UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Junio 2013, Vol. 6, No. 1

# Comparación de procedimientos FDR para la selección de parámetros en Regresión Poisson

## Comparison of FDR-based procedures to select parameters in Poisson Regression

Jorge Iván Vélez[a]
jorgeivanvelez@gmail.com

Juan Carlos Correa[b]
jccorrea@unal.edu.co

**Resumen**

La selección de variables significativas en modelos de regresión es un problema importante en el trabajo estadístico aplicado. El modelo de Regresión Poisson, útil para describir el número de ocurrencias de un evento particular como función de un conjunto de variables explicativas, ha sido recientemente empleado en biología, epidemiología, genética e ingeniería. En este trabajo se describen el modelo de Regresión Poisson y cuatro procedimientos para la selección de variables explicativas, todos basados en la tasa de falsos descubrimientos (FDR). Adicionalmente, estos procedimientos se comparan mediante un estudio de simulación y se dan algunas recomendaciones. Finalmente, se presenta una aplicación donde se modela el número de madres menores de edad en el Departamento de Antioquia.

***Palabras clave*:** comparaciones múltiples, FDR, Regresión Poisson, selección de variables.

**Abstract**

The selection of significant variables in regression models is an important problem in applied statistics. Poisson Regression, useful when it is of interest to describe the number of occurrences of a particular event as a function of exploratory variables, has recently been used for modeling purposes in biology, epidemiology, genetics and engineering. Here, the Poisson Regression model as well as four procedures to select variables, all of them based on the False Discovery Rate (FDR), are

[a]Translational Genomics Group, Genome Biology Department, John Curtin School of Medical Research, The Australian National University, Canberra, ACT, Australia. Grupo de Neurociencias de Antioquia, Universidad de Antioquia, Colombia. Grupo de Investigación en Estadística, Universidad Nacional de Colombia, sede Medellín.

[b]Grupo de Investigación en Estadística, Universidad Nacional de Colombia, sede Medellín. Profesor Asociado, Escuela de Estadística, Universidad Nacional de Colombia, sede Medellín.

described. In addition, these procedures are compared using a simulation study and some recommendations are given. As reference, the t-based and Bonferroni procedures were used. Finally, we model the number teenagers with children in the Department of Antioquia to illustrate these methods.

***Key words*:** multiple Testing, FDR, Poisson Regression, Variable Selection.

# 1. Introducción

Recientemente, gracias a los avances en genética y procesamiento de imágenes, es común encontrar aplicaciones en las que un conjunto de $m \to \infty$ variables predictoras influencian una única variable respuesta (Golub et al. 1999), Mootha03, por lo que la selección de cuáles variables deben incluirse en el modelo de regresión es fundamental. Una alternativa para seleccionar un subconjunto de $k < m$ variables explicativas consiste en realizar $m - 1$ pruebas de hipótesis independientes sobre los coeficientes del modelo ajustado. Sin embargo, el problema principal con esta metodología es que la tasa de falsos positivos (incluir una variable cuando su efecto es nulo) se incrementa drásticamente (Shaffer 1995). Algunas soluciones a este problema han sido propuestas, entre otros, por Efron (2004), Nguyen (2005) y Ghosh et al. (2006).

Efron (2004) propone utilizar la distribución empírica de los valores-$p$ e ilustra su metodología utilizando un modelo de regresión logística para determinar, entre seis tipos de drogas inhibidoras protéicas suministradas a 1391 pacientes con VIH, cuáles causan mutaciones en el genoma viral. Nguyen (2005), también, en regresión logística, compara, vía simulación, cuatro procedimientos basados en la tasa de falsos positivos o *False Discovery Rate* (FDR)(Benjamini & Hochberg 1995) para la selección de variables predictoras. Como ilustración, presenta datos en los que es de interés detectar rastros moleculares en 15 pacientes con cáncer de mama y durante la iniciación y progreso de cáncer de colon en 59 ratas. En ambos casos, se tienen más de 3000 variables predictoras (niveles de expresión de genes). Ghosh et al. (2006) proponen utilizar la FDR como un criterio para la selección de variables en un modelo de regresión lineal múltiple con variable respuesta normal.

La Regresión Poisson es útil en biología, epidemiología, genética e ingeniería para modelar variables respuesta que representan un conteo. Algunos ejemplos incluyen el número de sustituciones de nucleótidos que ocurren en un gen en un período de tiempo determinado, patrones de migración de especies, el número de casos con determinada enfermedad a lo largo del territorio colombiano o el número de fallas de un dispositivo electrónico (Correa & Castrillón 2006).

