Ingeniería y Ciencia
ISSN:1794-9165 | ISSN-e: 2256-4314
ing. cienc., vol. 11, no. 22, pp. 121–150, julio-diciembre. 2015.
http://www.eafit.edu.co/ingciencia
This article is licensed under a Creative Commons Attribution 4.0 By

Efecto de datos influyentes en el análisis de diseños factoriales de efectos fijos 3^{ω}

Oscar O. Melo [1],Carlos A. Falla [2] y José A. Jiménez [3]

Recepción: 10-02-2015 | Aceptación: 06-06-2015 | En línea: 31-07-2015

MSC: 62K15, 62E15, 62-07, 62J20

doi:10.17230/ingciencia.11.22.6

Resumen

En este trabajo se establece una metodología alternativa para la detección de observaciones influyentes en diseños factoriales de efectos fijos 3^{ω}, a través del planteamiento de la estadística de prueba (F_q) y la caracterización de los efectos de dichas observaciones sobre el análisis, las sumas de cuadrados y los estimadores del modelo que describe el diseño experimental.

Palabras clave: diseño factorial; datos influyentes; análisis de varianza; datos atípicos; sumas de cuadrados.

[1] Universidad Nacional de Colombia, Bogotá, Colombia, oomelom@unal.edu.co.
[2] BI Technical Consultant en Conexia SA, Bogotá, Colombia, cafallag@unal.edu.co.
[3] Universidad Nacional de Colombia, Bogotá, Colombia, josajimenezm@unal.edu.co.

Effect of Influential Data in 3^{ω} Fixed Factorial Designs

Abstract
This paper provides a methodology alternative for the detection of influential observations in factorial design of fixed effects 3^{ω}. Our proposal is developed through the approach of the test statistic (F_q), and the characterization of the impact of such observations on the analysis, the sums of squares and the estimators of the model that describes the experimental design.

Key words: factorial design; influential data; variance analysis; outliers data; sums of squares.

1 Introducción

El diseño experimental es usado frecuentemente en la investigación, principalmente en la industria, biología y ciencias agropecuarias, en las áreas del desarrollo de producción y control de calidad. Para la elaboración de un producto se deben tener en cuenta los ingredientes o componentes que éste requiera y las condiciones bajo las cuáles se fabrica. El objetivo de la experimentación es estudiar los efectos de la variación de los factores que se involucran en la elaboración y determinación de la mejor combinación de ellos.

Muchos experimentos tienen en cuenta dos o más factores, por lo que cada observación es respuesta de una de las posibles combinaciones de los niveles experimentales de dichos factores. Para estos casos, se recomienda la aplicación de un diseño con arreglo factorial como una alternativa más eficiente, que los métodos donde se estudian los factores en forma separada. Estos diseños investigan todas las posibles combinaciones de los niveles de los factores en cada ensayo completo o réplica del experimento. El efecto de un factor se define entonces como el cambio en la respuesta producido por un cambio en el nivel del factor.

En muchas áreas y procedimientos metodológicos de la estadística, el tema de observaciones influyentes es común y en cada uno de ellos existen elaboraciones teóricas para su tratamiento y análisis. Jiménez [1] dice que la presencia de estas observaciones puede distorsionar severamente la

interpretación del análisis de varianza, pues afecta directamente las sumas de cuadrados que permiten construir las estadísticas de prueba para rechazar o no las hipótesis planteadas, y por lo tanto, podrían tener una gran influencia sobre la decisión que se tome con respecto a ellas.

El objetivo de este trabajo es desarrollar un procedimiento de análisis de influencia en diseños factoriales 3^{ω}, acompañado de los métodos de análisis de varianza. Como los modelos estadísticos por lo general tienen algún grado de aproximación, es importante la evaluación de la influencia de la menor perturbación de un modelo hipotético [2]. Los resultados del análisis de influencia se pueden utilizar para identificar los problemas implícitos en un estudio con el fin de juzgar si una decisión es posiblemente engañosa y para tener una visión más completa de las conclusiones que se obtienen. Por lo tanto, el análisis de la influencia es considerado como un componente importante en el análisis de un diseño experimental 3^{ω}. Aunque el análisis de la influencia ha sido durante mucho tiempo un tema importante en varios modelos estadísticos (véase, [3],[4],[2],[5],[6],[7],[8],[9],[10],[11],[12],[13]), se ha trabajado muy poco en los diseños factoriales simétricos, y en particular, en los diseños 3^{ω}.

En este artículo se aborda los diseños factoriales de efectos fijos 3^{ω} con el fin de identificar los efectos de las observaciones influyentes sobre las hipótesis de interés, específicamente sobre las sumas de cuadrados y las estadísticas de prueba; planteando una metodología para su identificación y aplicándola en un caso de estudio, estableciendo patrones y características en su análisis.

El artículo esta organizado como sigue: en la sección 2 se presenta brevemente los principales temas relacionados con diseños factoriales, análisis de varianza, datos influyentes y sus métodos de detección. Además, se presenta la estadística F_q y su distribución, a partir de la cual se pueden identificar observaciones influyentes o conjuntos de observaciones influyentes en diseños factoriales de efectos fijos 3^{ω}. En la sección 3 se presenta la construcción teórica de la estadística F_q y se describe la forma de calcularla a partir de las sumas de cuadrado del diseño factorial de efectos fijos 3^{ω}.

En la sección 4 se caracterizan algunos de los posibles efectos que tienen las observaciones influyentes sobre las sumas de cuadrados utilizadas en el análisis de varianza, que soporta el experimento y sobre la estadística de

prueba; la sección 5 muestra un ejemplo de aplicación de la metodología propuesta y la sección 6 presenta las conclusiones correspondientes a la metodología propuesta.

2 Diseños factoriales y análisis detección de datos influyentes

Los diseños factoriales en general, se basan en el análisis de los diferentes factores que puedan intervenir en un experimento, encontrando la(s) mejor(es) combinación(es) de los niveles que éstos presentan. La selección de dicha(s) combinación(es), se realiza mediante la comprobación de hipótesis apropiadas con respecto a ellas, llegando así a una estimación de su efecto sobre el experimento.

Para probar las hipótesis, se plantea un modelo estadístico lineal que permita escribir cada una de las respuestas obtenidas en el experimento, a través de la suma de un parámetro común a las combinaciones de los niveles de los factores, un parámetro único para cada una de ellas (efecto de tratamiento) y una componente aleatoria de error, este modelo se denomina de "análisis de varianza" [14].

Sin perdida de generalidad, en este artículo se toma en particular los diseños factoriales simétricos 3^3 y posteriormente, se hace una generalización a los diseños 3^{ω}. Por lo tanto, se tienen tres factores cada uno con tres niveles, lo que genera un total de 27 combinaciones llamadas tratamientos. La respuesta observada en cada uno de los tratamientos es una variable aleatoria que depende de los niveles de los factores, por lo cual, resulta útil describir las observaciones mediante el siguiente modelo estadístico lineal:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl} \qquad (1)$$

donde $i = 1, 2, 3,\ j = 1, 2, 3,\ k = 1, 2, 3,\ l = 1, 2, \ldots n$, las y_{ijkl} son las respuestas al tratamiento ijk-ésimo en la replicación l-ésima, con n replicaciones de cada tratamiento, μ es un parámetro común a todos los tratamientos denominado media global, α_i es un parámetro del i-ésimo nivel del factor A, β_j es el parámetro del j-ésimo nivel del factor B, γ_k es el parámetro del k-ésimo nivel del factor C. Los términos en paréntesis son los respectivos efectos de la interacción entre los diferentes niveles de los

tres factores y ε_{ijkl} es la componente aleatoria del error, la cual se supone normal con media cero y varianza constante σ^2.

El procedimiento adecuado para probar las hipótesis de interés acerca de que los efectos de los tratamientos son cero o no, es el análisis de varianza (ANOVA). La denominación análisis de varianza resulta de descomponer la variabilidad total de los datos en sus componentes. La suma total de cuadrados corregida (SCT) se usa como medida de la variabilidad total de los datos, esta es:

$$SCT = SC_{Trata} + SC_E \tag{2}$$

donde (SC_{Trata}) es la suma de cuadrados de los tratamientos y (SC_E) es la suma de cuadrados del error. La forma usual de calcular dichas sumas está determinada por:

$$SCT = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n} y_{ijkl}^2 - \frac{y_{....}^2}{N} \quad SC_{Trata} = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} \frac{y_{ijk.}^2}{n} - \frac{y_{....}^2}{N} \tag{3}$$

donde N denota el total de observaciones en el diseño. Además, se tiene que

$$SC_E = SCT - SC_{Trata}$$

El procedimiento se prueba en una tabla de análisis de varianza para el modelo trifactorial de efectos fijos. Luego, se rechaza la hipótesis nula H_0 si su correspondiente valor F_0 (obtenido de los datos observados) es mayor que un valor tabulado F_{Tab}, con un valor crítico α.

Hasta el momento no se ha juzgado cuál de los niveles o combinación de niveles de los factores es el mejor. Para esto se desarrollan los estimadores de los parámetros del modelo dado en (1) mediante el método de mínimos cuadrados, partiendo de contrastes construidos con los promedios de los niveles de los factores. De esta forma, también se pueden determinar las estimaciones de los efectos de interacción.

2.1 Detección de datos influyentes

La veracidad de los modelos, se comprueba a través del análisis de los residuales. Este análisis permite identificar datos atípicos, observaciones por fuera del conjunto de datos, o de otra forma observaciones que no se comportan como lo hacen la mayoría de los datos, y que podrían afectar

los parámetros del modelo [15], es decir, que cambiarían notablemente las estimaciones de dichos parámetros si se realizara el análisis sin contar con ellas.

Hay que tener cuidado en la notación, pues ε_i es el i-ésimo error del modelo, mientras que e_i es el i-ésimo residual del mismo. La idea es identificar si los residuales se comportan como los errores del modelo ya que los e_i son valores observables y los ε_i son no observables, así los ε_i tienen distribución normal con media 0 y varianza σ^2, es decir que $\left(\frac{\varepsilon_i}{\sigma}\right)$ tienen distribución normal estándar con media 0 y varianza 1. Luego los $\frac{e_i}{\sqrt{(1-h_{ii})\sigma}}$ se deberían comportar normal estándar, con $0 \leq h_{ii} < 1$, el i-ésimo elemento diagonal de la matriz $\boldsymbol{H}$, las propiedades de esta matriz son dadas en Hoaglin y Welsch [16].

