativa. Implementando en R las librerías apropiadas generamos, a través de la función "**nb2_syn**", un conjunto de datos aleatorios con n observaciones de forma que la variables respuesta y explicativas quedan definida de acuerdo al modelo tradicional de regresión binomial negativa con parámetro de dispersión α y tal que los coeficientes de las variables explicativas del modelo sean $(\beta_0, \beta_1, ..., \beta_p)$.

En nuestro estudio particular generaremos $n = 500$ observaciones, 3 variables explicativas, siendo los coeficientes $(\beta_0, \beta_1, \beta_2, \beta_3)$ =(2,0.75,-0.3,2), y una variable respuesta, con parámetro de dispersión $\alpha = 0{,}5$.

Las instrucciones vienen dadas por:

```
> library(COUNT)
> library(MASS)
>
> sim.data <- nb2_syn(nobs = 500, alpha = 0.5,xv=c(2,0.75,-0.3,2))
```

Una vez generado el conjunto de datos, a continuación abordamos su estudio como un MRBN y como un MRP. Notamos que el lenguaje R permite un estudio del MRBN desde los dos enfoques planteados en el tercer capítulo.

Comenzamos analizando los datos como un modelo de regresión binomial negativa enfocado como una Poisson compuesta, usando el método de máxima verosimilitud.

A continuación establecemos las instrucciones y las salidas obtenidas:

```
> mynb <- ml.nb2(nby ~ . , data = sim.data)
> mynb
              Estimate          SE         Z         LCL         UCL
(Intercept)  2.0175351 0.06997534 28.83209  1.8803834  2.1546868
x1           0.7656145 0.06579962 11.63555  0.6366472  0.8945817
x2          -0.2588484 0.07644779 -3.38595 -0.4086861 -0.1090107
x3           1.9541939 0.07956690 24.56039  1.7982428  2.1101451
alpha        1.9298891 0.14179099 13.61080  1.6519788  2.2077995
```

En esta salida se obtiene para cada β_i su estimación puntual, el intervalo de confianza correspondiente a un nivel de confianza de un 95 %, además del error estandar de estimación.

Dado a que generamos los datos para ser modelados usando la regresión binomial negativa, no nos sorprende el buen ajuste del modelo. Notamos que mediante la función "**ml.nb2**", R realiza una estimación de los parámetros del MRBN, tanto de los parámetros β como del parámetro de dispersión α.

Hay que hacer constar que en esta salida el valor 1,929 no es la estimación de α sino de su inversa, siendo su valor 0,5181.

A continuación vamos a analizar el mismo conjunto de datos pero enfocando el modelo de regresión binomial negativa (MRBN) como un miembro de la familia de modelos lineales generalizados. Recogemos las siguientes instrucciones con sus salidas correspondientes:

```
> mynb2 <- glm.nb(nby ~ . , data = sim.data)
> summary(mynb2)

Call:
glm.nb(formula = nby ~ ., data = sim.data, init.theta = 0.5163514325,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4336  -1.1109  -0.5293   0.2135   2.4718
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.02329    0.07011  28.858  < 2e-16 ***
x1           0.76833    0.06582  11.674  < 2e-16 ***
x2          -0.25925    0.07278  -3.562 0.000368 ***
x3           1.95212    0.07960  24.523  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.5164) family taken to be 1)

    Null deviance: 1495.0  on 499  degrees of freedom
Residual deviance:  528.3  on 496  degrees of freedom
AIC: 3189.8

Number of Fisher Scoring iterations: 1
```

mediante la función "**glm.nb**" se obtiene una salida similar a la anterior salvo por el parámetro de dispersión (definido por theta, θ, en este algoritmo), el cual R lo estima previamente mediante el método de máxima verosimilitud para su posterior inclusión en el algoritmo de modelos lineales generalizados (GLM). Estos resultados aparecen compañados de unos estadísticos adicionales como el estadístico desviación, el criterio AIC,... . El estadístico desviación, bajo las hipótesis del modelo correcto, sigue una distribución chi-cuadrado con 496 grados de libertad. Este estadístico tiene un valor de 528.3, que evaluando la siguiente relación para detectar la sobredispersión en los datos, como señalamos en el segundo capítulo

$$\frac{D}{gl} \Rightarrow \frac{528{,}3}{496} = 1{,}065$$

observamos como el modelo acomoda la sobredispersión de los datos, es decir, el modelo se adapta a la heterogeneidad adicional que podría existir en los datos.