En este trabajo se presentan los conceptos fundamentales del modelo de Regresión Poisson y cuatro procedimientos basados en la FDR para la selección de variables. Asimismo, se realiza un estudio de simulación para comparar el desempeño de estos procedimientos en este modelo de regresión. Como referencia se utilizaron los procedimientos basados en la distribución t y Bonferroni (Bonferroni 1935).

Finalmente, como ilustración de estos procedimientos se presenta una aplicación en la que se modela el número de madres menores de edad en el departamento de Antioquia.

## 2. Regresión Poisson

El Modelo Lineal Generalizado (MLG) fue introducido por Nelder & Wedderburn (1972) como una generalización del modelo lineal clásico. A esta clase de modelos pertenecen, entre otros, el modelo lineal clásico, la Regresión Poisson, la Regresión Gamma y la Regresión Logística (Myers et al. 2001).

Si $Y$ es una variable aleatoria con media $\mu$, el uso del MLG está condicionado a que su función de masa (o densidad) de probabilidad pertenezca a la familia exponencial de distribuciones, es decir, que pueda escribirse como

$$f(y|\theta,\phi) = \exp\{\phi[y\theta - b(\theta) + c(y,\phi)]\} \tag{1}$$

donde $\theta$ es una función de $\mu$, $b(\theta)$ y $c(\theta)$ son funciones diferenciables y $\phi^{-1}$ es el parámetro de dispersión.

Este tipo de modelos permite ajustar una función de la media, cuya forma es:

$$g(\mu_i) = \mathbf{x}_i'\beta \quad i = 1, 2, \cdots, n \tag{2}$$

donde $g(\mu_i)$ es una función monótona y diferenciable, usualmente llamada función de enlace (Nelder & Wedderburn 1972), (McCullagh & Nelder 1983), (Myers et al. 2001), $\mu_i$ es la media de $y_i$, $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)$ es el vector de parámetros del modelo y $\mathbf{x}i = (x_{i1}, x_{i2}, \cdots, x_{ip})'$ es un conjunto de covariables.

Decimos que $Y$ tiene una distribución Poisson con parámetro $\lambda > 0$, $Y \sim Poisson(\lambda)$, si su función de masa de probabilidad está dada por (**?**)

$$P(Y = y|\lambda) = f(y\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} \quad y = 0, 1, \cdots \tag{3}$$

La distribución Poisson pertenece a la clase exponencial de distribuciones (1) con $\theta = \log(\lambda)$, $b(\cot) = \exp\{\lambda\}$, $c(\cdot) = -\log(y!)$ y $\phi = 1$. Una propiedad importante de esta distribución es que la media es igual a la varianza, esto es, $E_\lambda[Y] = V_\lambda[Y] = \lambda$. Algunas aplicaciones de la distribución Poisson incluyen la descripción de ocurrencias aleatorias en el tiempo o en el espacio (Myers et al. 2001), (Correa & Castrillón 2006), como por ejemplo el número de personas que no cumplen sus obligaciones en una entidad financiera (entran en mora), el número de llamadas que llegan a una central telefónica o el número de accidentes de tránsito por hora.

La Regresión Poisson es útil cuando estas ocurrencias se ven influenciadas por un conjunto de covariables $\mathbf{x} = (x_1, x_2, \cdots, x_k)'$ y es de interés cuantificar su impacto

sobre $\lambda$ (Myers et al. 2001, McCulloch & Searle 2001). El modelo de Regresión Poisson puede escribirse como:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_k x_{i,k} + \epsilon_i \quad i = 1, 2, \cdots, n \tag{4}$$

donde $\beta = (\beta_0, \beta_1, \cdots, \beta_k)$ es el vector de coeficientes y $\epsilon$ es el error aleatorio. Los estimadores de máxima verosimilitud para $\beta$ se obtienen al resolver la ecuación

$$y_i x_i - x_i \exp\{x_i' \beta\} = 0 \tag{5}$$

utilizando el método Newton-Raphson iterativo. La matriz de varianzas covarianzas estimada para $\hat{\beta}$ es $\hat{\Sigma} = (\mathbf{X}^T \hat{\mathbf{V}} \mathbf{X})^{-1}$, con $\hat{\mathbf{V}} = diag\hat{\mu}_i$. Con estos dos componentes es posible hacer inferencia sobre los parámetros del modelo (Paula 2009).

