Revista Colombiana de Estadística
Vol. **22** (1998), *págs. 9 - 16*

Identificación de celdas atípicas en experimentos factoriales mediante el uso de regresión robusta

JOSÉ ALBERTO VARGAS N. *

Resumen.

La identificación de celdas atípicas en diseños factoriales puede llevarse a cabo de una forma más eficaz mediante la utilización de estimadores M redescending. En regresión robusta se les asigna un peso pequeño a estas celdas. Se propone un método que permite identificar estas celdas cuando se ajusta un modelo de orden menor a las medianas de las observaciones recolectadas en cada celda.

PALABRAS CLAVES: Estimadores M redescending, valores ó celdas atípicos.

1. Introducción

En un análisis de varianza usual en el que la variabilidad total se descompone en efectos principales e interacciones de diferente orden, puede darse el caso que interacciones de orden alto, las que usualmente no son muy fáciles de interpretar, sean causadas por la presencia de una o unas pocas celdas atípicas. La identificación de estas celdas es entonces de gran utilidad para aquellos investigadores que trabajan con experimentos factoriales.

Adicional al uso de residuales mínimos cuadrados, la técnica corrientemente utilizada en diseño de experimentos para la identificación de valores atípicos, ver por ejemplo Barnett and Lewis (1984), varios autores han propuesto métodos alternos para la identificación de celdas atípicas. Carrol (1980) utilizó técnicas de regresión robusta para detectar valores atípicos en experimentos factoriales. Usó el estimador M de Huber y los estimadores M de Hampel y Andrews. Bradu and Hawkins (1982) propusieron un método robusto basado en las diferencias entre pares de celdas en una tabla de dos vías no replicada. Simonoff (1988)

*Departamento de Matemáticas y Estadística, Universidad Nacional de Colombia, Santafé de Bogotá, COLOMBIA

propone un procedimiento "backwards-stepping" para la identificación de celdas atípicas en tablas de contingencia. Oehlert (1994) también utilizó métodos robustos de regresión para la identificación de celdas atípicas. Utilizando el estimador M de Huber, identificó las denominadas celdas de interacción, en razón a que estas celdas podrían ser la causa de interacciones de orden superior. Posteriormente, mediante mínimos cuadrados ajustó el modelo combinado que incluía el modelo inicialmente propuesto más las celdas de interacción, con un grado de libertad para cada una de ellas.

El presente artículo propone el uso de regresión robusta, usando el estimador M de Huber y estimadores M redescending para la identificación de celdas atípicas en un modelo de orden menor, por ejemplo en un modelo de efectos principales únicamente. Luego, en contraste con Oehlert (1994) quien aisló estas celdas, se propone cambiar los valores iniciales de las celdas identificadas como atípicas por los valores ajustados que resultan de la regresión robusta. En caso de tener varias observaciones por celda, se propone utilizar la mediana de estas en la construcción del modelo. Finalmente se desarrolla el análisis de varianza usual con los datos completos, incluyendo las celdas modificadas.

La sección 2 presenta algunos aspectos generales de los estimadores M de regresión. En la sección 3 se explica el método propuesto, mientras que en la sección 4 se dan algunos ejemplos. Finalmente, las conclusiones al presente trabajo se presentan en la sección 5.

2. Estimadores M de regresión

Consideremos el modelo lineal,

$$y_i = \sum_{j=1}^{p} x_{ij}\theta_j + \epsilon_i \ , \ i = 1, 2, \ldots, n$$

donde los x_{ij} son coeficientes conocidos, los $\theta_j (j = 1, \ldots, p)$ son p parámetros desconocidos y los ϵ_i son variables aleatorias independientes e idénticamente distribuidas. Un estimador M de regresión se obtiene minimizando la suma de funciones,

$$\sum_{i=1}^{n} \rho \left(y_i - \sum_{j=1}^{p} x_{ij}\theta_j \right)$$

donde ρ es una función seleccionada adecuadamente.
Si ψ denota la derivada de ρ, entonces un estimador M de los parámetros del modelo, se encuentra resolviendo el sistema,

$$\sum_{i=1}^{n} \psi \left(y_i - \sum_{j=1}^{p} x_{ij}\theta_j \right) x_{ij} = 0 \ , \ \text{para} \ \ k = 1, 2, \ldots p.$$

