PROBABILIDADES POSTERIORES PARA IDENTIFICAR EFECTOS ACTIVOS EN EXPERIMENTOS FACTORIALES SIN Y CON POSIBLES DATOS ANÓMALOS

Román de la Vara Salazar

Comunicación Técnica No I-04-09/18-09-2004
(PE/CIMAT)

Probabilidades Posteriores Para Identificar Efectos Activos en Experimentos Factoriales Sin y Con Posibles Datos Anómalos

por

Román de la Vara Salazar
Ingeniería de Calidad
Centro de Investigación en Matemáticas
Apartado Postal 402
Guanajuato, Gto., 36000
MEXICO
delavara@cimat.mx

Resumen

Se presentan tres programas en SPLUS para calcular y graficar las probabilidades posteriores de efectos en experimentos factoriales. Dos de los programas son para el caso normal sin datos atípicos. El tercer programa es para el caso que sí contempla la posibilidad de observaciones anómalas. Una ventaja de estos programas es que están en código de SPLUS, lo que facilita su uso, dada la amplia difusión que este paquete comercial tiene.

Palabras y frases clave: Probabilidades Posteriores, Método Bayesiano, Experimentos Factoriales No-Replicados, Observaciones Anómalas, Factores Activos.

1 Introducción

Frecuentemente en la industria no es posible correr repeticiones en cada combinación de un experimento factorial completo, e incluso es común que solo se corra una réplica de una fracción de éste. En estas situaciones los datos consisten de una sola observación en cada combinación o punto experimental. Este hecho genera complicaciones para determinar cuáles efectos están activos al no contar con un estimador independiente de la varianza del error.

El primer método que se propuso para el análisis de factoriales no replicados, y quizás todavía el más utilizado, es el método de Daniel (1959), que consiste en graficar los efectos (coeficientes del modelo) en papel de probabilidad normal y considerar activos los efectos que no se alinean e inertes los que se alinean. La línea de referencia es la que señalan los puntos correspondientes a los efectos pequeños, ya que esa tendencia es la esperada si los efectos poblacionales fueran nulos. El problema es que en muchos experimentos no es fácil decidir visualmente si un efecto está lo suficientemente alejado de dicha línea como para concluir que es significativo. Esta subjetividad del gráfico de Daniel es una de las razones por las que en los últimos 25 años se hayan propuesto por diferentes autores no menos de 20 métodos que buscan ser objetivos al decidir cuáles efectos están activos en el experimento. Una revisión de la mayoría de los métodos propuestos hasta ahora para el caso normal se puede ver el trabajo de Hamada y Balakrishnan (H&B,1998). La mayoría de los métodos explotan de alguna manera el "principio de escasez" de efectos, de que solo algunos (digamos entre 20% y 30%) de los efectos

estarán activos, y construyen con los efectos más pequeños un "pseudoerror" o un error de referencia contra el cual evaluar la significancia de los efectos restantes.

Uno de tales métodos es el método bayesiano de Box y Meyer (B&M,1986) que busca subsanar la subjetividad del gráfico de Daniel al calcular las probabilidades a posteriori de todos los posibles grupos de efectos activos que se pueden formar con los factores considerados en el experimento, y de aquí, marginalizando, obtiene las probabilidades posteriores de que cada efecto sea activo. Una dificultad del método es la cantidad de modelos o grupos de efectos posibles que se pueden formar, cantidad que se incrementa geométricamente al aumentar el número de factores que se estudian. De aquí que en el mismo artículo B&M proponen un enfoque basado en la mezcla de normales para calcular las probabilidades posteriores de efectos como resultado de integrales que se puede resolver por métodos numéricos. Un programa en FORTRAN que aplica este último enfoque se puede ver en Stephenson y Hulting (1989).

Dado el incremento de la capacidad de cómputo, en la actualidad es posible realizar en cuestión de segundos el cálculo exhaustivo de las probabilidades posteriores de todos los modelos posibles para factoriales hasta con 15 efectos. Precisamente el primer programa que se presenta en este trabajo calcula las probabilidades posteriores de todos los posibles modelos, y de allí obtiene las probabilidades posteriores de los efectos.

En el segundo programa se muestra que las integrales del enfoque de mezcla de normales se pueden resolver numéricamente mediante una partición y las suma de las áreas de interés. Este programa se puede modificar fácilmente para analizar factoriales con más de 15 efectos.

Una situación mucho más complicada, y no considerada por Stephenson y Hulting (1989), es cuando se contempla la posibilidad de observaciones anómalas en el experimento (ver Aguirre-Torres y Pérez-Trejo 2001). El cálculo exhaustivo en esta situación consistiría en obtener las probabilidades posteriores de cada combinación de columnas y de renglones de la matriz $\mathbf{X}$, cálculos que rebasan la capacidad de cómputo usual actual para factoriales con 15 efectos o más. Por ello los mismos autores (Box y Meyer, 1987) proponen un procedimiento iterativo que es el que se realiza en el tercer programa.

Para ilustrar la operación de los programas, todos incluyen y analizan el mismo ejemplo tomado de Box y Meyer (1987), que es un factorial 2^4 no replicado. Se esperaría que con cualquiera de los tres programas se llegue a

la misma conclusión en un experimento sin datos atípicos.

Los tres programas están escritos para el caso particular de un experimento factorial con 15 efectos y 16 corridas experimentales. Esto es, se pueden usar sin mayores cambios para los factoriales 2^4, 2^{5-1}, 2^{6-2}, 2^{7-3} y 2^{8-4}. El método exhaustivo del primer programa se modifica fácilmente a factoriales más pequeños simplemente declarando el número de efectos de interés y ajustando los contrastes al tamaño deseado. Para factoriales más grandes la opción es modificar el Programa 2 que utiliza integración numérica: se declara el número de efectos de interés, se ajusta el tamaño de los contrastes y de las dos funciones que involucran los efectos de manera directa.

El caso que admite la posibilidad de datos anómalos (Programa 3) se puede modificar sin problemas para factoriales más pequeños, ajustando el tamaño de los contrastes; para factoriales más grandes la cantidad de cálculos es inmanejable para la capacidad de cómputo actual. Cabe decir que si se evita el cálculo del conjunto potencia usando en su lugar ciclos anidados, sí es posible analizar factoriales más grandes enfocándose a los efectos de mayor jerarquía y a una limitada cantidad de datos anómalos.