# 3. Pruebas de hipótesis múltiples y procedimientos basados en la FDR

Suponga que se realizan $m$ pruebas de hipótesis independientes de la forma

$$H_{0,i}: \ \theta_i \in \Theta \quad \text{vs.} \quad H_{1,i}: \ \theta_i \notin \Theta \quad i = 1, 2, \cdots, m \tag{6}$$

donde $\theta_i$ es algún parámetro de interés y $\Theta$ un espacio parametral. Sea $\alpha \in (0, 1)$ la probabilidad de error tipo I a la que se prueba la $i$-ésima hipótesis, es decir, la probabilidad de rechazar $H_{0,i}$ cuando en efecto es cierta, y $Pi \in (0, 1)$ el valor-p de la prueba dado por

$$P_i = 1 - G(T_i) \quad i = 1, 2, \cdots, m \tag{7}$$

donde $T_i$ el estadístico de prueba para la $i$-ésima hipótesis y $G$ su función de distribución acumulada.

Las pruebas de hipótesis múltiples hacen referencia a probar varias hipótesis independientes a la vez (Correa 2010), cada una con probabilidad de error tipo I $\alpha$. Uno de los principales problemas que se presentan es el incremento de la tasa de error a lo largo de la familia o Family-Wise ErrorRate (FWER), definida como la probabilidad de proporcionar al menos un falso positivo (rechazar $H_0$ cuando esta es cierta) en todas las hipótesis que se prueban (Shaffer 1995). Es fácil mostrar que esta probabilidad está dada por

$$\alpha_m = 1 - (1 - \alpha)^m \tag{8}$$

y que, independiente del valor de $\alpha$, esta incrementa considerablemente cuando $m \to \infty$ (ver Figura 1).

En la actualidad, en áreas como procesamiento de imágenes, genética y análisis funcional se realizan miles (y hasta millones) de pruebas de hipótesis a la vez, por lo

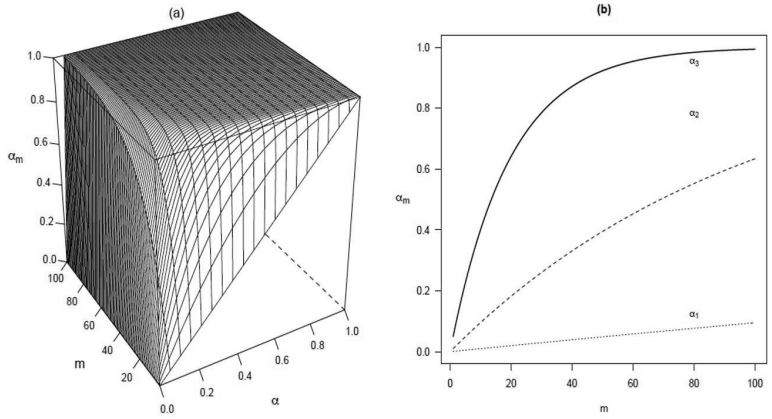


Figura 1: *(1)* $\alpha_m$ *como función de* $m$ *y* $\alpha$*. En (b),* $1 < m < 100$ *y* $0 < \alpha_1 < \alpha_2 < \alpha_3 < 1$*. Fuente: elaboracion propia.*

Tabla 1: *Posibles resultados cuando se prueban* $m$ *hipótesis. Fuente: elaboración propia.*

| | Acepto $H_0$ | Rechazo $H_0$ | Total |
|---|---|---|---|
| $H_0$ Verdadera | $U$ | $V$ | $m_0$ |
| $H_0$ Falsa | $T$ | $S$ | $m_1$ |
| Total | $W$ | $R$ | $m$ |

que el control de la FWER es fundamental. Si dicho control no se realiza, H0 sería rechazada muchas más veces y se obtendría un gran número de falsos positivos (descubrimientos que no son ciertos). Aunque en la literatura existen varias propuestas para resolver este problema (Bonferroni 1935, Wright 1992, (Benjamini & Hochberg 1995, Benjamini & Yekutieli 2001, **?**, **?**, **?**)), no existe una única solución para todas las situaciones (Correa 2010). Dos de los métodos más utilizados son el de Bonferroni (Bonferroni 1935) y la FDR (Benjamini & Hochberg 1995). En el primero, la probabilidad de error tipo I es $\alpha_B = \alpha/m$ en lugar de $\alpha$; en el segundo se controla la proporción $\alpha$ de hipótesis nulas falsamente rechazadas, relativo al número total de hipótesis rechazadas (Correa 2010). En ambos procedimientos, la FWER$\leqslant \alpha$.

## 3.1. Tasa de Falsos Positivos (FDR)

Cuando se realizan $m$ pruebas de hipótesis independientes, los posibles resultados se presentan en la Tabla 1, con $U$ no descubrimiento, $T$ no descubrimiento falso, $V$ falso descubrimiento, $S$ descubrimiento (rechazo de $H_0$).

