FACULTAD DE MATEMÁTICAS

DEPARTAMENTO DE ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

TRABAJO FIN DE GRADO EN MATEMÁTICAS

MODELOS DE DATOS DE CONTEO PARA EL ESTUDIO DE DATOS DE RNA-SEQ

YOLANDA CÓRDOBA CHAMIZO

24 de junio de 2015

Dirigido por:
Dra. Inmaculada Barranco Chamorro
Dr. Pedro Luis Luque Calvo

Abstract

In this work we will explain the knowledge and techniques which are necessaries to work with RNA-Seq data, a technology used in order to detect and quantify the quantity of DNA of a genome. Firstly, in Chapter 1, we will explain the aforementioned technique and its importance nowadays, emphasizing its application in medicine. We will also compare it with another technology called microarrays. In order to carry out our study, we will need statistical specific models for data count. For this reason, we will explain the Generalized Linear Models (GLM) in Chapter 2, along with other necessary algorithms which estimate the parameters of GLM. Later we will focus on the most useful models for count data. We will obtain their probability distributions, will estimate the parameters and will give the interpretation of them. So we will study the Poisson regression model in Chapter 3, and subsequently, the negative binomial regression model, which is the most appropriate one when we have to deal with RNA-seq data, since the Poisson regression model presents quite often the problem of overdispersion. That's the reason why we focus on the study of the negative binomial regression model in Chapter 4. Finally, in Chapter 5, we will give an application to a real dataset, obtained with RNA-seq which proceed from a biological study whose aim is to find Drosophila gens which are differentially expressed. We will carry out the statistical analysis with R , specifically with the software Bioconductor and the packages DESeq and DESeq2.

Índice general

Capítulo 1

Introducción.

En la actualidad, el análisis estadístico de datos génicos tiene una gran importancia debido a sus múltiples aplicaciones médicas, como por ejemplo la detección prematura de enfermedades. A lo largo de este capítulo se presentan las bases del método de secuenciación de ARN denominado RNA Sequencing (*RNA-seq*), una de las tecnologías que utiliza la ultrasecuenciación para detectar y cuantificar la cantidad de ADN de un genoma. A su vez, explicaremos las diferencias y semejanzas con la técnica de *microarrays*, y se explicarán los pasos previos a seguir con datos obtenidos de *RNA-seq*, en concreto las principales técnicas de normalización de datos.
Podemos encontrar más información sobre esto en Gonzalez (2014).

1.1. Datos de *RNA-seq*.

El término ciencias ómicas es muy amplio y recoge disciplinas como la genómica, la proteómica, la transcriptómica y la metabolómica. Todas ellas se basan en el análisis de un gran volumen de datos, y conllevan, entre otras cosas, la comprensión de ciertos mecanismos moleculares y la identificación de biomarcadores o de factores de exposición a ciertas enfermedades.

Desde el momento en el que dichas ciencias consiguen, mediante diversas técnicas, que las secuencias del genoma comiencen a estar disponibles, se empiezan a considerar nuevos tipos de estudios y surgen preguntas tales como la relación existente entre el perfil de expresión génica y la presencia de diversas patologías.

Una de las primeras técnicas que surgieron para llevar a cabo el estudio del nivel de expresión de todos los genes de forma simultánea, es la llamada técnica de *microarrays*.
Un *DNA microarray* consiste en una superficie sólida, a la que se le une una colección de fragmentos de ADN, y en la cual el nivel de expresión de cada gen se indica generalmente mediante fluorescencia y a través de un análisis de imagen.

A continuación mostramos una imagen de un análisis de este tipo.

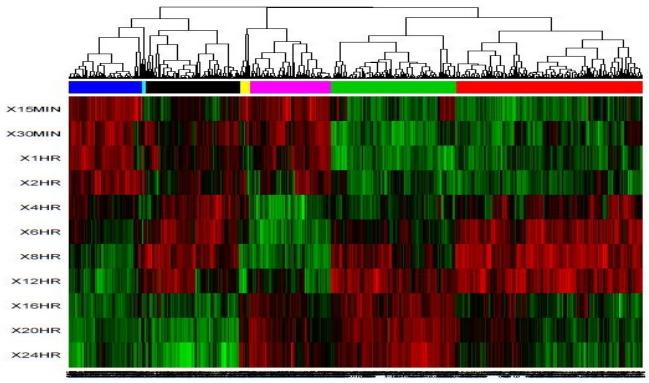

Figura 1.1: DNA microarray.

Puesto que esta técnica mostraba ciertas limitaciones, surgieron las denominadas *HTS* (*high-throughput sequencing* - técnicas de secuenciación de última generación), que son una contrapartida a la tecnología de microarrays para el estudio del nivel de expresión de genes.

Encontramos de esta forma que las técnicas de *RNA-seq*, son una aproximación reciente para realizar el análisis de perfiles de expresión utilizando tecnología HTS. Se trata de una tecnología que trabaja secuenciando cada una de las moléculas de *RNA* y que obtiene un perfil de expresión de un gen concreto. Más específicamente, fragmenta el ADN en pequeños trozos llamados *reads* o lecturas, los cuales se secuencian y se alinean frente a un genoma de referencia estándar. Para el estudio de la expresión génica, se alinean contra genes, de forma que si el número de *reads* alineados contra un determinado gen es mayor que contra otro, diremos que este gen muestra un nivel de expresión mayor.

Un esquema muy simplificado del procedimiento en RNA-seq es el siguiente:

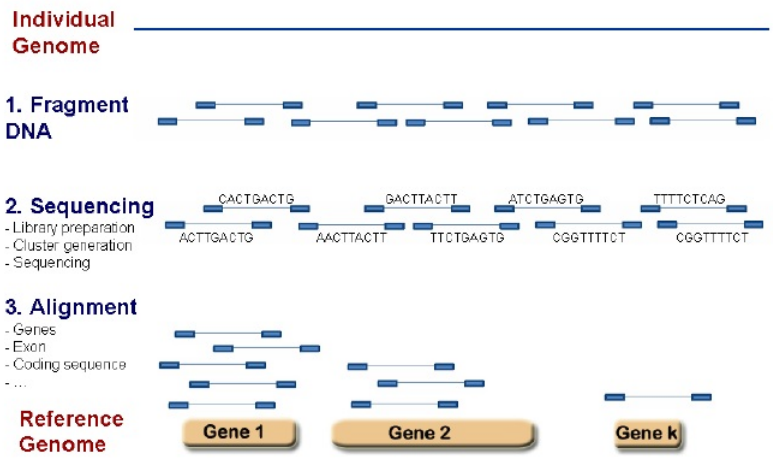

Figura 1.2: RNA-seq

Los resúmenes numéricos que obtenemos como datos al aplicar RNA-seq son datos de conteo. Para su estudio, dichos resultados se resumen en tablas. En estas tablas las columnas las constituyen dos grupos formados cada uno de ellos por un número determinado de individuos. Un ejemplo podría ser: Grupo A - individuos que padecen cierta enfermedad, Grupo B - individuos sanos o grupo control. Por otra parte, cada fila corresponde a un determinado gen. De esta forma se muestra cuantos conteos tenemos de cada gen en los distintos individuos. Estas tablas de conteo son semejantes a la que mostramos a continuación:

	$I_1^A, ..., I_{n_A}^A$	$I_1^B, ..., I_{n_B}^B$
Gen 1		
.		
.		
.		
Gen G		

Una vez recogidos los datos se pueden realizar diferentes estudios. Por ejemplo nos puede interesar fijar un determinado gen y hallar la media de conteos en cada uno de los grupos. Nuestro objetivo será comparar las medias. Llamemos X_g a la variable aleatoria que nos da el número de conteos en el gen g, y sean μ_A y μ_B las medias correspondientes de esta variable aleatoria en el grupo A y en el grupo B, respectivamente. En este caso estaríamos tratando con un problema de comparación de medias de dos grupos que son independientes. Los modelos estadísticos que utilizaremos serán modelos para datos de conteo, pero esto pertenece al estudio estadístico que se tratará en capítulos posteriores (Capítulos 2, 3 y 4). Otra cuestión añadida será que estaremos trabajando con miles de genes a la vez, por lo que algunas consideraciones metodológicas para tratar este problema se señalarán en el Capítulo 5.

1.2. Diferencias de los datos de *RNA-seq* con los datos obtenidos en el análisis de microarrays.

Ahora bien, como mencionamos al principio del capítulo, el análisis de *microarrays* también tiene como objetivo detectar genes que presenten distinto comportamiento en dos grupos de individuos, es decir, se encarga de la búsqueda de genes que presenten distintos niveles de expresión según el grupo en el que se esté estudiando. Pero a pesar de tener un objetivo común, ambas técnicas presentan varias diferencias, y en consecuencia, podemos observar ciertas ventajas de una técnica sobre la otra.

- En *RNA-seq*, los niveles de expresión de cada gen vendrán dados según el valor de conteo, mientras que en el análisis de *microarrays*, como se dijo anteriormente, mayores niveles de fluorescencia son los que nos indican que ese gen tiene un mayor nivel de expresión. Es decir, la principal diferencia entre ambos estudios es que mientras que en *RNA-seq* trabajamos con datos discretos, los niveles de fluorescencia en los *microarrays* son datos continuos.

- Otra diferencia importante es que al utilizar la técnica que da lugar a los datos de *RNA-seq* desaparece uno de los problemas que nos encontramos al trabajar con *microarrays*, y es que en determinados momentos los niveles de fluorescencia se saturan. Además los experimentos de *RNA-seq* se pueden ir actualizando a medida que se va obteniendo información nueva de la secuencia, mientras que los *microarrays* se limitan a la información de referencia de la que se dispone durante la producción.
- En *microarrays* tenemos una lista de genes candidatos para los que obtenemos su nivel de expresión. En cambio en *RNA-seq* obtenemos una lista de genes, y estudiamos las diferencias de nivel de expresión a partir de los datos.
- Una ventaja que presenta *RNA-seq* es que problemas de hibridación que se presentaban en *microarrays*, tales como la hibridación cruzada, entre otros, se pueden eliminar.
- *RNA-seq* tiene la capacidad de cuantificar un gran rango dinámico de los niveles de expresión, de hecho, no hay límite superior para dicha cuantificación y esta puede realizarse incluso con organismos que carecen de un genoma de referencia.
- En cuanto al análisis estadístico de los resultados, que es la parte que aquí más nos interesa, como ya dijimos, una diferencia obvia es que en una técnica se trabaja con datos discretos y en la otra con datos continuos (los niveles de fluorescencia), por lo que otra diferencia derivada de esto es que en *RNA-seq* trabajaremos con gráficos de barras de conteos, por ejemplo, un gráfico que represente el porcentaje de individuos con 0 conteos, 1 conteo, etc. Mientras que en los *microarrays* no podemos utilizar los mismo gráficos, en este caso podemos ayudarnos por ejemplo de un histograma suavizado que nos pueda aportar una idea de la densidad de los datos de los que disponemos.

Vemos a continuación una gráfica de cada una de las técnicas:

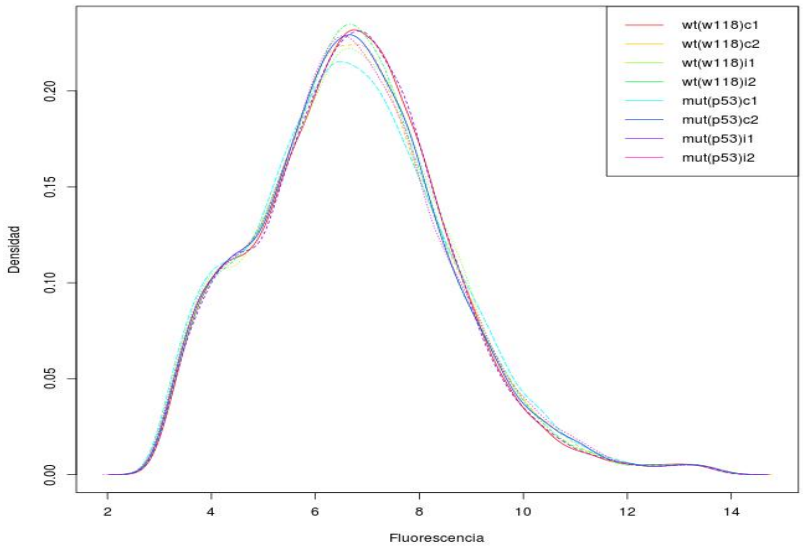

Figura 1.3: Histograma suavizado en datos de *microarrays*.

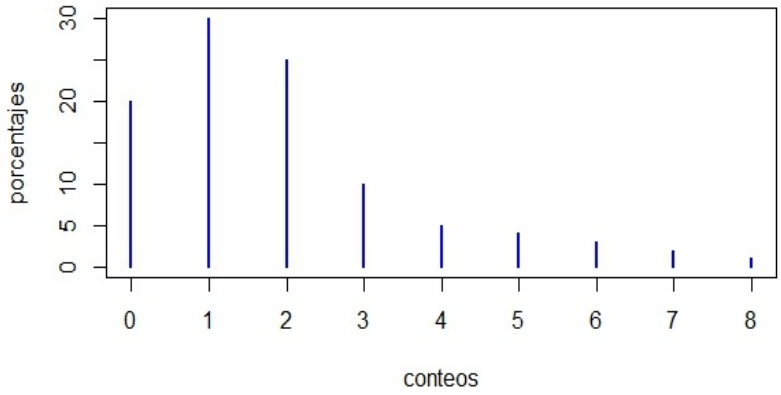

Figura 1.4: Gráfico de conteos en datos de *RNA-seq*.

1.3. Consideraciones previas al análisis de datos.

A la hora de estudiar los datos debemos de tener en cuenta 4 puntos importantes.

- En primer lugar los datos obtenidos en un experimento de *RNA-seq*, nos da cierto número de lecturas, los *reads*, cuya agrupación nos proporciona las llamadas librerías de secuenciación.
  Dichas librerías no sólo recogen las lecturas sino información particular de éstas. Por ejemplo, a parte de la expresión diferencial de los genes, puede aparecer información como la anotación de genes novedosos o la identificación de RNAs no codificantes. El número total de lecturas mapeadas en el genoma determina la profundidad de secuenciación del ensayo, determinando a su vez el tamaño de la librería.