2 Programa 1: Cálculo Exhaustivo de Probabilidades Posteriores

Sea $\mathbf{T}$ =(T_1, T_2,..., T_v) el vector de los v efectos estimados de la manera usual en un experimento factorial o arreglo ortogonal. Suponga que con probabilidad $1-\alpha$ el efecto T_i sigue una distribución normal con media cero y varianza σ^2 y con probabilidad α sigue una distribución normal con media cero y varianza $k^2\sigma^2$. Sea $a_{(r)}$ el evento que un particular conjunto de r efectos sea activo y sea $\mathbf{T}_{(r)}$ el correspondiente conjunto de efectos estimados. Box y Meyer (1986) muestran que la probabilidad posterior de que $\mathbf{T}_{(r)}$ sean los efectos activos está dada por

$$p\left(a_{(r)}\,|\mathbf{T},\alpha,k\right) \propto \left[\frac{\alpha k^{-1}}{1-\alpha}\right]^{r}\left[1-\varphi f_{(r)}\right]^{-v/2}, \qquad (1)$$

donde $\varphi = 1-1/k^2$ y $f_{(r)} = \mathbf{T}'_{(r)}\mathbf{T}_{(r)}/\mathbf{T}'\mathbf{T}$ es la fracción de la suma de cuadrados que se atribuye a los efectos en $\mathbf{T}_{(r)}$ (ver también Box y Tiao (1968) y Meyer (1987)). La probabilidad marginal p_i de que el efecto i está

activo, dados $\mathbf{T}$, α y k, es

$$p_i = \sum_{(r):i \ es \ activo} p\left(a_{(r)} \,|\mathbf{T}, \alpha, k\right). \tag{2}$$

Esto es, se suman las probabilidades posteriores de todos los conjuntos $a_{(r)}$ que contienen el efecto de interés.

En el Programa 1 que se enlista enseguida se obtienen de manera exhaustiva las probabilidades p_i de todos los grupos posibles de efectos usando la fórmula (1) y luego, normalizando y aplicando la fórmula (2), se obtienen las posteriores de cada efecto considerando para fines de ilustración el caso particular de un experimento factorial 2^4. En este factorial se tienen 15 efectos con los cuales se pueden construir $2^{15} = 32768$ modelos o conjuntos de efectos, contando desde el modelo constante (sin efectos activos) hasta el modelo con los 15 efectos (todos activos).

Listado de Programa 1

```
# MÉTODO DE BOX Y MEYER PARA EL ANÁLISIS DE FACTORIALES NO
# REPLICADOS CASO NORMAL SIN OBSERVACIONES ATÍPICAS
remove(ls()) # REMOVER OBJETOS DE LA MEMORIA
# INFORMACIÓN BÁSICA Y A PRIORI (nn=número de combinaciones
# experimentales, alfa=probabilidad a priori de que un efecto

# sea activo,k=factor de inflamiento de la desviación estándar
# de los efectos activos, y=vector de observaciones o datos)
nn1 <- 15; alf <- 0.2; k <- 10; fi <- 1-1/k^2
y <- c(47.46,49.62,43.13,46.31,51.47,48.49,49.34,46.10,
 46.76,48.56,44.83,44.45,59.15,51.33,47.02,47.90)
# CONSTRUCCIÓN DEL CONJUNTO POTENCIA QUE DEFINE TODOS LOS
# MODELOS POSIBLES
powerSet <- function(x)
{
    K <- NULL
    for(m in x)
        K <- rbind(cbind(K, F), cbind(K, T))
```

```
    apply(K, 1, function(x, s) s[x], s = x)
}
xx1 <- powerSet(1:nn1)
xx1 <- xx1[-1]
# CONTRASTES Y MATRIZ DE SIGNOS
X0 <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
A <- c(-1, 1,-1, 1,-1, 1,-1, 1,-1, 1,-1, 1,-1, 1,-1,1)
B <- c(-1,-1, 1, 1,-1,-1, 1, 1,-1,-1, 1, 1,-1,-1, 1,1)
C <- c(-1,-1,-1,-1, 1, 1, 1, 1,-1,-1,-1,-1, 1, 1, 1,1)
D <- c(-1,-1,-1,-1,-1,-1,-1,-1, 1, 1, 1, 1, 1, 1, 1,1)
E <- A*B; G <- A*C; H <- A*D; I <- B*C; J <- B*D; K <- C*D
L <- A*B*C; M <- A*B*D; N <- A*C*D; O <- B*C*D; P<-A*B*C*D

Cr <- cbind(A,B,C,D,E,G,H,I,J,K,L,M,N,O,P)
# CALCULANDO LAS SUMAS DE CUADRADOS DE EFECTOS
Efec <- rep(0,nn1)
for (i in 1:nn1)
{
    Efec[i] <- t(Cr[,i])%*%y/8
}
SC <- (t(Efec)%*%Efec)
# VECTOR Y VALOR DONDE SE GUARDAN LAS PROBABILIDADES POSTE-
# RIORES DE EFECTO ACTIVO Y SU SUMA
pEF <- rep(0,nn1+1); ppT <- 0
# PROBABILIDAD POSTERIOR DEL MODELO SIN TÉRMINOS
ppT <- 1; pEF[1] <- 1
# PROBABILIDAD POSTERIOR DE MODELOS CON UNO HASTA QUINCE
# TÉRMINOS
numdat1 <- length(xx1)
for(i in 1:numdat1)
    {
    ps <- xx1[[i]]
    te <- length(ps)
        Efec1 <- Efec[ps]
        pp <- (((alf*k^-1)/(1-alf))^te)
               *(1-fi*sum(Efec1^2)/SC)^-7.5
```

```
    pEF[ps+1] <- pEF[ps+1] + pp
    ppT <- ppT + pp
    }
# NORMALIZACIÓN Y GRAFICACIÓN DE PROBABILIDADES POSTERIORES
# DE EFECTO ACTIVO
names <- c("0","A", "B", "C", "D", "AB", "AC", "AD", "BC",
"BD", "CD", "ABC",
"ABD", "ACD", "BCD", "ABCD")
barplot(pEF/ppT, names=names,xlab="Efectos", ylab="Prob.Pos-
terior de Estar Activo")
### FIN DE PROGRAMA 1 ###
```

En la Sección 5 se muestran las salidas de todos los programas. En particular, la salida del primer programa es la Figura 1, donde se observan relativamente altas las probabilidades posteriores de los efectos B y C, pero con valores no contundentes de alrededor de 0.5.

