UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2010, Vol. 3, No. 2

Diagnósticos de Regresión Usando la FDR (Tasa de Descubrimientos Falsos)

Regression Diagnostics Using the FDR Technique

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

Resumen

Proponemos el uso de las pruebas Tasa de Descubrimientos Falsos (False Discovery Rate, FDR), test en lugar de los punto de cortes tradicionales utilizados en diagnósticos de regresión para detectar observaciones sospechosas. Este procedimiento disminuye la complejidad de los diagnósticos mediante la reducción del conjunto de puntos para ser considerados para análisis posteriores, manteniendo sólo aquéllos que son realmente extraños.

***Palabras clave*:** FDR, diagnósticos de regresión, pruebas múltiples.

Abstract

We use False Discovery Rate (FDR) tests instead of traditional cutoff values used in regression diagnostics to detect suspicious observations and control the rate of false discoveries. This method reduces the complexity of diagnostics by reducing the set of data points to be considered for further analysis, keeping only those that are really extraneous.

***Key words*:** FDR, Regression Diagnostics, Multiple Tests.

1. Introducción

En la construcción de modelos estadísticos, la determinación de observaciones que puedan tener efectos importantes, indeseables o no, es una parte fundamental. Esto ha dado origen a un área completa conocida como *diagnósticos*, en la cual se han desarrollado herramientas para medir el impacto que en la estimación del modelo, tanto global como a nivel de parámetros individuales, tiene cada observación

[a]Profesor Asociado. Universidad Nacional de Colombia - Sede Medellín.

(Belsley, Kuh y Welsch; 1980). Esta metodología consiste en realizar simultáneamente tantas pruebas de hipótesis como observaciones tengamos. El área de pruebas múltiples es conocida entre los usuarios de diseño experimental y allí son bien conocidos los problemas de éstas. Con respecto a lo anterior, Donoho y Jin (2004) presentan esta anécdota de Tukey:

> En sus notas de clase para Estadística 411 en Princeton University en 1976, Tukey introdujo la noción de *mayor criticismo* por medio de una historia. Un joven sicólogo administra muchas pruebas como parte de un proyecto, y encuentra que, de 250 pruebas 11 fueron significativas a un nivel del 5 %. El joven investigador se siente muy orgulloso de esto y está dispuesto a que todo el mundo se entere, hasta que un investigador mayor (Tukey mismo?) sugirió que uno esperaría 12.5 pruebas significativas aún en el caso nulo puro, simplemente por azar. En ese sentido, hallar solo 11 resultados significativos es algo descorazonador.

Pruebas múltiples hace referencia al concepto de probar más de una hipótesis a la vez. Es una sub-área de una mucho mayor conocida como inferencia múltiple, o inferencia simultánea, que incluye tanto la estimación como pruebas múltiples. Cuando muchas pruebas son verificadas, y cada una tiene una probabilidad preespecificada de error de Tipo I, la probabilidad de cometer algún error de Tipo I se incrementa, a menudo de manera drástica, a medida que se incrementa el número de hipótesis. En este campo hay numerosas aproximaciones para resolver este problema que han sido propuestas (Dudoit y van der Laan, 2008). Ninguna solución es aceptable para todas las situaciones. Algunos métodos de comparación múltiples controlan la tasa del error de Tipo I solo cuando todas las hipótesis nulas sean ciertas; otros controlan esta tasa para cualquier combinación de hipótesis ciertas y falsas. Éstas son llamadas de control débil y control fuerte.

Una prueba de hipótesis individual a un nivel de significancia α tiene una probabilidad α de rechazar la hipótesis nula cuando es en efecto cierta. Si se ejecutan n de tales pruebas cuando todas las hipótesis nulas son ciertas, entonces el número promedio de pruebas para las cuales la hipótesis nula es rechazada falsamente es $n\alpha$. Si se realizan 1000 pruebas, entonces esperamos rechazar 50 hipótesis nulas que son ciertas, lo que realmente es alto.

Tukey (Donoho y Jin, 2004) propuso el siguiente estadístico dentro de las llamadas pruebas de hipótesis de segundo nivel

$$HC_{0.05,n} = \sqrt{n}\frac{(\text{Fracción significativa a } 0.05) - 0.05}{\sqrt{0.05 \times 0.095}}$$

y sugirió que los valores de 2 o mayores indicaban un tipo de significancia del cuerpo total de pruebas. Donoho y Jin (2004) generalizan este estadístico

$$HC_n^* = \text{máx}_{0<\alpha\leq\alpha_0}\, \sqrt{n}\frac{(\text{Fracción significativa a } \alpha) - \alpha}{\sqrt{\alpha \times (1-\alpha)}}$$

Antes de realizar cualquier prueba, por los métodos tradicionales, debemos primero escoger la probabilidad nominal α de rechazar erróneamente cualquier hipótesis nula particular. A α lo podemos llamar la tasa de falsos positivos (FPR).