Es posible cuantificar el impacto que sobre los coeficientes tiene la eliminación de una observación, mediante diferentes métodos como: *Distancia de Cook*, *Distancia DFFITS* definida por Belsley [15], estadística *DFBETAS*, entre otros métodos; éstos se pueden consultar de manera detallada en Peña-Sánchez [17] ó en Draper y Smith [18].

Draper y John [19] desarrollaron una metodología para detectar un grupo de q observaciones influyentes o atípicas, equivalente a la propuesta por Bartlett [20] para estimar los parámetros del modelo de regresión lineal cuando existen observaciones faltantes en la variable respuesta. Jiménez [21] desarrolló una propuesta para imputar valores no influyentes en modelos de regresión lineal múltiple con información incompleta, con un modelo alterado que excluye del análisis el dato o conjunto de datos influyentes, de tal forma que la suma de cuadrados de los residuales del modelo modificado es:

$$SC_E^* = SC_E + \boldsymbol{\varphi}'(\boldsymbol{I} - \boldsymbol{H})[2\boldsymbol{Y} - \boldsymbol{\varphi}] \tag{4}$$

Así, la variación en las sumas de cuadrados, dada por la influencia de las observaciones es expresada como:

$$Q_q = SC_E - SC_E^* \tag{5}$$

Esta estadística, presentada en Draper y John [19], muestra a Q_q expresada en función de los residuales estimados. Partiendo de las expresiones (4) y (5), se llega a que la estadística $Q_{q=1}$, expresada de la siguiente manera:

$$\frac{\sqrt{Q_1}}{s} \sim t_{(N-r)} \tag{6}$$

la cual tiene una distribución t con $(N-r)$ grados de libertad, en donde s es la raíz cuadrada del estimador insesgado de σ^2 dado por $s^2 = \frac{SC_E}{N-r}$. Sin embargo, por teoría estadística se sabe que este cociente tiene una distribución t cuando las dos variables son independientes, pero en Jiménez [22] se prueba que no lo son.

2.2 Estadística F_q para los diseños factoriales de efectos fijos 3^3

Sin pérdida de generalidad, los resultados anteriores se pueden particularizar para los diseños factoriales de efectos fijos 3^3. De acuerdo a Jiménez [22] se encuentra que $\frac{Q_q}{\sigma^2}$ y $s^{*2} = \frac{SC_E^*}{N-27-q}$ son independientes, en donde

$$\frac{SC_E^*}{\sigma^2} = \frac{s^{*2}(N-27-q)}{\sigma^2} \sim \chi^2_{(N-27-q)} \tag{7}$$

Luego, se define:

$$F_q = \frac{\frac{Q_q}{q\sigma^2}}{\frac{SC_E^*}{(N-27-q)\sigma^2}} = \frac{Q_q}{qs^{*2}} \sim F_{(q,N-27-q)} \tag{8}$$

donde Q_q corresponde a la diferencia entre la suma de cuadrados del modelo (SC_E) planteado en (1) y la suma de cuadrados del modelo reducido SC_E^*, es decir, sin las q observaciones consideradas influyentes o atípicas.

Como se observa en (8), la estadística F_q depende del número de observaciones que se estén considerando como influyentes. En particular si $q = 1$, es decir, que se evalúe si la observación, notada por y^*_{ijkl}, es influyente o no, Q_q notada ahora como Q_1, resulta ser igual al cuadrado del error correspondiente a dicha observación

$$F_1 = \frac{Q_1}{s^{*2}} \sim F_{(1,N-28)}$$

donde $Q_1 = SC_E - SC_E^* = e^{*2}_{ijkl}$. Al tomar raíz cuadrada de F_1 se obtiene

$$\frac{|e^*_{ijkl}|}{s^*} \sim t_{(N-28)} \tag{9}$$

Ahora, el interés cuando se desea establecer si un grupo de observaciones es influyente o no, es probar la hipótesis:

H_0 : Ninguna de las q observaciones es influyente.
H_a : Por lo menos una de las q observaciones es influyente.

Luego, la hipótesis nula H_0 se rechaza a un nivel de significancia $\alpha\,\%$ si $F_q > F_{(q,N-27-q,\alpha)}$.

3 Estadística F_q a partir de las sumas de cuadrados del diseño factorial 3^3

Como se menciono anteriormente, el análisis de varianza se deriva de la partición de la variabilidad total en sus componentes (2). Partiendo de las ecuaciones planteadas en (3), se puede descomponer la suma de cuadrados de los tratamientos en la suma de cuadrados de cada uno de los factores principales y las interacciones.

Para el caso específico de los diseños factoriales de efectos fijos 3^3, se tiene las sumas de cuadrados de los factores principales y sus interacciones son:

$$
\begin{aligned}
SC_A &= \frac{1}{9n}\sum_{i=1}^{3} y_{i\ldots}^2 - \frac{y_{\ldots.}^2}{N}, \quad SC_B = \frac{1}{9n}\sum_{j=1}^{3} y_{.j.}^2 - \frac{y_{\ldots.}^2}{N},\\
SC_C &= \frac{1}{9n}\sum_{j=1}^{3} y_{..k.}^2 - \frac{y_{\ldots.}^2}{N},\\
SC_{AB} &= \frac{1}{3n}\sum_{i=1}^{3}\sum_{j=1}^{3} y_{ij.}^2 - \frac{1}{9n}\sum_{i=1}^{3} y_{i\ldots}^2 - \frac{1}{9n}\sum_{j=1}^{3} y_{.j.}^2 + \frac{y_{\ldots.}^2}{N},\\
SC_{AC} &= \frac{1}{3n}\sum_{i=1}^{3}\sum_{k=1}^{3} y_{i.k.}^2 - \frac{1}{9n}\sum_{i=1}^{3} y_{i\ldots}^2 - \frac{1}{9n}\sum_{k=1}^{3} y_{..k.}^2 + \frac{y_{\ldots.}^2}{N},\\
SC_{BC} &= \frac{1}{3n}\sum_{j=1}^{3}\sum_{k=1}^{3} y_{.jk.}^2 - \frac{1}{9n}\sum_{j=1}^{3} y_{.j.}^2 - \frac{1}{9n}\sum_{k=1}^{3} y_{..k.}^2 + \frac{y_{\ldots.}^2}{N},\\
SC_{ABC} &= \frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} y_{ijk.}^2 - \frac{1}{3n}\sum_{i=1}^{3}\sum_{j=1}^{3} y_{ij.}^2 - \frac{1}{3n}\sum_{i=1}^{3}\sum_{k=1}^{3} y_{i.k.}^2\\
&\quad - \frac{1}{3n}\sum_{j=1}^{3}\sum_{k=1}^{3} y_{.jk.}^2 + \frac{1}{9n}\sum_{i=1}^{3} y_{i\ldots}^2 + \frac{1}{9n}\sum_{j=1}^{3} y_{.j.}^2 + \frac{1}{9n}\sum_{k=1}^{3} y_{..k.}^2 - \frac{y_{\ldots.}^2}{N},
\end{aligned}
\tag{10}
$$

donde n es el número de repeticiones dentro de cada tratamiento, $y^2_{i...}$ es el cuadrado del total de los datos sobre el nivel i del factor A, $y^2_{.j..}$ es el cuadrado del total de los datos sobre el nivel j del factor B, $y^2_{..k.}$ es el cuadrado del total de los datos sobre el nivel k del factor C, $y^2_{ij..}$ es el cuadrado del total de los datos sobre los niveles i y j de la interacción AB, $y^2_{i.k.}$ es el cuadrado del total de los datos sobre los niveles i y k de la interacción AC, $y^2_{.jk.}$ es el cuadrado del total de los datos sobre los niveles j y k de la interacción BC, $y^2_{ijk.}$ es el cuadrado del total de los datos sobre los niveles i, j y k de la interacción ABC y $y^2_{....}$ es el cuadrado de la suma de todos los datos en el diseño. Para mayores detalles de estas sumas de cuadrado véase Montgomery [14]. Bajo estas expresiones, la suma de cuadrados de los errores está dada por:

$$SC_E = SCT - SC_{Subtotales}$$

La expresión del lado derecho es la diferencia entre la suma de cuadrados total presentada en (3) y la suma de cuadrados de los subtotales dada por:

$$SC_{Subtotales} = \frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} y^2_{ijk.} - \frac{y^2_{....}}{N}$$

Por lo tanto, otra expresión para la suma de cuadrados del error es:

$$SC_E = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n} y^2_{ijkl} - \frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} y^2_{ijk.} \tag{11}$$

3.1 Sumas de cuadrados del error para el diseño factorial de efectos fijos 3^3 reducido a una observación ($q = 1$)

Sea y^*_{ijkl} la $ijkl$-ésima observación del conjunto total de observaciones del diseño, que va a ser extraída para evaluar si es influyente o no. El subíndice $(ijkl)^*$ corresponde entonces a una de las $27n$ posibles observaciones de los niveles i, j, k de los factores A, B, C y las n réplicas en cada combinación, con $i = 1, 2, 3$, $j = 1, 2, 3$, $k = 1, 2, 3$ y $l = 1, 2, \ldots, n$.

Al eliminarse dicha observación del conjunto de datos, se tiene como resultado que el diseño se convierte en un diseño factorial desbalanceado de efectos fijos 3^3. En éste diseño sigue siendo posible aplicar el análisis de

varianza, pero deben hacerse ligeras modificaciones en las fórmulas de las sumas de cuadrados. Por lo tanto, las sumas de cuadrados del total, de los tratamientos y del error, son respectivamente:

$$SCT^* = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n_{ijk}} y_{ijkl}^2, SC_{Trata}^* = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} \frac{y_{ijk.}^2}{n_{ijk}} - \frac{y_{....}^{*2}}{N^*}, SC_E^* = \quad SCT^* - SC_{Trata}^*$$

donde n_{ijk} es el número de observaciones en el tratamiento $ijk-$ésimo, de modo que $N = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} n_{ijk}$. Cabe anotar que para todo caso $n_{ijk} = n$, excepto en el tratamiento del cual se extrajo la observación y_{ijkl}^*, en donde es igual a $n-1$. Las expresiones $y_{....}^*$ y $N^* = N-1$, corresponden respectivamente a la suma total y al número total de observaciones sin el dato y_{ijkl}^*, respectivamente[1].