3 Programa 2: Cálculo de Probabilidades Posteriores con Integración Numérica

Una alternativa al método exhaustivo, es resolver numéricamente las integrales que resultan del enfoque de mezcla de normales (ver Stephenson y Hulting, 1989). Bajo el supuesto de que los efectos T_1, $T_2, ...,$ T_v son una muestra independiente de la distribución $(1-\alpha)N(0,\sigma^2) + \alpha N(0, k^2\sigma^2)$, la probabilidad posterior de que el efecto T_i venga de la normal más amplia $N(0, k^2\sigma^2)$, dado σ, está dada por

$$P\left(i\ activo\,|T_i, \sigma\right) = \frac{\alpha\frac{1}{k}\exp\left\{\frac{-T_i^2}{2k^2\sigma^2}\right\}}{\alpha\frac{1}{k}\exp\left\{\frac{-T_i^2}{2k^2\sigma^2}\right\} + (1-\alpha)\exp\left\{\frac{-T_i^2}{2\sigma^2}\right\}}, \qquad (3)$$

donde α es la probabilidad a priori de que el efecto está activo, σ es la desviación estándar de un efecto inerte y k es el factor de inflamiento de la varianza debido a un efecto activo. Para que esta probabilidad no dependa

del parámetro σ este se integra respecto a su distribución posterior, es decir,

$$P\left(i\ activo\,|\mathbf{T}\right) = \int_0^\infty P\left(i\ activo\,|T_i,\sigma\right) P(\sigma\,|\mathbf{T})d\sigma \tag{4}$$

donde

$$\begin{aligned} P(\sigma\,|\mathbf{T}) &= \frac{P\left(\sigma,\mathbf{T}\right)}{P(\mathbf{T})} \\ &= \frac{\frac{1}{\sigma^{v+1}}\prod_{j=1}^{v}\left[\alpha\frac{1}{\sqrt{2\pi}k}\exp\left\{\frac{-T_i^2}{2k^2\sigma^2}\right\}+\frac{(1-\alpha)}{\sqrt{2\pi}}\exp\left\{\frac{-T_i^2}{2\sigma^2}\right\}\right]}{\int_0^\infty \frac{1}{\sigma^{v+1}}\prod_{j=1}^{v}\left[\alpha\frac{1}{\sqrt{2\pi}k}\exp\left\{\frac{-T_i^2}{2k^2\sigma^2}\right\}+\frac{(1-\alpha)}{\sqrt{2\pi}}\exp\left\{\frac{-T_i^2}{2\sigma^2}\right\}\right]d\sigma}. \end{aligned}$$

De aquí que las probabilidades posteriores de interés sean el cociente de integrales dado por

$$P\left(i\ activo\,|\mathbf{T}\right) = \frac{\int_0^\infty P\left(i\ activo\,|T_i,\sigma\right) P\left(\sigma,\mathbf{T}\right) d\sigma}{\left[\int_0^\infty P\left(\sigma,\mathbf{T}\right) d\sigma\right]}, \tag{5}$$

que se calculan con el siguiente programa usando sumas de Riemann con una partición y escala adecuadas. La probabilidad posterior de que ningún efecto está activo se calcula como

$$P\left(ninguno\ activo\,|\mathbf{T}\right) = \frac{\int_0^\infty \prod_{i=1}^{v}\left[1-P\left(i\ activo\,|T_i,\sigma\right)\right] P\left(\sigma,\mathbf{T}\right) d\sigma}{\left[\int_0^\infty P\left(\sigma,\mathbf{T}\right) d\sigma\right]}. \tag{6}$$

El Programa 2 que se enlista a continuación calcula los cocientes de integrales dados por las ecuaciones (5) y (6).

Listado de Programa 2

```
# METODO DE BOX Y MEYER, CASO NORMAL SIN OBSERVACIONES ATÍPI-
# CAS, USANDO SUMAS DE RIEMANN.
# DEFINIENDO No.  DE INTERVALOS, INCREMENTO, Y LUGARES DONDE
# GUARDAR LOS VALORES DE LAS FUNCIONES Y ÁREAS DE INTERÉS
nn <- 71; delta <- 1; part <- rep(0,nn+1);
y1 <- rep(0,nn+1); areas1 <- rep(0,nn)
y2 <- rep(0,nn+1); areas2 <- rep(0,nn)
y0 <- rep(0,nn+1); areas0 <- rep(0,nn)
# INFORMACIÓN A PRIORI
a <- 0.20; k <- 10
# DATOS DEL EJEMPLO DE BOX Y MEYER (1987)
y <- c(47.46,49.62,43.13,46.31,51.47,48.49,49.34,46.10,
       46.76,48.56,44.83,44.45,59.15,51.33,47.02,47.90)
# CONTRASTES Y MATRIZ DE SIGNOS
X0 <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
A <- c(-1, 1,-1, 1,-1, 1,-1, 1,-1, 1,-1, 1,-1, 1,-1,1)
B <- c(-1,-1, 1, 1,-1,-1, 1, 1,-1,-1, 1, 1,-1,-1, 1,1)
C <- c(-1,-1,-1,-1, 1, 1, 1, 1,-1,-1,-1,-1, 1, 1, 1,1)
D <- c(-1,-1,-1,-1,-1,-1,-1,-1, 1, 1, 1, 1, 1, 1, 1,1)
E <- A*B; G <- A*C; H <- A*D; I <- B*C; J <- B*D; K <- C*D
L <- A*B*C; M <- A*B*D; N <- A*C*D; O <- B*C*D; P<-A*B*C*D

Cr <- cbind(A,B,C,D,E,G,H,I,J,K,L,M,N,O,P)
# CALCULANDO LOS EFECTOS EN UNA ESCALA ADECUADA
Efec <- rep(0,nn1)
for (i in 1:nn1)
    {
        Efec[i] <- t(Cr[,i])%*%y/8
    }
C <- Efec/(mean(abs(Efec))/0.1)
# CALCULANDO EL NUMERADOR Y DENOMINADOR DE Pr[i activo dado
# Ti,sigma],DEFINIENDO PREVIAMENTE LA TRANSFORMACIÓN t=1/σ
pn <- function(t,C) {a*(1/k)*exp(-(t^2)*(C^2)/(2*k^2))}
pd <- function(t,C) {a*(1/k)*exp(-(t^2)*(C^2)/
              (2*k^2))+(1-a)*exp(-(t^2)*(C^2)/2)}
```

```
# CALCULANDO p[sigma,T]
pds <- function(t) {c(pd(t,C[1]),  pd(t,C[2]), pd(t,C[3]),
pd(t,C[4]), pd(t,C[5]), pd(t,C[6]), pd(t,C[7]), pd(t,C[8]),
pd(t,C[9]),pd(t,C[10]),pd(t,C[11]),pd(t,C[12]),pd(t,C[13]),
pd(t,C[14]), pd(t,C[15]))}
produc <- function(t){(t^(14))*prod(pds(t))}
# SE DEFINE LA PARTICIÓN
for (i in 2:(nn+1))
    {part[i]<-part[i-1] + delta}
# PROBABILIDADES POSTERIORES DE QUE LOS EFECTOS ESTÁN ACTI-
# VOS
posterior <- rep(0,nn1+1)
for (j in 1:nn1)
    {
    f1 <- function(t){(pn(t,C[j])/pd(t,C[j]))*produc(t)}
        for (i in 1:nn+1)
        { y1[i]<- f1(part[i]); y2[i] <- produc(part[i])}
        for (i in 2:(nn+1))
            {areas1[i-1] <- delta*((y1[i]+y1[i-1])/2)
            areas2[i-1]<- delta*((y2[i]+y2[i-1])/2)}
    integral1 <- sum(areas1)
    integral2 <- sum(areas2) # CONSTANTE DE NORMALIZACION
    posterior[j+1] <- integral1/integral2
    }
# PROBABILIDAD POSTERIOR DE ''TODOS LOS EFECTOS INACTIVOS''
f0 <- function(t,C){(a*(1/k)*exp(-(t^2)*(C^2)/(2*k^2)))/
((a/k)*exp(-(t^2)*(C^2)/(2*k^2))+(1-a)*exp(-(t^2)*(C^2)/2))}
f00 <- function(t) {c(1-f0(t,C[1]),1-f0(t,C[2]),1-f0(t,C[3]),
1-f0(t,C[4]), 1-f0(t,C[5]), 1-f0(t,C[6]), 1-f0(t,C[7]),
1-f0(t,C[8]), 1-f0(t,C[9]), 1-f0(t,C[10]), 1-f0(t,C[11]),
1-f0(t,C[12]), 1-f0(t,C[13]), 1-f0(t,C[14]), 1-f0(t,C[15]))}
produc0 <- function(t){prod(f00(t))}
f000 <- function(t){produc0(t)*produc(t)}
    for (i in 1:nn+1)
        {y0[i]<- f000(part[i])}
    for (i in 2:(nn+1))
    {areas0[i-1]<-(delta)*((y0[i]+y0[i-1])/2)}
```

```
    integral0 <- sum(areas0)
    posterior[1] <- integral0/integral2
# GRAFICACIÓN DE PROBABILIDADES POSTERIORES
names <- c("0", "A", "B", "C", "D", "AB", "AC", "AD", "BC",
"BD", "CD", "ABC", "ABD", "ACD", "BCD", "ABCD")
barplot(posterior, names=names,xlab="Efectos", ylab="Prob.
Posterior de Estar Activo")
# NOTA: PARA AJUSTAR LA PARTICIÓN PONGA AL INICIO DEL PROGRA-
# MA EL SIGUIENTE VALOR DE nn Y CÓRRALO OTRA VEZ
i<-1; while(produc(i) > 0.00001) {nn<-i; i<- i+1}
### FIN DE PROGRAMA 2 ###
```