La FDR está definida como la proporción de hipótesis nulas verdaderas que resul-

tan ser rechazadas dentro del total de hipótesis rechazadas (Benjamini & Hochberg 1995, Benjamini et al. 2006), esto es:

$$FDR = E\left(\frac{V}{R} I_{R>0}\right) = E\left(\frac{V}{R} \mid R > 0\right) Pr(R > 0) \tag{9}$$

donde $R$ es el número total de rechazos de $H_0$ e $I_A$ denota una función indicadora para el evento $A$ (rechazos de $H_0$). En (9) solo son observables $R$, $W$ y $m$; la FWER corresponde a $Pr(V > 0)$.

El procedimiento propuesto por Benjamini & Hochberg (1995) (BH-FDR de aquí en adelante) consiste en:

1. Probar $H_{0,1},\ H_{0,2}, \cdots, H_{0,m}$ y obtener los valores-$p$ $p_1,\ p_2, \cdots, p_m$.
2. Ordenar los valores-$p$ como $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
3. Estimar $\kappa$ como
$$\hat{\kappa} = \text{máx}\left\{i :\ p_{(i)} \leq \frac{i}{m}\alpha\right\} \tag{10}$$
y rechazar $H_{0,1}, H_{0,2}, \cdots, H_{0,\hat{\kappa}}$. Si no existe tal $i$, ninguna hipótesis nula podrá ser rechazada.

## 3.2. FDR menos conservadora (LC-FDR)

Si $\pi_0 = m0/m$ es la proporción de hipótesis nulas que son verdaderas, el procedimiento BH-FDR proporciona FDR$\leq \alpha$ puesto que $\pi_0 \in (0, 1)$. Ahora, si $\pi_0$ es conocido, la BH-FDR puede aplicarse a un nivel $\alpha' = \alpha/\pi_0$ en lugar de $\alpha$.

Nguyen (2005) define la LC-FDR y encuentra

$$\hat{\kappa}_{LC} = \text{máx}\left\{i :\ p_{(i)} \leq \frac{i}{m_0}\alpha\right\} \tag{11}$$

y rechazando las hipótesis nulas $H_{0,1}, H_{0,2}, \cdots, H_{0,\hat{\kappa}_L C}$.

Este procedimiento es poco utilizado en la práctica porque $\pi_0$ es desconocido. Sin embargo, mejora la precisión de la estimación de la FDR y la potencia para detectar hipótesis alternativas verdaderas (Nguyen 2005).

## 3.3. FDR en dos etapas (2S-FDR)

Benjamini et al. (2006) (BKY) proponen un procedimiento FDR de 2 etapas que consiste en:

1. Calcular $r_1$ como el número de rechazos al aplicar el procedimiento BH-FDR a un nivel $\alpha' = \alpha/(1 + \alpha)$.

2. Estimar $\pi_0$ como $\hat{\pi}_0(BKY) = (m - r_1)/m$.

3. Aplicar el procedimiento BH-FDR de nuevo a los mismos datos del paso 1, pero a un nivel $\alpha'/\hat{\pi}_0(BKY)$, de manera que

$$\hat{kappa}_{BKY} = \text{máx}\left\{i:\ p_{(i)} \leq \frac{i}{m - r_1}\alpha'\right\} \tag{12}$$

y se rechazan las hipótesis $H_{0,1}, H_{0,2}, \cdots, H_{0,\hat{\kappa}_{BKY}}$.

Observe que si $r_1 = 0$ no se rechaza ninguna hipótesis nula, mientras que si $r_1 = m$ todas las $m$ hipótesis son rechazadas y que, en ambos casos, el procedimiento termina en el paso 1. Benjamini & Hochberg (1995) y Benjamini et al. (2006) muestran que en la 2S-FDR, la FDR$\leq \alpha\ \forall \pi_0 \in [0, 1]$.

### 3.4. FDR en dos etapas modificada (2SM-FDR)

El procedimiento FDR modificado de 2 etapas (Benjamini et al. 2006) consiste en determinar

$$\hat{\kappa}_{BKY-M} = \text{máx}\left\{i:\ p_{(i)} \leq \frac{i}{m}\left(\frac{\alpha'}{\hat{\pi}_0(BKY - M)}\right)\right\} \tag{13}$$

y rechazar las hipótesis $H_{0,1}, H_{0,2}, \cdots, H_{0,\hat{\kappa}_{BKY-M}}$, donde $\hat{\pi}_0(BKY - M)$ es la estimación de $\pi_0$ en el paso 1 del procedimiento 2S-FDR a un nivel $\alpha$.

## 4. Estudio de simulación

Utilizando $\alpha = 0.05$ como la probabilidad de error tipo I nominal, se realizó un estudio de simulación para comparar la tasa de rechazos (el porcentaje de veces que el parámetro es diferente de cero) de los cuatro procedimientos basados en la FDR. Los procedimientos basados en la distribución $t$ y Bonferroni se utilizaron como referencia.