Igual que en el diseño balanceado, la suma de cuadrados de los tratamientos puede descomponerse en sumas de cuadrados de efectos principales e interacciones; por lo tanto, las expresiones de las sumas de cuadrados para el modelo desbalanceado son:

$$\begin{aligned}
SC_A^* =& \frac{1}{9n}\sum_{\substack{i'=1\\ i'\neq i}}^{3} y_{i'...}^2 + \frac{y_{i....}^{*2}}{9n-1} - \frac{y_{....}^{*2}}{N^*} \\
SC_B^* =& \frac{1}{9n}\sum_{\substack{j'=1\\ j'\neq j}}^{3} y_{.j'..}^2 + \frac{y_{.j..}^{*2}}{9n-1} - \frac{y_{....}^{*2}}{N^*} \\
SC_C^* =& \frac{1}{9n}\sum_{\substack{k'=1\\ k'\neq k}}^{3} y_{..k'.}^2 + \frac{y_{..k.}^{*2}}{9n-1} - \frac{y_{....}^{*2}}{N^*} \\
SC_{AB}^* =& \frac{1}{3n}\sum_{\substack{i'=1\\ i'\neq i}}^{3}\sum_{\substack{j'=1\\ j'\neq j}}^{3} y_{i'j'..}^2 + \frac{y_{ij..}^{*2}}{3n-1} - \frac{y_{....}^{*2}}{N^*} - SS_A^* - SS_B^* \\
SC_{AC}^* =& \frac{1}{3n}\sum_{\substack{i'=1\\ i'\neq i}}^{3}\sum_{\substack{k'=1\\ k'\neq k}}^{3} y_{i'.k'.}^2 + \frac{y_{i.k.}^{*2}}{3n-1} - \frac{y_{....}^{*2}}{N^*} - SS_A^* - SS_C^* \\
SC_{BC}^* =& \frac{1}{3n}\sum_{\substack{j'=1\\ j'\neq j}}^{3}\sum_{\substack{k'=1\\ k'\neq k}}^{3} y_{.j'k'.}^2 + \frac{y_{.jk.}^{*2}}{3n-1} - \frac{y_{....}^{*2}}{N^*} - SS_B^* - SS_C^*
\end{aligned} \tag{12}$$

[1]Cuando se desea evaluar no solo una, sino q observaciones a fin de comprobar si son o no influyentes, $y_{...}^*$ es la suma total de las observaciones del modelo sin las q evaluadas y $N^* = N - q$.

$$SC^*_{ABC} = \frac{1}{n}\sum_{\substack{i'=1\\ i'\neq i}}^{3}\sum_{\substack{j'=1\\ j'\neq j}}^{3}\sum_{\substack{k'=1\\ k'\neq k}}^{3} y^2_{i'j'k'.} + \frac{y^{*2}_{ijk.}}{n-1} - \frac{y^{*2}_{....}}{N^*} - SS^*_A$$
$$- SS^*_B - SS^*_C - SS^*_{AB} - SS^*_{AC} - SS^*_{BC}$$

Bajo éstas expresiones, la suma de cuadrados de los errores está dada por

$$SC^*_E = SCT^* - SC^*_{Subtotales}$$

La expresión del lado derecho es la diferencia entre la suma de cuadrados total y la suma de cuadrados de los subtotales esta dada por:

$$SC^*_{Subtotales} = \frac{1}{n}\sum_{\substack{i'=1\\ i'\neq i}}^{3}\sum_{\substack{j'=1\\ j'\neq j}}^{3}\sum_{\substack{k'=1\\ k'\neq k}}^{3} y^2_{i'j'k'.} + \frac{y^{*2}_{ijk.}}{n-1} - \frac{y^{*2}_{....}}{N^*}$$

Otra expresión para la suma de cuadrados del error es entonces:

$$SC^*_E = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n_{ijk}} y^2_{ijkl} - \frac{1}{n}\sum_{\substack{i'=1\\ i'\neq i}}^{3}\sum_{\substack{j'=1\\ j'\neq j}}^{3}\sum_{\substack{k'=1\\ k'\neq k}}^{3} y^2_{i'j'k'.} - \frac{y^{*2}_{ijk.}}{n-1} + \frac{y^{*2}_{....}}{N^*} \qquad (13)$$

en donde $\frac{y^{*2}_{ijk.}}{n-1}$ es el total del tratamiento, elevado al cuadrado, de donde se extrajo la observación y^*_{ijkl} sobre el número actual de observaciones que ahora existe allí. Analizando el término de la derecha, se tiene que el primer sumando puede escribirse como el primer sumando de la ecuación (11) menos la observación y^*_{ijkl}, es decir:

$$\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n_{ijk}} y^2_{ijkl} = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n} y^2_{ijkl} - y^{*2}_{ijkl} \qquad (14)$$

Por otro lado, el segundo término de la ecuación, se puede expresar como el segundo término de la ecuación (11) menos el total del tratamiento en donde se encuentra la observación y^*_{ijkl} (notado por y_Q), en el modelo balanceado, elevado al cuadrado y dividido por n, es decir:

$$\frac{1}{n}\sum_{\substack{i'=1\\ i'\neq i}}^{3}\sum_{\substack{j'=1\\ j'\neq j}}^{3}\sum_{\substack{k'=1\\ k'\neq k}}^{3} y^2_{i'j'k'.} = \frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} y^2_{ijk.} - \frac{y^2_{Q.}}{n} \qquad (15)$$

Al sustituir (14) y (15) en (13), se tiene

$$SC^*_E = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n} y^2_{ijkl} - y^{*2}_{ijkl} - \frac{1}{n}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} y^2_{ijk.} + \frac{y^2_{Q.}}{n} - \frac{y^{*2}_{ijk.}}{n-1} \qquad (16)$$

3.2 Cálculo de la estadística F_q

Recordando que la estadística Q_1 es la diferencia entre la suma de cuadrados residuales y la suma de cuadrados residuales del diseño excluyendo la observación y^*_{ijkl} (diseño reducido), puede encontrarse una nueva expresión para ella a partir de las ecuaciones de las sumas de cuadrados del diseño factorial.

A través de la diferencia $SC_E - SC^*_E$, entre las expresiones (11) y (16), se obtiene:

$$Q_1 = y^{*2}_{ijkl} - \frac{y^2_{Q\cdot}}{n} + \frac{y^{*2}_{ijk\cdot}}{n-1}$$

Es decir que la estadística se puede encontrar a partir de la diferencia entre el cuadrado de la observación y^*_{ijkl} y el total elevado al cuadrado y dividido por n del tratamiento en donde se encuentra dicha observación en el diseño balanceado, más el total elevado al cuadrado y dividido por $n-1$ del tratamiento del diseño desbalanceado en donde se encontraba y^*_{ijkl}.

Una vez obtenida Q_1, el cálculo de F_1 puede hacerse a partir de:

$$F_1 = \frac{e^{*2}_{ijkl}}{s^{*2}} = (N-28)\frac{e^{*2}_{ijkl}}{SC^*_E} \sim F_{(1,N-28)}$$

Sin pérdida de generalidad, el estadístico F_1 para probar si una observación es influyente o no en un diseño factorial de efectos fijos 3^{ω}, es decir, un diseño con ω factores cada uno a tres niveles es:

$$F_1 = (N - 3^{\omega} - 1)\,\frac{e^{*2}_{ijk\ldots wl}}{SC^*_E} \sim F_{(1,N-3^{\omega}-1)}$$

donde $(ijk\ldots wl)^*$ hace referencia a un punto específico, ubicado en el nivel i del factor A, j del factor B, k del factor C, y así hasta el nivel w del factor ω, en la replicación l, con $i = 1,2,3$, $j = 1,2,3$, $k = 1,2,3,\ldots$, $w = 1,2,3$ y $l = 1,2,\ldots,n$.

Por otra parte, puede demostrarse que para el caso de q observaciones consideradas influyentes, es posible calcular la estadística Q_q a partir de la siguiente ecuación:

$$Q_q = \sum_{t=1}^{q} y_t^2 - \sum_{s=1}^{S} \frac{y^2_{s\cdot}}{n} + \sum_{s=1}^{S} \frac{y^{*2}_{s\cdot}}{n_s}$$

es decir, la diferencia entre la suma de los cuadrados de las q observaciones consideradas influyentes, indicadas por el subíndice $t = 1,2,\ldots,q$, y

la suma de los totales de los S tratamientos ($s = 1, 2, \ldots, S$) en donde se encuentran distribuidas dichas observaciones en el diseño balanceado, elevados al cuadrado y divididos por n; más la suma de los totales de los S tratamientos del diseño desbalanceado, en donde se encontraban las q observaciones, elevados al cuadrado y divididos por su correspondiente tamaño n_s. Luego, el cálculo de F_q puede hacerse a partir de (8).

Sin pérdida de generalidad, puede decirse que la estadística F_q para evaluar la influencia de q observaciones en un diseño factorial de efectos fijos 3^ω, es:

$$F_q = \frac{N-3^\omega-q}{q}\left(\frac{SC_E}{SC_E^*}-1\right) \sim F_{(q,N-3^\omega-q)}$$

4 Efecto de datos influyentes en las sumas de cuadrados y el análisis de varianza

A partir de las sumas de cuadrados descritas en la sección anterior, pueden construirse las Tablas 1 y 2 de análisis de varianza para el modelo balanceado y el modelo desbalanceado resultante de la extracción de la observación y^*_{ijkl} del conjunto de datos.

Tabla 1: Análisis de varianza para el modelo balanceado.