La salida del Programa 2 es también la Figura 1 de la Sección 5.

4 Programa 3: Probabilidades Posteriores de Efectos Activos y de Datos Anómalos

Box y Meyer (1987) extienden el método bayesiano para incluir la posibilidad de observaciones anómalas que pueden influir en las probabilidades posteriores de que los efectos sean activos. En este caso los conjuntos de interés $a_{(r_1,r_2)}$ son todas las posibles combinaciones de efectos activos y de observaciones anómalas, que en el caso por ejemplo de un factorial 2^4, son tantos como $2^{15} \times 2^{16}$, que es una cantidad enorme como para hacer los cálculos de manera exhaustiva. Los mismos autores proponen un procedimiento iterativo para aproximar las probabilidades posteriores tanto de que los efectos sean activos como de que los datos sean anómalos.

Más específicamente, ellos encuentran (Meyer y Box, 1992) que la probabilidad posterior de que un conjunto particular de r_1 efectos está activo y de que al mismo tiempo las observaciones r_2 sean anómalas (evento $a_{(r_1,r_2)}$)

está dada por

$$
\begin{aligned}
p\left(a_{(r_1,r_2)}\,|\mathbf{y}\right) \;\propto\; & \left(\frac{\alpha_1}{1-\alpha_1}\right)^{r_1}\left(\frac{\alpha_2}{1-\alpha_2}\right)^{r_2}\gamma^{-r_1}k^{-r_2} \\
& \times\frac{\left|\mathbf{X}'_{(0)}\mathbf{X}_{(0)}\right|^{1/2}}{\left|\Gamma_{r_1}+\mathbf{X}'_{(r_1)}\mathbf{X}_{(r_1)}-\varphi\mathbf{X}'_{(r_1,r_2)}\mathbf{X}_{(r_1,r_2)}\right|^{1/2}} \\
& \times\left(\frac{\mathbf{S}\left(\widehat{\tau}_{(r_1,r_2)}\right)+\widehat{\boldsymbol{\tau}}'_{(r_1,r_2)}\Gamma_{(r_1)}\widehat{\boldsymbol{\tau}}_{(r_1,r_2)}}{\mathbf{S}\left(\widehat{\tau}_{(0)}\right)}\right),
\end{aligned}
\tag{7}
$$

donde $\mathbf{y}$ es el vector de observaciones y $\mathbf{X}_{(r_1,r_2)}$ es la matriz de columnas y renglones de $\mathbf{X}$ que corresponden a los efectos activos y observaciones anómalas, y además

$$
\begin{aligned}
\varphi &= 1-1/k^2 \\
\widehat{\boldsymbol{\tau}}_{(r_1,r_2)} &= \left(\Gamma_{r_1}+\mathbf{X}'_{(r_1)}\mathbf{X}_{(r_1)}-\varphi\mathbf{X}'_{(r_1,r_2)}\mathbf{X}_{(r_1,r_2)}\right)^{-1}\left(\mathbf{X}'_{(r_1)}\mathbf{y}-\varphi\mathbf{X}'_{(r_1,r_2)}\mathbf{y}_{(r_2)}\right) \\
\mathbf{S}\left(\widehat{\boldsymbol{\tau}}_{(r_1,r_2)}\right) &= \left(\mathbf{y}-\mathbf{X}'_{(r_1)}\widehat{\boldsymbol{\tau}}_{(r_1,r_2)}\right)'\left(\mathbf{y}-\mathbf{X}'_{(r_1)}\widehat{\boldsymbol{\tau}}_{(r_1,r_2)}\right)- \\
&\quad \varphi\left(\mathbf{y}_{(r_2)}-\mathbf{X}'_{(r_1,r_2)}\widehat{\boldsymbol{\tau}}_{(r_1,r_2)}\right)'\left(\mathbf{y}_{(r_2)}-\mathbf{X}'_{(r_1,r_2)}\widehat{\boldsymbol{\tau}}_{(r_1,r_2)}\right)
\end{aligned}
$$

donde $\mathbf{y}_{(r_2)}$ son las supuestas observaciones anómalas.