- En determinadas ocasiones, el hecho de alcanzar una profundidad de secuenciación efectiva puede ser obstaculizado por la cantidad limitante de RNA. Por esta razón se suelen amplificar las librerías mediante diversas técnicas, aunque actualmente sigue siendo objeto de debate cuál es la mejor manera de obtener suficiente RNA a partir de un número limitado de muestras.

- El tercer factor a tener en cuenta es la longitud de un gen, dado por su número de bases (recordamos que un gen es una secuencia de nucleótidos y a cada nucleótido le corresponde una base).
  Es obvio que aquellos genes con longitudes reducidas pueden mostrar debido a ello un número de lecturas menor y parecer que poseen un bajo nivel de expresión, por el contrario genes con una longitud grande mostrarán mayor número de lecturas.

- La última observación que haremos antes de comenzar con el análisis de los datos hace referencia a los tipos de réplicas con las que trabajar. Podemos encontrarnos con dos tipos:

Réplicas biológicas: Son las que necesitamos para poder hacer inferencia estadística, es decir, para poder generalizar las conclusiones obtenidas del estudio de una muestra a la población de la que procede.

Réplicas técnicas: Son medidas repetidas. A la hora de trabajar con ellas, nos quedamos con la media. Por ejemplo si tenemos 3 medidas de un individuo, tomadas con 3 técnicas diferentes, trabajaremos con la media de los 3 casos.
De este tipo de réplicas no se pueden extraer conclusiones generales, sólo son válidas para muestras que puedan ser directamente comparables. Son útiles para reducir la variabilidad y mejorar la potencia estadística.

Veamos en qué casos seleccionamos un tipo de réplica u otra para entender mejor el concepto de ambas.
Si por ejemplo se quiere estudiar una enfermedad rara, nos interesa coger 3 réplicas técnicas para cada individuo que tengamos en el estudio. Por el contrario, supongamos que buscamos genes diferencialmente expresados en individuos enfermos de asma. En este estudio queremos extraer conclusiones generales que sean válidas para la población de enfermos de asma. Sería conveniente tener muchas réplicas biológicas, o lo que es equivalente, muestras de muchos individuos.

1.4. Análisis de los datos.

Para el análisis de los datos de *RNA-seq* tenemos dos posibilidades:

1. **Transformación de los datos de conteo en datos continuos.** De esta forma podemos aplicar los métodos de análisis clásicos empleados para el caso de *microarrays*.
   Para llevar a cabo este procedimiento seguimos los siguientes pasos:
   - Tomar logaritmos de los datos. Este primer paso presenta el problema de qué hacer cuando tenemos entre nuestros resultados datos iguales a 0.
   - Aplicar una transformación que estabilice la varianza. Una posible transformación es la del arco-seno. Sea x el número observado de conteos en un gen particular, y n el número de lecturas en la muestra, tenemos x' los datos transformados de la siguiente manera:

   $$x' = \sqrt{n}\ \ arcsin\left(\sqrt{\frac{x}{n}}\right)$$

   Nota: Después de aplicar la transformación tendríamos que hacer también el preprocesamiento de los datos continuos con los algoritmos de microarrays clásicos para eliminar el ruido de fondo y normalizar.

2. **Utilizar modelos estadísticos específicos para datos de conteo.** Es la opción con la que trabajaremos a partir de ahora, pues es la más adecuada metodológicamente. Como modelos para datos de conteo utilizaremos los Modelos Lineales Generalizados que se explicarán en el siguiente capítulo.

Para poder aplicar métodos estadísticos, lo primero que debemos hacer, una vez creadas las tablas de conteo, es la normalización.

1.4.1. Normalización.

Como ya se ha indicado anteriormente, los datos que obtenemos en un experimento de *RNA-seq* son los números de fragmentos secuenciados que se mapean frente a la unidad de ARN de interés, generalmente un gen. A estos datos se les denomina "conteos brutos". Pero con las tablas de conteos brutos no podemos sacar conclusiones fiables, ni realizar comparaciones entre grupos de muestras, ya que debido a la gran complejidad de la técnica utilizada en los datos de secuenciación se puede presentar una gran diferencia en el número total de lecturas obtenidas en cada experimento. La normalización de los datos pretende resolver este problema.
Como ejemplo, un factor a tener en cuenta para comparar el número de conteos entre genes, es la longitud del gen, como ya se explicó en 1.3. Otros factores son la profundidad de secunciación, la longitud del tránscrito, el nivel de expresión del *mRNA* (RNA mensajero) y el contenido de secuencias GC (puesto que estas secuencias son raras de obtener). El objetivo principal de la normalización es eliminar o disminuir de alguna forma los efectos que todos esos factores puedan tener sobre los datos, y así intentar que tengan una influencia mínima en los resultados obtenidos.

Existen diversos métodos para normalizar los datos. Veremos a continuación algunos de los más importantes.

- Uno de los procedimientos de normalización, que ha sido y sigue siendo muy utilizado a pesar de ser considerado no sofisticado, es el de dividir los conteos que tenemos para cada gen en una muestra por el número total de lecturas en dicha muestra.

- El segundo procedimiento con el que nos encontramos, también un procedimiento no sofisticado, es el método **RPKM** (Reads Per Kilobase of transcript and Million mapped reads - reads esperadas por kilobase de transcrito por millón de lecturas mapeadas). Este método normaliza por la longitud de las regiones en cuestión, por ejemplo, por la longitud de los genes.

$$RPKM = \frac{\frac{número\ \ de\ \ lecturas\ \ en\ \ la\ \ región}{amplitud\ \ de\ \ la\ \ región\ \ \times 10^3}}{número\ \ total\ \ de\ \ lecturas\ \ \times 10^6}$$

  Podemos encontrar más información sobre este método en Mortazavi et al. (2008).

- El método **TMM** (Trimmed Mean of M-values - media truncada de M-valores), fue propuesto por Robinson and Oshlack (2010).Estos autores indican que:

  " La proporción de lecturas atribuidas a un gen dependen de las propiedades de expresión de la muestra completa y no sólo del nivel de expresión de ese gen".

  A continuación explicamos brevemente la base de este método introduciendo alguna notación:

  1. Sea Y_{gk} el número de conteos observado para el gen g en la librería k.

2. Sea μ_{gk} el nivel de expresión del gen g en la librería k. Será un valor real y desconocido, y vendrá dado en términos del número de transcripciones.
3. Sea L_g la longitud del gen g.
4. Sea N_k el número total de lecturas en la librería k.

El valor esperado de Y_{gk} es

$$E[Y_{gk}] = \frac{\mu_{gk} L_g}{S_k} N_k$$

siendo

$$S_k = \sum_{g=1}^{G} \mu_{gk} L_g,$$

donde S_k representa la salida de RNA total de la muestra. La dificultad que presenta este método es que el valor S_k, que se estima a partir de la muestra, es desconocido y puede variar drásticamente de una muestra a otra, dependiendo de la composición de RNA.
Basándose en este modelo se calculan medias truncadas de medidas de la abundancia relativa de cada gen entre dos muestras. Pueden verse más detalles en Robinson and Oshlack (2010).

- Los métodos **CQN** (normalización de cuantil condicional - *Conditional Quantile Normalization*) y **EDAseq** se emplean sobre todo para la normalización de los datos en función del contenido de cadenas GC, aunque también corrige para el tamaño de la librería, la longitud del gen, etc. Para poder aplicar estos métodos se considera el siguiente modelo:

$$Y_{g,i} | \mu_{g,i} \backsim Poisson(\mu_{g,i}),$$

$$\mu_{g,i} = exp\left\{ h_i(\theta_{g,i}) + \sum_{j=1}^{p} f_{i,j}(X_{g,j}) + log(m_i) \right\}$$

donde

1. $Y_{i,j}$ denota los conteos observados para el gen i en la librería j.
2. $X_g = (X_{g,1}, ..., X_{g,p})$ son covariables tales como el contenido de GC o la longitud del gen.
3. h_i anota la variabilidad técnica (sesgo diferencial).
4. $f_{i,j}$ son funciones que dependen de covariables, y tienen en cuenta sesgos sistemáticos en las muestras, como por ejemplo, que haya un *"bach effect"* (efecto lote).
5. m_i es la profundidad de secuenciación para cada muestra.

Detalles adicionales pueden verse en Risso et al. (2011).

Puede decirse que la normalización es la primera fase del análisis estadístico de unos datos de *RNA-seq*. Dependiendo del software utilizado será necesario realizarla antes de aplicar el método o lo realizará internamente el programa (como veremos que ocurre con DESeq en el Capítulo 5).

Una vez vistos los principales métodos de normalización, el siguiente paso consiste en estudiar los modelos lineales generalizados más usados para el estudio de datos de *RNA-seq*.

Capítulo 2

Modelos Lineales Generalizados.

Como se ha indicado en el capítulo anterior, para poder realizar un estudio estadístico de datos procedentes de *RNA-seq* lo más conveniente es utilizar modelos estadísticos específicos para datos de conteo. En este capítulo presentamos los Modelos Lineales Generalizados como el marco teórico apropiado para realizar este estudio. En estos modelos se propone una distribución de la familia exponencial, destacando el papel de la función vínculo. Los parámetros se estiman por máxima verosimilitud. Tendremos que las ecuaciones de verosimilitud son no lineales por lo que técnicas numéricas del tipo Newton-Raphson han de utilizarse para obtener las estimaciones de los parámetros y los errores estándar de los estimadores. Finalmente, se incluyen medidas de ajuste del modelo, en particular, la función *deviance* y el estadístico $\mathcal{X}^2$ de Pearson generalizado.
A parte de la información aquí recogida, podemos encontrar un desarrollo más amplio de los Modelos Lineales Generalizados en Hardin and Hilbe (2007) y Hilbe (2014)

2.1. Introducción.

Un Modelo Lineal Generalizado o *GLM* (Generalized Linear Model) es una generalización de los modelos de regresión lineal estándar. Puesto que estos últimos se basan en una serie de supuestos, vamos a recordar algunos de ellos antes de profundizar en los *GLMs*.

Modelo de regresión lineal estándar.

- Cada observación de la variable respuesta puede modelizarse por una distribución normal o Gaussiana, lo cuál denotamos como $Y_i \backsim N(\mu_i, \sigma_i^2)$.
- Las distribuciones para todas las observaciones tienen una varianza común, es decir, $\sigma_i^2 = \sigma^2$ para cualquier i.
- Existe una relación directa entre el predictor lineal $\mathbf{X}\boldsymbol{\beta}$ (combinación lineal de los valores de las covariables y los parámetros asociados) y el valor esperado del modelo, de tal manera que

$$E[Y] = \mu$$

$$\mu = \mathbf{X}\boldsymbol{\beta}$$

siendo $\mathbf{X}$ la matriz que recoge las covariables relacionadas con la variable respuesta Y, y $\boldsymbol{\beta}$ el vector de parámetros desconocidos del modelo.

Al igual que en los modelos de regresión lineal estándar, el propósito de *GLMs* es encontrar la relación entre la variable respuesta observada y cierto número de covariables. La aparición de *GLMs* surge tras la idea de que mediante la reestructuración de la relación entre el predictor lineal y el ajuste, podemos linealizar relaciones que inicialmente parecen ser no lineales y por lo tanto no válidas para los modelos estándar.

Considerando que los modelos lineales describen el resultado de una variable respuesta Y como la suma de su media y una variable aleatoria ϵ, Nelder and Wedderburn (1972) linealizaron cada miembro de la familia *GLM* por medio de una función link. Más tarde modificaron el método de mínimos cuadrados ponderados, el cuál se usaba en los modelos de regresión estándar, y proporcionaron el método de mínimos cuadrados ponderados iterativos. Aparte de la introducción de la función link, también introdujeron la función varianza como un elemento en la ponderación de la regresión. Las iteraciones de los algoritmos implementados actualizan las estimaciones de los parámetros para producir predictores lineales apropiados, valores ajustados, y errores estándar de los estimadores.

A pesar del hecho de que se piensa que las raíces históricas de los *GLMs* están basadas en la metodología IRLS (Iterative Recursive Least Squares), muchas generalizaciones de los modelos lineales requieren técnicas numéricas del tipo Newton-Raphson comunes a metodologías basadas en la función de verosimilitud. Veremos todos estos algoritmos con mayor detalle en la Sección 2.4.

Veamos a continuación las componentes de un *GLM*:

- La primera de ellas es la componente aleatoria, que identifica la variable respuesta Y. Para esta variable se supone una distribución perteneciente a la familia exponencial.
- Otra componente es la sistemática, que especifica las variables explicativas que se utilizan en la función predictora lineal. Recordamos que el predictor lineal es una combinación lineal de k variables explicativas de la forma $\eta = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$, con $\beta_i \in \mathbb{R}$ y x_i variables explicativas.
- Y la tercera y última componente es la función link, una función del valor esperado de la variable aleatoria Y, que se expresará como una combinación lineal de las variables predictoras. Es decir, la función link relaciona las componentes sistemática y aleatoria.
  Suponemos además que existe la inversa de esta función, que relaciona la respuesta media esperada con el predictor lineal de tal manera que $\mu = g^{-1}(\eta)$.

Además de las componentes, es importante mencionar que en *GLMs* la varianza debe cambiar con las covariables sólo como función de la media.

En resumen, la clase de *GLMs* extiende los modelos lineales tradicionales de tal manera que un predictor lineal se asigna a través de una función de enlace para modelizar la media de una respuesta caracterizada por cualquier miembro de la familia exponencial de distribuciones. El modelo lineal tradicional no es apropiado cuando no podemos suponer que los datos siguen una distribución normal o si la variable respuesta tiene un conjunto de resultados limitado. Además, en otros casos en los que la homocedasticidad es un requisito que no se puede mantener, los modelos lineales vuelven a ser de nuevo inapropiados. Los *GLMs* permiten tratar estas nuevas situaciones.