- La *tasa de error por hipótesis (PCE: error rate per hypothesis ó comparison-wise error rate)* se define para cada hipótesis como la probabilidad del error de Tipo I o, cuando el número de hipótesis es finito, el PCE promedio puede definirse como el valor esperado del

$$\alpha_C = \frac{\text{Número de Rechazos Falsos}}{\text{Número de Hipótesis}}.$$

  Esta tasa hace referencia desde el punto de vista frecuentista a todas las repeticiones del experimento.

- La *tasa de error por familia (PFE: error rate per family ó within experiment error rate)* es el número esperado de rechazos falsos.

$$\alpha_W = \frac{\text{Número de Rechazos Falsos en un Experimento}}{\text{Número de Hipótesis en un Experimento}}.$$

  α_W es una variable aleatoria. $\alpha_{PE} = E(\alpha_W)$ es llamada la tasa de error por experimento.

- La *tasa de error a lo largo de la familia (FWE: error rate familywise, experimentwise error rate)* es definida como la probabilidad de al menos un error en la familia o la probabilidad de cometer al menos un error en un experimento cuando no hay diferencias entre los tratamientos.

Para prevenir hallar demasiadas pruebas significativas por el simple azar en un experimento individual, a menudo se intenta fijar el experimentwise error rate en algún nivel prescrito, tal como 0.05.

Si realizamos m pruebas independientes, cada una con un nivel α es fácil mostrar que la probabilidad de cometer uno o más errores entre las m pruebas es

$$\alpha_p = 1 - (1 - \alpha)^m$$

α_p puede ser llamada la tasa de un error compuesto.

El α_p sirve como una cota superior para la tasa de error compuesto de pruebas dependientes. La siguiente tabla presenta la gravedad del problema para el caso de m pruebas.

Tabla 1. *Probabilidad de rechazo erróneo de una o más hipótesis*

m	α_p
1	0.0500000
2	0.0975000
3	0.1426250
4	0.1854938
5	0.2262191
6	0.2649081
7	0.3016627
8	0.3365796
9	0.3697506
10	0.4012631

Procedimientos populares son:

- Método de Bonferroni: las m pruebas tienen una experimentwise error rate $\leq \alpha$ y una comparisonwise error rate $<< \alpha$. Desafortunadamente no es posible determinar cuánto menos. Para medias el estadístico de prueba usualmente es $t_{\alpha/(2m);\nu}$. Es útil solo para un número fijo de m.
- Método de Scheffé: este procedimiento se recomienda cuando se realiza un número grande de pruebas no planeadas.

2. La tasa de descubrimientos falsos (FDR)

Una alternativa reciente es la *Tasa de Descubrimientos Falsos* (FDR: False Discovery Rate) de Benjamini (Benjamini y Hockberg, 1995; Benjamini y Yekutiele, 2001; Chi, 2008; Scott y Bellala, 2009), la cual controla la proporción q de hipótesis nulas falsamente rechazadas relativo al número total de hipótesis rechazadas. Efron y Tibshirani (2002) muestran la relación entre la metodología Bayes empírica y la teoría de la FDR. La metodología FDR ha probado ser útil en áreas importantes tales como la genética (van den Oord, 2008).

El problema general es verificar H_0. Sea T el estadístico de prueba. Se van a realizar m pruebas. Asumamos que valores grandes $|T|$ proveen evidencia contra H_0. Los valores-p correspondientes son

$$p_i = P\left(|T| \geq |t_i| \,|H_0 \text{ es cierto}\right),$$

donde H_0 cierto significa que la ecuación anterior es calculada con respecto a la distribución de T bajo el supuesto que H_0 es cierta.

La ecuación anterior da la probabilidad que $|T|$ pueda tener un valor mayor que $|t_i|$ cuando H_0 es cierta. Una alta probabilidad significa que t_i, el valor observado

de T, es perfectamente esperable cuando H_0 sea cierta. Una baja probabilidad indica que el valor observado t_i es inusual si H_0 es cierta o que H_0 no es cierta. Por lo tanto un valor-p p_i pequeño proporciona una fuerte evidencia, aunque no certeza absoluta, contra H_0.

Un procedimiento tradicional FPR rechaza H_{0_i} si $p_i < \alpha$ y el mismo valor es aplicado a todas las pruebas.

El procedimiento FDR de Benjamini rechaza H_{0_i} para los cuales $p_i \leq p_k$, donde

$$k = \text{máx}_{0 \leq i \leq n} \left\{ i \ : \ p_{(i)} \leq q \cdot \frac{i}{n} \right\}$$

con $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(n)}$, y con $p_{(0)} = 0$.

Para el FDR la tasa nominal de descubrimientos falsos q, esto es, la tasa de falsos rechazos que queremos permitir.