El procedimiento iterativo que proponen los autores comienza por suponer que no hay observaciones anómalas, esto es, en la fórmula (7) se toma $r_2=0$, $\mathbf{X}_{(r_1,r_2)}=\mathbf{0}_{(r_1)}$ y $\mathbf{y}_{(r_2)}=0$, de manera que ésta se reduce a los casos de los primeros dos programas donde no se suponen observaciones anómalas. En un segundo paso y con los efectos que hayan resultado más importantes, digamos con probabilidades posteriores de 0.5 o más, se supone temporalmente el modelo correspondiente y se calculan las probabilidades posteriores de que los datos sean anómalos; así, en este segundo paso, r_1 y las columnas de las matrices $\mathbf{X}_{(r_1)}$ y $\mathbf{X}_{(r_1,r_2)}$ se mantienen fijos. Luego, en un tercer paso y suponiendo como datos anómalos los que tengan probabilidades posteriores altas (>0.5) en el paso anterior, se vuelven a calcular las probabilidades posteriores de que los efectos están activos. En este tercer paso r_2 vuelve a ser fijo al igual que los renglones de la matriz $\mathbf{X}_{(r_1,r_2)}$ y de $\mathbf{y}_{(r_2)}$. Y así sucesivamente. Este procedimiento generalmente converge en una o dos iteraciones, es decir, en la segunda o tercera vez que se calculan las probabilidades posteriores de efecto activo.

A continuación se presenta el programa paso a paso a lo largo de una iteración, usando el mismo ejemplo que se trabaja en los dos programas anteriores. Se debe correr el programa paso por paso, puesto que la salida de uno se convierte en información de entrada del siguiente. Incluso se recomienda correr de manera separada cada uno de los bloques donde se obtienen y ordenan los conjuntos potencia de los 15 efectos y 16 observaciones.

Listado de Programa 3

```
# PASO 1: SUPONIENDO QUE NO HAY DATOS ANÓMALOS SE CALCULAN
# LAS PROBABILIDADES POSTERIORES DE QUE LOS EFECTOS SEAN AC-
# TIVOS.
# VECTOR Y VALOR DONDE SE GUARDA LA PROBABILIDAD POSTERIOR
# DE CADA EFECTO Y SU ACUMULADO
nn1<-15; ppEF1 <- rep(0,nn1+1); ppT1 <- 0
# INFORMACIÓN A PRIORI TOMADA DE BOX Y MEYER (1987)
alf1 <- 0.2; alf2 <- 0.05; g <- 2.5; h <- 5; fi <- 1-1/h^2
re <- 0 # SE SUPONE QUE NO HAY DATOS ANORMALES
# SE OBTIENE EL CONJUNTO POTENCIA DE LOS nn1=15 EFECTOS
powerSet <- function(x)
{
    K <- NULL
    for(m in x)
        K <- rbind(cbind(K, F), cbind(K, T))
    apply(K, 1, function(x, s) s[x], s = x)
}
xx1 <- powerSet(1:nn1)
xx1 <- xx1[-1]
# SE ORDENA EL CONJUNTO POTENCIA DE LOS nn1=15 EFECTOS
numdat1 <- length(xx1); ps <- rep(0,numdat1)
for (i in 1:numdat1) {ps[i] <- length(xx1[[i]])}
xx1 <- xx1[sort.list(ps)]
# DATOS DE EJEMPLO
y <- c(47.46,49.62,43.13,46.31,51.47,48.49,49.34,46.10,
 46.76,48.56,44.83,44.45,59.15,51.33,47.02,47.90)
# CONTRASTES Y MATRIZ DE SIGNOS
```

```
X00 <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
A <- c(-1, 1,-1, 1,-1, 1,-1, 1,-1, 1,-1, 1,-1, 1,-1,1)
B <- c(-1,-1, 1, 1,-1,-1, 1, 1,-1,-1, 1, 1,-1,-1, 1,1)
C <- c(-1,-1,-1,-1, 1, 1, 1, 1,-1,-1,-1,-1, 1, 1, 1,1)
D <- c(-1,-1,-1,-1,-1,-1,-1,-1, 1, 1, 1, 1, 1, 1, 1,1)
E <- A*B; G <- A*C; H <- A*D; I <- B*C; J <- B*D; K <- C*D
L <- A*B*C; M <- A*B*D; N <- A*C*D; O <- B*C*D; P<-A*B*C*D

Cr <- cbind(A,B,C,D,E,G,H,I,J,K,L,M,N,O,P)
# PROBABILIDADES POSTERIORES PARA MODELO SIN TÉRMINOS
ppEF1[1] <- 1; ppT1 <- 1
# PROBABILIDADES POSTERIORES PARA MODELOS CON HASTA 7 EFEC-
# TOS ACTIVOS (16383 MODELOS)
for(i in 1:16383)
    {
    ps <- xx1[[i]]; te <- length(ps)
    Xr1 <- cbind(X00,Cr[,ps]); Xr1r2 <- rep(0,te+1);
    yr2 <- 0; gg <- matrix(0,te+1,te+1)
    for (ii in 1:te) {gg[ii+1,ii+1] <- 1}; gg1 <- (1/g^2)*gg

    tr1r2 <- solve(gg1+t(Xr1)%*%Xr1-fi*t(t(Xr1r2))%*%Xr1r2)
                %*%(t(Xr1)%*%y-fi*t(t(Xr1r2))%*%yr2)
    Sr1r2 <- t(y-Xr1%*%tr1r2)%*%(y-Xr1%*%tr1r2)
            -fi*t(yr2-Xr1r2%*%tr1r2)%*%(yr2-Xr1r2%*%tr1r2)
    pp1 <-((alf1*g^-1)/(1-alf1))^te*((alf2*h^-1)/(1-alf2))^re
    pp2 <- 4/det(gg1+t(Xr1)%*%Xr1-fi*t(t(Xr1r2))%*%Xr1r2)^0.5
    pp3 <- ((Sr1r2+t(tr1r2)%*%gg1%*%tr1r2)
            /t(y-mean(y))%*%(y-mean(y)))^-7.5
    ppEF1[ps+1] <- ppEF1[ps+1] + pp1*pp2*pp3
    ppT1 <- ppT1+ pp1*pp2*pp3
    }
# NORMALIZACIÓN Y GRAFICACIÓN DE LAS PROBABILIDADES POSTE--
# RIORES DE EFECTO ACTIVO
posterior <- ppEF1/ppT1
names <- c("0", "A", "B", "C", "D", "AB", "AC", "AD", "BC",
                "BD", "CD", "ABC", "ABD", "ACD", "BCD", "ABCD")
```

```
barplot(posterior,names=names,xlab="Efectos, Sin Suponer
Datos Anómalos", ylab="Prob.  Posterior de Estar Activo")
### TERMINA PASO 1 ###
```

