UNA ESTRATEGIA PARA LA SELECCION DE MODELOS DE REGRESION CON DOS REGRESORES CUALITATIVOS

Ernestina Castell Gil, Facultad de Matemática y Computacion, Universidad de La Habana

RESUMEN

En este artículo se propone una estrategia para la selección de modelos de regresión con respuesta cuantitativa y dos regresores cualitativos. La propuesta se basa en la minimización de un estimador del Cuadrado Medio del Error de Predicción ($\widehat{C\ M\ E\ P}$) dentro de una clase predefinida de modelos. Se consideran algunas estructuras para las interacciones entre las que se encuentran las propuestas por Mandel (1971) y Johnson y otros (1972), etc. Se proponen algunos gráficos para diagnosticar la forma de la interacción.

Palabras clave: variables cualitativas, clasificación doble, selección de modelos.

ABSTRACT

In this methodological paper, a strategy for selection of regression models is proposed. The application is posible in situations with quantitative regressand and two qualitative regressors. The proposal is based on a minimization of an estimator of the Mean Square Error of Prediction ($\widehat{C\ M\ E\ P}$) in certain predefined class of models. Some structures for interaction (like in Mandel (1971) Johnson and others (1972)) are considered. Some graphical displays for the form of interaction are also proposed.

Key words: qualitative variable, two way clasification, selection of model.

1. INTRODUCCION

La forma tradicional de abordar el problema de regresión con dos regresores cualitativos es el de la Clasificación Doble, donde desde el comienzo mismo se asume un modelo como adecuado y a posteriori se prueban determinadas hipótesis para los parámetros con el objetivo de comparar los efectos para los diferentes niveles. En esa forma de abordar el problema no está presente la idea de la selección del modelo.

Sean X_1 y X_2 dos variables cualitativas que pueden tomar $\bar{n}_1$ y $\bar{n}_2$ valores categóricos diferentes respectivamente. Estas categorías serán denotadas por $i_1, i_2,..., i_{\bar{n}1}$ e $i'_1, i'_2,..., i'_{\bar{n}_2}$. Supóngase que se han realizado observaciones sobre una variable aleatoria cuantitativa Y en $m = n_1 \times n_2$ puntos del conjunto,

$$\left\{ \left(i_j, i'_{j'} \right) : 1 \le j \le \bar{n}_1, 1 \le j' \le \bar{n}_2 \right\} .$$

Sin pérdida de generalidad asumiremos $n_1 = \bar{n}_1$ y $n_2 = \bar{n}_2$, además en el punto $\left(i_j, i'_{j'} \right)$ se realizan $n_{i_j i'_{j'}}$ observaciones.

Tradicionalmente se asume un modelo como el siguiente:

$$E\left(Y_{i_j i'_{j'} k}\right) = \mu_{i_j i'_{j'}} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \gamma_{i_j i'_{j'}} \qquad \text{con} \qquad (1.1)$$

$$\sum_{j=1}^{n_1} \alpha_{i_j} = \sum_{j'=1}^{n_2} \beta_{i'_{j'}} = \sum_{j=1}^{n_1} \gamma_{i_j i'_{j'}} = \sum_{j'=1}^{n_2} \gamma_{i_j i'_{j'}} = 0$$

$$\text{Var}\left(Y_{i_j i'_{j'} k}\right) = \sigma^2 \qquad 1 \le j \le n_1, \quad 1 \le j' \le n_2, \quad 1 \le k \le n_{i_j i'_{j'}}$$

Las observaciones se asumen no correlacionadas.

Un modelo como el anterior está sobreparametrizado (contiene el número máximo de parámetros, esto es, $n_1 + n_2 + (n_1 \times n_2) + 1$ parámetros), lo que implica una reducción en la precisión de las estimaciones, pues la varianza de los estimadores crece cuando crece el número de parámetros en el modelo.