Causas de Variación	**Grados de Libertad**	**Sumas de Cuadrados**	**Cuadrado Medio (CM)**	F_0
Tratamientos	$27-1$	SC_{Trata}	$\frac{SC_{Trata}}{27-1}$	$\frac{CM_{Trata}}{CM_E}$
Error	$N-27$	SC_E	$\frac{SC_E}{N-27}$	
Total	$N-1$	SCT		

Los valores tabulados F_{Tab} y F^*_{Tab}, utilizados para determinar el resultado de las pruebas de hipótesis para realizar el análisis de varianza planteado en las Tablas 1 y 2, tienen $(26, N-27)$ y $(26, N-28)$ grados de libertad, respectivamente. Como los grados de libertad del numerador para ambos valores en los diseños factoriales de efectos fijos son 26, entonces la diferencia entre los valores que tomen F_{Tab} y F^*_{Tab} dependerá exclusivamente de los grados de libertad del denominador.

No importa el número de observaciones a evaluar como influyentes, los grados de libertad para F^*_{Tab} serán menores que los grados de libertad de F_{Tab} en el modelo completo $(N-27) > (N-27-q)$ para el caso de q observaciones, $q \geq 1$. Luego, al observar en una tabla de distribución F, el

Tabla 2: Análisis de varianza para el modelo desbalanceado.

Causas de Variación	**Grados de Libertad**	**Sumas de Cuadrados**	**Cuadrado Medio (CM)**	F_0
Tratamientos	$27-1$	SC^*_{Trata}	$\frac{SC^*_{Trata}}{27-1}$	$\frac{CM^*_{Trata}}{CM^*_E}$
Error	$N-28$	SC^*_E	$\frac{SC^*_E}{N-28}$	
Total	$N-2$	SCT^*		

comportamiento para dichos grados de libertad con un valor α determinado, se puede concluir que:

$$F_{(26,N-27,\alpha)} < F^*_{(26,N-27-q,\alpha)} \tag{17}$$

para el caso de q observaciones, $q \geq 1$.

Como el objetivo es rechazar la hipótesis nula, de manera que se compruebe la diferencia de los efectos generados por los tratamientos, para en un siguiente nivel del experimento, poder seleccionar la mejor combinación de ellos, se busca que $F_0 > F_{Tab}$. Si el caso es buscar si una observación y^*_{ijkl} resulta ser influyente, al aislarla del análisis, el valor tabulado a usar en la prueba es F^*_{Tab}, es decir, para rechazaría la hipótesis nula de no influencia de un grupo de observaciones si $F_0 > F^*_{Tab}$. Por lo tanto, uno de los efectos que tendría una observación influyente (y^*_{ijkl}) sobre el análisis, es que si $F_{Tab} < F_0 < F^*_{Tab}$ entonces puede rechazarse la hipótesis nula cuando en realidad no hay evidencia suficiente para hacerlo.

Por otra parte, cabe anotar que a medida que el valor de q tiende a ser muy grande, la diferencia entre los valores tabulados también aumenta. Sin embargo, es de esperar que el número de observaciones consideradas influyentes en un experimento no sea muy grande en relación al número total de observaciones.

4.1 Efecto sobre las sumas de cuadrados del modelo

Una forma clara de observar el efecto que tendrían las observaciones influyentes sobre las sumas de cuadrados del modelo, es analizar la diferencia entre las ecuaciones de éstas en el modelo balanceado y sus ecuaciones en el modelo desbalanceado. Partiendo de este punto, si se hace la diferencia entre la suma de cuadrados de los tratamientos planteada mediante la

ecuación presentada (3) y la suma de cuadrados de los tratamientos del modelo reducido, en una observación, se tiene que:

$$SC_{Trata} - SC^{*}_{Trata} = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\frac{y^{2}_{ijk.}}{n} - \frac{y^{2}_{....}}{N} - \left(\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\frac{y^{2}_{ijk.}}{n_{ijk}} - \frac{y^{*2}_{....}}{N^{*}}\right) \tag{18}$$

pero la primera expresión de la suma de cuadrados de los tratamientos del modelo desbalanceado, se puede expresar en términos del primer sumando de la ecuación de la suma de cuadrados de los tratamientos del modelo balanceado, es decir:

$$\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\frac{y^{2}_{ijk.}}{n_{ijk}} = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\frac{y^{2}_{ijk.}}{n} - \frac{y^{2}_{(ijk.)}}{n} + \frac{y^{*2}_{ijk.}}{n-1} \tag{19}$$

el término $y^{2}_{ijk.}$ corresponde al total del tratamiento ijk-ésimo en donde se encuentra la observación considerada influyente en el modelo balanceado, elevado al cuadrado y dividido por n, y el termino $y^{*2}_{ijk.}$ corresponde al total del mismo tratamiento ijk-ésimo pero en el modelo desbalanceado, notado por $(ijk)^{*}$ y dividido por el número de observaciones resultantes allí $(n-1)$.

Reemplazando (19) en (18) y luego, desarrollando y despejando se llega a que:

$$SC_{Trata} = \left(\frac{y^{2}_{ijk.}}{n} - \frac{y^{*2}_{ijk.}}{n-1}\right) - \left(\frac{y^{2}_{....}}{N} - \frac{y^{*2}_{....}}{N^{*}}\right) + SC^{*}_{Trata}$$

Por consiguiente, se tiene que la suma de cuadrados de los tratamientos en el modelo completo, es decir bajo la influencia de la observación y^{*}_{ijkl}, se puede expresar como la suma de cuadrados del modelo desbalanceado (modelo sin influencia) más el efecto de dicha observación sobre el tratamiento que la contiene (dado por la diferencia entre el cuadrado del total del tratamiento que la contiene en el modelo balanceado y el cuadrado del total del mismo tratamiento en el modelo desbalanceado, cada uno sobre el número de observaciones que contiene), corregido por el efecto de la observación sobre el total general del modelo dado por la diferencia entre el cuadrado del total del modelo balanceado, dividido por el número total de observaciones (N), y el cuadrado del total del modelo desbalanceado, dividido por su número total de observaciones $(N-1)$.

De esta manera, es claro que el efecto que tiene la observación influyente sobre las sumas de cuadrados de los tratamientos, es significativa a medida que la influencia sobre el total del tratamiento en donde se encuentre la

observación, sea grande, es decir si la observación guarda amplia diferencia con los valores de las demás réplicas en el mismo tratamiento.

Un resultado muy importante es la generalización de este hecho, para q observaciones influyentes en un diseño de efectos fijos con ω factores a tres niveles cada uno:

R1. El efecto de observaciones influyentes sobre la suma de cuadrados de los tratamientos de un diseño factorial 3^{ω}, está dado por la siguiente ecuación:

$$SC_{Trata} = \left(\sum_{s=1}^{S} \frac{y_s^2}{n} - \sum_{s=1}^{S} \frac{y_s^{*2}}{n_s}\right) - \left(\frac{y_{....}^2}{N} - \frac{y_{....}^{*2}}{N^*}\right) + SC^*_{Trata}$$

El primer término corresponde a la diferencia entre la suma de los S totales de los tratamientos ($s = 1, 2, \ldots, S$), donde se encuentren distribuidas las q observaciones influyentes en el modelo balanceado, elevados al cuadrado y dividido por el número de réplicas hechas en ellos; y la suma de los mismos S totales de los tratamientos pero del modelo desbalanceado, es decir, sin las q observaciones; elevados al cuadrado y ponderados por el número de observaciones en cada uno. El segundo término corresponde al efecto de las observaciones al nivel de los totales del modelo y el último a la suma de cuadrados del modelo reducido.

El subíndice s entonces, hace referencia a una combinación de los niveles de los ω factores involucrados, que conforman un tratamiento específico, es decir $s = (i, j, k, \ldots, w)$.

Al igual que con la suma de cuadrados de los tratamientos, puede analizarse el efecto de las observaciones influyentes, sobre las sumas de cuadrados de los efectos principales, las sumas de cuadrados de los efectos de las interacciones dobles y triples, realizando la diferencia entre las ecuaciones (10) y las ecuaciones (12). Los resultados se presentan a continuación:

- Efectos sobre sumas de cuadrados de efectos principales:

$$SC_A = \left(\frac{y_{i...}^2}{9n} - \frac{y_{i...}^{*2}}{9n-1}\right) - \left(\frac{y_{....}^2}{N} - \frac{y_{....}^{*2}}{N^*}\right) + SC^*_A$$

$$SC_B = \left(\frac{y_{.j..}^2}{9n} - \frac{y_{.j..}^{*2}}{9n-1}\right) - \left(\frac{y_{....}^2}{N} - \frac{y_{....}^{*2}}{N^*}\right) + SC^*_B$$

$$SC_C = \left(\frac{y_{..k.}^2}{9n} - \frac{y_{..k.}^{*2}}{9n-1}\right) - \left(\frac{y_{....}^2}{N} - \frac{y_{....}^{*2}}{N^*}\right) + SC_C^*$$

– Efectos sobre sumas de cuadrados de interacciones dobles:

$$SC_{AB} = \left(\frac{y_{ij..}^2}{3n} - \frac{y_{ij..}^{*2}}{3n-1}\right) - \left(\frac{y_{i...}^2}{9n} - \frac{y_{i...}^{*2}}{9n-1}\right) - \left(\frac{y_{.j..}^2}{9n} - \frac{y_{.j..}^{*2}}{9n-1}\right) + \left(\frac{y_{....}^2}{N} - \frac{y_{....}^{*2}}{N^*}\right) + SC_A^*$$

$$SC_{AC} = \left(\frac{y_{i.k.}^2}{3n} - \frac{y_{i.k.}^{*2}}{3n-1}\right) - \left(\frac{y_{i...}^2}{9n} - \frac{y_{i...}^{*2}}{9n-1}\right) - \left(\frac{y_{..k.}^2}{9n} - \frac{y_{..k.}^{*2}}{9n-1}\right) + \left(\frac{y_{....}^2}{N} - \frac{y_{....}^{*2}}{N^*}\right) + SC_{AC}^*$$

$$SC_{BC} = \left(\frac{y_{.jk.}^2}{3n} - \frac{y_{.jk.}^{*2}}{3n-1}\right) - \left(\frac{y_{.j..}^2}{9n} - \frac{y_{.j..}^{*2}}{9n-1}\right) - \left(\frac{y_{..k.}^2}{9n} - \frac{y_{..k.}^{*2}}{9n-1}\right) + \left(\frac{y_{....}^2}{N} - \frac{y_{....}^{*2}}{N^*}\right) + SC_{BC}^*$$