El algorítmo en R (R Development Core Team 2011) utilizado opera de la siguiente manera:

1. Defina el tamaño de muestra $n$ y el número de variables $k$ con $n > k$ y genere la matriz de diseño $X_{n\times k}$ (fija) de columnas ortogonales.

2. Para cada fila de $X_{n\times k}$, genere $B = 10000$ observaciones de una distribución Poisson con media $\mu_i = \exp\left(\sum_{j=1^k} x'_{ij}\beta_j\right)$ y construya la matriz de respuestas $\mathbf{Y}_{ib}$, $b = 1, 2, \cdots, B$. Los coeficientes $\beta_j$, $j = 1, 2, \cdots, k$, son fijos.

3. Ajuste, para la matriz $X_{n\times k}$ y cada vector $\mathbf{Y}_{n\times b}$, una Regresión Poisson. Extraiga el valor-$p$ asociado al coeficiente $\beta_j$ y denótelo $p_j$, $j = 1, 2, \cdots, k$.

4. Calcule las tasas de rechazo (TR) de la prueba de hipótesis $H_{0,j}: \beta_i = 0$, para los procedimientos $t$, Bonferroni, BH-FDR, LC-FDR, 2S-FDR y 2SM-FDR. En los dos primeros, se rechaza $H_{0,j}$ si $p_j \leq \alpha$ y $p_j \leq \alpha/k$, respectivamente. En los procedimientos FDR, se rechaza $H_{0,k}$ de acuerdo con el valor de $\kappa$ (ver Sección 3.1-3.4). Las TR se calcularon como el porcentaje de veces que $H_0$ fue rechazada en las $B$ simulaciones.

Cuando en el modelo (4) se tienen dos covariables (ver Figura 2), las TR de todos los procedimientos son similares excepto para FDR-LC y FDR-2SM; en el primero, las TR corresponden a las más altas, en el segundo, estas son mucho mayores que las demás cuando $\beta > 0$. Por otro lado, el tamaño muestral $n$ y las TR parecen estar inversamente relacionados, por lo que en muestras grandes se necesitaría un mayor efecto de $x_k$ $(k = 1, 2)$ sobre la respuesta para incluirla (rechazar $H_{0,k}$) en el modelo final. La FWER $\leq \alpha$ excepto para el procedimiento FDR-LC.

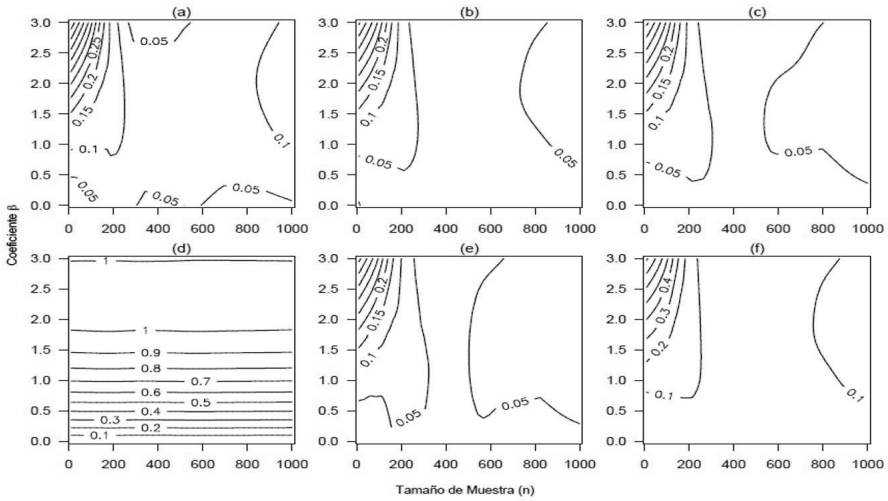


Figura 2: *Tasa de rechazos de los procedimientos (a) t, (b) Bonferroni, (c) FDR-BY, (d) FDR-LC, (e) FDR-2S y (f) FDR-2SM para $m = 2$ variables. Fuente: elaboracion propia.*

En la Figura 3 se presentan los resultados para $m = 5$ variables. Independiente de los valores de $\beta$ y $n$, los procedimientos Bonferroni, FDR-BY y FDR-2S presentan TR similares y controlan la FWER al nivel nominal. Sin embargo, estos presentan TR bajas cuando $\beta \neq= 0$. Si bien estos resultados son esperados (ver Sección 3), llama la atención que en estos procedimientos el tamaño de muestra no tenga ningún efecto sobre la TR. Por otro lado, los procedimientos FDR-LC, FDR-2SM y $t$, en ese orden, presentan las TR más altas; las TR de estos últimos aumentan conforme aumenta el valor de $\beta$.