Se sabe que las estimaciones mínimos cuadráticas en el modelo 1.1. son:

$$\hat{\mu} = \overline{Y}\ldots = \frac{1}{n}\sum_{j=1}^{n_1}\sum_{j'=1}^{n_2}\sum_{k=1}^{n_{i_j i'_{j'}}} Y_{i_j i'_{j'} k} \tag{1.2}$$

$$\hat{\alpha}_{i_j} = \overline{Y}_{i_j ..} - \overline{Y}_{\ldots} \quad \text{con} \quad \overline{Y}_{i_j ..} = \frac{1}{\sum_{j'=1}^{n_2} n_{i_j i'_{j'}}} \sum_{j'=1}^{n_2}\sum_{k=1}^{n_{i_j i'_{j'}}} Y_{i_j i'_{j'} k}$$

$$\hat{\beta}_{i'_{j'}} = \overline{Y}_{\cdot i'_{j'} \cdot} - \overline{Y}_{\ldots} \quad \text{con} \quad \overline{Y}_{\cdot i'_{j'} \cdot} = \frac{1}{\sum_{j=1}^{n_1} n_{i_j i'_{j'}}} \sum_{j=1}^{n_1}\sum_{k=1}^{n_{i_j i'_{j'}}} Y_{i_j i'_{j'} k}$$

$$\hat{\gamma}_{i_j i'_{j'}} = \overline{Y}_{i_j i'_{j'} \cdot} - \overline{Y}_{i_j ..} - \overline{Y}_{\cdot i'_{j'} \cdot} + \overline{Y}\ldots \quad \text{con} \quad \overline{Y}_{i_j i'_{j'} \cdot} = \frac{1}{n_{i_j i'_{j'}}} = \sum_{k=1}^{n_{i_j i'_{j'}}} Y_{i_j i'_{j'} k}$$

Estos estimadores poseen propiedades interesantes pues son estimadores mínimos cuadráticos, (Scheffé, (1982)). La consideración de diferentes ecuaciones de regresión es equivalente a considerar diferentes particiones sobre el espacio de valores de los regresores. En el caso bien sencillo de $n_1 = 3$ y $n_2 = 2$ y las interacciones iguales a cero (modelo aditivo). Las ecuaciones que definen los modelos posibles son:

I. *Modelo con un parámetro.*

$g_1(X_1; X_2) = \mu$

II. *Modelos con dos parámetros.*

$$g_2(X_1; X_2) = \begin{cases} \mu + \alpha & \text{si} \quad X_1 = i_1, i_2; \quad X_2 = i'_1, i'_2. \\ \mu - \alpha & \text{si} \quad X_1 = i_3, \quad X_2 = i'_1, i'_2. \end{cases}$$

$$g_3(X_1; X_2) = \begin{cases} \mu + \alpha & \text{si} \quad X_1 = i_1, i_3; \quad X_2 = i'_1, i'_2. \\ \mu - \alpha & \text{si} \quad X_1 = i_2; \quad X_2 = i'_1, i'_2. \end{cases}$$

$$g_4(X_1; X_2) = \begin{cases} \mu + \alpha & \text{si} \quad X_1 = i_3, i_2; \quad X_2 = i'_1, i'_2. \\ \mu - \alpha & \text{si} \quad X_1 = i_1; \quad X_2 = i'_1, i'_2. \end{cases}$$

$$g_5(X_1;X_2)=\begin{cases}\mu+\beta & \text{si} \quad X_1=i_1,i_2,i_3; \quad X_2=i'_1.\\ \mu-\beta & \text{si} \quad X_1=i_1,i_2,i_3; \quad X_2=i'_2.\end{cases}$$