A continuación analizamos el conjunto de datos binomiales negativos a través de la regresión Poisson que, como caso particular del GLM, vienen dados a través de la función "**glm**[..., $family =$ "$poisson$", ..]". Obtenemos:

```
> m3 <- glm(nby ~ . , family = "poisson", data = sim.data)
> summary(m3)

Call:
glm(formula = nby ~ ., family = "poisson", data = sim.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-48.946   -3.826   -1.139    0.958   96.395

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.942383   0.015715  123.60   <2e-16 ***
x1           0.845718   0.006690  126.42   <2e-16 ***
x2          -0.631961   0.007666  -82.44   <2e-16 ***
x3           1.943694   0.007969  243.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

    Null deviance: 136900  on 499  degrees of freedom
Residual deviance:  38379  on 496  degrees of freedom
AIC: 39856

Number of Fisher Scoring iterations: 6
```

Entre los resultados obtenidos aparecen las estimaciones de los coeficientes para el MRP. También obtenemos el estadístico desviación que, a diferencia del caso anterior, éste indica que los datos presentan una fuerte sobredispersión, pues en este caso:

$$\frac{D}{gl} \Rightarrow \frac{38379}{496} = 77{,}377$$

Los modelos de regresión Poisson sobredispersos algunas veces conducen a la confusión de que algunas variables explicativas contribuyen significativamente al modelo, cuando no esto no ocurre.

Debemos tener en cuenta que cuando nos referimos de forma general a un conjunto de datos sobredispersos Poisson no tiene por qué ser sinónimo de datos binomiales negativos, pues existen distintos tipos de datos de conteo sobredispersos los cuales no pueden ser explicados mediante un modelo de regresión binomial negativo.

4.2. Aplicación a una base de datos particular.

En el siguiente ejemplo abordaremos de una forma más detallada, a través de R, un modelo de ajuste que se adecue al siguiente caso:

Estudio, realizado por los administradores escolares, del comportamiento de asistencia de los estudiantes de tercero de secundaria. Las variables explicativas, que usaremos para el estudio de la cantidad de días de ausencia, vienen dadas por el tipo de programa en el que está matriculado el estudiante y la nota alcanzada en un examen de matemáticas.

Descripción de los datos

Disponemos de 314 observaciones de estudiantes de secundaria en el archivo "*nb_data.dta*". La variable respuesta de interés es la cantidad de días de ausencia, **daysabs**, que viene dada en función de las variables **math**, que da la puntuación estandarizada en un examen de matemáticas realizado por cada estudiante, y la variable **prog**, que es una variable nominal de tres niveles que indica el tipo de programa de instrucción en el que está matriculado el estudiante, estos son el general, vocacional y académico.

Instrucciones y resumen descriptivo de la base de datos:

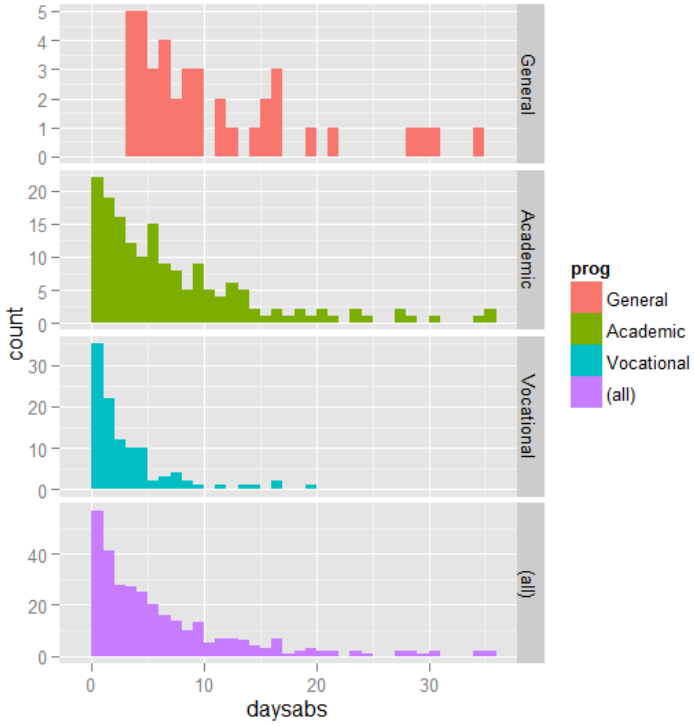
```
> library(MASS)
> library(ggplot2)
> library(foreign)
```

```
> dat <- read.dta("nb_data.dta")
> dat <- within(dat, {
+     prog <- factor(prog, levels = 1:3, labels = c("General", "Academic", "Vocational"))
+     id <- factor(id)
+ })
>
> summary(dat)
       id          gender          math           daysabs              prog
 1001   :  1   female:160   Min.   : 1.00   Min.   : 0.000   General    : 40
 1002   :  1   male  :154   1st Qu.:28.00   1st Qu.: 1.000   Academic   :167
 1003   :  1                Median :48.00   Median : 4.000   Vocational:107
 1004   :  1                Mean   :48.27   Mean   : 5.955
 1005   :  1                3rd Qu.:70.00   3rd Qu.: 8.000
 1006   :  1                Max.   :99.00   Max.   :35.000
 (Other):308
>
> var(dat$daysabs)
[1] 49.51877
>
> mean(dat$daysabs)
[1] 5.955414
```