Para que estos estimadores sean invariantes de escala, se introduce algún estimativo s de escala. Así que un estimador M de regresión también se puede definir como la solución al sistema,

$$\sum_{i=1}^{n} \psi\left(\frac{y_i - \sum_{j=1}^{p} x_{ij}\theta_j}{cs}\right) x_{ik} = 0\ ,\ k = 1, 2, \ldots, p,$$

donde c es una constante llamada constante de ajuste.
Una función muy conocida es la denominada función ψ de Huber, definida como

$$\psi(x) = \max(-c, \min(c, x)).$$

La curva de influencia del estimador de Huber, que nos indica como una proporción infinitesimal de contaminación afecta al estimador en muestras grandes, Hoaglin, Mosteller and Tukey (1982), es constante para todas las observaciones localizadas más allá de cierto punto. Con el objeto de obtener estimadores M más resistentes a valores atípicos, se propusieron estimadores cuya función ψ y por lo tanto la respectiva curva de influencia, retornara a cero en algún punto. Estos son los denominados estimadores M "redescending", entre los que podemos mencionar los siguientes:
Función ψ "biweight" de Tukey:

$$\psi(x) = \begin{cases} x(1-x^2)^2 & \text{si } |x| \leq 1 \\ 0 & \text{en otro caso.} \end{cases}$$

Función ψ de Andrews:

$$\psi(x) = \begin{cases} \sin(x) & \text{si } -\pi \leq x \leq \pi \\ 0 & \text{en otro caso.} \end{cases}$$

Función ψ de Hampel:

$$\psi(x) = \begin{cases} x & \text{si } 0 \leq |x| \leq a \\ a \cdot \text{sign}(x) & \text{si } a \leq |x| \leq b \\ a\frac{c-|x|}{c-b}\text{sign}(x) & \text{si } b \leq |x| \leq c \\ 0 & \text{si } c \leq |x|, \end{cases}$$

con a, b y c constantes, tales que $0 \leq a \leq b < c < \infty$.
Los estimadores M "redescending" son más resistentes a los valores atípicos, en razón a que estos no tienen ningún efecto sobre los estimadores a partir de cierto punto. Aunque el punto de colapso (breakdown point) de los estimadores M redescending es menor que el del estimador M de Huber, tal como lo mostraron Yand and Van Ness (1995), nuestro principal interés es la detección de valores

atípicos. Teniendo en cuenta este objetivo, mostraremos que el uso de los estimadores M redescending nos conduce a mejores resultados.

3. Identificación de celdas atípicas

Consideremos el arreglo factorial como una tabla con N celdas y r replicaciones en cada una de ellas. r puede ser igual a 1. Vamos a suponer que los datos siguen un modelo de orden menor, y que es de la forma

$$y_{mi} = \sum_{j=1}^{p} x_{ij}\theta_j + \delta_i + \epsilon_i \ , i = 1, 2, \ldots, N,$$

y_{mi} es la mediana de los r valores de la i-ésima celda y $\delta_i = 0$ para la mayoría de las celdas. Aquella o aquellas celdas para las que $\delta_i \neq 0$ serán denominadas celdas atípicas. Asumimos además que $\epsilon_i \sim$ iid $N(0, \sigma^2)$.

Para determinar las celdas en las que $\delta_i \neq 0$, se ajusta el modelo anterior usando regresión robusta. Aquellas celdas a las que el ajuste robusto les asigne pesos bien pequeños, serán las celdas identificadas como posibles atípicas. Oehlert (1994) calculó valores críticos para varios diseños factoriales comunes, mediante simulación. Sin embargo, el hecho de tener una amplia gama de posibles arreglos factoriales, hace impractico su método. Mediante el uso de estimadores M redescending las celdas atípicas son tan claramente identificadas, que no se hace necesario el cálculo de valores críticos para determinar cuando un peso es "pequeño"

Para el ajuste robusto se han utilizado las funciones ψ de Huber, Tukey, Andrews y Hampel. Todos los cálculos fueron hechos utilizando S-PLUS versión Windows. Debe tenerse cuidado en la selección de las constantes de ajuste. Aun cuando no existe una medida que nos indique el efecto que tiene el cambio de estas constantes, Kelly (1992) advierte que la selección de las constantes de ajuste es importante y debe hacerse antes de entrar a resolver el problema en el que se van a utilizar estos estimadores.