De la Figura 1 de la Sección 5, que es la salida de este primer paso, se toman temporalmente como importantes los efectos con probabilidades posteriores altas, digamos mayores a 0.5, sin que esto implique que esta regla se deba aplicar de manera estricta. En este ejemplo solo el efecto B tiene probabilidad posterior arriba 0.5, pero el efecto C está arriba de 0.4, y junto con B se observa relativamente importante en relación al resto de los efectos. De aquí que para el segundo paso se decida trabajar con el modelo que tiene los términos B y C.

```
# PASO 2: CALCULANDO LAS POSTERIORES DE QUE LOS DATOS SEAN
# ANÓMALOS, SUPONIENDO POR EL MOMENTO EL MODELO IDENTIFICADO
# EN EL PASO ANTERIOR.

# CONSTRUCCIÓN DE MATRICES Xr1 y gg PARA EL MODELO SELECCIONA-
# DO EN PASO 1
Xr1 <- cbind(X00,Cr[,2],Cr[,3])
gg <- matrix(0, 3, 3); gg[2,2] <- 1; gg[3,3] <- 1;
te <- 2 # DOS TÉRMINOS ACTIVOS SE DETECTARON EN EL PASO 1

# CALCULANDO EL CONJUNTO POTENCIA DE LOS nn2=16 DATOS
powerSet <- function(x)
{
    K <- NULL
    for(m in x)
        K <- rbind(cbind(K, F), cbind(K, T))
    apply(K, 1, function(x, s) s[x], s = x)
}
nn2 <- 16
xx2 <- powerSet(1:nn2)
xx2 <- xx2[-1]

# SE ORDENA EL CONJUNTO POTENCIA DE LOS nn2=16 DATOS
numdat2 <- length(xx1); ps <- rep(0,numdat2)
for (i in 1:numdat2) {ps[i] <- length(xx2[[i]])}
xx2 <- xx2[sort.list(ps)]

# VECTOR Y VALOR DONDE SE GUARDAN LAS PROB. POSTERIORES DE
# DATO ANÓMALO
```

```
pYY <- rep(0,nn2); ppyy <- 0
# INICIA CÁLCULO DE PROBABILIDADES POSTERIORES USANDO 14892
# MODELOS, QUE SON LOS CONJUNTOS CON A LO MÁS 6 DATOS ANÓMALOS
for(i in 1:14892)
    {
    ps <- xx2[[i]]
    re <- length(ps)
        if (re == 1)
            {
                Xr1r2 <- Xr1[ps,]
                yr2 <- as.matrix(y[ps])

        tr1r2 <- solve(gg+t(Xr1)%*%Xr1-fi*t(t(Xr1r2))%*%Xr1r2)
                   %*%(t(Xr1)%*%y-fi*t(t(Xr1r2))%*%yr2)
        Sr1r2 <- t(y-Xr1%*%tr1r2)%*%(y-Xr1%*%tr1r2)
               -fi*t(yr2-Xr1r2%*%tr1r2)%*%(yr2-Xr1r2%*%tr1r2)

pp1 <- ((alf1*g^-1)/(1-alf1))^te*((alf2*h^-1)/(1-alf2))^re
pp2 <- 4.0/sqrt(det(gg+t(Xr1)%*%Xr1-fi*t(t(Xr1r2))%*%Xr1r2))
pp3 <- ((Sr1r2+t(tr1r2)%*%gg%*%tr1r2)/
       t(y-mean(y))%*%(y-mean(y)))^(-7.5)
            }
            else
            {
            Xr1r2 <- Xr1[ps,]
            yr2 <- as.matrix(y[ps])

        tr1r2 <- solve(gg+t(Xr1)%*%Xr1-fi*t(Xr1r2)%*%Xr1r2)
                   %*%(t(Xr1)%*%y-fi*t(Xr1r2)%*%yr2)
        Sr1r2 <- t(y-Xr1%*%tr1r2)%*%(y-Xr1%*%tr1r2)
           -fi*t(yr2-Xr1r2%*%tr1r2)%*%(yr2-Xr1r2%*%tr1r2)

pp1 <- ((alf1*g^-1)/(1-alf1))^te*((alf2*h^-1)/(1-alf2))^re
pp2 <- 4.0/sqrt(det(gg+t(Xr1)%*%Xr1-fi*t(Xr1r2)%*%Xr1r2))
pp3 <- ((Sr1r2+t(tr1r2)%*%gg%*%tr1r2)/
       t(y-mean(y))%*%(y-mean(y)))^(-7.5)
            }
    pYY[ps] <- pYY[ps] + pp1*pp2*pp3
    ppyy <- ppyy + pp1*pp2*pp3
    }
```

```
# NORMALIZACIÓN Y GRAFICACIÓN DE PROBABILIDADES POSTERIORES
posterior <- pYY/ppyy
names <- paste("Ren",1:nn2,sep=".")
barplot(posterior,names=names,srt=90,xlab="Observaciones,
Suponiendo Modelo B y C", ylab="Prob.  Posterior de Ser
Anómalo")
### TERMINA PASO 2 ###
```