Observamos como la varianza muestral (49.52) es bastante mayor a la media (5.955), lo que nos hace empezar a cuestionarnos sobre el problema de sobredispersión en los datos.

Realizamos el siguiente gráfico que representa, dependiendo del tipo de programa en el que está matriculado el estudiante, el número de estudiantes que se ausentan una determinada cantidad de días.

```
> ggplot(dat, aes(daysabs, fill = prog)) + geom_histogram(binwidth = 1) + facet_grid(prog
+
+   ., margins = TRUE, scales = "free")
```



La siguiente tabla muestra las medias y varianzas condicionadas de los dias de ausencia por cada tipo de programa. Se aprecia que el valor medio del número de días de ausencia, nuestra variable respuesta, parece variar según la variable **prog**. Y vemos que las varianzas dentro de cada nivel de **prog** son mayores a la media. Esta diferencia sugiere que el exceso de dispersión está presente y que un modelo binomial negativo sería apropiado para su análisis.

```
> with(dat, tapply(daysabs, prog, function(x) {
+
+     sprintf("M (SD) = %1.2f (%1.2f)", mean(x), sd(x))
+
+ }))
               General                 Academic              Vocational
"M (SD) = 10.65 (8.20)"  "M (SD) = 6.93 (7.45)"  "M (SD) = 2.67 (3.73)"
```

Como hemos señalado en el capítulo anterior, podemos hacer uso del modelo de regresión binomial negativo para el estudio de datos de recuento que presentan sobredispersión, cosa que sospechamos que ocurra en estos datos debido a las claras diferencias media-varianza obtenidas.

El análisis de regresión binomial negativo

A continuación se utiliza la función **glm.nb** de la librería **MASS** para estimar los parámetros del modelo de regresión binomial negativa tradicional, es decir, los parámetros del modelo que obtendremos serán los del modelo binomial negativo NB2 que hemos estudiado. Mediante esta función R realiza una estimación previa del parámetro dispersión θ mediante el método de máxima verosimilitud, para posteriormente incluirlo en el algoritmo GLM.

```
> summary(m1 <- glm.nb(daysabs ~ math + prog, data = dat))

Call:
glm.nb(formula = daysabs ~ math + prog, data = dat, init.theta = 1.032713156,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1547  -1.0192  -0.3694   0.2285   2.5273

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.615265   0.197460  13.245  < 2e-16 ***
math           -0.005993   0.002505  -2.392   0.0167 *
progAcademic   -0.440760   0.182610  -2.414   0.0158 *
progVocational -1.278651   0.200720  -6.370 1.89e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.615265   0.197460  13.245  < 2e-16 ***
math           -0.005993   0.002505  -2.392   0.0167 *
progAcademic   -0.440760   0.182610  -2.414   0.0158 *
progVocational -1.278651   0.200720  -6.370 1.89e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En esta salida se recoge la estimación de los coeficientes de regresión para cada una de las variables, junto a los errores estándar y p-valores. A través de los p-valores vemos que todos los β_i son no nulos, es decir, que todas las variables explicativas (**math, progAcademic** y **progVocational**) influyen en la variable objetivo.

A nivel interpretativo, tenemos que un aumento de la variable **math** produce una disminución del valor esperado del número de días de ausencia, de forma análoga ocurre para las variables dummy definidas (**progAcademic** y **progVocational**), que tienen como referencia el tipo de instrucción categórica General.

Nos muestra también la desviación nula calculada a partir del modelo con solo el coeficiente β_0, la desviación del modelo completo, el resíduo desviación, y nos muestra el criterio AIC y la $2 * log$ verosimilitud.

Hay que hacer constar que en esta salida el valor 1,0327 no es la estimación de θ sino de su inversa, siendo su valor 0,968.

Comprobación del modelo seleccionado

Como hemos mencionado anteriormente, los modelos binomiales negativos consideran que la medias condicionadas no son iguales a las varianzas condicionadas. Esta desigualdad media-varianza es capturada por la estimación del parámetro de dispersión. Al contrario de la binomial negativa,el modelo de Poisson el cual si establece la igualdad, conocida como equidispersión. Luego, el modelo Poisson puede relacionarse con el modelo binomial negativo. Podemos usar una prueba de razón de verosimilitud para comparar estos dos modelos y probar el modelo seleccionado.