– Efectos sobre sumas de cuadrados de interacciones triples:

$$SC_{ABC} = \left(\frac{y_{ijk.}^2}{n} - \frac{y_{ijk.}^{*2}}{n-1}\right) - SC_A - SC_B - SC_C - SC_{AB} - SC_{AC} - SC_{BC} - \left(\frac{y_{....}^2}{N} - \frac{y_{....}^{*2}}{N^*}\right) + SC_{ABC}^*$$

Similarmente, resulta muy importante la generalización de este hecho para q observaciones influyentes en un diseño de efectos fijos con ω factores a tres niveles cada uno:

R2. El efecto de q observaciones influyentes sobre la suma de cuadrados de los efectos principales, de un diseño factorial 3^ω, está dado por la siguiente ecuación:

$$SC_{\omega_t} = \left(\sum_{s=1}^{S} \frac{y_s^2}{3^{\omega_t - 1} n} - \sum_{s=1}^{S} \frac{y_s^{*2}}{3^{\omega_t - 1} n - q}\right) - \left(\frac{y_{....}^2}{N} - \frac{y_{....}^{*2}}{N^*}\right) + SC_{\omega_t}^*$$

El primer termino corresponde a la diferencia de la suma de los totales de los S tratamientos donde se encuentran distribuidas las observaciones influyentes, elevados al cuadrado y dividido por el número

de observaciones en ellos; y la suma de los totales de los mismos S tratamientos, pero en el modelo desbalanceado. Esta diferencia es corregida por el efecto de las observaciones sobre los totales del modelo. El subíndice $t = 1, 2, \ldots, \omega$ denota el factor al que se hace referencia.

La adición de estos términos a las sumas de cuadrado de los efectos principales del modelo desbalanceado, es la influencia significativa que resulta en las sumas de cuadrado del modelo balanceado (bajo la influencia).

Las ecuaciones para mostrar los efectos sobre las sumas de cuadrados de las interacciones dobles, triples y las demás combinaciones de los factores, no se presentan por ser compleja su escritura. Sin embargo, se aclara que la interpretación y los resultados básicamente son los mismos encontrados para las sumas de cuadrados de los efectos principales.

4.2 Efecto sobre las estimaciones de los parámetros del modelo

Las estimaciones de los parámetros del modelo planteado en (1) están dadas en términos del promedio general, de los promedios de los tratamientos y de las interacciones entre los mismos [23]. Los estimadores para los parámetros del modelo desbalanceado se calculan de igual forma, como se muestra a continuación:

Estimaciones modelo balanceado	Estimaciones modelo desbalanceado
$\cdot\ \hat{\mu} = \bar{y}_{....}$	$\cdot\ \hat{\mu}^* = \bar{y}^*_{....}$
$\cdot\ \widehat{\alpha}_i = \bar{y}_{i...} - \bar{y}_{....}$	$\cdot\ \widehat{\alpha}^*_i = \bar{y}^*_{i...} - \bar{y}^*_{....}$
$\cdot\ \widehat{\beta}_j = \bar{y}_{.j..} - \bar{y}_{....}$	$\cdot\ \widehat{\beta}^*_j = \bar{y}^*_{.j..} - \bar{y}^*_{....}$
$\cdot\ \widehat{\gamma}_k = \bar{y}_{..k.} - \bar{y}_{....}$	$\cdot\ \widehat{\gamma}^*_k = \bar{y}^*_{..k.} - \bar{y}^*_{....}$
$\cdot\ \widehat{\alpha\beta}_{ij} = \bar{y}_{ij..} - \bar{y}_{i...} - \bar{y}_{.j..} + \bar{y}_{....}$	$\cdot\ \widehat{\alpha\beta}^*_{ij} = \bar{y}^*_{ij..} - \bar{y}^*_{i...} - \bar{y}^*_{.j..} + \bar{y}^*_{....}$
$\cdot\ \widehat{\alpha\gamma}_{ik} = \bar{y}_{i.k.} - \bar{y}_{i...} - \bar{y}_{..k.} + \bar{y}_{....}$	$\cdot\ \widehat{\alpha\gamma}^*_{ik} = \bar{y}^*_{i.k.} - \bar{y}^*_{i...} - \bar{y}^*_{..k.} + \bar{y}^*_{....}$
$\cdot\ \widehat{\beta\gamma}_{jk} = \bar{y}_{.jk.} - \bar{y}_{.j..} - \bar{y}_{..k.} + \bar{y}_{....}$	$\cdot\ \widehat{\beta\gamma}^*_{jk} = \bar{y}^*_{.jk.} - \bar{y}^*_{.j..} - \bar{y}^*_{..k.} + \bar{y}^*_{....}$
$\cdot\ \widehat{\alpha\beta\gamma}_{ijk} = \bar{y}_{ijk.} - \bar{y}_{i...} - \bar{y}_{.j..} - \bar{y}_{..k.} + \bar{y}_{....}$	$\cdot\ \widehat{\alpha\beta\gamma}^*_{ijk} = \bar{y}^*_{ijk.} - \bar{y}^*_{i...} - \bar{y}^*_{.j..} - \bar{y}^*_{..k.} + \bar{y}^*_{....}$

Por lo tanto, el efecto que sobre el estimador de la media general del modelo μ, pudieran causar las observaciones influyentes está dado por:

$$\hat{\mu} - \hat{\mu}^* = \bar{y}_{....} - \bar{y}^*_{....} \tag{20}$$

Teniendo en cuenta que las expresiones correspondientes a estos dos sumandos son:

$$\bar{y}_{....} = \frac{1}{N}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n} y_{ijkl} \qquad \text{y} \qquad \bar{y}^{*}_{....} = \frac{1}{N^{*}}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n_{ijk}} y_{ijkl}$$

con $\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3} n_{ijk} = N^{*} = N - q$, en donde q es el número de observaciones consideradas influyentes.

La media estimada del modelo reducido se puede escribir en términos del modelo completo partiendo de la siguiente igualdad:

$$\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n_{ijk}} y_{ijkl} = \sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n} y_{ijkl} - \sum_{s=1}^{q} y_s$$

con $s = 1, 2, \ldots, q$, y $\sum_{s=1}^{q} y_s$ la suma total de los valores de las observaciones influyentes en el modelo completo. Luego, la diferencia planteada en (20), puede expresarse como:

$$\hat{\mu} - \hat{\mu}^{*} = \frac{1}{N}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n_{ijk}} y_{ijkl} - \frac{1}{N-q}\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{n} y_{ijkl} + \frac{1}{N-q}\sum_{s=1}^{q} y_s$$

Desarrollando algebraicamente y despejando $\hat{\mu}$ se llega a que:

$$\hat{\mu} = \frac{1}{N}\sum_{s=1}^{q} y_s + \left(1 - \frac{q}{N}\right)\hat{\mu}^{*}$$

Lo que dice esta expresión es que la estimación del parámetro de la media global, involucrando las observaciones influyentes, resulta igual a la estimación del parámetro sin ellas presentes, ponderada por la proporción de observaciones no influyentes, más la suma de los valores de las observaciones influyentes, divididas por el número total de observaciones.

4.2.1 Efecto sobre las estimaciones de los efectos principales

Se puede demostrar que al igual que en el caso del parámetro de la media general del modelo, la estimación del parámetro de un efecto principal en el modelo completo, es igual a la estimación del parámetro para el mismo efecto en el modelo reducido, ponderada por la proporción de observaciones no influyentes que contienen los tratamientos en donde se involucra dicho factor; más la suma de los valores de las observaciones influyentes en el nivel considerado, divididas por número total de observaciones de dicho nivel.

Adicionalmente, se le resta el efecto dado por la diferencia entre las medias globales de ambos modelos, ponderada por el porcentaje de observaciones no influyentes en el grupo considerado.

$$\hat{\alpha}_i = \frac{1}{9n}\sum_{s=1}^{q_i} y_s + \left(1 - \frac{q_i}{9n}\right)\hat{\alpha}_i^* - \left(1 - \frac{q_i}{9n}\right)(\hat{\mu} - \hat{\mu}^*),$$

$$\hat{\beta}_j = \frac{1}{9n}\sum_{s=1}^{q_j} y_s + \left(1 - \frac{q_j}{9n}\right)\hat{\beta}_j^* - \left(1 - \frac{q_j}{9n}\right)(\hat{\mu} - \hat{\mu}^*)$$

$$\hat{\gamma}_k = \frac{1}{9n}\sum_{s=1}^{q_k} y_s + \left(1 - \frac{q_k}{9n}\right)\hat{\gamma}_k^* - \left(1 - \frac{q_k}{9n}\right)(\hat{\mu} - \hat{\mu}^*)$$

4.2.2 Efecto sobre las estimaciones de los efectos dobles y triples

De manera general, tomando la diferencia entre cualquier estimador de un parámetro del modelo completo, y el estimador del mismo parámetro en el modelo reducido, puede verse que la estimación del primero está dada en términos del segundo, más un término que corresponde a la suma de las observaciones influyentes, y sustrayendo la diferencia entre las medias globales de ambos modelos, ponderada por la proporción de observaciones no influyentes.

$$\widehat{\alpha\beta}_{ij} = \frac{1}{3n}\sum_{s=1}^{q_{ij}} y_s + \left(1 - \frac{q_{ij}}{9n}\right)\widehat{\alpha\beta}_{ij}^* - \left(1 - \frac{q_{ij}}{9n}\right)(\hat{\mu} - \hat{\mu}^*)$$

$$\widehat{\alpha\beta\gamma}_{ijk} = \frac{1}{n}\sum_{s=1}^{q_{ijk}} y_s + \left(1 - \frac{q_{ijk}}{n}\right)\widehat{\alpha\beta\gamma}_{ijk}^* - \left(1 - \frac{q_{ijk}}{n}\right)(\hat{\mu} - \hat{\mu}^*)$$

Es decir, que si existen observaciones influyentes, los estimadores de los parámetros del modelo, serán las estimaciones de los parámetros del modelo que excluye dichas observaciones más unos términos correspondientes al peso del número de observaciones influyentes y a su efecto en la media general; que modifican significativamente el valor del estimador si no se consideraran las observaciones influyentes.