Para $m = 10$, los TR se presentan en la Figura 4. En este caso, el tamaño de muestra parece tener poca (o ninguna) influencia sobre los valores de la TR y solo los

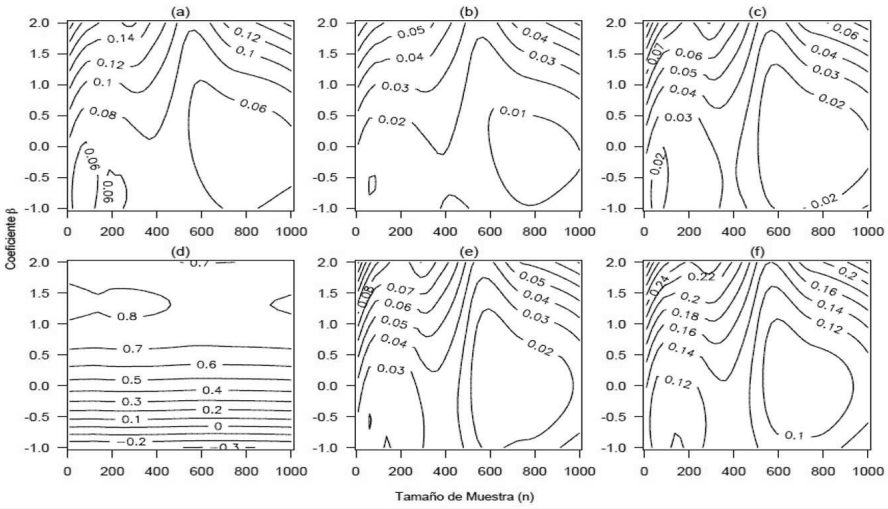


Figura 3: *Tasa de rechazos de los procedimientos (a) t, (b) Bonferroni, (c) FDR-BY, (d) FDR-LC, (e) FDR-2S y (f) FDR-2SM para* $m = 5$ *variables. Fuente: elaboracion propia.*

procedimientos Bonferroni, FDR-BY y FDR-2S parecen controlar adecuadamente la FWER cuando $\beta = 0$. Llama particularmente la atención que para $\beta > 1$ las TR más altas se obtengan para los procedimientos $t$ y FDR-2S y las más bajas con Bonferroni y FDR-BY; este último presenta TR ligeramente mayores. En general, los procedimientos FDR-LC y FDR-2SM presentan las TR más altas.

Al considerar $m = 50$ covariables en el modelo 4 (ver Figura 5), se obtuvo que los procedimientos FDR-BY y FDR-2S se comportan de manera similar, Bonferroni presenta las TR más bajas, y FDR-2SM, FDRLC y $t$, en ese orden, presentan las TR más altas. En cuanto al control de la FWER, solo los procedimientos Bonferroni, FDR-BY y FDR-2S lo realizan satisfactoriamente.

Al comparar el efecto de $m$ sobre las TR (ver Sección 3), este es mayor en los procedimientos $t$, FDR-LC y FDR-2SM. Como consecuencia, la FWER no se controla adecuadamente y las TR aumentan conforme aumenta $m$.

# 5. Ilustración

A partir de información del *Anuario Estadístico de Antioquia* de 1993, se estimó un modelo de Regresión Poisson para el número de madres menores de edad en el departamento de Antioquia. El modelo propuesto es:

$$\log(Y_i) = \beta_0 + \beta_1 I_{mi} + \beta_2 NBI_i + \beta_3 Region_i + \log(PM_i/1000) + \epsilon_i \qquad (14)$$