```
> m3 <- glm(daysabs ~ math + prog, family = "poisson", data = dat)
> pchisq(2 * (logLik(m1) - logLik(m3)), df = 1, lower.tail = FALSE)
'log Lik.' 2.157298e-203 (df=5)
```

El valor del chi-cuadrado asociado es de $2{,}157 \cdot 10^{-203}$, con 5 grados de libertad. Esto sugiere fuertemente que el modelo binomial negativo, con la estimación del parámetro de dispersión,es más apropiado que el modelo de Poisson.

Podemos obtener los intervalos de confianza para los coeficientes a través de la función de verosimilitud.

```
> (est <- cbind(Estimate = coef(m1), confint(m1)))
Waiting for profiling to be done...
                   Estimate       2.5 %       97.5 %
(Intercept)     2.615265446  2.24205576  3.012935926
math           -0.005992988 -0.01090086 -0.001066615
progAcademic   -0.440760012 -0.81006586 -0.092643481
progVocational -1.278650721 -1.68348970 -0.890077623
```

Dado que se trata de un modelo cuya función enlace es la logarítmica, a la hora de estimar los coeficientes, tenemos que exponenciar nuestros coeficientes del modelo. Lo mismo se aplica a los intervalos de confianza.

```
> exp(est)
                  Estimate     2.5 %     97.5 %
(Intercept)    13.6708448 9.4126616 20.3470498
math            0.9940249 0.9891583  0.9989340
progAcademic    0.6435471 0.4448288  0.9115184
progVocational  0.2784127 0.1857247  0.4106239
```

La ecuación de regresión viene dada por:

$$\widehat{daysabs}_i = e^{\beta_0+\beta_1(prog_i=2)+\beta_2(prog_i=3)+\beta_3 math_i} = e^{\beta_0} e^{\beta_1(prog_i=2)} e^{\beta_2(prog_i=3)} e^{\beta_3 math_i}$$

Donde los coeficientes tienen un efecto multiplicativo. El parámetro dispersión no tiene efecto sobre los resultados esperados en la regresión binomial negativa , pero si lo tiene sobre la varianza estimada de los valores esperados.

Los valores pronosticados

Una ayuda para entender mejor el modelo, consiste en mirar los valores esperados para los distintos niveles de las variables explicativas. A continuación creamos nuevos conjuntos de datos con valores de **math** y **prog**, y entonces usaremos el comando **predict** para calcular el número previsto de eventos.

En primer lugar podemos mirar los valores predichos para cada valor de **prog** mientras que tomamos como valor de **math** la media. Para ello, creamos un nuevo conjunto de datos con las combinaciones de **prog** y **math** para la cual nos gustaría encontrar los valores, y usamos entonces el comando **predict**.