En el presente trabajo se seleccionaron las siguientes constantes de ajuste. Para el estimador de Huber se usó $c = 0.75$, valor seleccionado por Oehlert (1994). Para el estimador de Tukey se usó $c = 6$. Para el estimador de Andrews se escogió $c = 3$. Para el estimador de Hampel $a = 3, b = 4$ y $c = 10$, constantes que satisfacen la desigualdad $c - b \leq 2a$ sugerida en Hoaglin, Mosteller and Tukey (1982).

Una vez identificadas las celdas atípicas, es decir aquellas cuyo peso en la regresión robusta es bien pequeño, se remplazarán los valores iniciales de esta celda por aquellos que indique el ajuste robusto. Se utilizará el ajuste obtenido mediante el uso del estimador M de Tukey, aunque bien se hubiera podido utilizar cualquiera de los otros, pues los valores ajustados varían muy poco de un caso a otro.

4. Ejemplos

El primer ejemplo es el factorial de estructura 2^4 replicado de Oehlert (1994). Este conjunto de datos simulados fue construido específicamente para que se ajustara a un modelo de efectos principales más una celda de interacción. Al desarrollarse el análisis de varianza usual con estos datos, se encontró que todos los efectos principales así como todas las interacciones eran altamente significantes. Se ajustó entonces un modelo de efectos principales mediante regresión robusta. En la tabla 1 se presentan los valores de los pesos obtenidos para la primera celda, que en todos los casos fue el menor, así como también el siguiente peso más pequeño. La tabla nos muestra que los estimadores M redescending le asignan un peso de 0.0 a la primera celda, identificándola claramente como una posible atípica.

Tabla 1. Pesos obtenidos en el ajuste robusto para el factorial 2^4

Peso	Huber	Tukey	Andrews	Hampel
Primera celda	0.03	0.00	0.00	0.00
Siguiente más pequeña	0.16	0.14	0.66	0.49

Una vez identificada esta celda, se procedió a substituir las dos observaciones originales de la celda, por el valor ajustado obtenido a partir del método de Tukey, que en este caso fue de 14.4. Debe anotarse que los valores ajustados para los casos considerados aquí, oscilaron entre 14.26 y 14.54. Por lo tanto, cualquier otra selección nos hubiera conducido a resultados similares.
El análisis de varianza usual con la primera celda modificada, mostró que los efectos principales eran altamente significantes. El valor p de la interacción $C * D$ fue de 0.075 y el de $A * C * D$ fue de 0.015. Ninguna otra interacción resultó significante.

El segundo ejemplo ha sido tomado de Mason, Gunst and Hess (1989, pág. 351). Consiste en un experimento llevado a cabo para investigar los efectos de tres factores, A, M y P, en la fragmentación de un dispositivo explosivo. Se hicieron tres pruebas para cada combinación de los niveles de los factores. Los datos se reproducen en la tabla 2.
En la tabla 4 se muestran los valores P del análisis de varianza realizado con estas observaciones. Como se observa, la interacción $A * M * P$ es significante al cinco por ciento. Aunque la presencia de una sola interacción no justificaría el uso del método propuesto aquí, se decidió ajustar un modelo de efectos principales mediante regresión robusta, solamente por propósito de ilustración.

La tabla 3 presenta los pesos asignados a la primera celda $(A_1 - M_1 - P_1)$, que en todos los casos fue el menor, así como el siguiente peso más pequeño. Los valores de esta tabla pueden ser algo desconcertantes a primera vista. El estimador de Huber identifica dos celdas atípicas, Tukey lo hace con la primera celda, mientras que Andrews y Hampel no identifican ninguna. Varias observaciones se deben hacer al respecto. Esta situación, que se presenta frecuentemente, muestra la gran sensibilidad de estos métodos para identificar valores atípicos. Se dan casos en que se identifican más valores atípicos de los que realmente existen, hasta el caso contrario, en el que no se identifica ninguno, aunque estos puedan existir.

Tabla 2. Factorial 2^3 en el estudio de un dispositivo explosivo.