III. *Modelos con tres parámetros.*

$$g_6(X_1;X_2)=\begin{cases}\mu+\alpha_{i'_1} & \text{si} \quad X_1=i_1; \quad X_2=i'_1,i'_2.\\ \mu+\alpha_{i'_2} & \text{si} \quad X_1=i_2; \quad X_2=i'_1,i'_2.\\ \mu-\alpha_{i'_1}-\alpha_{i'_2} & \text{si} \quad X_1=i_3; \quad X_2=i'_1,i'_2.\end{cases}$$

$$g_7(X_1;X_2)=\begin{cases}\mu+\alpha+\beta & \text{si} \quad X_1=i_1,i_2; \quad X_2=i'_1.\\ \mu+\alpha-\beta & \text{si} \quad X_1=i_1,i_2; \quad X_2=i'_2.\\ \mu+\beta-\alpha & \text{si} \quad X_1=i_3; \quad X_2=i'_1.\\ \mu-\alpha-\beta & \text{si} \quad X_1=i_3; \quad X_2=i'_2.\end{cases}$$

$$g_8(X_1;X_2)=\begin{cases}\mu+\alpha+\beta & \text{si} \quad X_1=i_1,i_3; \quad X_2=i'_1.\\ \mu+\alpha-\beta & \text{si} \quad X_1=i_1,i_3; \quad X_2=i'_2.\\ \mu+\beta-\alpha & \text{si} \quad X_1=i_2; \quad X_2=i'_1.\\ \mu-\alpha-\beta & \text{si} \quad X_1=i_2; \quad X_2=i'_2.\end{cases}$$

$$g_9(X_1;X_2)=\begin{cases}\mu+\alpha+\beta & \text{si} \quad X_1=i_2,i_3; \quad X_2=i'_1.\\ \mu+\alpha-\beta & \text{si} \quad X_1=i_2,i_3; \quad X_2=i'_2.\\ \mu+\beta-\alpha & \text{si} \quad X_1=i_1; \quad X_2=i'_1.\\ \mu-\alpha-\beta & \text{si} \quad X_1=i_1; \quad X_2=i'_2.\end{cases}$$

IV. *Modelo con cuatro parámetros.*

$$g_{10}(X_1;X_2)=\begin{cases}\mu+\alpha_{i'_1}+\beta & \text{si} \quad X_1=i_1; \quad X_2=i'_1.\\ \mu+\alpha_{i'_1}-\beta & \text{si} \quad X_1=i_1; \quad X_2=i'_2.\\ \mu+\alpha_{i'_2}+\beta & \text{si} \quad X_1=i_2; \quad X_2=i'_1.\\ \mu+\alpha_{i'_2}-\beta & \text{si} \quad X_1=i_2; \quad X_2=i'_2.\\ \mu+\beta-\alpha_{i'_1}-\alpha_{i'_2} & \text{si} \quad X_1=i_3; \quad X_2=i'_1.\\ \mu-\beta-\alpha_{i'_1}-\alpha_{i'_2} & \text{si} \quad X_1=i_3; \quad X_2=i'_2.\end{cases}$$

El número de parámetros en el modelo crece al crecer el número de clases en la partición que el modelo induce en el espacio de valores $X = X_1 \times X_2$.

Contar el número de modelos con p parámetros sería equivalente a contar las particiones donde la suma de los números de clases de las particiones en los espacios de valores de X_1 y X_2 sea igual a p + 1.

Proposición 1.1.

Sean X_1 y X_2 dos variables cualitativas con n_1 y n_2 valores diferentes ($n_1 \le n_2$). Entonces existen:

$$\sum_{a=p+1-\text{Min}(n_2,p)}^{\text{Min}(n_1,p)}\left\{\left[\sum_{j=1}^{a}\frac{(-1)^{a-j}}{j!(a-j)!}j^{n_1}\right]\left[\sum_{j=1}^{p+1-a}\frac{(-1)^{p+1-a-j}}{j!(p+1-a-j)!}j^{n_2}\right]\right\} \tag{1.3}$$

modelos aditivos diferentes con p parámetros.

Demostración:

Se sabe que si X es una variable cualitativa con m categorías diferentes, el número de modelos con p parámetros (número de particiones con p clases) es:

$$\sum_{j=1}^{p} \frac{(-1)^{p-j}}{j!(p-j)!} j^m$$

((Kovacs (1980); Bunke y Castell (1998)).