R3 Cuando se tienen ω factores, la forma general de los estimadores bajo el modelo completo y el efecto de observaciones influyentes, está dada por:

$$\hat{\alpha}_i = \frac{1}{3^{\omega-1}n}\sum_{s=1}^{q_i} y_s + \left(1 - \frac{q_i}{3^{\omega-1}n}\right)\hat{\alpha}_i^* - \left(1 - \frac{q_i}{3^{\omega-1}n}\right)(\hat{\mu} - \hat{\mu}^*)$$

$$\widehat{\alpha\beta}_{ij} = \frac{1}{3^{\omega-2}n}\sum_{s=1}^{q_{ij}} y_s + \left(1 - \frac{q_{ij}}{3^{\omega-2}n}\right)\widehat{\alpha\beta}^*_{ij} - \left(1 - \frac{q_{ij}}{3^{\omega-2}n}\right)(\hat{\mu} - \hat{\mu}^*)$$

y así sucesivamente, hasta la interacción de todos los ω factores:

$$\hat{Z}_{ijk\ldots w} = \frac{1}{n}\sum_{s=1}^{q_{ijk\ldots w}} y_s + \left(1 - \frac{q_{ijk\ldots w}}{n}\right)\hat{Z}^*_{ijk\ldots w} - \left(1 - \frac{q_{ijk\ldots w}}{n}\right)(\hat{\mu} - \hat{\mu}^*)$$

donde los $y_s (s = 1, 2, \ldots, q_{ijk\ldots w})$ corresponden a las observaciones influyentes consideradas en el tratamiento dado por la combinación $(ijk\ldots w)$ de los niveles de los ω factores.

5 Aplicación

A continuación se presenta un ejemplo de un diseño factorial 3^3, citado por Melo, López y Melo [24] y estudiado por Méndez [25]. En una planta industrial se estudió el efecto de los factores días, operadores y concentraciones de solventes en el rendimiento de la planta. Días y operadores eran efectos cualitativos y las concentraciones fueron 0.5, 1.0 y 2.0, que aunque no son igualmente espaciadas, sus logaritmos si son igualmente espaciados, y éstos se usan si se desea observar la forma de la respuesta a través de este factor.

El diseño experimental fue completamente aleatorizado y los factores se consideraron fijos. Se hicieron tres repeticiones de cada uno de los 27 tratamientos. Los datos codificados, a los que se les restó 20 para simplificar los cálculos se presentan en la Tabla 3.

Tabla 3: Datos para el ejemplo de un diseño factorial 3^3.

Concentraciones (C)	**Días (D)**								
	5/14			5/15			5/16		
	Operadores (O)								
	O1	O2	O3	O1	O2	O3	O1	O2	O3
0.5	1.0	0.2	0.2	1.0	1.0	1.2	1.7	0.2	0.5
	1.2	0.5	0.0	0.0	0.0	0.0	1.2	0.7	1.0
	1.7	0.7	0.3	0.5	0.0	0.5	1.2	1.0	1.7
1.0	5.0	3.2	3.5	0.4	3.2	3.7	4.5	3.7	3.7
	4.7	3.7	3.5	3.5	3.0	4.0	5.0	4.0	4.5
	4.2	3.5	3.2	3.5	4.0	4.2	4.7	4.2	3.7
2.0	7.5	6.0	7.2	6.5	5.2	7.0	6.7	7.5	6.2
	6.5	6.2	6.5	6.0	5.7	6.7	7.5	6.0	6.5
	7.7	6.2	6.7	6.2	6.5	6.8	7.0	6.0	7.0

Como primera medida, se hizo una revisión gráfica de la información con el fin de observar si existen interacciones entre los niveles de los factores. En la Figura 1 se observa que los tres factores interactúan entre sí. Por otro lado, el interés es evaluar la existencia de observaciones influyentes dentro del conjunto de datos, de tal forma que si es afirmativa, pueda verse algunos de los efectos causados sobre las sumas de cuadrados, las estimaciones y las hipótesis a probar. Para el ejemplo, después de realizar una PROC GLM en el software estadístico SAS, se realizaron las pruebas mencionadas en la sección 3, sobre detección de datos influyentes, con el fin de comparar los resultados con los arrojados por la estadística F_q. El procedimiento para evaluar si una observación es influyente o no a través de la estadística F_1, es similar al utilizado en la distancia de Cook. Es decir, que deben evaluarse las N observaciones de modo que resultan N estadísticas F_1.

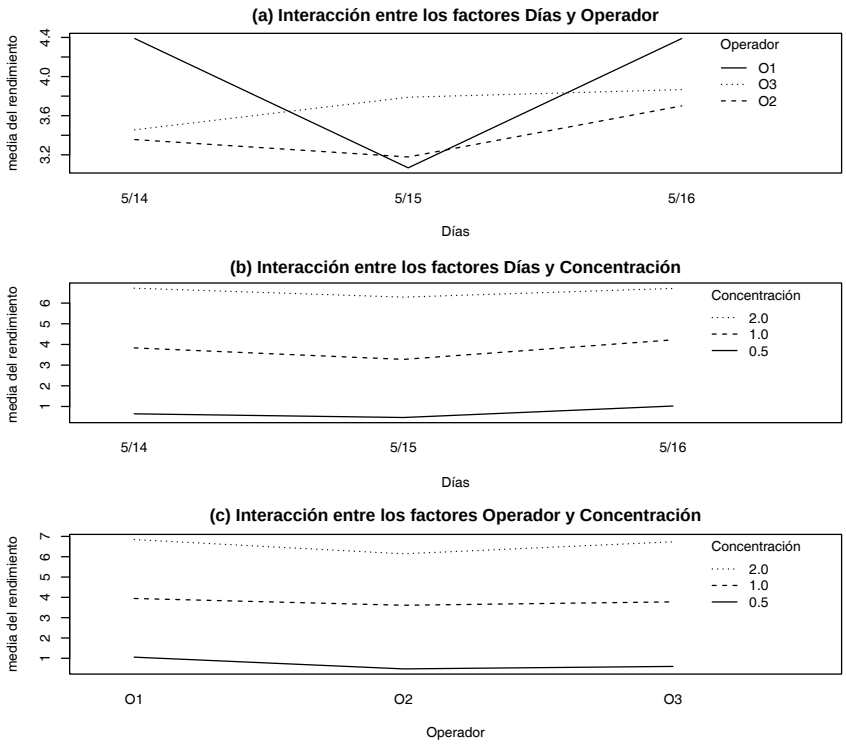

Figura 1: Interacción entre combinaciones de dos factores

Luego de realizar las evaluaciones para cada una de las observaciones, podría pensarse en realizarla para grupos de observaciones, pero este caso va sujeto al conocimiento del experimentador o un análisis más detallado de la información. Para el caso del ejemplo, se evaluaron una a una cada observación obteniendo los resultados presentados en la Tabla 4, en donde se aprecian los valores de la distancia de Cook, los Dffits y los valores para la estadística F_1, calculados para los datos del ejemplo; a través de los cuales se puede observar la validez de la metodología propuesta.

Tabla 4: Datos para el ejemplo de un diseño factorial 3^3.

Obs	D	O	C	Cook's	Dffits	F_1	Obs	D	O	C	Cook's	Dffits	F_1
1	5/14	O1	0.5	0.0084	-0.473	0.447	42	5/15	O2	1.0	0.0335	0.958	1.835
2	5/14	O1	0.5	0.0009	-0.157	0.049	43	5/15	O2	2.0	0.0335	-0.958	1.835
3	5/14	O1	0.5	0.0149	0.633	0.800	44	5/15	O2	2.0	0.0009	-0.157	0.049
4	5/14	O1	1.0	0.0125	0.579	0.678	45	5/15	O2	2.0	0.0456	112.47	2.530
5	5/14	O1	1.0	0.0004	0.105	0.022	46	5/15	O3	0.5	0.0373	101.32	0.293
6	5/14	O1	1.0	0.0175	-0.686	0.942	47	5/15	O3	0.5	0.0299	-0.903	1.631
7	5/14	O1	2.0	0.0066	0.420	0.353	48	5/15	O3	0.5	0.0004	-0.105	0.022
8	5/14	O1	2.0	0.0500	-118.1	2.789	49	5/15	O3	1.0	0.0066	-0.420	0.353
9	5/14	O1	2.0	0.0203	0.740	1.095	50	5/15	O3	1.0	0.0001	0.052	0.005
10	5/14	O2	0.5	0.0066	-0.420	0.353	51	5/15	O3	1.0	0.0051	0.367	0.270
11	5/14	O2	0.5	0.0001	0.052	0.005	52	5/15	O3	2.0	0.0026	0.262	0.137
12	5/14	O2	0.5	0.0051	0.367	0.270	53	5/15	O3	2.0	0.0017	-0.209	0.877
13	5/14	O2	1.0	0.0066	-0.420	0.353	54	5/15	O3	2.0	0.0001	-0.052	0.005
14	5/14	O2	1.0	0.0051	0.367	0.270	55	5/16	O1	0.5	0.0103	0.526	0.553
15	5/14	O2	1.0	0.0001	0.052	0.005	56	5/16	O1	0.5	0.0026	-0.262	0.137
16	5/14	O2	2.0	0.0017	-0.210	0.088	57	5/16	O1	0.5	0.0026	-0.262	0.137
17	5/14	O2	2.0	0.0004	0.105	0.022	58	5/16	O1	1.0	0.0051	-0.368	0.270
18	5/14	O2	2.0	0.0004	0.105	0.022	59	5/16	O1	1.0	0.0066	0.420	0.353
19	5/14	O3	0.5	0.0001	0.052	0.005	60	5/16	O1	1.0	0.0001	-0.052	0.005
20	5/14	O3	0.5	0.0026	-0.262	0.137	61	5/16	O1	2.0	0.0125	-0.579	0.671
21	5/14	O3	0.5	0.0017	0.210	0.877	62	5/16	O1	2.0	0.0175	0.686	0.942
22	5/14	O3	1.0	0.0009	0.157	0.049	63	5/16	O1	2.0	0.0004	-0.105	0.022
23	5/14	O3	1.0	0.0009	0.157	0.049	64	5/16	O2	0.5	0.0175	-0.686	0.942
24	5/14	O3	1.0	0.0037	-0.315	0.198	65	5/16	O2	0.5	0.0004	0.105	0.022
25	5/14	O3	2.0	0.0149	0.633	0.800	66	5/16	O2	0.5	0.0125	0.579	0.671
26	5/14	O3	2.0	0.0084	-0.473	0.447	67	5/16	O2	1.0	0.0066	-0.420	0.353
27	5/14	O3	2.0	0.0009	-0.157	0.049	68	5/16	O2	1.0	0.0001	0.052	0.005
28	5/15	O1	0.5	0.0232	0.794	1.261	69	5/16	O2	1.0	0.0051	0.367	0.270
29	5/15	O1	0.5	0.0232	-0.794	1.261	**70**	**5/16**	**O2**	**2.0**	**0.0930**	**164.82**	**5.433**
30	5/15	O1	0.5	0.0000	0.000	0.000	71	5/16	O2	2.0	0.0232	-0.794	1.261
31	**5/15**	**O1**	**1.0**	**0.3971**	**-417.8**	**34.91**	72	5/16	O2	2.0	0.0232	-0.794	1.261
32	**5/15**	**O1**	**1.0**	**0.0993**	**170.9**	**5.842**	73	5/16	O3	0.5	0.0299	-0.903	1.631
33	**5/15**	**O1**	**1.0**	**0.0993**	**170.9**	**5.842**	74	5/16	O3	0.5	0.0004	-0.105	0.022
34	5/15	O1	2.0	0.0066	0.420	0.353	75	5/16	O3	0.5	0.0373	101.32	2.053
35	5/15	O1	2.0	0.0051	-0.367	0.270	76	5/16	O3	1.0	0.0066	-0.420	0.353
36	5/15	O1	2.0	0.0001	-0.052	0.005	77	5/16	O3	1.0	0.0264	0.848	1.440
37	5/15	O2	0.5	0.0413	106.9	2.284	78	5/16	O3	1.0	0.0066	-0.420	0.353
38	5/15	O2	0.5	0.0103	-0.526	0.553	79	5/16	O3	2.0	0.0125	-0.579	0.671
39	5/15	O2	0.5	0.0103	-0.526	0.553	80	5/16	O3	2.0	0.0004	-0.105	0.022
40	5/15	O2	1.0	0.0037	-0.315	0.198	81	5/16	O3	2.0	0.0175	0.686	0.942
41	5/15	O2	1.0	0.0149	-0.633	0.800							