En *RNA-seq* los modelos de *regresión de Poisson* y *regresión binomial negativa* son los *GLMs* más usados, por lo que vamos a estudiarlos con mayor profundidad. Antes de ello, y debido a que en un *GLM* la variable dependiente está generada por una función de distribución de la familia exponencial, recordamos en primer lugar los conceptos básicos de dicha familia.

2.2. Familia exponencial.

Definición 2.1 *Una familia uniparamétrica de distribuciones, con espacio paramétrico* Θ *un intervalo de* $\mathbb{R}$*, y en la que la función de densidad, (o de probabilidad), puede expresarse de la forma*

$$f(y;\theta,\phi) = e^{\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)} \tag{2.1}$$

con a(), b(), c() funciones arbitrarias de sus argumentos, y soporte $\{y\text{: } f(y;\theta,\phi) > 0\}$ *independiente de parámetros desconocidos, se denomina familia exponencial.*

En 2.1

- θ es el parámetro natural. (Posteriormente dará lugar a la *función link* o *función vínculo*).
- $b(\theta)$ es la función cumulante. A partir de ella obtendremos la media y la varianza del modelo.
- ϕ es el parámetro de dispersión (que suponemos conocido y que es necesario para que en *GLM* tengamos errores estandar con distribución en la familia exponencial).

Observación: Para facilitar cálculos posteriores, en muchas ocasiones vamos a tomar el logaritmo de la función de densidad o de probabilidad, según nos encontremos con una distribución continua o discreta respectivamente, trabajando con la expresión

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

que resulta más sencilla para identificar las componentes del *GLM*.

A la familia exponencial de distribuciones pertenecen tanto modelos discretos como continuos. Como ejemplos podemos citar la distribución binomial, Poisson, binomial negativa, normal, gamma, e inversa gaussiana.

Función de verosimilitud

Debido a que consideramos observaciones independientes, la función de densidad(o probabilidad) conjunta de la muestra completa de observaciones y_i, dados los parámetros θ y ϕ, se define como el producto de la función de densidad sobre las observaciones individuales. Es decir,

$$f(y_1, y_2, ..., y_n; \theta, \phi) = \prod_{i=1}^{n} e^{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi)}$$

Esta función puede expresarse como una función de θ y ϕ dadas las observaciones y_i. En ese caso la llamamos función de verosimilitud y viene dada por la siguiente expresión:

$$L(\theta, \phi) = L(\theta, \phi; y_1, y_2, ..., y_n) = \prod_{i=1}^{n} e^{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i;\phi)} \tag{2.2}$$

Más adelante desearemos obtener estimaciones de (θ, ϕ), y estas vendrán dadas como los valores en que se alcanza el máximo de 2.2. Para maximizar la función de verosimilitud, nos resulta más fácil trabajar con la función log-verosimilitud puesto que los resultados que la maximizan son los mismos en ambas funciones.

$$\mathcal{L}(\theta, \phi) = lnL(\theta, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi)$$

La función de log-verosimilitud para la familia exponencial vemos que tiene una forma relativamente básica, lo que permite cálculos simples de primera y segunda derivada para obtener las estimaciones de máxima verosimilitud.

Proposición 2.1 *En la familia exponencial definida en 2.1, se tiene que la primera y la segunda derivada de la función cumulante con respecto de θ nos proporciona, respectivamente, las funciones media y varianza. Es decir*

$$b'(\theta) = E[Y]$$

$$b''(\theta) = var[Y]$$

2.3. Esquema general en GLM.

Explicamos los pasos que seguiremos en los capítulos posteriores para la obtención de resultados en los diversos modelos. Para ello consideramos Y la variable aleatoria de la que partimos y que queremos explicar.

- Denotamos por $\mu = \mathrm{E}[Y]$.

- Hallamos la función vínculo. Esta expresa el parámetro natural, θ, como una función de la media: $\theta = \theta(\mu)$. A su vez $\theta(\mu)$ se expresará como un modelo de regresión lineal:

$$\theta(\mu) = \beta_0 + \beta_1 \, X_1 + ... + \beta_k \, X_k \, .$$

- Hallamos la inversa de la función vínculo, que expresa la media como función del parámetro natural. Esta se denomina función respuesta. Nos permite predecir valores de μ a partir de las variables explicativas.

2.4. Algoritmos de estimación de parámetros.

Los parámetros del modelo se estiman por máxima verosimilitud.
Comenzamos considerando la función de log-verosimilitud para la familia exponencial.

$$\mathcal{L}(\theta, \phi; y_1, y_2, ..., y_n) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i; \phi)$$

Los elementos de esta función, θ, $b(\theta)$, ϕ, a() y c(), son los términos ya definidos en la Sección 2.2.

Recordamos que:

- **Término offset:** Estamos parametrizando una función de la media μ, en términos de covariables conocidas $\mathbf{X}$ con coeficientes asociados desconocidos $\boldsymbol{\beta}$. En la función vínculo se puede incluir un término offset. En este caso tendremos una expresión de la forma

$$\theta(\mu) = \mathbf{X}\boldsymbol{\beta} + \textit{offset}.$$

  Antes de continuar aclaramos el significado del término *offset*.
  Un *offset* es una componente dada en un problema de estimación. Su idea es simple y para entenderla veamos el siguiente ejemplo. Especificamos que $\boldsymbol{\theta}$ es una función de las covariables especificadas, $\mathbf{X}$, y sus coeficientes asociados, $\boldsymbol{\beta}$. Dentro de la combinación lineal de las covariables y sus coeficientes $\boldsymbol{X\beta}$, podemos desear la restricción de un subconjunto particular de los coeficientes β_i a valores particulares. Supongamos que queremos incluir que $\beta_3 = 2$ en un modelo con una constante, X_0, y tres covariables X_1, X_2 y X_3. Entonces el predictor lineal viene dado por

$$\eta = \widehat{\beta_0} + \widehat{\beta_1}X_1 + \widehat{\beta_2}X_2 + 2X_3$$

  en cada paso. Así el predictor lineal se compone de una combinación lineal de los parámetros libres, más dos veces la covariable X_3. Podríamos generar una nueva variable igual a dos veces la variable que contiene las observaciones de X_3 y especificar esta nueva variable como el *offset*. De esta forma escribimos el predictor lineal como

$$\eta = \mathbf{X}\boldsymbol{\beta} + \textit{offset}.$$

- **Base para el cálculo de los estimadores de máxima verosimilitud de los parámetros**: Puesto que $\eta = \sum_{j=1}^{k} x_j\ \beta_j + \textit{offset}$, tenemos la igualdad

$$\frac{\partial \eta}{\partial \beta_j} = x_j.$$

Por propiedades de la familia exponencial y teniendo en cuenta los dos puntos anteriores, se tiene que

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{a(\phi)V(\mu_i)} \left(\frac{\partial \mu}{\partial \eta}\right)_i x_{ji}$$

donde i=1,...,n son los índices de las observaciones y x_{ji} es la observación i-ésima de la covariable X_j, con j=1,...,k.

Para hallar las estimaciones $\widehat{\boldsymbol{\beta}}$ en las ecuaciones anteriores, se pueden utilizar algoritmos numéricos basados en el método de Newton-Raphson cuya base teórica exponemos a continuación.

2.4.1. Método de Newton-Raphson

Este método es una aproximación lineal de la serie de Taylor, donde desarrollamos la derivada de la función de log-verosimilitud, (tambien llamado el gradiente o la ecuación de estimación) en una serie de Taylor. Para los cálculos siguientes usaremos $\mathcal{L}' = \partial \mathcal{L}/\partial \boldsymbol{\beta}$ para el gradiente y $\mathcal{L}'' = \partial^2 \mathcal{L}/(\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T)$.
Se desea resolver

$$\mathcal{L}'(\boldsymbol{\beta}) = 0.$$

El resultado de esta ecuación nos proporciona las estimaciones de $\boldsymbol{\beta}$. Por esta razón la llamamos ecuación de estimación.

El desarrollo en serie de Taylor de $\mathcal{L}'(\boldsymbol{\beta})$ en $\boldsymbol{\beta}^{(0)}$ es

$$0 = \mathcal{L}'(\boldsymbol{\beta}^{(0)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})\mathcal{L}''(\boldsymbol{\beta}^{(0)}) + \frac{(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^2}{2!}\mathcal{L}'''(\boldsymbol{\beta}^{(0)}) + ...$$

Para reducir la ecuación de estimación a una ecuación lineal, nos quedamos sólo con los dos primeros términos, es decir, nos quedamos con la ecuación

$$0 = \mathcal{L}'(\boldsymbol{\beta}^{(0)}) + (\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})\mathcal{L}''(\boldsymbol{\beta}^{(0)}).$$

de tal manera que la podemos reescribir resolviendo para $\boldsymbol{\beta}$ como se muestra a continuación:

$$\boldsymbol{\beta} \approx \boldsymbol{\beta}^{(0)} - \left\{\mathcal{L}''(\boldsymbol{\beta}^{(0)})\right\}^{-1} \mathcal{L}'(\boldsymbol{\beta}^{(0)}).$$

Debemos iterar esta estimación usando

$$\boldsymbol{\beta}^{(r)} \approx \boldsymbol{\beta}^{(r-1)} - \left\{\mathcal{L}''(\boldsymbol{\beta}^{(r-1)})\right\}^{-1} \mathcal{L}'(\boldsymbol{\beta}^{(r-1)})$$

para r = 1,2,... y un vector razonable de vectores iniciales $\boldsymbol{\beta}^{(0)}$.

Observación: Esta aproximación de serie de Taylor linealizada es exacta si la función de verosimilitud es realmente cuadrática.

Podemos ver que la matriz de las segundas derivadas de la log-verosimilitud (primera derivada de la ecuación de estimación) es necesaria para que podamos obtener una estimación actualizada del vector de parámetros $\boldsymbol{\beta}$. Esta matriz es, numéricamente hablando, difícil de obtener, por lo que una aproximación consiste en tomar las derivadas segundas a partir de una especificación analítica o a partir de una aproximación numérica usando la ecuación de estimación. A la matriz de derivadas segundas se le llama matriz Hessiana, y viene dada por

$$\left(\frac{\partial^2 \mathcal{L}}{\partial\beta_j \partial\beta_k}\right) = \sum_{i=1}^{n} \frac{1}{a(\phi)} \left(\frac{\partial}{\partial\beta_k}\right) \left\{ \frac{y_i - \mu_i}{V(\mu_i)} \left(\frac{\partial\mu}{\partial\eta}\right)_i x_{ji} \right\}$$

$$= -\sum_{i=1}^{n} \frac{1}{a(\phi)} \left[\frac{1}{V(\mu_i)} \left(\frac{\partial\mu}{\partial\eta}\right)_i^2 - (\mu_i - y_i) \left\{ \frac{1}{V(\mu_i)^2} \left(\frac{\partial\mu}{\partial\eta}\right)_i^2 \frac{\partial V(\mu_i)}{\partial\mu} - \frac{1}{V(\mu_i)} \left(\frac{\partial^2\mu}{\partial\eta^2}\right)_i \right\} \right] x_{ji} x_{ki}$$

La matriz Hessiana es una matriz cuadrada de dimensiones $(k+1)$x$(k+1)$. Hallaremos la inversa de la opuesta de esta matriz, esta es la matriz de varianzas-covarianzas de los estimadores. Los errores estándar son las raíces cuadradas de los elementos en la diagonal de esta última matriz.
El cálculo de la matriz de varianzas-covarianzas puede presentar problemas numéricos, por lo que los programas estadísticos suelen tener implementados algoritmos de optimización que evitan dichos problemas.

En resumen podemos decir que el método de N-R (Newton-Raphson) nos proporciona

1. Un algoritmo para estimar los coeficientes de todos los miembros *GLM* para la familia exponencial uniparamétrica.

2. Estimaciones de los errores estándar de los coeficientes estimados: estos son las raíces cuadradas de los elementos de la diagonal de la inversa de la matriz Hessiana negativa estimada.

En cuanto a la ejecución del algoritmo, necesitamos tener unos valores iniciales. Para dar estos valores no hay un mecanismo global pero hay una solución razonable para obtenerlos cuando hay una constante en el modelo. Es decir, se aconseja considerar el modelo con sólo un término constante $\eta = \beta_0$ y comenzar las iteraciones con $\widehat{\beta_0}$ que suele tener una solución analítica explícita.

Ejemplo: Como ilustración, veamos el ejemplo del modelo de *Poisson*. La función de log-verosimilitud viene dada en este caso por

$$\mathcal{L} = \sum_{i=1}^{n} \{ y_i(\mathbf{x}_i\boldsymbol{\beta}) - e^{\mathbf{x}_i\boldsymbol{\beta}} - ln\Gamma(y_i + 1) \}. \tag{2.3}$$

Si asumimos que el modelo tiene una única covariable constante, 2.3 se reduce a

$$\mathcal{L} = \sum_{i=1}^{n} \{y_i \beta_0 - e^{\beta_0} - ln\Gamma(y_i + 1)\}.$$

El estimador de máxima verosimilitud de β_0 lo obtenemos igualando la siguiente expresión a cero y resolviéndola en β_0:

$$\frac{\partial \mathcal{L}}{\partial \beta_0} = \sum_{i=1}^{n} \{y_i - exp(\beta_0)\}.$$

Tenemos por lo tanto

$$0 = \sum_{i=1}^{n} \{y_i - exp(\widehat{\beta}_0)\},$$

$$\widehat{\eta} = \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n},$$

$$\widehat{\beta_0} = ln(\overline{y}).$$

Es decir, tomaremos como vector de parámetros iniciales $\widehat{\boldsymbol{\beta}_{\mathbf{0}}} = (\widehat{\beta_0}, 0, ..., 0)'$ con $\widehat{\beta_0} = \ln(\overline{y})$en el algoritmo de Newton-Raphson.