Por tanto, las cantidades,

$$H_1 = \sum_{j=1}^{a} \frac{(-1)^{a-j}}{j!(a-j)!} j^{n_1}$$

$$H_2 = \sum_{j=1}^{p+1-a} \frac{(-1)^{p+1-a-j}}{j!(p+1-a-j)!} j^{n_2}$$

representan el número de particiones con a clases en el espacio de valores de X_1 y con p + 1 - a clases en el espacio de valores de X_2; siendo su suma igual a p + 1.

Si M es un modelo aditivo definido por la ecuación que induce particiones como las consideradas, este tendrá p parámetros.

Para a fijo, cada partición con a clases en el espacio de valores de X_1 se combina con cada una de las H_2 particiones con p + 1 - a clases en el espacio de valores de X_2, es decir, para a fijo el número de modelos con p parámetros será $H = H_1.H_2$. Pero el número de clases de las particiones en el espacio de valores de X_1 puede ser como mínimo igual a p + 1 - Min(n_2,p) y a lo sumo Min(n_1,p). Sumando ahora para todo valor de a posible se obtiene la fórmula dada en 1.3.

La cantidad de modelos aditivos servirá como referencia para valorar cuan grande es la cantidad de modelos posibles.

2. MODELACION DE LAS INTERACCIONES

Cuando se trabaja con dos regresores cualitativos y se quiere realizar un análisis completo es necesario determinar la presencia o no del término de interacción y estimar la varianza. El modelo usual con interacciones está sobreparametrizado y como ya se mencionó esto es un gran inconveniente. La consideración de determinadas formas de interacciones (Mandel (1961) y (1971); Johnson y Graybill (1972); Darroch y Speed (1983); Wilhelm (1985); Huet (1991)) produce una disminución del número de parámetros. Esta consideración es tremendamente importante cuando se tiene una sola observación por combinaciones de los tratamientos, porque entonces no se puede utilizar la teoría clásica de los modelos lineales, ya que la estimación de la varianza del error tiene que obtenerse de la suma de cuadrados de las interacciones.

Los diferentes modelos que se proponen considerar (Mandel (1961) y (1971); Johnson y Graybill (1972); Milliken y Graybill (1970); Wilhelm (1985) y otros) son:

1. *Modelo aditivo*

$$Y_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \varepsilon_{i_j i'_{j'} k}$$

2. *Modelo concurrente*

$$Y_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \lambda \alpha_{i_j} \beta_{i'_{j'}} + \varepsilon_{i_j i'_{j'} k}$$

3. *Modelo de regresión por columna*.

$$Y_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \theta_{i_j} \beta_{i'_{j'}} + \varepsilon_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} (1 + \theta_{i_j}) + \varepsilon_{i_j i'_{j'} k}$$

4. *Modelo de regresión por fila*.

$$Y_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \vartheta_{i'_{j'}} \alpha_{i_j} + \varepsilon_{i_j i'_{j'} k} = \mu + (\vartheta_{i'_{j'}} + 1)\alpha_{i_j} + \beta_{i'_{j'}} + \varepsilon_{i_j i'_{j'} k}$$

5. *Quinto modelo.*

$$Y_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \lambda\theta\alpha_{i_j}\vartheta_{i'_{j'}} + \varepsilon_{i_j i'_{j'} k}$$

6. *Modelo de Mandel.*

$$Y_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \sum_{l=1}^{t} \lambda_l U_{li_j} V_{li'_{j'}} + \varepsilon_{i_j i'_{j'} k} \qquad 0 \le t \le \text{Min}(n_1, n_2)$$

7. *Séptimo modelo.*

$$Y_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \lambda\alpha_{i_j}\beta_{i'_{j'}} + \theta_{i_j}\beta_{i'_{j'}} + \varepsilon_{i_j i'_{j'} k}$$