	P	P_0		P_1	
	M	M_0	M_1	M_0	M_1
A	A_0	0.0698 0.0698 0.0686	0.0659 0.0651 0.0676	0.0625 0.0615 0.0619	0.0699 0.0620 0.0602
	A_1	0.0618 0.0613 0.0620	0.0658 0.0635 0.0633	0.0589 0.0601 0.0621	0.0612 0.0598 0.0594

Tabla 3. Pesos obtenidos en el ajuste robusto - ejemplo del dispositivo.

Peso	Huber	Tukey	Andrews	Hampel
Primera celda	0.00	0.00	0.93	1.00
Siguiente más pequeña	0.00	0.91	0.97	1.00

La sensibilidad del método está muy relacionada con la selección de la constante de ajuste. Si cambiáramos el valor de nuestras constantes por ejemplo, podríamos obtener resultados diferentes a los que aparecen en la tabla 3, para el mismo conjunto de datos. Aunque la constante de ajuste puede seleccionarse

de tal forma, que la eficiencia relativa asimptótica de los estimadores obtenidos tenga un valor mínimo, es posible escogerla bajo otras condiciones. Por ejemplo, seleccionarla de tal manera, que el método resulte altamente sensible a la presencia de valores atípicos. Este último fue el criterio de Oehlert (1994) para usar $c = 0.75$ con el estimador M de Huber. En conclusión, aunque debe tenerse mucho cuidado en la selección de las constantes de ajuste, no hay una regla fija que nos indique cuál es el mejor valor para todos los casos.
Otra posible causa de los resultados obtenidos en la tabla 3, es que la celda en cuestión, es decir la primera, aunque extrema, no necesariamente es atípica. Por esta razón, mientras unos métodos la identifican como atípica, otros no lo hacen.

Tabla 4. Valores p obtenidos en el análisis de varianza - ejemplo del dispositivo.

Fuente de V.	Grados de L.	Datos originales	Datos modificados
A	1	0.00	0.00
M	1	0.74	0.11
P	1	0.00	0.11
$A * M$	1	0.33	0.75
$A * P$	1	0.23	0.93
$M * P$	1	0.47	0.57
$A * M * P$	1	0.03	0.31
Residual	16		

Sin embargo, para ilustrar el método completo, se sustituyeron los valores de la primera celda por 0.0649, que corresponde al valor ajustado de Tukey, y se realizó el análisis de varianza usual. Los valores p se representan en la tabla 4. Observamos que en el análisis de varianza de los datos modificados ninguna interacción es significante.

5. Conclusiones

Los estimadores M redescending son una herramienta muy poderosa en la identificación de valores atípicos. El método propuesto hace uso de ellos para identificar celdas atípicas en un arreglo factorial. Este método se aconseja cuando al hacer el análisis de varianza con los datos originales se encuentra un número grande de interacciones altamente significantes. Se ha visto que en varios casos, estas interacciones pueden ser causadas por una o unas pocas celdas atípicas. Su identificación y posterior modificación usando ajustes robustos, nos conduce generalmente a un análisis de varianza mucho más fácil de interpretar.

Referencias

1. Barnett, V., and Lewis, T., *Outliers in Statistical Data*, (2^{nd} ed.), John Wiley, 1984.
2. Bradu, D., and Hawkins, D. M., *Location of Multiple outliers in Two-Way Tables, Using Tetrads*, Technometrics no. 24 (1982), 103-108.
3. Carrol, R. J. (1980), *Robust Methods for Factorial Experiments with outliers*, Applied Statistics no. 29, 246-251.
4. Hoaglin, D. C., Mosteller, F., and Tukey, J. W., *Understanding Robust and Exploratory Data Analysis*, John Wiley, 1982.
5. Kelly, G. E., *Robust Regression Estimators - The Choice of Tuning Constants*, The Statistician no. 41 (1992), 303-314.
6. Mason, R. L., Gunst, R., and Hess J., *Statistical Design and Analysis of Experiments*, John Wiley, 1989.
7. Oehlert, G. W., *Isolating One-Cell Interactions*, Technometrics no. 36 (1994), 403-408.
8. Simonoff, J. S., *Detecting Outlying Cells in Two-Way Contingency Tables Via Backwards - Stepping*, Technometrics no. 30 (1988), 339-345.
9. Yang, J. J.m and Van Ness, J. W., *Breakdoum Points For Redescending M-Estimates of Location*, Communications in Statistics-Theory and Methods no. 24 (1995), 1769-1787.