Seguir este método tiene dos ventajas. En primer lugar comenzamos nuestras iteraciones a partir de un $\widehat{\beta}_0$ que está en el espacio paramétrico y es sencillo de obtener. En segundo lugar, permite hacer posteriormente tests de razón de verosimilitudes basados en comparar$\mathcal{L}(\widehat{\beta}_0)$ y $\mathcal{L}(\widehat{\boldsymbol{\beta}})$. En estas funciones se basarán algunos de los diagnósticos de bondad de ajuste del modelo que trataremos posteriormente.

2.4.2. Bondad de ajuste del modelo.

Deseamos obtener valores ajustados $\widehat{\mu}$ que sean cercanos a los datos $\mathbf{y}$.
En primer lugar recordemos que existen diversos posibles modelos. Si tenemos n observaciones y tomamos n covariables, llamaremos a este modelo, modelo saturado, puesto que no prescinde de ninguna de las covariables. Es un modelo de la forma

$$\eta = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n.$$

El modelo saturado incluye un parámetro para cada observación y como resultado $\widehat{\mu_i} = y_i$, pero este modelo presenta el problema de que no supone ningún resumen de los datos.
Es por esto que definimos una medida de ajuste del modelo a los datos como el doble de la diferencia entre la log-verosimilitud del modelo de interés y la del modelo saturado.
Esta medida recibe el nombre de *deviance*, D, y viene dada por

$$D = \sum_{i=1}^{n} 2\left[y\{\theta(y_i) - \theta(\mu_i)\} - b\{\theta(y_i)\} + b\{\theta(\mu_i)\}\right]$$

O lo que es equivalente,

$$D = 2\{ \underbrace{\mathcal{L}(\widehat{\beta}_0, ..., \widehat{\beta}_k)}_{Modelo \quad propuesto} - \underbrace{\mathcal{L}(\widehat{\beta}_0, ..., \widehat{\beta}_n)}_{Modelo \quad saturado} \}$$

En el ajuste de un modelo en particular a unos datos, buscamos los valores de los parámetros que minimicen D. Estos valores son los mismos que maximizan la verosimilitud.

Otra medida de bondad de ajuste de un modelo ampliamente utilizada es el estadístico $\mathcal{X}^2$ de Pearson generalizado, definido como

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

donde $V(\mu_i)$ es la función de varianza estimada para el modelo que se esté considerando. Veremos cómo se aplica este estadístico en capítulos posteriores.

Capítulo 3

Regresión de Poisson.

Una vez explicado en qué consisten los Modelos Lineales Generalizados, nos centraremos en dos de ellos, el modelo de regresión de Poisson y el modelo de regresión binomial negativa, pues son los más apropiados para trabajar con datos de conteo como los que obtenemos a través de *RNA-seq*. En este capítulo en particular trataremos la regresión de Poisson. En primer lugar recordaremos la distribución de Poisson, y de acuerdo con lo visto en el capítulo anterior, identificaremos sus componentes como miembro de la familia exponencial. A continuación explicaremos el modelo de regresión siguiendo el esquema explicado en la Sección 2.3, veremos cómo se estiman los parámetros de dicho modelo y también como interpretar estos parámetros. Para finalizar el capítulo explicaremos uno de los problemas que presenta la regresión de Poisson, el problema de sobredispersión, y que por lo tanto, convierte al modelo de regresión binomial negativa en el más apropiado para el estudio de datos de *RNA-seq*.
Podemos encontrar más información acerca de la regresión de Poisson en Hardin and Hilbe (2007).

3.1. Distribución de Poisson.

Cuando hablamos del modelo de regresión de Poisson, nos encontramos con un *GLM* en el cuál la variable respuesta Y sigue una distribución de Poisson de media μ, con $\mu > 0$, lo que denotamos como $Y \backsim \text{Po}(\mu)$.

Tradicionalmente, esta distribución se utiliza para modelizar cuál es la probabilidad de que durante un determinado periodo de tiempo ocurra un cierto número de eventos, todo esto a partir de una frecuencia de ocurrencia media. Por ejemplo, podemos estudiar el número de personas que tienen un infarto, número de personas que llaman a una centralita de teléfono, etc, siempre evaluando todo esto en una unidad de tiempo determinada. También podemos encontrarnos que este tipo de distribución se emplee para el estudio de conteos en el espacio tales como número de accidentes de tráfico que se producen en el cruce de 2 carreteras, o número de células sanguíneas en una muestra de sangre (en este caso el espacio sería igual al volumen en centímetros cúbicos).

Recordemos que la función de probabilidad de este modelo es:

$$P[\mathrm{Y}= y] = f(y;\mu) = e^{-\mu}\frac{\mu^y}{y!}\ , \qquad\qquad y \in \mathbb{Z}_0^+, \quad \mu > 0,$$

y además se verifica que

$$\mathrm{E}[Y]{=}\mathrm{var}[Y]{=}\mu.$$

Aquí podemos ver que tenemos una distribución de probabilidad discreta que cuenta con la característica de equidispersión, es decir, se tiene la igualdad entre las funciones media y varianza. O sea, es una distribución que cuanto más grande es el valor esperado más dispersión tienen los valores que puede tomar la variable que se distribuya así.

Podemos tomar logaritmo en la función de probabilidad y ver que claramente se trata de un modelo perteneciente a la familia exponencial:

$$lnP[Y = y] = -\mu + yln(\mu) - ln(y!) = yln(\mu) - \mu - ln(y!)$$

Identifiquemos las componentes dadas en 1.1. En este caso:

- La *función vinculo* o *función link*:

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

- El parámetro de dispersión:

$$\phi = 1$$

- Función cumulante:

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

  La función cumulante la expresamos en función del parámetro natural:

$$b(\theta) = e^{\theta}$$

Una vez identificados estos elementos pasamos a estudiar en detalle el modelo de regresión.

3.2. Modelo de regresión de Poisson.

A continuación, aplicamos el esquema dado en la Sección 2.3 al caso en el que nos encontramos.

- En primer lugar, la media es conocida, puesto que si $Y \backsim \mathrm{Po}\ (\mu)$ entonces $\mathrm{E}[Y]{=}\mu$. Ahora bien, nos resulta de utilidad conocer μ en función de θ, y para ello podemos seguir dos métodos diferentes.
  El primero de ellos consiste en despejar directamente μ de la función link, por lo que obtenemos rápidamente que $\mu = e^{\theta}$.
  Por otra parte, en la Sección 2.2 en la cual tratamos la familia exponencial, se

mencionó como una de sus características la posibilidad de obtener las funciones media y varianza mediante la primera y la segunda derivada de la función cumulante con respecto de θ.
Siguiendo este método llegaríamos al mismo resultado de la siguiente forma:

$$E[Y] = b'(\theta) = \frac{d}{d\theta}e^{\theta} = e^{\theta} = \mu$$

De la misma forma obtenemos la varianza:

$$var[Y] = b''(\theta) = \frac{d^2}{d\theta^2}e^{\theta} = \frac{d}{d\theta}e^{\theta} = e^{\theta} = \mu$$

Comprobamos así la propiedad de equidispersión de la que hablamos al principio de la sección.

Resumimos los datos obtenidos hasta este momento para tener una visión más clara a la hora de desarrollar el modelo.

función vínculo: $\theta(\mu)$	$b(\theta)$	ϕ	inversa de la función vínculo	media	varianza
$ln(\mu)$	μ	1	e^{θ}	$\mu = e^{\theta}$	$\mu = e^{\theta}$

- Puesto que la función vínculo ya está calculada, lo que tenemos que ver es cómo se expresa $\theta(\mu)$ como un modelo de regresión lineal.
  Puesto que $\theta(\mu) = ln(\mu)$, para una única variable independiente X tenemos un modelo de la forma:

  $$ln(\mu|X = x\ ) = \beta_0 + \beta_1\ x$$

  O escrito de una forma más simplificada:

  $$ln(\mu) = \beta_0 + \beta_1\ x$$

  donde β_0 y β_1 son constantes mientras que X hace referencia a una variable que puede ser aleatoria o no, continua, discreta o cualitativa.
  Podemos generalizar este modelo para el caso en el que aparezcan k variables independientes, quedando el modelo de la siguiente forma:

  $$\ln(\mu) = \beta_0 + \beta_1\ x_1 + ... + \beta_k\ x_k$$

- Volvemos al caso en el que aparece una única variable independiente. La función de enlace canónico para la distribución de Poisson es $\theta = ln(\mu)$. Tenemos entonces:

  $$\theta(\mu) = \beta_0 + \beta_1\ x$$

y de esta forma obtenemos (sin más que despejar) que la función respuesta de nuestro modelo, es decir la función que predice la media, es:

$$\mu = e^{\beta_0+\beta_1 x}$$

O lo que es igual si denotamos por $\eta = \beta_0 + \beta_1\, x$:

$$\mu = e^{\eta}$$

El uso de la función exponencial asegura que la respuesta esperada (E $[Y] = \mu$) será siempre positiva.

De esta forma ya tenemos identificadas las diferentes componentes de todo *GLM*: la componente aleatoria Y que sigue una distribución de Poisson; la componente sistemática que es el predictor lineal que expresa la combinación lineal de las variables explicativas y proporciona el valor predicho ($\eta_i = \beta' x_i$); y la función de enlace, aquella que relaciona η con μ ($\theta(\mu_i) = \ln(\mu_i)$).

Para los algoritmos numéricos nos resulta importante tener la derivada de la función vínculo:

$$\theta'(\mu) = \frac{d}{d\mu}\ \theta(\mu) = \frac{d}{d\mu}\ ln(\mu) = \frac{1}{\mu}$$

Puesto que $\mu = e^{\eta}$, esto nos quedaría:

$$\frac{1}{\mu} = \frac{1}{e^{\beta_0+\beta_1 x}}$$

A partir de ahora vamos a emplear la notación vectorial, es decir:

$$\mathbf{x}' = (1, x), \qquad \boldsymbol{\beta} = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \qquad \beta_0 + \beta_1 x = \mathbf{x}'\boldsymbol{\beta}$$

Parametrizado en términos de esta notación, la log-verosimilitud (logaritmo de la función de verosimilitud) para nuestro modelo es:

$$\mathcal{L}(\boldsymbol{\beta}; y_i) = \sum_{i=1}^{n}\{y_i(\mathbf{x}_i'\boldsymbol{\beta}) - e^{\mathbf{x}_i'\boldsymbol{\beta}} - ln(y_i!)\}$$

donde $\mathbf{x}_i' = (1, x_i)$ es el valor de la variable explicativa en el individuo iésimo.

Teniendo en cuenta que $y_i \in \mathbb{Z}_0^+$ podemos considerar $y_i! = \Gamma(y_i + 1)$ por lo que la expresión final de la log-verosimilitud obtenida es la siguiente:

$$\mathcal{L}(\boldsymbol{\beta}; y_i) = \sum_{i=1}^{n}\{y_i(\mathbf{x}_i'\boldsymbol{\beta}) - e^{\mathbf{x}_i'\boldsymbol{\beta}} - ln\Gamma(y_i + 1)\}$$

Con esta expresión tenemos que para aquellos valores $y_j = 0$, el término correspondiente de la suma anterior se reduce a:

$$-e^{\mathbf{x}_j'\boldsymbol{\beta}}$$

3.3. Estimación de los parámetros.

Como se ha comentado en el capítulo anterior, el método más utilizado para estimar el vector de parámetros $\boldsymbol{\beta}$, es el método de máxima verosimilitud iterativo.
Dicho método necesita calcular derivadas de la función log-verosimilitud, y necesitaremos el gradiente y la matriz hessiana como observamos a continuación.

El vector gradiente (o gradiente) de la función de log-verosimilitud respecto a los parámetros de $\underline{\beta}$ es:

$$\frac{\partial(\mathcal{L}(\boldsymbol{\beta}; y_i))}{\partial(\boldsymbol{\beta})} = \Sigma(y_i - exp(x_i\beta))x_i$$

Igualar esta expresión a cero, nos proporciona las soluciones de los parámetros estimados.

$$\Sigma(y_i - exp(x_i\beta))x_i = 0$$

Por otra parte la matriz hessiana se calcula como la inversa negativa de la derivada segunda de la función log-verosimilitud:

$$\frac{\partial(\mathcal{L}(\boldsymbol{\beta}; y_i))}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = [-\Sigma(exp(x_i\boldsymbol{\beta}))x_i x_j]^{-1}$$

Teniendo esta expresión tenemos también los valores de los errores estándar de los parámetros, que vienen dados por la raíz cuadrada de los respectivos términos de la diagonal de la inversa negativa de la matriz Hessiana.

Por último en cuanto a la estimación de máxima verosimilitud de los parámetros, vemos que puede utilizarse un algoritmo de Newton-Raphson como se indica a continuación:

$$\beta_{i+1} = \beta_i - H^{-1}g$$

3.4. Interpretación de los coeficientes estimados.

Seguimos trabajando con una sola variable explicativa, es decir, considerando un modelo de regresión simple con un único predictor x. Veamos como interpretar los coeficientes a partir de las observaciones en dos individuos.
En primer lugar denotamos por x_1 y x_2 los valores de la variable X para los individuos 1 y 2 respectivamente, y tenemos por consiguiente dos predicciones:

$$\theta_1^* = b_0 + b_1 x_1$$

$$\theta_2^* = b_0 + b_1 x_2$$

A partir de estas dos expresiones obtenemos una nueva expresión:

$$\theta_2^* - \theta_1^* = b_1(x_2 - x_1)$$

Puesto que en nuestro modelo hemos definido $\theta = ln(\mu)$, la relación anterior es equivalente a:

$$ln(\mu_2^*) - ln(\mu_1^*) = b_1(x_2 - x_1) \Longleftrightarrow ln(\frac{\mu_2^*}{\mu_1^*}) = b_1(x_2 - x_1) \Longleftrightarrow \frac{\mu_2^*}{\mu_1^*} = e^{b_1(x_2 - x_1)}$$

Esta expresión puede escribirse como:

$$\mu_2^* = \mu_1^* e^{b_1(x_2 - x_1)}$$

Cuando consideramos el incremento de una unidad en el predictor x tenemos una función media:

$$E[Y|x+1] = \mu^*(x+1) = \mu^*(x)e^{b_1}$$

Vemos de esta manera que el impacto de una unidad de cambio en x es un múltiplo de la media anterior. En la predicción anterior, el factor de proporcionalidad es e^{b_1}.
Se conoce como cociente de razones de incidencia a:

$$\frac{\mu^*(x+1)}{\mu^*(x)} = e^{b_1}$$

3.5. Problema de la sobredispersión.

En ciertas ocasiones el modelo de regresión de Poisson puede resultar inapropiado debido a que no se cumplen ciertos supuestos. El problema más común es la ausencia de equidispersión. Teóricamente podríamos encontrarnos con subdispersión o sobredispersión, pero en la práctica la que aparece con mayor frecuencia es la sobredispersión. Por este mismo hecho las pruebas para evaluar equidispersión reciben el nombre habitualmente de pruebas de sobredispersión.