La estadística F_1 en este caso, debe probarse contra una $F_{(1,53,0.05)} = 4.023$, de tal forma que las observaciones con valores F_1 mayores a este valor, pueden considerarse influyentes. Según las distancias de Cook, las

observaciones que resultan influyentes son: 31, 32, 33 y 70. Esto concuerda totalmente con el criterio de la metodología propuesta, en donde las observaciones con valores F_1 mayores a 4.023, corresponden a las mismas detectadas por Cook. Incluso al observar el criterio de Cook y el de los Dffits con el de la estadística F_1 en la observación 30, es claro que coinciden al determinar como nulos esos valores (ver Figura 2).

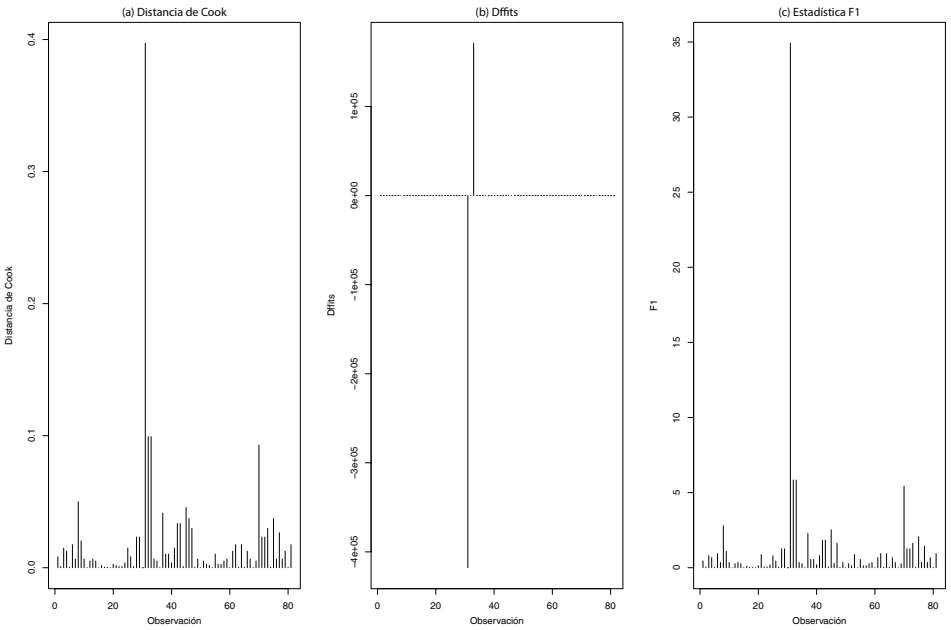

Figura 2: Distancia de Cook, Dffits y estadística F_1 para los datos del ejemplo

Como se observa en el gráfico, comparados a escala, los resultados son similares. La observación número 31, correspondiente al valor 7.0 del rendimiento de la planta, es la observación con mayor influencia, según lo dice la estadística F_1. Las estimaciones de los parámetros realizadas con el modelo completo se presentan en la Tabla 5, junto a las estimaciones del modelo reducido (excluyendo la observación 31). Es claro que hay diferencias significativas en las estimaciones del modelo cuando se excluye la observación influyente. Sin embargo, el efecto de ésta se ve reflejado en las características del modelo.

Tabla 5: Estimaciones de los parámetros del modelo para los datos del ejemplo.

	Parámetro	
Efecto	**Est**	**Est***
μ	3.688	3.729
α_1	0.046	0.005
α_2	-0.343	-0.271
α_3	0.298	0.256
β_1	0.261	0.356
β_2	-0.277	-0.318
β_3	0.016	-0.025
γ_1	-2.977	-3.018
γ_2	0.090	0.179
γ_3	2.886	2.845
$\alpha\beta_{11}$	7.770	7.757
$\alpha\beta_{12}$	7.274	7.397
$\alpha\beta_{13}$	7.082	7.205
$\alpha\beta_{21}$	6.837	7.044
$\alpha\beta_{22}$	7.485	7.495
$\alpha\beta_{23}$	7.804	7.814
$\alpha\beta_{31}$	10.756	10.879
$\alpha\beta_{32}$	7.000	6.993
$\alpha\beta_{33}$	4.370	4.494
$\alpha\gamma_{11}$	7.263	7.386
$\alpha\gamma_{12}$	7.385	7.379
$\alpha\gamma_{13}$	7.478	7.601

	Parámetro	
Efecto	**Est**	**Est***
$\alpha\gamma_{21}$	7.474	7.484
$\alpha\gamma_{22}$	7.219	7.458
$\alpha\gamma_{23}$	7.433	7.443
$\alpha\gamma_{31}$	7.389	7.512
$\alpha\gamma_{32}$	7.522	7.516
$\alpha\gamma_{33}$	7.215	7.338
$\beta\gamma_{11}$	7.459	7.446
$\beta\gamma_{12}$	7.282	7.581
$\beta\gamma_{13}$	7.385	7.372
$\beta\gamma_{21}$	7.419	7.542
$\beta\gamma_{22}$	7.485	7.479
$\beta\gamma_{23}$	7.222	7.346
$\beta\gamma_{31}$	7.248	7.371
$\beta\gamma_{32}$	7.359	7.353
$\beta\gamma_{33}$	7.519	7.642
$\alpha\beta\gamma_{111}$	7.658	7.686
$\alpha\beta\gamma_{112}$	7.925	7.823
$\alpha\beta\gamma_{113}$	7.728	7.756
$\alpha\beta\gamma_{121}$	7.362	7.526
$\alpha\beta\gamma_{122}$	7.295	7.330
$\alpha\beta\gamma_{123}$	7.165	7.330
$\alpha\beta\gamma_{131}$	6.769	6.934

	Parámetro	
Efecto	**Est**	**Est***
$\alpha\beta\gamma_{132}$	6.936	6.970
$\alpha\beta\gamma_{133}$	7.540	7.704
$\alpha\beta\gamma_{211}$	7.247	7.162
$\alpha\beta\gamma_{212}$	6.147	6.965
$\alpha\beta\gamma_{213}$	7.117	7.032
$\alpha\beta\gamma_{221}$	7.617	7.668
$\alpha\beta\gamma_{222}$	7.617	7.539
$\alpha\beta\gamma_{223}$	7.221	7.272
$\alpha\beta\gamma_{231}$	7.558	7.609
$\alpha\beta\gamma_{232}$	7.891	7.813
$\alpha\beta\gamma_{233}$	7.962	8.013
$\alpha\beta\gamma_{311}$	7.473	7.501
$\alpha\beta\gamma_{312}$	7.773	7.671
$\alpha\beta\gamma_{313}$	7.310	7.338
$\alpha\beta\gamma_{321}$	7.277	7.441
$\alpha\beta\gamma_{322}$	7.543	7.578
$\alpha\beta\gamma_{323}$	7.280	7.445
$\alpha\beta\gamma_{331}$	7.417	7.582
$\alpha\beta\gamma_{332}$	7.251	7.285
$\alpha\beta\gamma_{333}$	7.054	7.219

Tabla 6: Procedimiento GLM de SAS, para evaluar el modelo completo (balanceado)

Fuente	**DF**	**Suma de Cuadrados**	**Cuadrado Medio**	F	$Pr > F$
Modelo	26	485.49	18.67	62.5	<.0001
Error	54	16.13	0.30		
Total	80	501.62			

R-cuadrado	Coef Var	Raiz MSE	Y Media
0.9678	14.82	0.55	3.69

Fuente	**DF**	**Tipo I SS y Tipo III SS**	**Cuadrado Medio**	F	$Pr > F$
D	2	5.63	2.81	9.42	0.000
O	2	3.90	1.95	6.53	0.003
C	2	464.38	232.19	777.17	<.0001
D*O	4	6.99	1.75	5.85	0.001
D*C	4	0.98	0.24	0.82	0.520
O*C	4	0.81	0.20	0.68	0.609
D*O*C	8	2.80	0.35	1.17	0.333

Este resultado dice que el modelo rechaza la hipótesis nula de que los efectos de los tratamientos son iguales un nivel de significancia del 5 % y una $F_{Tab} = 1.701636$.