```
> newdata1 <- data.frame(math = mean(dat$math), prog = factor(1:3, levels = 1:3,
+     labels = levels(dat$prog)))
> newdata1$phat <- predict(m1, newdata1, type = "response")
> newdata1
```

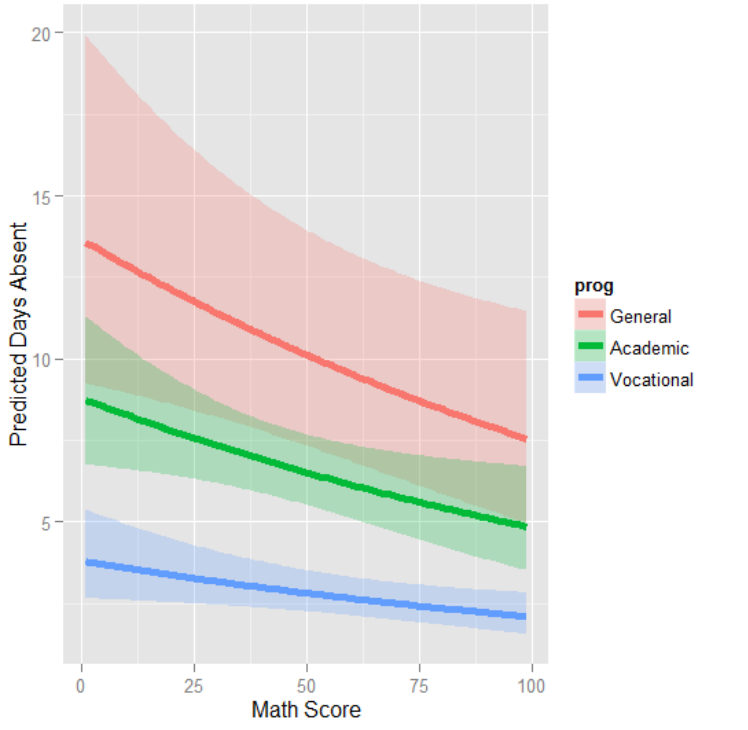
```
      math       prog      phat
1 48.26752    General 10.236899
2 48.26752   Academic  6.587927
3 48.26752 Vocational  2.850083
```

En la salida anterior, vemos que el número predicho de eventos (días de ausencia) para un programa General, es de apróximadamente 10.24, tomando como valor de **math** la media.

Y vemos que el número previsto de eventos para un programa Académico es inferior a 6.59, y el número previsto para un programa Profesional es de 2.85.

A continuación, vamos a obtener el número medio pronosticado de eventos para los valores de **math** a través de toda su área de distribución para cada nivel de **prog**.