Aunque ya se mencionó el concepto de equidispersión al explicar la distribución de Poisson vamos a explicarlo con mayor detalle.

La equidispersión se trata de un supuesto básico en el que se supone que $var[Y] = \alpha^2 E[Y]$, siendo el parámetro de dispersión $\alpha^2 = 1$. La sobredispersión aparece cuando $var[Y] > E[Y]$, o lo que es igual, $\alpha^2 > 1$. Cuando existe este 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 sobredispersión puede aparecer por diversos motivos. Entre los más comunes podemos citar:

- Los datos no provienen de una distribución de Poisson.
- Alta variabilidad en los datos.
- Falta de estabilidad, es decir, la probabilidad de ocurrencia de un evento puede ser independiente de la ocurrencia de un evento previo pero éste no es constante.

- Los eventos no ocurren independientemente a través del tiempo.
- Errores de especificación de la media, bien que se hayan omitido variables explicativas relevantes o que deban entrar al modelo a través de alguna transformación en lugar de linealmente.
- Errores al elegir la función de enlace, es decir, tal vez no fue apropiado escoger la función link log-lineal.

Existen diversos métodos para detectar la sobredispersión aunque generalmente se detecta evaluando la relación entre los estadísticos $\mathcal{X}^2$ de Pearson o la función *deviance* D y sus respectivos grados de libertad (gl), es decir evaluamos:

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

En el caso en que estos valores sean mayor que 1, nos indican la existencia de sobredispersión.

Ambos estadísticos están definimos anteriormente en la Sección 2.4.2.

Capítulo 4

Regresión binomial negativa.

Tradicionalmente el modelo de distribución binomial negativa se ha estudiado como un modelo de la familia exponencial de distribuciones, puede decirse que fue utilizado por primera vez como un modelo de regresión por Anscombe en 1949. Otros autores, como Lawless (1987), Breslow (1990), o McCullagh and Nelder (1989), hacen referencia a él como un posible modelo de regresión. Pero no es hasta mediados de los años 90, con los textos de Hilbe, que este modelo se ha estudiado en profundidad desde el punto de vista de los GLMs.

En este capítulo presentamos diversos métodos para hallar el modelo de regresión binomial negativa. Dos de ellos se basan en proponer una mixtura de Poisson-Gamma, el tercer método propuesto se basa en la expresión clásica de la función de probabilidad de la binomial negativa. En todos ellos obtenemos una densidad marginal de la variable explicativa, su media y su varianza, destacando el papel del índice de sobredispersión, obtenido como cociente de la varianza a la media. Este índice nos permite proponer este modelo como una generalización del modelo de Poisson, que resuelve el problema de la sobredispersión.

El resultado obtenido con los métodos anteriormente mencionados son los llamados modelos de regresión con sobredispersión constante, denotado como NB-1, y el modelo de regresión con sobredispersión variable, denotado como NB-2. A continuación explicamos ambos modelos.

4.1. Sobredispersión constante.

En este primer caso tenemos, como se dijo en apartados anteriores, una mixtura Poisson-Gamma en la cual consideramos el parámetro de Poisson λ_i como una variable aleatoria que se distribuye según una Gamma. Es decir, $Y_i \backsim Po(\lambda_i)$ donde:

$$P[Y_i = y_i] = e^{-\lambda_i}\frac{\lambda_i^{y_i}}{y_i!}, \qquad y_i \in \mathbb{Z}_0^+, \quad \lambda_i > 0.$$

La variable λ_i suponemos que sigue una distribución $Ga(\delta, \mu_i)$ con función de densidad

$$f_{\lambda_i}(\lambda_i) = \frac{1}{\Gamma(\mu_i)}\delta^{\mu_i}\lambda_i^{\mu_i - 1}e^{-\delta\lambda_i}$$

donde

$$\mu_i = e^{\mathbf{x_i'}\beta + offset_i}$$

El valor esperado del parámetro de la variable de Poisson y su varianza vienen dados por

$$E[\lambda_i] = \frac{e^{\mathbf{x_i'}\beta + offset_i}}{\delta} = \frac{\mu_i}{\delta},$$

$$var[\lambda_i] = \frac{e^{x_i\beta + offset_i}}{\delta^2} = \frac{\mu_i}{\delta^2}.$$

Como hemos visto, tenemos dos variables aleatorias, Y_i/λ_i y λ_i, por lo que la distribución resultante de la mezcla (la marginal de Y_i) ha de calcularse a partir de la distribución conjunta de ambas variables.

Proposición 4.1 *En las condiciones descritas anteriormente, la función de densidad marginal de Y_i es*

$$f_{Y_i}(y_i) = \frac{\Gamma(y_i + \mu_i)}{\Gamma(\mu_i)\Gamma(y_i + 1)} \left(\frac{\delta}{1+\delta}\right)^{\mu_i} \left(\frac{1}{1+\delta}\right)^{y_i}, \qquad y_i \in \mathbb{Z}_0^+.$$

Demostración:

$$\begin{aligned}
f_{Y_i}(y_i|x_i) &= \int_0^\infty P[Y_i = y_i|\lambda_i]\ \ f_{\lambda_i}(\lambda_i)\ \ d\lambda_i \\
&= \int_0^\infty e^{-\lambda_i} \frac{\lambda_i^{y_i}}{y_i!} \frac{1}{\Gamma(\mu_i)} \delta^{\mu_i} \lambda_i^{\mu_i - 1} e^{-\delta\lambda_i}\ \ d\lambda_i \\
&= \frac{\delta^{\mu_i}}{\Gamma(y_i+1)\Gamma(\mu_i)} \int_0^\infty \lambda_i^{(y_i+\mu_i)-1} e^{-\lambda_i(\delta+1)}\ \ d\lambda_i \\
&= \frac{\delta^{\mu_i}}{\Gamma(y_i+1)\Gamma(\mu_i)} \frac{\Gamma(y_i+\mu_i)}{(\delta+1)^{y_i+\mu_i}} \underbrace{\int_0^\infty \frac{(\delta+1)^{y_i+\mu_i}}{\Gamma(y_i+\mu_i)} \lambda_i^{(y_i+\mu_i)-1} e^{-\lambda_i(\delta+1)}\ \ d\lambda_i}_{=1 \ \ usando \ \ \frac{\Gamma(p)}{a^p} = \int_0^\infty x^{p-1}e^{-ax}dx} \\
&= \frac{\delta^{\mu_i}}{\Gamma(y_i+1)\Gamma(\mu_i)} \frac{\Gamma(y_i+\mu_i)}{(\delta+1)^{y_i+\mu_i}} \\
&= \frac{\Gamma(y_i+\mu_i)}{\Gamma(\mu_i)\Gamma(y_i+1)} \left(\frac{\delta}{1+\delta}\right)^{\mu_i} \left(\frac{1}{1+\delta}\right)^{y_i}
\end{aligned}$$

■

Una vez obtenida la distribución sus momentos son:

$$E[Y_i] = \frac{e^{\mathbf{x_i}\boldsymbol{\beta}+offset_i}}{\delta},$$

$$var[Y_i] = \frac{e^{\mathbf{x_i}\boldsymbol{\beta}+offset_i}(1+\delta)}{\delta^2}.$$

Si calculamos el cociente entre la varianza y la media observamos que éste es constante para todas las observaciones, de ahí que este modelo se conozca como de sobredispersión constante:

$$\frac{var[Y_i]}{E[Y_i]} = \frac{1+\delta}{\delta}$$

Este cociente recibe el nombre de índice de sobredispersión, y considerando $\alpha = \frac{1}{\delta}$ podemos reparametrizar este modelo obteniendo

$$\frac{var[Y_i]}{E[Y_i]} = 1+\alpha$$

$$f_{Y_i}(y_i) = \frac{\Gamma(y_i+\mu_i)}{\Gamma(\mu_i)\Gamma(y_i+1)}\left(\frac{1}{1+\alpha}\right)^{\mu_i}\left(\frac{\alpha}{1+\alpha}\right)^{y_i} \tag{4.1}$$

Tomando límite en 4.1 cuando $\alpha \to \infty$, se tiene una distribución de Poisson.

En este caso no podemos considerar la distribución resultante como miembro de la familia exponencial de distribuciones. Veamos a continuación la regresión con sobredispersión variable(NB-2).

4.2. Sobredispersión variable.

A su vez, en el marco de *GLM*, podemos obtener la binomial negativa siguiendo dos métodos diferentes. El primero de ellos consiste en concebirla como un modelo de Poisson con heterogeneidad Gamma, donde el ruido Gamma tiene una media igual a 1. El segundo método consiste en tratar la binomial negativa como una función de probabilidad por derecho propio, independientemente de la distribución de Poisson. Lo que hacemos de esta forma es ver su función de probabilidad como la probabilidad de observar y fallos antes del i-ésimo éxito en una serie de pruebas independientes e idénticamente distribuidas (i.i.d.) Bernoulli.
A pesar de sus diferencias, ambos métodos convergen en la misma función de log-verosimilitud y con ambas obtenemos la varianza como una función cuadrática de la media. Por esto último este modelo recibe el nombre de sobredispersión variable o tipo NB-2. Vamos a estudiar ambos métodos con más detalle a continuación.

4.2.1. Derivación en términos de una mixtura Poisson-Gamma.

En este método consideramos $Y_i \backsim Po(\mu_i)$ con $\mu_i = \lambda_i u_i$, donde $ln\mu_i$ se modeliza como un modelo lineal de la siguiente manera:

$$ln\mu_i = x_i\beta + \epsilon_i$$

$$= ln\lambda_i + lnu_i$$

siendo u_i un efecto no observado (ruido) y $\lambda_i > 0$ constante. Suponemos además que $u_i \backsim Ga(\nu, \nu)$ con $\nu > 0$.

Con esta notación tenemos $Y_i|\lambda_i, u_i \equiv Y_i|x_i, u_i \backsim Po(\mu_i)$ con $\mu_i = \lambda_i u_i$ cuya expresión es

$$f(y_i|x_i, u_i) = \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} \quad con \quad E[Y_i|x_i, u_i] = \lambda_i u_i.$$

La densidad de Y condicionada sólo a las variables explicativas x_i quedaría

$$f(y_i|x_i) = \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} g(u_i)\, du_i$$

donde el resultado que obtengamos dependerá de la función g(·) que elijamos para modelizar el error o ruido u_i.

Se supone que la media de la derivación Gamma es 1. Utilizando esta normalización se tiene el siguiente resultado:

Proposición 4.2 *En las condiciones descritas anteriormente, la función de densidad marginal de Y_i es*

$$f_{Y_i}(y_i) = \frac{\Gamma(y_i + \nu)}{\Gamma(y_i + 1)\Gamma(\nu)} \left(\frac{1}{1 + \frac{\lambda_i}{\nu}}\right)^\nu \left(1 - \frac{1}{1 + \frac{\lambda_i}{\nu}}\right)^{y_i}$$

Demostración:

$$f_{Y_i}(y_i) = \int_0^\infty \frac{e^{-\lambda_i u_i}(\lambda_i u_i)^{y_i}}{y_i!} \frac{\nu^\nu}{\Gamma(\nu)} u_i^{\nu-1} e^{-\nu u_i}\, du_i$$

$$= \frac{\lambda_i^{y_i}}{\Gamma(y_i + 1)} \frac{\nu^\nu}{\Gamma(\nu)} \int_0^\infty e^{-(\lambda_i+\nu)u_i} u_i^{(y_i+\nu)-1}\, du_i$$

$$= \frac{\lambda_i^{y_i}}{\Gamma(y_i + 1)} \frac{\nu^\nu}{\Gamma(\nu)} \frac{\Gamma(y_i + \nu)}{(\lambda_i + \nu)^{y_i+\nu}} \underbrace{\int_0^\infty \frac{(\lambda_i + \nu)^{y_i+\nu}}{\Gamma(y_i + \nu)} e^{-(\lambda_i+\nu)u_i} u_i^{(y_i+\nu)-1}\, du_i}_{=1 \quad usando \quad \frac{\Gamma(p)}{a^p} = \int_0^\infty x^{p-1}e^{-ax}dx}$$

$$= \frac{\lambda_i^{y_i}}{\Gamma(y_i + 1)} \frac{\nu^\nu}{\Gamma(\nu)} \frac{\Gamma(y_i + \nu)}{(\lambda_i + \nu)^{y_i+\nu}}$$

$$= \frac{\lambda_i^{y_i}}{\Gamma(y_i+1)} \frac{\nu^\nu}{\Gamma(\nu)} \Gamma(y_i+\nu) \left(\frac{\nu}{\lambda_i+\nu}\right)^\nu \frac{1}{\nu^\nu} \left(\frac{\lambda_i}{\lambda_i+\nu}\right)^{y_i} \frac{1}{\lambda_i^{y_i}}$$

$$= \frac{\Gamma(y_i+\nu)}{\Gamma(y_i+1)\Gamma(\nu)} \left(\frac{\nu}{\lambda_i+\nu}\right)^\nu \left(\frac{\lambda_i}{\lambda_i+\nu}\right)^{y_i}$$

$$= \frac{\Gamma(y_i+\nu)}{\Gamma(y_i+1)\Gamma(\nu)} \left(\frac{1}{1+\frac{\lambda_i}{\nu}}\right)^\nu \left(1-\frac{1}{1+\frac{\lambda_i}{\nu}}\right)^{y_i}$$

■

Reescribimos el resultado considerando $\nu = \frac{1}{\alpha}$:

$$f_{Y_i}(y_i) = \frac{\Gamma(y_i+\frac{1}{\alpha})}{\Gamma(y_i+1)\Gamma(\frac{1}{\alpha})} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(1-\frac{1}{1+\alpha\mu_i}\right)^{y_i}.$$

con $\mu_i = \lambda_i = e^{\mathbf{x_i}\beta + offset_i}$.