8. *Octavo modelo.*

$$Y_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \lambda\alpha_{i_j}\beta_{i'_{j'}} + \vartheta_{i'_{j'}}\alpha_{i_j} + \varepsilon_{i_j i'_{j'} k}$$

9. *Noveno modelo.*

$$Y_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \lambda\alpha_{i_j}\beta_{i'_{j'}} + \theta_{i_j}\beta_{i'_{j'}} + \vartheta_{i'_{j'}}\alpha_{i_j} + \varepsilon_{i_j i'_{j'} k}$$

10. *Décimo modelo.*

$$Y_{i_j i'_{j'} k} = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \gamma_{i_j i'_{j'}} + \varepsilon_{i_j i'_{j'} k}$$

Sujetos a las restricciones

$$\sum_{j=1}^{n_1} \alpha_{i_j} = \sum_{j'=1}^{n_2} \beta_{i'_{j'}} = \sum_{j=1}^{n_1} \theta_{i_j} = \sum_{j'=1}^{n_2} \vartheta_{i'_{j'}} = 0$$

$$\sum_{j=1}^{n_1} \theta_{i_j}^2 = \sum_{j'=1}^{n_2} \vartheta_{i'_{j'}}^2 = \sum U_{li_j}^2 = \sum V_{li'_{j'}}^2 = 1$$

$$\sum_{j=1}^{n_1} \gamma_{i_j i'_{j'}} = \sum_{j'=1}^{n_2} \gamma_{i_j i'_{j'}} = 0$$

El número de parámetros en cualquier modelo de los enumerados anteriormente, será igual al número de parámetros del modelo aditivo contenido en él, más el número de parámetros que aportan las interacciones.

3. UNA ESTRATEGIA PARA LA SELECCION DEL MODELO

Se considera que se realizan observaciones $Y_{i_j i'_{j'} k}$ sobre una variable aleatoria Y, que satisfacen la ecuación de regresión:

$$Y_{i_j i'_{j'} k} = f\left(i_j, i'_{j'}\right) + \varepsilon_{i_j i'_{j'} k} \qquad (3.1)$$

Se supone que $\varepsilon_{i_j i'_{j'} k}$ son errores aleatorios (no observables) con esperanza cero y varianza σ^2 y que están incorrelacionados.

La función $f(i_j, i'_{j'})$ pertenece a un conjunto M definido por:

$$M = \{g(X_1, X_2, \beta): \beta \in \beta\} \tag{3.2}$$

con

$$g(x_1, x_2, \beta) = \sum_r 1_{C_r^g}(x_1, x_2)\mu_{i_j i'_{j'}} \tag{3.3}$$

$$\beta^t = \left(\mu, \alpha_{i_1}, \alpha_{i_2}, ..., \alpha_{i_{p_1}} \vdots \beta_{i'_1}, \beta_{i'_2}, ..., \beta_{i'_2} \vdots \gamma_1, \gamma_2, ..., \gamma_{p_3}\right)$$

donde $\forall l\ \gamma_l = \gamma_{i_j i'_{j'}}$ para algún $(i_j; i'_{j'})$ y es cualquiera de los tipos de interacciones descritos y $\mu_{i_j i'_{j'}}$ está dada por una de las expresiones de la (1) a la (10) dadas anteriormente.

$$N = \bigcup_r C_r^g$$

Así que la selección del modelo viene dada por la selección de una función g en M para aproximar f. El conjunto de observaciones será utilizado para calcular un estimador $\hat{\beta}$ del vector de parámetros β y seleccionar una función $g(X_1, X_2, \hat{\beta})$.

En la selección de la función $g(X_1, X_2, \hat{\beta})$ se utilizará el criterio de minimización de un estimador del CMEP o del CME. El vector de parámetros β será estimado por el Método de los Mínimos Cuadrados.

En la primera etapa de la estrategia se analizarán los modelos con un número de parámetros inferior o igual a un número p_0 fijado por el usuario apropiadamente y dentro de los límites permisibles. Posteriormente hay que encontrar una manera de reducir el número de modelos a comparar y la idea sería hacer una reducción de forma tal que entre un paso y otro de la estrategia los modelos analizados no cambien bruscamente en cuanto al número de parámetros. En la consecución de este último objetivo juega un rol fundamental el concepto de modelo vecino que se da más adelante.