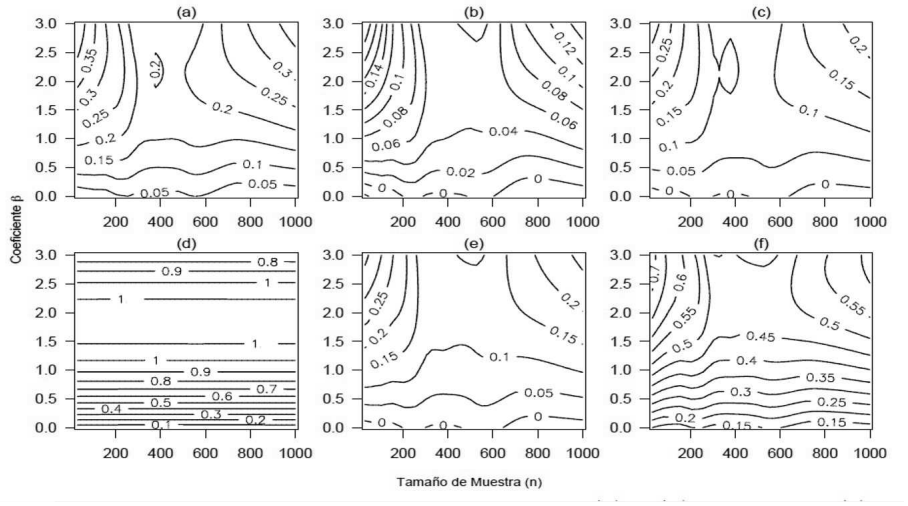


Figura 4: *Tasa de rechazos de los procedimientos (a) t, (b) Bonferroni, (c) FDR-BY, (d) FDR-LC, (e) FDR-2S y (f) FDR-2SM para* $m = 10$ *variables. Fuente: elaboracion propia.*

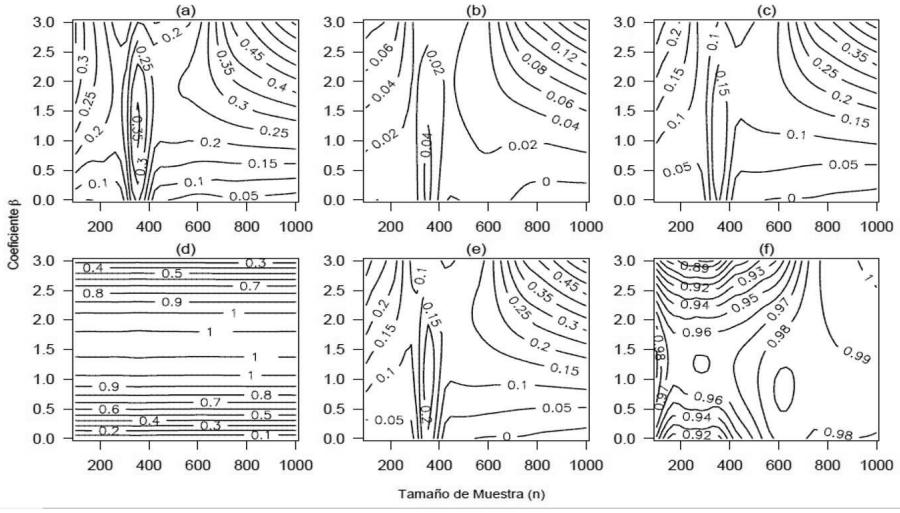


Figura 5: *Tasa de rechazos de los procedimientos (a) t, (b) Bonferroni, (c) FDR-BY, (d) FDR-LC, (e) FDR-2S y (f) FDR-2SM para* $m = 50$ *variables. Fuente: elaboracion propia.*

donde $Y$ es el número de mujeres menores de edad con hijos nacidos vivos, $I_m$ es

Tabla 2: *Parámetros estimados para el modelo de Regresión Poisson. Fuente: elaboracion propia.*

| Variable | $\hat{\beta}$ | $SE(\hat{\beta})$ | z | Valor-$p$ |
|---|---|---|---|---|
| Intercepto | 2.5404 | 0.2623 | 9.68 | 0.0000 |
| $I_m$ | 0.8644 | 0.2813 | 3.07 | 0.0021 |
| NBI | 0.0106 | 0.0008 | 13.79 | 0.0000 |
| Magdalena | -0.1555 | 0.0603 | -2.58 | 0.0099 |
| Norte | -0.4897 | 0.0424 | -11.55 | 0.0000 |
| Nordeste | -0.2362 | 0.0411 | -5.74 | 0.0000 |
| Occidente | -0.4476 | 0.0435 | -10.30 | 0.0000 |
| Oriente | -0.6036 | 0.0395 | -15.27 | 0.0000 |
| Suroriente | -0.3625 | 0.0385 | -9.42 | 0.0000 |
| Urabá | 0.2200 | 0.0303 | 7.26 | 0.0000 |
| Valle de Aburrá | -0.5244 | 0.0535 | -9.80 | 0.0000 |

el índice de masculinidad definido como

$$I_m = \frac{\text{Número de Hombres}}{\text{Número de Mujeres}}, \tag{15}$$

NBI es el índice de necesidades básicas insatisfechas, Región es la región a la que pertenece el municipio, PM es la población de mujeres en el municipio y $\epsilon$ es el error aleatorio. El término $\log(PMi/1000)$ en (14) es una variable *offset* que cuantifica, por cada mil, el número de madres en el municipio $i$. En total se consideraron 124 municipios. Como nivel de referencia se seleccionó la región del Bajo Cauca.