Notación dentro de la familia exponencial.

Tomando logaritmo de la densidad que hemos obtenido anteriormente podemos ver claramente que pertenece a la familia exponencial e identificar cada una de las componentes. En notación de esta familia de distribuciones tenemos:

$$lnf(y;\mu,\alpha) = yln\left(\frac{\alpha\mu}{1+\alpha\mu}\right) + \frac{1}{\alpha}ln\left(\frac{1}{1+\alpha\mu}\right) + ln\Gamma\left(y+\frac{1}{\alpha}\right) - ln\Gamma\left(y+1\right) - ln\Gamma\left(\frac{1}{\alpha}\right)$$

Tenemos de esta forma las principales componentes que necesitamos para la construcción del *GLM* binomial negativo.

- La función vínculo o función link:

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

- Parámetro de dispersión:

$$\phi = 1$$

- Función cumulante:

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

- Media:

$$E[Y] = b'(\theta) = \frac{\partial b}{\partial \mu}\frac{\partial \mu}{\partial \theta} = \left(\frac{1}{1+\alpha\mu}\right)\mu(1+\alpha\mu) = \mu$$

- Varianza:

$$\begin{aligned} var[Y] = V[\mu] = b''(\theta) &= \frac{\partial^2 b}{\partial \mu^2}\left(\frac{\partial \mu}{\partial \theta}\right)^2 + \frac{\partial b}{\partial \mu}\frac{\partial^2 \mu}{\partial \theta^2} \\ &= \frac{-\alpha}{(1+\alpha\mu)^2}\mu^2(1+\alpha\mu)^2 + \left(\frac{1}{1+\alpha\mu}\right)(1+\alpha\mu)(\mu+2\alpha\mu^2) \\ &= -\alpha\mu^2 + \mu + 2\alpha\mu^2 \\ &= \mu + \alpha\mu^2 \end{aligned}$$

- Derivada de la varianza en función de la media:

$$\frac{\partial V(\mu)}{\partial \mu} = 1 + 2\alpha\mu$$

De esta forma tenemos que

$$\mu_i^* = \lambda_i = exp(x_i\beta + offset_i).$$

y el índice de sobredispersión es

$$\frac{Var[Y_i]}{E[Y_i]} = 1 + \alpha\mu_i.$$

En esta parametrización hemos considerado α positivo y a medida que aumenta α también lo hace la sobredispersión del modelo. El caso límite en el que $\alpha = 0$ se corresponde con el modelo de Poisson. Puesto que la probabilidad a veces la expresamos en términos de $\tau = ln(\alpha)$, la condición frontera para el modelo de Poisson se corresponde con $\tau = -\infty$.

Para finalizar el estudio de este método vemos la función de log-verosimilitud y la función *deviance*.
Puesto que

$$\mathcal{L}(\mu; y, \alpha) = \sum_{i=1}^{n}\left\{y_i ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}ln(1+\alpha\mu_i) + ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - ln\Gamma\left(y_i+1\right) - ln\Gamma\left(\frac{1}{\alpha}\right)\right\}$$

se tiene

$$D = 2\sum_{i=1}^{n}\left\{y_i ln\left(\frac{y_i}{\mu_i}\right) - \left(y_i + \frac{1}{\alpha}\right) ln\left(\frac{1+\alpha y_i}{1+\alpha\mu_i}\right)\right\}.$$

4.2.2. Derivación en términos de la función de probabilidad de la binomial negativa.

En este método, al igual que el anterior, tratamos la binomial negativa dentro del marco *GLM*. Lo primero que debemos hacer es ver la distribución como la probabilidad de observar y fallos antes del r-ésimo éxito en una serie de pruebas de Bernoulli i.i.d. Si

Y es nuestra variable, denotamos este modelo como $Y \backsim$ BN(r,p). En este caso la función de probabilidad es

$$f(y;r,p) = \binom{y+r-1}{r-1} p^r (1-p)^y, \qquad con \;\; r \in \mathbb{N}, \;\; 0 < p < 1, \;\; y \in \mathbb{N} \cup \{0\}.$$

Expresamos esta misma función como un miembro de la familia exponencial.

$$lnf(y;r,p) = yln(1-p) + rln(p) + ln\binom{y+r-1}{r-1} \tag{4.2}$$

En este caso

- Función link:

$$\theta = ln(1-p)$$

  De aquí podemos obtener:

$$e^\theta = 1 - p \quad \Rightarrow \quad p = 1 - e^\theta$$

- Parámetro de dispersión:

$$\phi = 1$$

- Función cumulante:

$$b(\theta) = -rln(p) = -rln(1-e^\theta)$$

Al igual que en los métodos anteriores, calculamos la media y la varianza de esta distribución derivando b(θ) respecto del parámetro θ:

- Media:

$$\begin{aligned} b'(\theta) = \frac{\partial b}{\partial p}\frac{\partial p}{\partial \theta} &= \left(-\frac{r}{p}\right)(-(1-p)) \\ &= \left(-\frac{r}{p}\right)(-(1-p)) \\ &= \frac{r(1-p)}{p} \\ &= \mu \end{aligned}$$

- Varianza:

$$\begin{aligned} b''(\theta) &= \frac{\partial^2 b}{\partial p^2}\left(\frac{\partial p}{\partial \theta}\right)^2 + \frac{\partial b}{\partial p}\frac{\partial^2 p}{\partial \theta^2} \\ &= \left(\frac{r}{p^2}\right)(1-p)^2 + \frac{-r}{p}(1-p) \\ &= \frac{r(1-p)^2 + rp(1-p)}{p^2} \\ &= \frac{r(1-p)}{p^2} \end{aligned}$$

Podemos reescribir la varianza en términos de la media de la siguiente forma:

$$var[Y] = V(\mu) = b''(\theta) = \mu + \frac{\mu^2}{r}$$

Ahora vamos a realizar una reparametrización en términos de α para que de esta forma la varianza sea directamente (en vez de inversamente) proporcional a la media. Consideramos por lo tanto $\alpha = \frac{1}{r}$, por lo que reescribiendo lo anterior con esta nueva parametrización obtenemos:

$$p = 1 - e^{\theta} = \frac{1}{1+\alpha\mu} \tag{4.3}$$

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

$$b(\theta) = -\frac{1}{\alpha}ln(p) = \frac{1}{\alpha}ln(1+\alpha\mu)$$

$$b'(\theta) = \frac{1-p}{\alpha p} = \mu = \frac{1}{\alpha(e^{-\theta}-1)}$$

$$b''(\theta) = \frac{1-p}{\alpha p^2} = \mu + \alpha\mu^2$$

$$g'(\theta) = \frac{\partial\theta}{\partial\mu} = \frac{\partial}{\partial\mu}ln\left(\frac{\alpha\mu}{1+\alpha\mu}\right) = \frac{1}{\mu+\alpha\mu^2}$$

$$V(\mu) = b''(\theta) = \mu + \alpha\mu^2$$

Teniendo en cuenta esta reparametrización y substituyendo la expresión (4.3) en (4.2), obtenemos las siguientes expresiones de la función de log-verosimilitud, y *deviance*:

$$\mathcal{L}(\mu; y, \alpha) = \sum_{i=1}^{n} \left\{ y_i ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha} ln(1+\alpha\mu_i) + ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - ln\Gamma\left(y_i + 1\right) - ln\Gamma\left(\frac{1}{\alpha}\right) \right\}$$

$$D = 2\sum_{i=1}^{n} \left\{ y_i ln\left(\frac{y_i}{\mu_i}\right) - \left(y_i + \frac{1}{\alpha}\right) ln\left(\frac{1+\alpha y_i}{1+\alpha\mu_i}\right) \right\}$$

Vemos que obtenemos los mismos resultados que en la subsección anterior, por lo que ambos métodos son válidos para la construcción de nuestro modelo.

4.3. Interpretación de los coeficientes.

Esta interpretación es similar al caso de Regresión de Poisson.
Vamos a trabajar con una sola variable explicativa, es decir, considerando un modelo de regresión simple con un único predictor x. Veamos como interpretar los coeficientes a partir de las observaciones en dos individuos.
En primer lugar denotamos por x_1 y x_2 los valores de la variable X para los individuos 1 y 2 respectivamente, y tenemos por consiguiente dos predicciones:

$$\theta_1^* = b_0 + b_1 x_1$$

$$\theta_2^* = b_0 + b_1 x_2$$

A partir de estas dos expresiones obtenemos:

$$\theta_2^* - \theta_1^* = b_1(x_2 - x_1)$$

Cuando consideramos el incremento de una unidad en la variable explicativa, es decir, vamos a considerar $x_1 = x$ y $x_2 = x + 1$, tenemos

$$\frac{\widehat{\mu^*}_2}{\widehat{\mu^*}_1} = e^{\beta_1(x_2 - x_1)}$$

Vemos de esta manera que el impacto de una unidad de cambio en x es un múltiplo de la media anterior.

Se conoce como cociente de razones de incidencia a:

$$\frac{\mu^*(x+1)}{\mu^*(x)} = e^{\beta_1}.$$

Veamos como se interpretaría esto con un ejemplo. Si $e^{\beta_1} = 2$, significa que al aumentar el valor de X_1 en una unidad, la predicción $\mu^*(x+1)$ es el doble de la que se tenía para x. Una vez estudiado este modelo, podemos aplicarlo a datos reales como veremos a continuación.

Capítulo 5

Aplicación a datos reales.

En este capítulo realizamos el estudio de un conjunto de datos de *RNA-seq* reales. Utilizamos para ello el Bioconductor y los paquetes *DESeq* y *DESeq2*. Ambos paquetes están especialmente indicados para el estudio de datos de *RNA-seq* utilizando el modelo de regresión binomial negativa. Indicaremos las librerías necesarias para el análisis, cómo proceder a la entrada de datos y cómo realizar contrastes para detectar genes diferencialmente expresados. Añadiremos también en este capítulo interpretaciones de las salidas de R, gráficos con los resultados y técnicas de evaluación de los datos y del modelo.
La mayor parte de la información aquí recogida podemos consultarla en Gentleman et al. (2004), Anders and Huber (2010) y I Love et al. (2014).

5.1. Marco de trabajo.

En primer lugar vamos a presentar el programa con el que trabajaremos, *Bioconductor*. Se trata de un software desarrollado para el análisis de datos genómicos, y herramientas relacionadas con este tipo de estudios. Destacamos que es software libre. Comenzó en 2001 en Harvard y actualmente se desarrolla en uno de los institutos líderes a nivel mundial en la investigación del cáncer, Fred Hutchinson Cancer Research Center, en Seatle. Podemos encontrar *Bioconductor* en www.bioconductor.org

A su vez, dentro de Bioconductor, utilizaremos los paquetes *DESeq* y *DESeq2*. Estos paquetes se han de descargar desde la página web de *Bioconductor*.

- *DESeq* es un paquete que se emplea para el análisis de la expresión diferencial de genes. De una forma general, se encarga de estimar la dependencia de la media y la varianza en datos de conteo procedentes de ensayos de secuanciación de alto rendimiento, y prueba la expresión diferencial basándose en un modelo que sigue una distribución binomial negativa.
- *DESeq2* es una mejora del paquete *DESeq*.

El paquete de datos con el que trabajaremos se llama *pasilla*, y se trata de un conjunto de

datos presentados en Brooks et al. (2011). Los datos proceden de un experimento sobre el cultivo de células de la mosca *Drosophila melanogaster*, a partir del cuál se investiga la reducción en la expresión de los genes ortólogos para mamíferos NOVA1 y NOVA2, utilzando el RNAi (ARN interferente - interfering RNA). El paquete de datos proporciona el recuento de lecturas por exón y por gen calculadas para los genes seleccionados a partir de datos de *RNA-seq*.

5.2. Entrada de datos.

El paquete *DESeq2* espera como entrada de datos una matriz de conteos donde las filas corresponden a los genes y las columnas a las muestras o individuos. Así, por ejemplo, el valor en la i-ésima fila y la j-ésima columna de la matriz de conteos nos dice las lecturas que tenemos para el gen i en la muestra j. Los conteos deben ser "conteos brutos" de valores secuenciados.