En este segundo paso se encuentra que la observación 13 tiene una alta probabilidad de ser anómala (ver Figura 2). Luego, lo que procede es suponer que es anómala, y bajo este supuesto, volver a calcular las probabilidades posteriores de que los efectos están activos. Esto se hace en el tercer paso.

```
# PASO 3: SE VUELVEN A CALCULAR LAS PROBABILIDADES POSTE-
# RIORES DE QUE LOS EFECTOS ESTÁN ACTIVOS SUPONIENDO LAS
# OBSERVACIONES ANÓMALAS DETECTADAS EN EL PASO ANTERIOR.

# SE DEFINEN LAS OBSERVACIONES ANÓMALAS ENCONTRADAS EN EL
# PASO 2, QUE EN ESTE EJEMPLO ES SOLO LA OBSERVACIÓN 13
YA1 <- 13; re <- 1

# SE DEFINE EL VECTOR Y EL ACUMULADO QUE GUARDA LAS PROBABI-
# LIDADES POSTERIORES SIN NORMALIZAR
ppEF3 <- rep(0,nn1+1)
ppT3 <- 0

# PROBABILIDADES POSTERIORES PARA MODELO SIN TÉRMINOS
# (MODELO CONSTANTE)
    te <- 0
    Xr1 <- cbind(X00); Xr1r2 <- Xr1[YA1,] ; yr2 <- y[YA1]
    gg <- matrix(0,1,1); gg1 <- (1/g^2)*gg

    tr1r2 <- solve(gg1+t(Xr1)%*%Xr1-fi*t(t(Xr1r2))
           %*%Xr1r2)%*%(t(Xr1)%*%y-fi*t(t(Xr1r2))%*%yr2)
    Sr1r2 <- t(y-Xr1%*%tr1r2)%*%(y-Xr1%*%tr1r2)
           -fi*t(yr2-Xr1r2%*%tr1r2)%*%(yr2-Xr1r2%*%tr1r2)

    pp1 <- ((alf1*g^-1)/(1-alf1))^te*((alf2*h^-1)/(1-alf2))^re
    pp2 <- 4/det(gg1+t(Xr1)%*%Xr1-fi*t(t(Xr1r2))%*%Xr1r2)^0.5
    pp3 <- ((Sr1r2+t(tr1r2)%*%gg1%*%tr1r2)/
           t(y-mean(y))%*%(y-mean(y)))^-7.5

    ppEF3[1] <- pp1*pp2*pp3
```

```
    ppT3 <- pp1*pp2*pp3
# PROBABILIDADES POSTERIORES PARA MODELOS HASTA CON 7 EFECTOS
# ACTIVOS(16383 MODELOS)
for(i in 1:16383)
    {
    ps <- xx1[[i]]
    te <- length(ps)

    Xr1 <- cbind(X00,Cr[,ps]); Xr1r2 <- Xr1[YA1,] ;
    yr2 <- y[YA1]; gg <- matrix(0,te+1,te+1)
    for (ii in 1:te) {gg[ii+1,ii+1] <- 1}
        gg1 <- (1/g^2)*gg

    tr1r2 <- solve(gg1+t(Xr1)%*%Xr1-fi*t(t(Xr1r2))%*%Xr1r2)
              %*%(t(Xr1)%*%y-fi*t(t(Xr1r2))%*%yr2)
    Sr1r2 <- t(y-Xr1%*%tr1r2)%*%(y-Xr1%*%tr1r2)
           -fi*t(yr2-Xr1r2%*%tr1r2)%*%(yr2-Xr1r2%*%tr1r2)

    pp1 <- ((alf1*g^-1)/(1-alf1))^te*((alf2*h^-1)/(1-alf2))^re
    pp2 <- 4/det(gg1+t(Xr1)%*%Xr1-fi*t(t(Xr1r2))%*%Xr1r2)^0.5
    pp3 <- ((Sr1r2+t(tr1r2)%*%gg1%*%tr1r2)/
           t(y-mean(y))%*%(y-mean(y)))^-7.5

    ppEF3[ps+1] <- ppEF3[ps+1] + pp1*pp2*pp3
    ppT3 <- ppT3 + pp1*pp2*pp3
    }
# NORMALIZACIÓN Y GRAFICACIÓN DE LAS PROBABILIDADES POSTERIORES
posterior3 <- ppEF3/ppT3
names <- c("0", "A", "B", "C", "D", "AB", "AC", "AD", "BC",
                "BD", "CD", "ABC", "ABD", "ACD", "BCD", "ABCD")
barplot(posterior3,names=names,xlab="Efectos, Suponiendo
Observación 13 Anómala", ylab="Prob.  Posterior de Estar
Activo")
### TERMINA PASO 3 ###
```

La salida del paso 3 es la Figura 3 de la siguiente sección, en la cual se observan claramente importantes los efectos B y C, que ahora tienen

posteriores cercanas a uno. Y los efectos AC y ACD que no habían destacato ahora tienen probabilidades posteriores por arriba de 0.5.

Lo que procede a continuación es suponer el modelo con los términos B, C, AC y ACD y volver a calcular las probabilidades posteriores de los datos anómalos (Paso 2), declarando las matrices $X_{(r_1)}$ y $gg = \Gamma_{r_1}$ al comienzo de este paso como

```
Xr1 <- cbind(X00,Cr[,2],Cr[,3],Cr[,6],Cr[,13])
gg <- matrix(0, 5, 5);
gg[2,2] <- 1; gg[3,3] <- 1; gg[4,4] <- 1; gg[5,5] <- 1
te <- 4 # CUATRO TÉRMINOS ACTIVOS SE DETECTARON EN EL PASO 1
```

La salida que resultaría es otra vez la Figura 2, con lo cual el procedimiento ha convergido, y se concluye que la observación 13 es anómala y los efectos B, C, AC y ACD están activos.

5 Salidas de los Programas

Se presentan las salidas de los tres programas. En primera instancia se muestra la Figura 1 que es la salida de los primeros dos programas y del primer paso del Programa 3. Esta figura constituye la misma salida para tres maneras diferentes de calcular las probabilidades posteriores de que los efectos sean activos, suponiendo la ausencia de observaciones anómalas.

En la Figura 1 se observan relativamente importantes los efectos B y C, pero sus probabilidades posteriores de estar activos son de alrededor de 0.5. En tercera instancia aparece el modelo constante con probabilidad posterior un poco arriba de 0.2.

Tomando por el momento como activos los efectos B y C, se calcula la probabilidad posterior de que cada dato sea anómalo con el Programa 3 Paso 2. En la Figura 2 se destaca claramente la alta probabilidad de ser anómala que tiene la la observación 13.
Considerando la observación 13 como anómala se recalculan las probabilidades posteriores de efecto activo, y el resultado se muestra en la Figura 3. Ahora son cercanas a 0.9 las probabilidades posteriores de que B y C sean activos, y aparecen también con probabilidades superiores a 0.5 los efectos AC y ACD.
Como se menciona al final de la sección anterior, se supone ahora que los términos B, C, AC y ACD están activos y se calcula con el Paso 2 la proba-

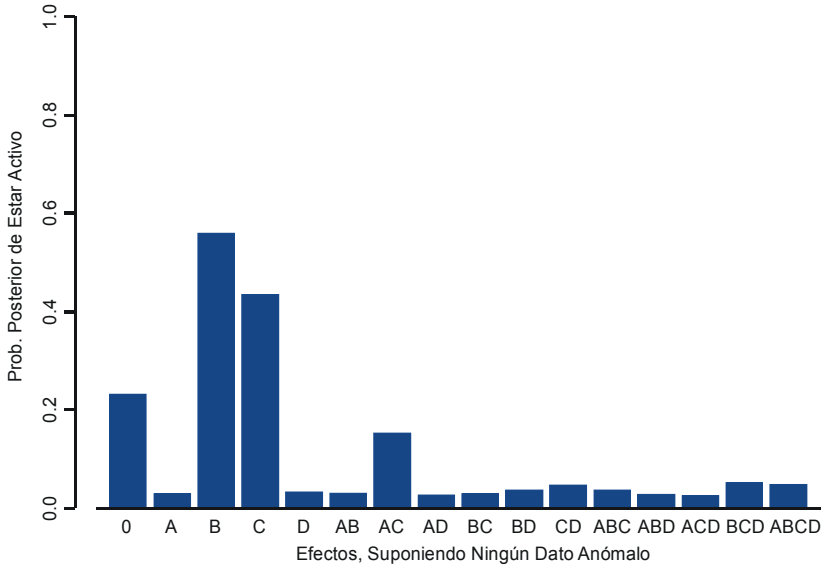

Figura 1: Probabilidad posteriores de efecto activo, suponiendo que no hay datos anómalos.