La tasa de descubrimientos falsos (FDR) es definida como la proporción de pruebas que bajo la hipótesis nula siendo verdaderas dentro del conjunto de pruebas rechazadas.

Tabla 2. *Situación real vs. Decisión*

	H_0 Retenidas	H_0 Rechazadas	Total
H_0 Verdadera	TN	FD	T_0
H_0 Falsa	FN	TD	T_1
Total	N	D	m

donde

- T significa cierto,
- F Falsa,
- D Descubrimiento, y
- N No descubrimiento.

Aquí un descubrimiento es el rechazo de una hipótesis nula y un falso descubrimiento es el rechazo de una hipótesis nula verdadera.

La FDR ofrece la posibilidad de balancear los descubrimientos falsos y verdaderos. Benjamini y Hockberg (1995) la definen como

$$FDR = E\left(\frac{FD}{D} \,|D > 0\right) P\left(D > 0\right)$$

Un procedimiento de pruebas múltiples usa el criterio FDR si controla la proporción esperada de falsos positivos. Sea p_i el valor-p asociado con la i-ésima hipótesis nula, $i = 1, 2, \cdots, m$. Sean

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$$

los valores-p ordenados y $H_{(1)}, \cdots, H_{(m)}$ las correspondientes hipótesis nulas.

Definamos los m valores críticos como

$$d_i = \text{mín}\left(1, \frac{m}{(m-i+1)^2}\alpha\right)$$

para $1 \leq i \leq m$. La prueba rechaza $H_{(1)}, \cdots, H_{(k-1)}$ donde k es la más pequeña i para la cual $p_{(i)} < d_i$.

3. FDR y Diagnósticos de Regresión

El modelo más usado en estadística es el modelo de regresión $y = X\beta + \epsilon$. A partir del trabajo de Belsey, Kuh y Welch (1980), para el modelo de regresión lineal se dió origen a un área en estadística conocida como diagnósticos. Ellos propusieron gran cantidad de técnicas para determinar cuándo una observación tiene un impacto tal que sea capaz de modificar significativamente uno o más resultados del modelo, por ejemplo los estimadores del modelo o de la varianza, etc. Otros han extendido esta metodología a áreas tales como la regresión logística (Pregibon, 1981; Fowlkes, 1986), en ecuaciones estimadoras (Preisser y Qaqish, 1996), etc. Es común utilizar las reglas propuestas inicialmente para la detección de casos influyentes. Estos diagnósticos son realmente pruebas múltiples; ya que ellas son de carácter exploratorio, poca atención se ha puesto en los tipos de erros que pueden ser cometidos por el usuario: Por ejemplo, ellos definen los $DFBETAS$ como

$$DFBETAS_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_j(i)}{s(i)\sqrt{(X^TX)^{-1}_{jj}}}$$

donde

- X es la matriz de diseño,
- $\hat{\beta}_j$ es la estimada de β_j,
- $\hat{\beta}_j(i)$ es la estimada de β_j habiendo removido la i-ésima observación,
- $s(i)$ es una estimada de σ habiendo removido la i-ésima observación, y
- $\left(X^TX\right)^{-1}_{jj}$ es el j-ésimo elemento de la diagonal principal de $\left(X^TX\right)^{-1}$.

Esta es una medida del impacto relativo que la i-ésima observación tiene en el j-ésimo coeficiente estimado.

$$H_0 : E\left(DFBETAS_{ij}\right) = 0$$

El $DFBETAS_i$ muestral es aproximadamente normal con media cero y varianza $1/m$, donde m es el tamaño muestral usado para estimar el modelo de regresión.

Por ejemplo, la regla típica dice que debemos estar alerta con aquellos $DFBETA$'s mayores que $2/\sqrt{m}$ en valor absoluto corresponde a un problema de pruebas múltiples en el cual la posibilidad de tomar una o más decisiones incorrectas entre las m pruebas es

$$\alpha_m = 1 - (1 - \alpha)^m$$

α_m, que es llamada la *tasa de error compuesta*, es una tasa de error para m pruebas independentes. En el caso de los $DFBETA$'s el nivel es aproximadamente α =5 % para cada prueba.

4. Ejemplo

Consideremos la aplicación acerca de ahorros en toda la vida presentados por Belsley, Kuh y Welsch (1980), Sección 2.2, Tabla 2.6. La respuesta fue la tasa promedio de ahorros personales agregados (SR), y las variables explicativas fueron el porcentaje promedio de la población que está por debajo de los 15 años (POP15), el porcentaje promedio de la población que está por encima de los 75 años (POP75), el nivel real de ingreso per cápita disponible promedio en dólares (DPI), y la tasa promedio de crecimiento del DPI (ΔDPI). La teoría económica dice que la tasa de ahorros es menor en países donde el porcentaje de personas que no pertencen al mercado laboral es mayor. Se ajustó el modelo utilizando información proveniente de 50 países y los resultados respaldan la teoría planteada. Ellos presentaron los $DFBETAS$ para el modelo

$$SR = \beta_0 + \beta_1 POP15 + \beta_2 POP75 + \beta_3 DPI + \beta_4 \Delta DPI$$

El modelo ajustado fue

$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
28.56	-0.4611	-1.691	-0.000337	0.4096
(7.34)	(0.14)	(1.08)	(0.0009)	(0.19)

En paréntesis aparecen los errores estándares de los estimadores.