```
> newdata2 <- data.frame(
+   math = rep(seq(from = min(dat$math), to = max(dat$math), length.out = 100), 3),
+   prog = factor(rep(1:3, each = 100), levels = 1:3, labels =
+   levels(dat$prog)))
>
> newdata2 <- cbind(newdata2, predict(m1, newdata2, type = "link", se.fit=TRUE))
> newdata2 <- within(newdata2, {
+   DaysAbsent <- exp(fit)
+   LL <- exp(fit - 1.96 * se.fit)
+   UL <- exp(fit + 1.96 * se.fit)
+ })
>
> ggplot(newdata2, aes(math, DaysAbsent)) +
+   geom_ribbon(aes(ymin = LL, ymax = UL, fill = prog), alpha = .25) +
+   geom_line(aes(colour = prog), size = 2) +
+   labs(x = "Math Score", y = "Predicted Days Absent")
```



El gráfico muestra los valores esperados en todo el rango de calificaciones en matemáticas, para cada tipo de programa junto con los intervalos de confianza del 95 %. Tengamos en cuenta que las líneas no son rectas porque es un modelo log lineal, y lo que representa son los valores esperados, no el logaritmo de ellos.

Algunas consideraciones.

- No se recomienda el uso de modelos binomiales negativos para muestras pequeñas ya que puede conducir al sobreajuste del modelo.

- Una causa común de la sobredispersión es el exceso de ceros por un proceso de generación de datos adicional.
- La variable respuesta en una regresión binomial negativa no puede tomar valores negativos.

Bibliografía

[Abdel y Radwan, 2000] Abdel-Aty MA y Radwan AE., *Modeling traffic accident occurrence and involvement*, 2000.

[Ascombe, 1972] Ascombe, F.J., « Contribution to the discusión of H. Hotelling's paper», *Journal of the Royal Statistical Society*, 1972.

[Boswell y Patil, 1970] Boswell, M.T. y Patil, G.P., «Chance mechanisms generating the negative binomial distributions», *Random Counts in Scientific Work* Vol. 1: *Random Counts in Models and Structures*, G.P. Patil (editor), University Park: Pennsylvania State University Press 1970.

[Böhning, 1994] Böhning, D., «A note on a test for Poisson overdispersion», *Biometrika*, 1994.

[Breslow 1984] Breslow, N.E., «Extra-Poisson variation in log-linear models», *Applied Statistics*, 1984.

[Byers, Allore, Gill y Peduzzi 2003] Amy L. Byers, Heather Allore, Thomas M. Gill y Peter N. Peduzzi, «Application of negative binomial modeling for discrete outcomes A case study in aging research», *Journal of Clinical Epidemiology 56*, 2003.

[Cameron y Trivedi, 1986] Cameron, A.C. y Trivedi, P.K., «Econometric models based on count data: comparisons and applications of some estimators and tests», *Journal of Applied Econometrics*, 1986.

[Cameron y Trivedi, 1998] Cameron, A.C. y Trivedi, P.K., «Econometric Society Monographs», *Regression Analysis of Count Data*, New York: Cambridge University Press, 1998.

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

[Finch y Chen, 1999] Finch S.J. y Chen J.B., «Process control procedures to augment quality control of leukocyte-reduced red cell blood products», *Stat Med*, 1999.

[GTM, 1984] Gourieroux, C., Monfort, A. y Trognon, A., «Pseudo maximum likelihood methods: applications to Poisson models», *Econometrica*, 1984.

[Hardin y Hilbe, 2001] Hardin, J. y Hilbe, J., *Generalized Linear Models and Extensions*, College Station, TX: Stata Press, 2001.

[Hauer, 2001] Hauer, E., « Overdispersion in modelling accidents on road sections and in empirical bayes estimation [Versión electrónica].», *Accident Analysis and Prevention*, 2001.

[Hausman, Hall y Griliches, 1984] Hausman, J., B. Hall y Z. Griliches, «Econometric models for count d an application to the patents», *Econometrica*, 1984.

[Hilbe 1993] Hilbe, J.M., Log negative binomial regression as a generalized linear Model, technical Report, Department of Sociology, Arizona State University, 1993.

[Hilbe, 1994] Hilbe, J., « Generalized linear models», *The American Statistician*, 1994.

[Johnson,Kotz y Balakrishnan 1994] Johnson, N.L., Kotz, S. y Balakrishnan, N., *Continuous univariate distributions*, New York: John Wiley, 1994.

[Hilbe, 2007] Joseph M. Hilbe, *Negative Binomial Regression*, Cambridge University Press, 2007.

[King, G., 1989] King, G., «Variance specification in event count models: From restrictive assumptions to a generalized estimator», *American Journal of Political Science*, 1989.

[Lindsey, 1995] Lindsey J.K., *Modelling Frequency and Count Data*, Oxford: Oxford University Press. 1995.

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

[Lindsey, 2001] Lindsey J.K., *Nonlinear Models in Medical Statistics*, Oxford: Oxford University Press, 2001.

[Long, 1997] Long. J.S., *Regression models for categorical and limited dependent variables*, Thousand Oaks, CA: Sage, 1997.

[McCullagh y Nelder, 1989] P. McCullagh y J.A. Nelder, *Generalized Linear Models* London, 1989.

[Meliciani, 2000] Meliciani, V., «The relationship between RyD , investment and patents: a panel data analysis» *Applied Economics* 2000.

[Merklesson y Roth, 2000] Melkersson, M. y D. Roth, «Modeling of household fertility using inflated count data models» *Journal of Population Economics* London, 2000.

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

[Osgood, 2000] Osgood, D.W.,« Poisson-based regression analysis of aggregate crime rates [Versión electrónica]», *Journal of Quantitative Criminology*, 2000.

[Poortema, 1999] Poortema, K., «On modelling overdispersion of counts» *Statistica Neerlandica*, 1999.

[Sormani, Bruzzi y Miller, 1999] Sormani MP, Bruzzi P y Miller DH, «Modelling MRI enhancing lesion counts in multiple sclerosis using a negative binomial model: implications for clinical trials.», 1999.

[Stigler, 1981] Stigler S.M., *Gauss and the invention of least squares*, 1981.

[Sturman, 1999] Sturman, M.C., «Multiple approaches to analysing count data in studies of individual differences: The propensity for Type I errors, illustrated with the case of absenteeism prediction», *Educational and Psychological Measurement, 59*, 1999.

[Vuong, 1989] Vuong, Q.H., «Likelihood ratio tests for model selection and non-nested hypotheses», *Econometrica*, 1989.

[Welsh, Cunningham y Chambers, 2000] Welsh AH, Cunningham RB y Chambers RL., «Methodology for estimating the abundance of rare animals: seabirds nesting on North East Herald Cay», *Biometrics* , 2000.

[Winkelmann, 2000] Winkelmann, R., *Econometric Analysis of Count Data.* Berlín, 2000.

[Winkelmann y Zimmermann, 1991] Winkelmann, R. y Zimmermann, K., «A new approach for modeling economic count data», *Economics Letters*, 1991.