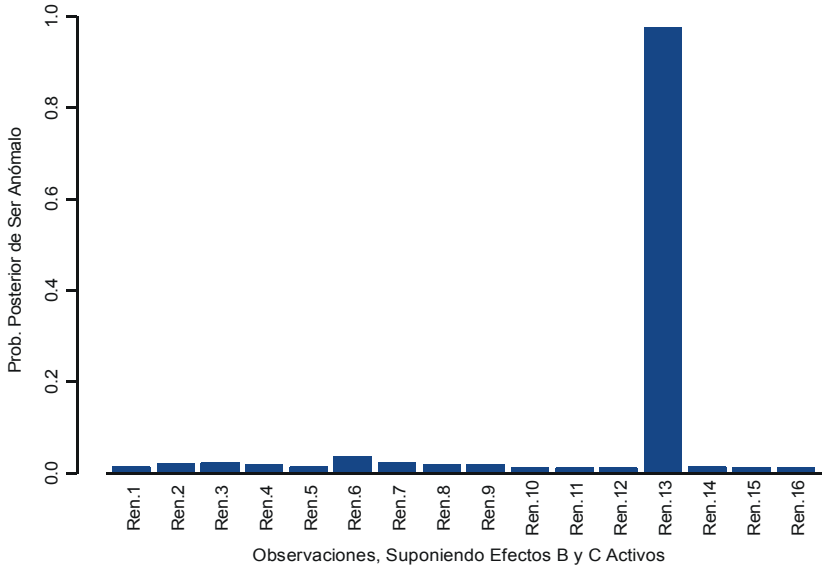

Figura 2: Probabilidad posterior de dato anómalo, suponiendo efectos B y C activos.

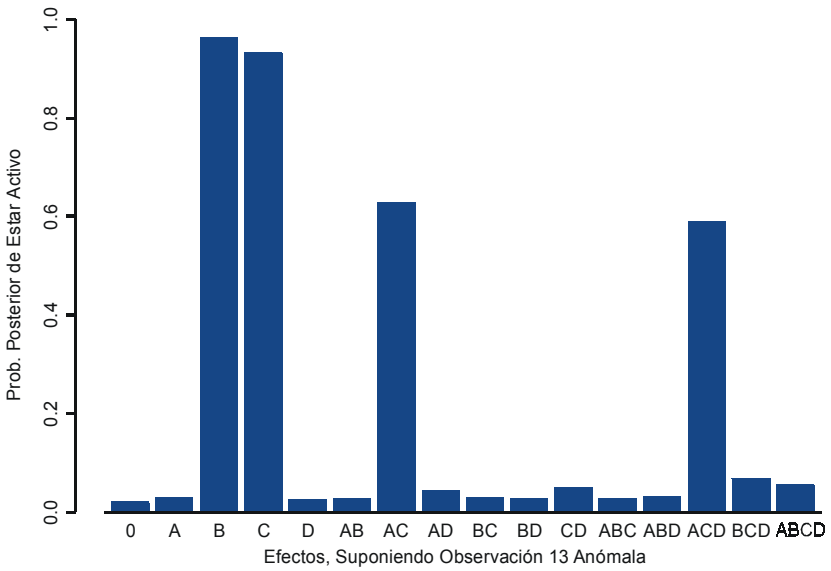

Figura 3: Probabilidades posteriores de efecto activo, suponiendo dato 13 anómalo.

bilidad posterior de dato anómalo. El resultado es otra vez la Figura 2, con lo cual se da por terminado en análisis y se concluye que la observación 13 es anómala y que los efectos B, C, AC y ACD están activos.

6 Conclusiones

En este trabajo se presentan tres programas que calculan probabilidades posteriores de que los efectos están activos, y en el tercero se obtienen también probabilidades posteriores de que los datos sean anómalos.

Si bien estos métodos bayesianos tienen alrededor de 18 años que se propusieron por Box y Meyer (1986, 1987) su aplicación se ha visto limitada por no encontrarse todavía en los paquetes comerciales. Y es que como ocurre con muchas técnicas bayesianas, el aspecto computacional puede ser complicado (o tardado) aún en nuestros días, en particular el del método que contempla la posibilidad de datos anómalos.

Stephenson y Hulting (1989) con su programa en FORTRAN reportan tiempos bastante buenos de 2 y 2.5 minutos para el caso que no contempla

datos anómalos en una computadora IBM PC para un factorial 2^4, considerando las soluciones analítica y numérica, respectivamente. Ellos no programan el caso con posibles datos atípicos. Los Programas 1 y 2 aquí reportados usando también un factorial 2^4 tardan a lo más un minuto en una computadora Pentium 4 con CPU de 2.26 GHz. El caso que contempla datos anómalos (Programa 3) tarda alrededor de 45 minutos en correr una sola pasada de los tres pasos, enfocándose solamente a los conjuntos conformados por a lo más 7 efectos activos y 6 datos anómalos.

Una ventaja de los programas aquí reportados es que están en código de SPLUS y no en código FORTRAN, lo que los hace más fáciles de usar, dada la amplia difusión que tiene este paquete comercial. Si un lector interesado quiere evitar teclear los programas, puede solicitarlos sin costo alguno al autor.(delavara@cimat.mx).

Referencias

Aguirre-Torres, V. and E. Pérez-Trejo (2001). Outliers and the use of the rank transformation to detect active effects in unreplicated 2^f experiments. *Communications in Statistics: Simulation and Computation* 30, pp. 637-663.

Box, G. E. P. and R. Meyer (1987). Analysis of unreplicated factorials allowing for possibly faulty observations. Design, Data and Analysis, C. Mallows (ed.), Wiley, New York.

Box, G. E. P. and R. Meyer (1986). An analysis of unreplicated fractional factorials. Technometrics 28, pp. 11-18.

Box, G. and G. C. Tiao (1968). A bayesian approach to some outlier problems. *Biometrika* 55, pp. 119-129.

Daniel, C. (1959). Use of half normal plots in interpreting factorial two-level experiments. *Technometrics*, 1, 311-341.

Hamada M. and N. Balakrishnan (1998). Analyzing unreplicated factorial experiments: a review with some new proposals. *Statistica Sinica* 8, pp. 1-41.

Meyer, R. (1987). Further details of an analysis for unreplicated fractional factorials. *CQPI Report* No. 80. University of Wisconsin, Madison.

Meyer, R y G. E. P. Box (1992). Finding the active factors in fractionated screenig experiments. *CQPI Report* No. 80, University of Wisconsin, Madison.

Stephenson, R. W. and F. L. Hulting (1989). Posterior probabilities for identifying active effects in unreplicated experiments. *Journal of Quality Technology* 21, pp. 202-212.