Lo primero que hacemos es cargar las librerías y el paquete *pasilla* en el que están almacenados los datos con los que vamos a trabajar. Este paquete como ya dijimos incluye el número de conteos por gen y por exón. Puesto que nos interesa solamente estudiar la expresión diferencial para los genes, seleccionamos los datos guardados en *pasillaGenes*.

```
## Instalación de librerías.
source("http://bioconductor.org/biocLite.R")
biocLite()
biocLite("Biobase")
biocLite("DESeq")
biocLite("DESeq2")
biocLite("pasilla")
## Carga de librerías.
library(Biobase)
library(DESeq)
library(DESeq2)
# Cargamos los datos.
library("pasilla")
#Nos quedamos con los datos pertenecientes a los genes.
data("pasillaGenes")
head(pasillaGenes)
```

A continuación damos la tabla de conteos. En ella apreciamos que el experimento se ha realizado para 14470 genes, que son las filas, y 7 individuos (moscas en nuestro caso) que corresponden a las columnas.

```
countData <- counts(pasillaGenes)
> dim(countData)
[1] 14470     7
```

Podemos ver también los primeros datos de la tala de conteos de la siguiente forma:

```
> head(countData)
```

	treated1fb	treated2fb	treated3fb	untreated1fb	untreated2fb	untreated3fb	untreated4fb
FBgn0000003	0	0	1	0	0	0	0
FBgn0000008	78	46	43	47	89	53	27
FBgn0000014	2	0	0	0	0	1	0
FBgn0000015	1	0	1	0	1	1	2
FBgn0000017	3187	1672	1859	2445	4615	2063	1711
FBgn0000018	369	150	176	288	383	135	174

Para poder realizar el análisis, debemos tener los datos en formato DESeqDataSet (la clase utilizada por el paquete DESeq2 para almacenar los recuentos de lectura), y para ello debemos tener la información de las columnas de la matriz de conteos. Veremos que tenemos dos factores para cada individuo, además, cada factor está formado por dos niveles. El primer factor es la condición, que diferencia entre individuos no tratados (es el nivel de control) e individuos tratados. El segundo factor indica el tipo de secuenciación empleada, cuyos niveles son "single-read" y "paired-end".

```
colData <- pData(pasillaGenes)[, c("condition","type")]
> colData
             condition        type
treated1fb     treated single-read
treated2fb     treated  paired-end
treated3fb     treated  paired-end
untreated1fb untreated single-read
untreated2fb untreated single-read
untreated3fb untreated  paired-end
untreated4fb untreated  paired-end
```

Una vez que tenemos esto ya podemos construir un DESeqDataSet. Además, como dijimos, cada individuo está identificado por dos factores. A continuación señalaremos que queremos realizar el estudio teniendo en cuenta el factor *condition* (condición), que diferencia entre individuos no tratados e individuos tratados. En nuestro ejemplo, como podemos ver en los resultados de *colData*, tenemos 3 individuos tratados y 4 no tratados.

```
dds <- DESeqDataSetFromMatrix(countData=countData,
                              colData=colData,
                              design=~condition)
colData(dds)$condition <- factor(colData(dds)$condition,
                                 levels=c("untreated", "treated"))
dds
```

Hemos aplicado la función *factor* a la columna que nos interesa en *colData*, en este caso a la correspondiente al factor condición que es con el que queremos realizar el análisis. Con esta función le hemos asignado un orden a los distintos niveles, considerando los

individuos no tratados como el nivel de control. Si no hacemos esto el programa elegiría los niveles según el orden alfabético, mientras que con el orden que hemos asignado nos resultará más fácil la interpretación de los resultados posteriormente.

5.3. Estudio estadístico y análisis diferencial.

El análisis de expresión diferencial en *DESeq2* usa un modelo lineal generalizado de la forma:

$$K_{ij} \sim NB(\mu_{ij}, \alpha_i), \quad \mu_{ij} > 0, \quad \alpha_i > 0$$

$$\mu_{ij} = s_j q_{ij}$$

donde K_{ij} es el número de conteos del gen i en la muestra j. Se propone una distribución binomial negativa con una media ajustada, μ_{ij}, y un parámetro de dispersión específico para cada gen, α_i. La media ajustada se considera como el producto de un factor de tamaño específico para cada muestra s_j y un parámetro q_{ij} proporcional a la concentración de los fragmentos real esperada en la muestra j.

Recordemos que al estar interesados en encontrar aquellos genes que presenten diferentes niveles de expresión según se trate de un individuo control o un individuo tratado, el contraste que realizamos es el siguiente:

$$\begin{cases} H_0 : \mu_{iA} = \mu_{iB} \\ \\ H_1 : \mu_{iA} \neq \mu_{iB} \end{cases}$$

donde μ_{iA} es la media de conteos del gen i en el grupo A (individuos no tratados) y μ_{iB} la media de conteos del mismo gen pero en el grupo B (tratados).
El contraste anterior es equivalente al siguiente:

$$\begin{cases} H_0 : \dfrac{\mu_{iA}}{\mu_{iB}} = 1 \\ \\ H_1 : \dfrac{\mu_{iA}}{\mu_{iB}} \neq 1 \end{cases}$$

donde $\dfrac{\mu_{iA}}{\mu_{iB}}$ se denomina *fold-change*. En términos del logaritmo del *fold-change*, el contraste sería

$$\begin{cases} H_0 : log_2 \dfrac{\mu_{iA}}{\mu_{iB}} = 0 \\ \\ H_1 : log_2 \dfrac{\mu_{iA}}{\mu_{iB}} \neq 0 \end{cases}$$

Este contraste se debe realizar para cada uno de los genes de forma independiente.

Ahora podemos comenzar con el análisis de expresión diferencial. Con el programa que trabajamos, la función *DESeq* recoge todos los pasos a seguir para un análisis de expresión diferencial estándar, y además, para acceder a los resultados basta con utilizar la función *results*. Ordenaremos los resultados por orden creciente del p-valor ajustado, de esta forma los primeros genes que aparecen en la tabla serán aquellos que muestren una mayor diferencia en los niveles de expresión entre los dos grupos.

```
analisisdds <- DESeq(dds)
respordefecto <- results(analisisdds)
res <- respordefecto[order(respordefecto$padj),]
```

Para ver los primeros datos de esta matriz basta con aplicar la función *head*.

```
> head(res)
log2 fold change (MAP): condition treated vs untreated
Wald test p-value: condition treated vs untreated
DataFrame with 6 rows and 6 columns
             baseMean log2FoldChange     lfcSE      stat        pvalue          padj
            <numeric>      <numeric> <numeric> <numeric>     <numeric>     <numeric>
FBgn0039155  453.2753      -3.714214 0.1600580 -23.20543 4.013332e-119 3.089463e-115
FBgn0029167 2165.0445      -2.082793 0.1035963 -20.10491  6.684462e-90  2.572849e-86
FBgn0035085  366.8279      -2.227243 0.1369744 -16.26028  1.888619e-59  4.846196e-56
FBgn0029896  257.9027      -2.206780 0.1586969 -13.90563  5.854592e-44  1.126716e-40
FBgn0034736  118.4074      -2.565002 0.1847628 -13.88268  8.067450e-44  1.242065e-40
FBgn0040091  610.6035      -1.430433 0.1201539 -11.90501  1.114552e-32  1.429971e-29
```

La interpretación de las columnas de la tabla de resultados obtenida es la siguiente:

- baseMean : media de los conteos normalizados (de todas las muestras de ambas condiciones).
- log2FoldChange :

$$log_2 \frac{\mu_{iA}}{\mu_{iB}}.$$

  Un valor nulo indicará que el gen en cuestión presenta los mismos niveles de expresión en ambos grupos, mientras que de lo contrario, un valor no nulo, indica que el gen se expresa de forma diferente en cada grupo, y según el signo de dicho valor, veremos si presenta un mayor nivel de expresión en el grupo de individuos no tratados o en el grupo de los tratados.
- pvalue : p-valor para el contraste descrito anteriormente.
- padj : p-valor ajustado. Como dijimos anteriormente, tenemos que realizar el contraste para cada uno de los genes, por lo tanto podemos encontrarnos con el problema de comparaciones múltiples. Este problema consiste en que al realizar un gran número de contrastes, se produce un aumento de la probabilidad de obtener falsos positivos y cometer un error de tipo I (rechazar la hipótesis nula, siendo verdadera). El programa ajusta los p-valores mediante el procedimiento de Benjamin-Hochberg.

Para un estudio biológico, lo que interesa es conocer todos los genes diferencialmente expresados. Si consideramos un nivel de significación $\alpha = 0{,}01$, podemos dar la tabla sólo con aquellos genes para los que se ha rechazado la hipótesis nula, es decir, aquellos genes que presentan diferentes niveles de expresión en los grupos. Obtenemos dicha tabla de la siguiente manera:

```
resDF = as.data.frame(res)
difexpDF = resDF[resDF$padj<0.01,]
> dim(difexpDF)
[1] 7190    6
```

Vemos, por la dimensión de la nueva tabla, que para el nivel de significación escogido, hay 7190 genes diferencialmente expresados.

Podemos realizar también una gráfica con la función *plotMA*, la cuál representa los valores de *log2FoldChange* sobre la media de los conteos. Los puntos en rojo pertenecen a aquellos genes cuyo contraste ha rechazado la hipótesis nula, es decir, son los genes diferencialmente expresados.

```
plotMA(res, ylim=c(-2,2), main="Gráfico MA")
```

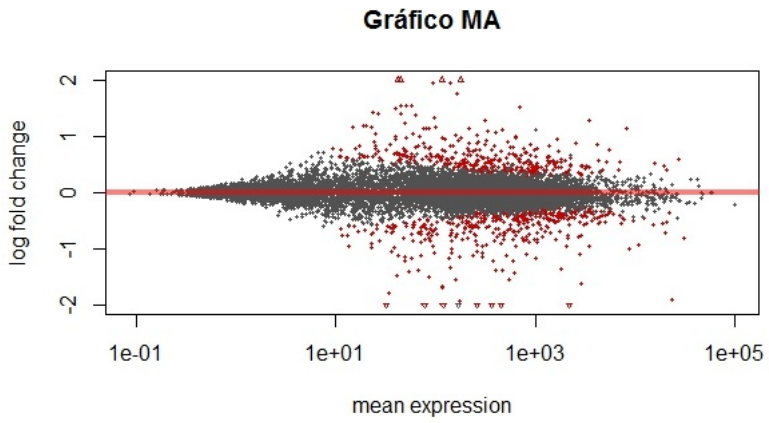

Figura 5.1: plotMA.

Todo esto se ha realizado teniendo en cuenta sólo el factor condición, pero también podemos crear el modelo teniendo en cuenta, además de la condición, el tipo de secuenciación. Aun así, la condición seguirá siendo nuestra variable de interés, y por ello debemos de situarla al final de la fórmula del modelo, como hacemos a continuación:

```
design(dds)<-formula(~ type + condition)
dds<-DESeq(dds)
res<-results(dds)
```

```
> head(res)
log2 fold change (MAP): condition treated vs untreated
Wald test p-value: condition treated vs untreated
DataFrame with 6 rows and 6 columns
               baseMean log2FoldChange      lfcSE        stat     pvalue      padj
              <numeric>      <numeric>  <numeric>   <numeric>  <numeric> <numeric>
FBgn0000003    0.1594687    0.032696497 0.04372641  0.74775164 0.45460998        NA
FBgn0000008   52.2256776    0.012190023 0.20784124  0.05865064 0.95323037 0.9878051
FBgn0000014    0.3897080    0.009706837 0.05629229  0.17243633 0.86309451        NA
FBgn0000015    0.9053584   -0.035666912 0.09317872 -0.38277961 0.70188318        NA
FBgn0000017 2358.2434078   -0.256745586 0.11004554 -2.33308492 0.01964369 0.1338745
FBgn0000018  221.2415562   -0.066692352 0.14172828 -0.47056488 0.63795149 0.8835854
```

5.4. Evaluación de calidad de los datos.

El control de calidad de los datos es un paso fundamental en este tipo de análisis. Nuestro objetivo es la búsqueda de genes diferencialmente expresados, y en particular hemos de ver para muestras cuyo tratamiento experimental ha sufrido alguna anomalía que pueda transformar los datos obtenidos en perjudiciales para nuestro objetivo. Es por ello que realizamos un control de calidad.

5.4.1. Mapa de calor de la tabla de conteos.

A partir de la tabla de conteos podemos crear un mapa de calor que nos proporcionará información sobre los genes que muestran un mayor nivel de expresión en nuestro experimento. Podemos realizar el mapa de calor a partir de los conteos brutos o con los conteos transformados. En primer lugar vemos como obtener el mapa de calor a partir de los datos sin transformar.

```
library("RColorBrewer")
install.packages("gplots")
library("gplots")
select <- order(rowMeans(counts(dds,normalized=TRUE)),
                decreasing=TRUE)[1:30]
hmcol <- colorRampPalette(brewer.pal(9,"GnBu"))(100)

heatmap.2(counts(dds,normalized=TRUE)[select,], col=hmcol,
          Rowv=FALSE, Colv=FALSE, scale="none",
          dendrogram="none", trace="none", margin=c(10,6))
```

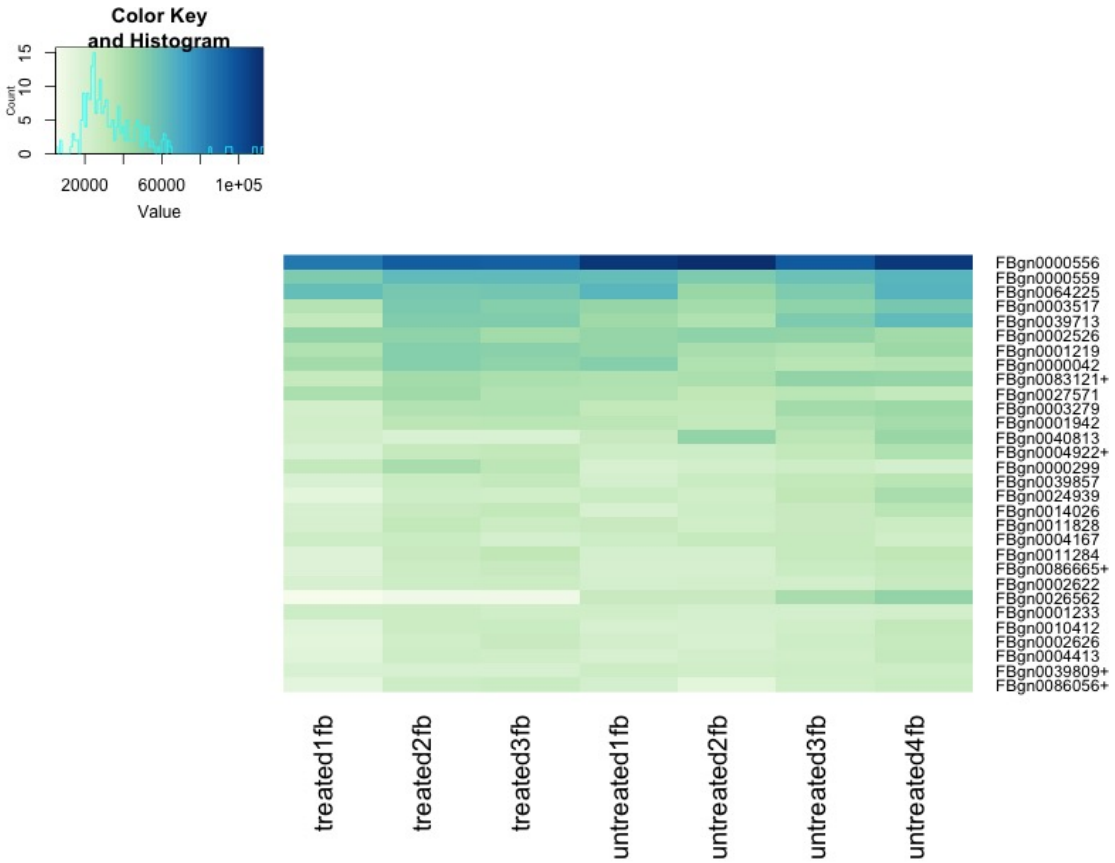

Figura 5.2: Heatmap.