Tabla 7: Procedimiento GLM de SAS para evaluar el modelo reducido (desbalanceado)

Fuente	**DF**	**Suma de Cuadrados**	**Cuadrado Medio**	F	$Pr > F$
Modelo	26	480.96	18.50	100.8	<.0001
Error	53	9.73	0.18		
Total	79	490.69			

R-cuadrado	Coef Var	Raiz MSE	Y Media
0.9802	11.49	0.43	3.73

Fuente	**DF**	**Tipo I SS**	**Cuadrado Medio**	F	$Pr > F$
D	2	3.69	1.84	10.04	0.000
O	2	5.85	2.92	15.93	<.001
C	2	465.32	232.66	1267.75	<.001
D*O	4	3.90	0.98	5.31	0.001
D*C	4	0.48	0.12	0.65	0.628
O*C	4	0.72	0.18	0.98	0.425
D*O*C	8	1.00	0.13	0.68	0.704

Fuente	**DF**	**Tipo III SS**	**Cuadrado Medio**	F	$Pr > F$
D	2	3.63	1.82	9.90	0.000
O	2	5.59	2.80	15.24	<.001
C	2	465.14	232.57	1267.25	<.001
D*O	4	3.92	0.98	5.34	0.001
D*C	4	0.48	0.12	0.65	0.628
O*C	4	0.70	0.18	0.95	0.440
D*O*C	8	1.00	0.13	0.68	0.704

Este resultado dice que el modelo rechaza la hipótesis nula de que los efectos de los tratamientos son iguales un nivel de significancia del 5 % y un $F^*_{Tab} = 1.71$. Aunque el resultado es el mismo, es decir, ambos modelos rechazan la hipótesis nula superando los valores $F_{Tab} = 1.70$ y $F^*_{Tab} = 1.71$, con $F_0 = 62.5$ y $F^*_0 = 100.8$. Sin embargo, la diferencia entre estos dos últimos es grande, es decir, F^*_0 casi duplica a F_0.

El efecto de la observación 31, influyente en este caso, sobre el estadístico de prueba del análisis de varianza, si fuese el valor tabulado más pequeño

que el del ejemplo, podría llevar a conclusiones erróneas. Es decir, si el valor tabulado estuviese entre 62.5 y 100.8, el análisis del modelo completo no rechazaría la hipótesis nula y diría que los efectos de los tratamientos son los mismos, mientras que al excluir la influencia, sería claro que no es así.

Por otra parte, cabe anotar que las características de los modelos varían significativamente. El modelo balanceado ajustado (ver Tabla 6) tiene un $R^2 = 0.9678$ y un coeficiente de variación de 14.822, mientras que el modelo desbalanceado ajustado (ver Tabla 7) tiene un $R^2 = 0.9801$ y un coeficiente de variación de 11.488; lo que indica que el modelo ajustado, excluyendo la influencia, es mucho mejor que cuando ésta se tiene en cuenta.

6 Conclusiones

Se ha cerrado un primer paso en la construcción de una herramienta estadística que es útil como monitor de alerta a la influencia de observaciones, en conjuntos de datos con un diseño factorial de efectos fijos 3^ω. Esta herramienta se refiere a la estadística F_q, que a un nivel de significancia $\alpha\,\%$ sigue una distribución F con q y $N - 27 - q$ grados de libertad y rechaza la hipótesis nula si $F_q > F_{(q,N-27-q,\alpha)}$.

Por otra parte, se ha mostrado que la eficiencia de la estadística, presenta un criterio igualmente objetivo y sólido como el de otras alternativas y técnicas reconocidas en la detección de observaciones influyentes. Sin embargo, al momento del análisis, se sugiere no excluir el uso de dichas técnicas, de tal manera que en la aplicación siempre se comparen sus resultados con el estadístico propuesto en este artículo.

Como este trabajo constituye una propuesta teórica, se sugiere que trabajos posteriores sometan los resultados a estudios comparativos y de aplicación. Adicionalmente, el desarrollo teórico y práctico de la estadística F_q para diseños factoriales 3^ω, no resulta complicada, por lo que puede realizarse el mismo desarrollo para diseños factoriales con otro número de niveles, con la garantía de obtener resultados como los logrados en este trabajo.

Agradecimientos

Los autores agradecen los comentarios y sugerencias de los evaluadores anónimos por sus valiosas contribuciones. Este trabajo fue parcialmente apoyado por el grupo de Estadística Aplicada en la Investigación Experimental, Industria y Biotecnología de la universidad Nacional de Colombia.

Referencias

[1] M. T. Jiménez, "Ajuste de factoriales 2^k con presencia de observaciones influyentes y valores faltantes mediante modelos de regresión," Master's thesis, Universidad Nacional de Colombia, Bogotá, 2000. 122

[2] C. R., "Assessment of Local Influence (with discussion)," *Journal of the Royal Statistical Society, Series B*, vol. 48, pp. 133–169, 1986. 123

[3] ——, "Robust Test for the Equality of Variantes," *Technometrics*, vol. 19, pp. 15–18, 1977. 123

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the EM Algorithm (with discussion)," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977. 123

[5] Thomas W. and R. D. Cook, "Assessing Influence on Regression Coefficients in Generalized Linear Models," *Biometrika*, vol. 76, no. 4, pp. 741–749, 1989. [Online]. Available: http://www.jstor.org/stable/2336634 123

[6] ——, "Assessing Influence on Predictions from Generalized Linear Models," *Technometrics*, vol. 32, no. 1, pp. 59–65, 1990. [Online]. Available: http://dx.doi.org/10.2307/1269845 123

[7] A. J. Lawrence, "Local and Deletion Influence," in *Directions in Robust Statistics and Diagnostics, Part I*, W. S. . S. Weisberg, Ed. Berlin: Springer, 1991, pp. 141–157. 123

[8] E. B. Andersen, "Diagnostics in Categorical Data Analysis," *Journal of the Royal Statistical Society, Series B*, vol. 54, no. 3, pp. 784–791, 1992. [Online]. Available: http://www.jstor.org/stable/2345858 123

[9] F. Critchley, R. A. Atkinson, G. Lu, and E. Biazi, "Influence Analysis Based on the Case Sensitivity Function," *Journal of the Royal Statistical Society, Series B*, vol. 63, no. 2, pp. 307–323, 2001. [Online]. Available: http://dx.doi.org/10.1111/1467-9868.00287 123

[10] H. T. Zhu and S. Y. Lee, "Local Influence for Incomplete Data Models," *Journal of the Royal Statistical Society, Series B*, vol. 63, no. 1, pp. 111–126, 2001. [Online]. Available: http://www.jstor.org/stable/2680637 123

[11] R. Tsai and U. Böckenholt, "Two-Level Linear Paired Comparison Models: Estimation and Identifiable Issues," *Mathematical Social Science*, vol. 43, no. 3, pp. 429–449, 2002. [Online]. Available: http://dx.doi.org/10.1016/S0165-4896(02)00019-7 123

[12] S. Y. Lee and N. S. Tang, "Local Influence Analysis of Nonlinear Structural Equation Models," *Psychometrica*, vol. 69, no. 4, pp. 573–592, 2004. [Online]. Available: http://dx.doi.org/10.1007/BF02289856 123

[13] L. Xu, W. Y. Poon, and S. Y. Lee, "Influence Analysis for the Factor Analysis Model with Ranking Data," *British Journal of Mathematical and Statistical Psychology*, vol. 61, no. 1, pp. 133–161, 2008. [Online]. Available: http://dx.doi.org/10.1348/000711006X169991 123

[14] D. C. Montgomery, *Design and Analysis of Experiments*, 8th ed. New York: John Wiley & Sons, 2012. 124, 129

[15] D. A. Belsley, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* New York: John Wiley & Sons, 1980. 126

[16] D. C. Hoaglin and R. E. Welsch, "The Hat Matrix in Regression and ANOVA," *The American Statistician*, vol. 32, no. 1, pp. 17–22, 1978. 126

[17] D. Peña-Sánchez, *Estadística Modelos y Métodos.* Madrid: Alianza Editorial, 1995. 126

[18] N. R. Draper and H. Smith, *Applied Regresión Analysis*, 3rd ed. New York: John Wiley & Sons, 1998. 126

[19] N. Draper and J. A. John, "Influential Observations and Outliers in Regression," *Technometrics*, vol. 23, no. 1, pp. 21–26, 1981. 126

[20] M. S. Bartlett, "Some Examples of Statistical Methods of Research in Agriculture y Applied Botany," *Journal Royal of the Statistical Society B*, vol. 4, pp. 137–170, 1937. 126

[21] J. A. Jiménez, "Propuesta metodológica para imputar valores no influyentes en modelos de regresión lineal múltiple con información incompleta," Master's thesis, Universidad Nacional de Colombia, Bogotá, 1999. 126

[22] ——, "Un criterio para identificar datos atípicos," *Revista Colombiana de Estadística*, vol. 27, no. 2, pp. 109–121, 2011. [Online]. Available: http://www.revistas.unal.edu.co/index.php/estad/article/view/28709 127

[23] D. C. Montgomery, *Introduction to Linear Regression Analysis*. New York: John Wiley & Sons, 1992. 138

[24] O. O. Melo, L. A. López, and S. E. Melo, *Diseño de Experimentos: Métodos y Aplicaciones*, 1st ed. Bogotá: Facultad de Ciencias, Universidad Nacional de Colombia, 2007. 141

[25] I. Méndez, "Diseño de Experimentos," in *Memorias del X Coloquio Distrital de Matemáticas y Estadística*, Bogotá, 1993. 141