Tabla 3. *Detección de observaciones influyentes*

	País detectado como significativamente mayor $DFBETAS$ usando BKW	País detectado como significativamente mayor $DFBETA$ usando FDR
$DFBETAS_0$	21, 23, 49	23, 49
$DFBETAS_1$	10, 21, 23, 49	23, 49
$DFBETAS_2$	21, 23, 46, 49	21, 23
$DFBETAS_3$	Ninguno	Ninguno
$DFBETAS_4$	23, 33, 47, 49	49

Como observamos, de la forma original tenemos 15 observaciones que son marcadas como influyentes y en teoría nos tocaría determinar el impacto de cada una de ellas y si realmente deben estar incluídas en el modelo. Con las pruebas FDR hemos reducido el número de puntos candidatos a la mitad, 7 observaciones, lo cual nos aliviana el trabajo en forma sustancial.

5. Conclusiones

La detección de observaciones influyentes en el modelo lineal es un problema de múltiples pruebas de hipótesis. La metodología de pruebas FDR es una alternativa para limitar el número de falsas alarmas, o sea para reducir el número de falsos candidatos. Esto puede reducir sustancialmente el trabajo que se realiza posterior al ajuste del modelo. Hemos ilustrado esto con uno de los criterios como lo es el DFBETA. Es fácil extender esta metodología a otras áreas de modelación tales como el modelo lineal generalizado, modelos mixtos, etc.

Recibido: 11 de mayo de 2010
Aceptado: 23 de septiembre de 2010

Referencias

Belsey, D. A., K. E. y. W. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, John Wiley & Sons, New York.

Benjamini, Y. y Hochberg, Y. (1995), 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing ', *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Benjamini, Y. y Yekutiele, D. (2001), ' The control of the False Discovery Rate in Multiple Testing Under Dependency ', *The Annals of Statistics* **29**(4), 1165–1188.

Chi, Z. (2008), ' False discovery rate control with multivariate p-values ', *Electronic Journal of Statistics* **2**, 368–411.

Donoho, D. y Jin, J. (2004), 'Higher Criticism for Detecting Sparse Heterogeneous Mixtures ', *The Annals of Statistics* **32**(3), 962–994.

Dudoit, S. y van der Laan, M. J. (2008), *Probability and Statistical Inference*, Springer-Verlag, New York.

Efron, B, y. T. R. (2002), ' Empirical Bayes Methods and False Discovery Rates for Microarrays ', *Genetic Epidemiology* **23**, 70–86.

Fowlkes, E. B. (1986), ' Some Diagnostics for Binary Logistic Regression Via Smoothing', *Proceedings of the Statitical Computing Section:ASA* .

Pregibon, D. (1981), ' Logistic Regression Diagnostics ', *The Annals of Statistics* **9**(4), 705–724.

Preisser, J. S. y Qaqish, B. F. (1996), ' Deletion Diagnostics for Generalized Estimating Equations ', *Biometrika* **83**, 551–562.

Scott, C. y Bellala, G. (2009), ' The false discovery rate for statistical pattern recognition ', *Electronic Journal of Statistics* **3**, 651–677.

van den Oord, E. J. C. G. (2008), 'Controlling False Discoveries in Genetics Studies', *American Journal of Medical Genetics Part B(Neuropsychiatic Genetics)* **147B**, 637–644.

Apéndice

La siguiente función en R permite la determinación de las observaciones que son influyentes mediante el uso del criterio establecido usando FDR. Como argumento se tiene p, un vector, los $valores - p$ correspondientes a los estadísticos calculados para cada observación, y q, el nivel de significancia global, preterminado en el 5 %.

```
FDR<-function(p,q=0.05){
m<-length(p)
orden<-rank(p)
di<-m*q/(m-1:m+1)^2
di<-ifelse(di>1,1,di)
p.ordenado<-sort(p)
dif<-ifelse(p.ordenado-di>0,1,0)
i.min<-which(dif[2:m]-dif[1:(m-1)]==1)+1
or<-1:m
resultado<-or[orden<i.min]
resultado
}
```

Como resultado, la función entrega un vector con las posiciones de las observaciones influyentes. En caso de no encontrar, entrega el mensaje "integer(0)".