El mapa de calor de la figura 5.2 muestra los conteos de los 30 genes con mayor nivel de expresión. Los genes aparecen en las filas y los individuos en las columnas. Mas específicamente, a mayor valor del conteo más intensidad del color. Aquí vemos claramente que el gen con un mayor número de conteos es el gen identificado como *FBgn0000556*, y en concreto, muestra una mayor expresión en el individuo *untreated2fb*.

Para el análisis diferencial trabajamos con los conteos brutos y usamos distribuciones discretas, sin embargo para análisis posteriores a veces nos resulta más útil trabajar con una transformación de los datos. En este caso elegiremos la función *rlogTransformation* para crear los nuevos datos.

```
rld<-rlogTransformation(dds, blind=TRUE)
```

Al igual que hicimos con los datos brutos, podemos obtener un mapa de calor de los datos transformados tal y cómo mostramos a continuación:

```
heatmap.2(assay(rld)[select,], col=hmcol,
          Rowv=FALSE, Colv=FALSE, scale="none",
          dendrogram="none", trace="none", margin=c(10,6))
```

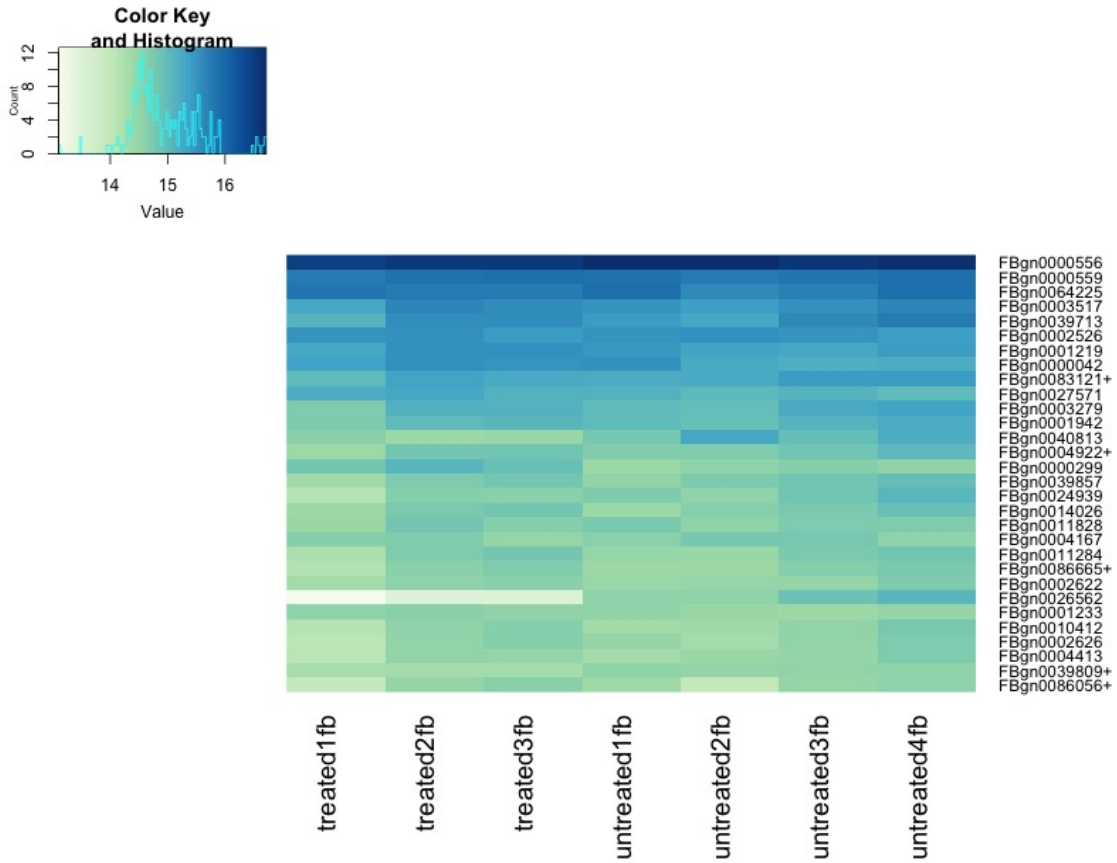

Figura 5.3: Heatmap 2.

La interpretación es la misma que para el caso de los conteos brutos.

5.4.2. Mapa de calor de las distancias entre muestras.

Otro de los usos que podemos dar a los datos transformados es la agrupación de la muestra, es decir, realizar un análisis de conglomerados (cluster). En primer lugar tenemos que aplicar la función *dist* a la transpuesta de la matriz de conteos transformados para obtener las distancias euclídeas entre las muestras.

```
#Distancias.
distsRL<-dist(t(assay(rld)))
#Mapa de calor.
mat <- as.matrix(distsRL)
```

```
> mat
             treated1fb treated2fb treated3fb untreated1fb untreated2fb untreated3fb
treated1fb      0.00000  16.065502  17.783173     18.24376     17.30473     21.43254
treated2fb     16.06550   0.000000   8.735605     19.32895     20.18690     16.77521
treated3fb     17.78317   8.735605   0.000000     20.81629     21.33046     17.16133
untreated1fb   18.24376  19.328951  20.816289      0.00000     15.88635     17.50223
untreated2fb   17.30473  20.186898  21.330465     15.88635      0.00000     15.04018
untreated3fb   21.43254  16.775214  17.161326     17.50223     15.04018      0.00000
untreated4fb   20.94729  15.603203  15.407050     15.58567     15.79314     11.02833
             untreated4fb
treated1fb       20.94729
treated2fb       15.60320
treated3fb       15.40705
untreated1fb     15.58567
untreated2fb     15.79314
untreated3fb     11.02833
untreated4fb      0.00000
```

Tenemos ya la matriz de distancias, por lo que podemos crear a partir de ella el mapa de calor, y además los resultados obtenidos en dicho mapa deben coincidir con los que observamos en la matriz. Es decir, deberemos ser capaces de observar en el mapa que los individuos *untreated1fb* y *untreated3fb* son los que presentan mayor diferencia en los datos, mientras que de lo contrario, *treated2fb* y *treated3fb* son los que muestran datos más semejantes.

```
rownames(mat)<-colnames(mat)<-with(colData(dds),
                    paste(condition,type,sep=":"))
heatmap.2(mat,trace="none",col=rev(hmcol),margin=c(13,13))
```

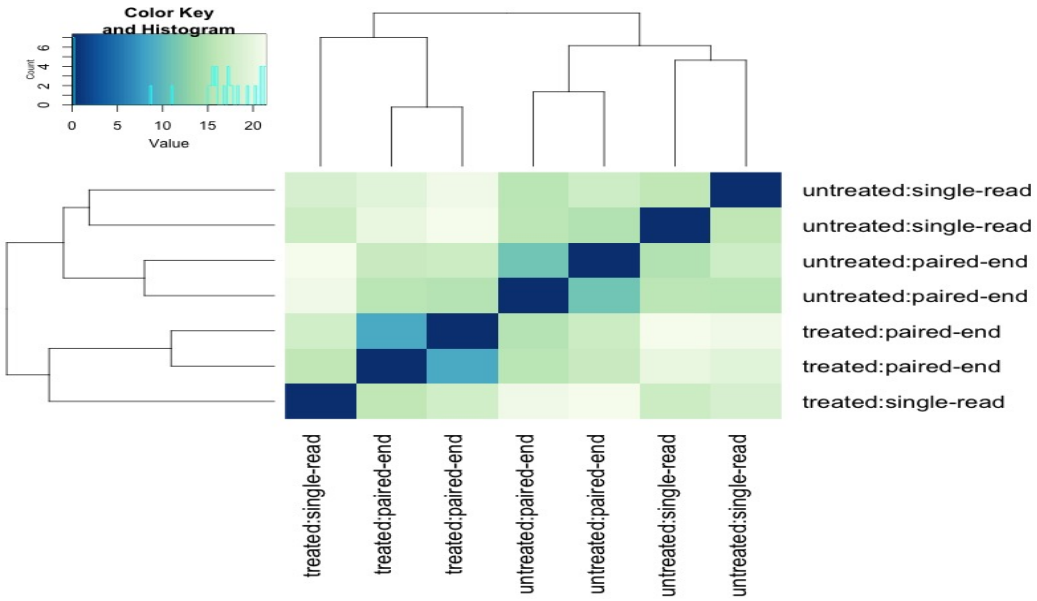

Figura 5.4: Distancias entre muestras.

Como habíamos explicado a partir de la matriz de distancias, vemos en la figura 5.4 los individuos que presentan mayor diferencia entre sus datos según el color que se les otorga en el mapa. Como vemos en el histograma, colores más fuertes indican mayor similitud entre los datos. Además, en el mismo histograma podemos observar los valores de las distancias y cuántos pares de muestras presentan ese valor (están contadas dobles al ser una matriz simétrica).

5.4.3. Componentes principales de las muestras.

Una gráfica de las componentes principales de las muestras resulta útil para ver el efecto total de las covariables así como la posible existencia de efectos lote (batch effects). En la figura 5.5 se observan las 7 muestras representadas utilizando la primera y segunda componente principal, para lo que se ha utilizado el siguiente código:

```
print(plotPCA(rld,intgroup=c("condition","type")))
```

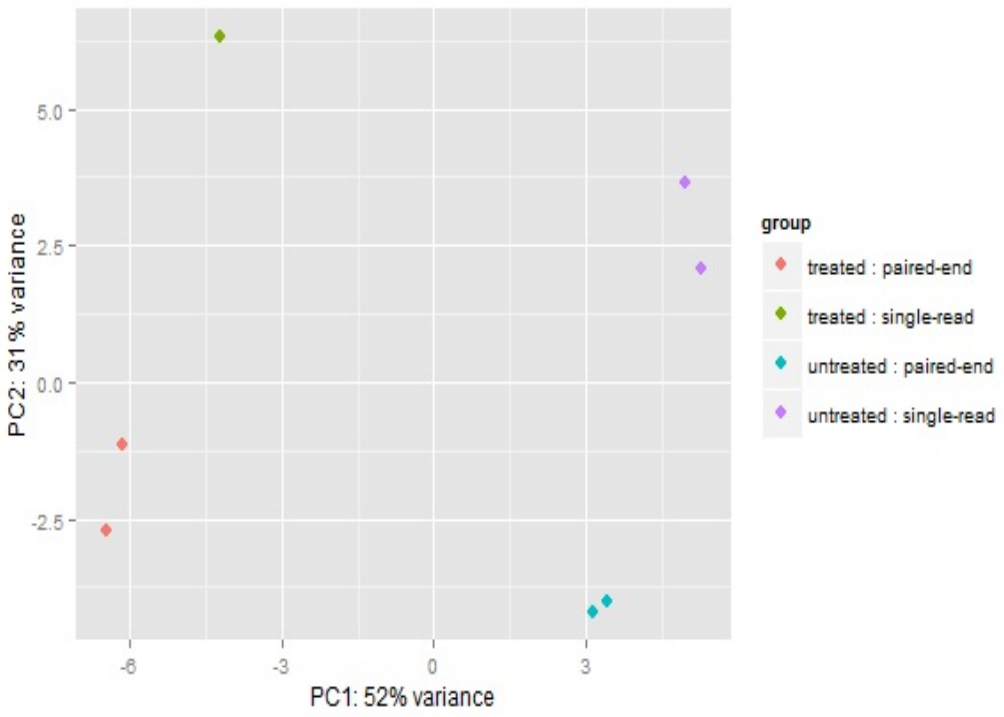

Figura 5.5: Componentes principales.

Para finalizar, destacamos que tras el estudio estadístico se debe realizar por un experto en bioquímica o biomedicina, el llamado análisis de enriquecimiento. Este consiste en ver los procesos biológicos en que participan los genes que se han detectado como diferencialmente expresados en el análisis estadístico.

Bibliografía

S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010. doi: 10.1186/gb-2010-11-10-r106. URL `http://genomebiology.com/2010/11/10/R106/`.

N.E. Breslow. Tests of hypotheses in overdispersed poisson regression and other quasi-likelihood models. *Journal of the American Statistical Association*, 85, 1990.

A.N. Brooks, L. Yang, M.O. Duff, K.D. Hansen, J.W. Park, S. Dudoit, S.E. Brenner, and B.R. Graveley. Conservation of an rna regulatory map between drosophila and mammals. *Genome Research*, 21, 2011.

R.C. Gentleman, V.J. y Care, D.M. Bates, and others. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80, 2004. URL `http://genomebiology.com/2004/5/10/R80`.

J.R. Gonzalez. Course on 'omic'data analysis with R. April 2014.

J.W. Hardin and J.M. Hilbe. *Generalized Linear Models and Extensions*. Stata Press, 2 edition, 2007.

J.M. Hilbe. *Modeling Count Data*. Cambridge University Press, 1 edition, 2014.

M. I Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550, 2014. doi: 10.1186/s13059-014-0550-8. URL `http://dx.doi.org/10.1186/s13059-014-0550-8`.

J.F. Lawless. Negative binomial and mixed poisson regression. *The Canadian Journal of Statistics*, 15, 1987.

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman Hall, 2 edition, 1989.

A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5, 2008.

J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, 135, 1972.

MD. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome Biol*, 11, 2010.