Los coeficientes del modelo ajustado se presentan en la Tabla 2. De acuerdo con nuestros resultados, el número de madres menores de edad por cada 1000 madres en las regiones del Bajo Cauca y Urabá es mayor que en otras regiones del departamento de Antioquia; dicho número está positivamente correlacionado con los indicadores $I_m$ y NBI.

Cuando se utiliza un nivel de significancia $\alpha = 0.05$, las pruebas $t$, FDR-BY y FDR-LC seleccionan todos los coeficientes, mientras Bonferroni no considera el efecto de la región del Magdalena; esto indica que los modelos para esta región y la región de referencia serían equivalentes. Los procedimientos FDR-2S y FDR-2SM no rechazan ninguno de los coeficientes.

# 6. Conclusiones

Independientemente del valor del coeficiente $\beta$, las tasas de rechazo en los procedimientos Bonferroni, FDR-BY y FDR-2S no aumentan considerablemente con el número de variables en el modelo de Regresión Poisson. Desafortunadamente lo opuesto ocurre con el procedimiento $t$, uno de los más utilizados.

En situaciones prácticas donde sea necesaria la estimación y posterior selección de parámetros en modelos de Regresión Poisson, se recomienda utilizar los procedimientos FDR-BY y FDR-LC en paralelo con Bonferroni, independientemente del número de variables regresoras. El procedimiento $t$ no debería considerarse.

Futuros trabajos podrían estar direccionados a determinar el desempeño de los procedimientos aquí evaluados cuando existe algún grado de dependencia entre las covariables o cuando el número de variables sea mayor que el número de observaciones.

**Recibido: 30 de noviembre de 2012**
**Aceptado: 11 de marzo de 2013**

## Referencias

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistial Society* **57**(1), 289–300.

Benjamini, Y., Krieger, A. M. & Yekutieli, D. (2006), 'Adaptive linear stepup procedures that control the false discovery rate', *Biometrika* **93**(3), 491–507.

Benjamini, Y. & Yekutieli, D. (2001), 'The control of the false discovery rate in multiple testing under dependency', *Annals of Statistics* **29**, 1165–1188.

Bonferroni, C. E. (1935), 'Il calcolo delle asicurazioni su gruppi di teste', *Studi in Onore del Professore Salvatore Ortu Carboni,* pp. 13–60.

Casella, G. & Berger, R. (2001), *Statistical Inference*, Duxbury Press.

Correa, J. C. (2010), 'Diagnósticos de regresión usando la fdr (tasa de descubrimientos falsos)', *Comunicaciones en Estadística* **3**(2), 109–118.

Correa, J. C. & Castrillón, F. (2006), 'Comparación por intervalos entre diferentes métodos de estimación de la media de la distribución poisson', *Revista Universidad EAFIT* **42**(144), 81–98.

Efron, B. (2004), 'Large-scale simultaneous hypothesis testing: The choice of a null hyphotesis', *Journal of the American Statistical Association* **99**(465), 96–104.

Ghosh, D., Chen, W. & Raughunathan, T. (2006), 'The false discovery rate: A variable selection perspective', *Journal of Statistical Inference* **136**(8), 2668–2684.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. & Lander, E. (1999), 'Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring', *Science* **286**, 531–537.

McCullagh, P. & Nelder, J. A. (1983), *Generalized Linear Models*, Chapman & Hall, London.

McCulloch, C. & Searle, S. (2001), *Generalized, Linear and Mixed Models*, John Wiley & Sons, United States.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstroale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. & Groop, L. C. (2003), 'Pgc-1$\alpha$-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes', *Nat. Genet.* **34**(3), 267–273.

Myers, R., Montgomery, D. & Vining, G. (2001), *Generalized Linear Models: With Applications in Engineering and the Sciences*, John Wiley & Sons, Netherlands.

Nelder, J. A. & Wedderburn, R. M. (1972), 'Generalized linear models', *Journal of the Royal Statistical Society, Series A* **135**, 370–384.

Nguyen, D. V. (2005), 'A unified computational framework to compare direct and sequential false discovery rate algorithms for exploratory dna microarray studies', *Journal of Data Scie* **3**, 331–352.

Paula, G. (2009), *Modelos de Regressão com Apoio Computacional em Splus e R.*, Instituto de Matemática e Estadística Universidade de Sao Paulo, IME-USP.

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. *http://www.R-project.org

Shaffer, J. P. (1992), 'Adjusted p-values for simultaneous inference', *Biometrics* **48**, 1005–1013.

Shaffer, J. P. (1995), 'Multiple hypothesis testing', *Annual Review of Psychology* **46**(2), 561–576.