Definición 3.1. (modelo vecino)

Sea,

$$\delta \in \{\alpha; \beta; \gamma(X_1, X_2)\}$$

Sea M_0 un modelo. Sea π_0 la partición determinada por la función $g_0(X_1, X_2, \beta)$ que define al modelo M_0. Sean C_{ok} las clases de esta partición. Se dice que M_v^δ es un modelo vecino de M_0 según δ, si se cumple:

1. Existe un y sólo un δ y una y solamente una clase C_{ok} tal que:

   $C_{ok} = (C_{vk'})^\delta \cup (C_{vk''})^\delta$

   donde se ha utilizado el supraíndice δ para indicar que se afecta la clase sólo según ese parámetro.

2. Para todo $r \neq k$ existe r', tal que $k' \neq r' \neq k''$, para el cual se cumple:

   $C_{or} = (C_{vr})^\delta$

3. Además la forma de la interacción del modelo M_v^{δ} es la misma que la del modelo M_0.

Por la forma en que se ha definido el modelo vecino, es intuitivamente claro que, este tiene un parámetro más que M_0.

En un problema con dos variables cualitativas existen tres nominaciones de parámetros, que se han venido representando así:

$\alpha \rightarrow$ niveles de la primera variable

$\beta \rightarrow$ niveles de la segunda variable

$\gamma \rightarrow$ interacciones

Para construir una estrategia de selección, basándose en los modelos vecinos hay que decidir en qué orden se toman las nominaciones para ir aumentando la cantidad de parámetros en el modelo. Como no hay preferencia entre un factor u otro, este orden es irrelevante. Como hay que decidirse por un orden, se propone: $\alpha \rightarrow \beta \rightarrow \gamma$.

3.1. Estrategia para la selección del modelo

Sea $\gamma_l(X_1,X_2)$ la forma de interacción en el modelo $M_l(\cdot)$, l = 1,2,...,10. Sea p(l) el número máximo de parámetros permisibles para el modelo $M_l(\cdot)$.

Sea

$$M_l^0 = \{(M_l(p): p \leq p_0;\ 1 \leq p_0 \leq p(l)\}$$

Sea

$$\hat{r}(M_l(p)) = \hat{r}(l,p(l))$$

un estimador del CMEP para el modelo $M_l(p)$.

1. Calcular el número de modelos aditivos posibles para cada p = 1,2,...,p(l). Siendo l un número fijo.
2. De acuerdo a lo determinado en el punto anterior, a las facilidades de cómputo y al tiempo y esfuerzo que se esté dispuesto a emplear, seleccionar un valor p_0 y así quedará determinado el conjunto M_l^0 que llamaremos conjunto de modelos básicos.

En lo que sigue se utilizará la siguiente notación, por ejemplo $\left(M_l^0\right)_v^{\alpha}$ es un modelo vecino de M_l^0 según las particiones correspondientes al parámetro α. Es decir, se ha aumentado un parámetro con respecto al número que contenía M_l^0 pero el aumento se hace en la denominación α. La clase formada por los vecinos de M_l^0 según el parámetro α, se denotará por $M_l^{v,\alpha}$.

Realizar los siguientes pasos:

1. Determinar

$$M_l^0 = \underset{M \in M_l^0}{\text{ArgMin}}\ \hat{r}(l,p)$$

Sea

$$M_l^0 = M_l^0(p')$$

donde p' representa la cantidad de parámetros en M_l^0.

2. Sea

$$M_l^{v,\alpha} = \left\{ \left(M_l^0\right)_v^{\alpha} \right\}$$

Determinar

$$M_l^{1,\alpha} = \underset{M \in M_l^{v,\alpha}}{\text{ArgMin}}\ \hat{r}(l, p'+1)$$

entonces

$$M_l^{1,\alpha} = M_l^{1,\alpha}(p'+1)$$

3. Sea

$$M_l^{v,\beta} = \left\{ \left(M_l^{1,\alpha}\right)_v^{\beta} \right\}$$

Determinar

$$M_l^{2,\beta} = \underset{M \in M_l^{v,\beta}}{\text{ArgMin}}\ \hat{r}(l, p'+2)$$

4. Sea

$$M_l^{v,\gamma(X_1,X_2)} = \left\{ \left(M_l^{2,\beta}\right)_v^{\gamma(X_1,X_2)} \right\}$$

Determinar

$$M_l^{3,\gamma(X_1,X_2)} = \underset{M \in M_l^{v,\gamma(X_1,X_2)}}{\text{ArgMin}}\ \hat{r}(l, p'+3)$$

5. Sea

$$M_l^{v,\alpha} = \left\{ \left(M_l^{3,\gamma(X_1,X_2)}\right)_v^{\alpha} \right\}$$

Determinar

$$M_l^{4,\alpha} = \underset{M \in M_l^{v,\alpha}}{\text{ArgMin}}\ \hat{r}(l, p'+4)$$

6. Sea

$$M_l^{v,\beta} = \left\{ \left(M_l^{4,\alpha}\right)_v^{\beta} \right\}$$

Determinar

$$M_l^{5,\alpha} = \underset{M \in M_l^{v,\alpha}}{\text{ArgMin}}\ \hat{r}(l, p'+5)$$

7. Sea

$$M_l^{v,\gamma(X_1,X_2)} = \left\{ \left(M_l^{5,\beta}\right)_v^{\gamma(X_1,X_2)} \right\}$$

Determinar

$$M_l^{6,\gamma(x_1,X_2)} = \underset{M \in M_l^{v,\gamma(x_1,X_2)}}{\text{ArgMin}}\ \hat{r}(l, p'+6)$$

8. Repetir los pasos 5, 6 y 7. De esta forma, para $q \in \{z: 0 \leq z\}$, los pasos $2 + 3q$, $3 + 3q$ y $4 + 3q$, quedarían determinados por:

⋮

⋮

$2 + 3q.-$

$$M_l^{v,\alpha} = \left\{ \left(M_l^{3q,\gamma(X_1,X_2)} \right)_v^{\alpha} \right\}$$

Determinar

$$M_l^{1+3q,v} = \underset{M \in M_l^{v,\alpha}}{\text{ArgMin}}\ \hat{r}(l, p' + 1 + 3q)$$

3 + 3q.- Sea

$$M_l^{v,\beta} = \left\{ \left(M_l^{1+3q,\alpha} \right)_v^{\beta} \right\}$$

Determinar

$$M_l^{2+3q,v} = \underset{M \in M_l^{v,\beta}}{\text{ArgMin}}\ \hat{r}(l, p' + 2 + 3q)$$

4 + 3q. - Sea

$$M_l^{v,\gamma(X_1,X_2)} = \left\{ \left(M_l^{3q+2,\beta} \right)_v^{\gamma(X_1,X_2)} \right\}$$

Determinar

$$M_l^{3q+3,v} = \underset{M \in M_l^{v,\gamma(X_1,X_2)}}{\text{ArgMin}}\ \hat{r}(l, p' + 3 + 3q)$$

Cuando en un paso la denominación correspondiente no pueda ser aumentada se sustituye esta por la siguiente en el orden de afectación.

4. ALGUNOS GRAFICOS UTILES PARA EXPLORAR LA MODELACION DE LAS INTERACCIONES

Dos modelos han sido considerados vecinos cuando las clases que ellos definen se mantienen invariantes excepto una de ellas que se divide en dos nuevas clases. Sin embargo, de acuerdo al número de parámetros que aportan las interacciones se podrían formar grupos bastante homogéneos. Un modelo que resulte seleccionado con un tipo de interacción en el grupo k, puede ser examinado cambiando la forma de su interacción por otra del grupo k+1.

$$\begin{array}{ccccccc} & & \vdots\ a)\ \alpha_i d_j & & \vdots\ a)\ \lambda c_i d_j & & \vdots\ a)\ \lambda_1 U_{1i} V_{1i} + \lambda_2 U_{2i} V_{2i} \\ 0 \rightarrow \lambda\alpha_i\beta_j \rightarrow & & \vdots\ b)\ c_i\beta_j & \rightarrow & \vdots\ b)\ \lambda\alpha_i\beta_j + \alpha_i d_j + c_i\beta_j & \rightarrow & \vdots\ b)\ \gamma_{ij} \\ & & \vdots\ c)\ \lambda\alpha_i\beta_j + c_i\beta_j & & \vdots & & \\ & & \vdots\ d)\ \lambda\alpha_i\beta_j + \alpha_i d_j & & \vdots & & \\ 1 \quad 2 & & 3 & & 4 & & 5 \end{array} \qquad (4.1)$$

Sea el i-j-ésimo residuo

$$r_{ij} = \hat{Y}_{ij} - Y_{ij}$$

Se sabe que Zwanzig (1979); Humak (1983)

$$r_{i,i'_{j'}} \xrightarrow{L} N\left(f\left(i_j, i'_{j'}\right)\right) - g\left(i_j, i'_{j'}, \beta; (\cdot)\right) \qquad (4.2)$$

Como la distribución Normal es una distribución simétrica su media coincide con la mediana y se puede aprovechar este hecho para pensar en algunas situaciones gráficas que permitan hacer una mayor exploración. Si el modelo es correcto entonces la mediana de la distribución límite es cero.

Con el objetivo de ganar claridad, se considerará el caso donde ambas variables cualitativas sólo pueden tomar tres valores.

Caso 1. Supóngase que resultó seleccionado un modelo con una forma de interacción como la del tipo descrita en la columna 2 de 4.1, y que el modelo (la interacción) correcta es de la forma en a) de la columna 3:

Med $r_{lm} = \alpha_l d_m - \lambda\alpha_l\beta_m$

$= \alpha_l(d_m - \lambda\beta_m) = \alpha_l k_m$

Es decir, la mediana del l-m-ésimo residuo es función de α_l. Como $\sum_{l=1}^{3} \alpha_l = 0$.

Se tiene que:

Med $r_{lm} = -\alpha_2 k_m - \alpha_3 k_m$

$=$ - Med r_{2m} - Med r_{3m}

Un gráfico de residuos donde se muestren estos contra los $\hat{\alpha}_i$ o contra sus índices, puede aportar información valiosísima que permita orientar las exploraciones subsiguientes.

En el Anexo se dan algunas ilustraciones de gráficos, un gráfico como el número 2, no sugiere esa relación y, por tanto, tampoco cambios en ese sentido. Mientras que los gáficos números 1, 3 y 4, si sugieren cambios, y entonces se debe explorar el modelo haciendo dicho cambio en la modelación de la interacción.

Caso 2. Considérese la situación del caso 1 pero ahora el modelo verdadero es el dado en b) de la columna 3 en 4.1. El resultado es similar pero la dependencia surge respecto de β_j:

Med $r_{ij} = k_i\beta_j$.

Los residuos deben graficarse contra los $\hat{\beta}_j$ o contra sus índices.

Estos gráficos pueden resultar útiles en casi todas las situaciones con la excepción de aquella donde esté involucrado el modelo dado en a) de la columna 5 en 4.1.

Consideraciones finales:

Todos los modelos referidos anteriormente excepto el 1 y el 10; son no lineales, por tanto, parece razonable desestimar la hipótesis de varianza constante. Esto ha sido considerado en un programa confeccionado en Turbo Pascal 7.0.

Se simularon veinte juegos de datos con los modelos No. 2 y No. 4 (diez con cada uno) y se obtuvo que la estrategia seleccionó al modelo optimal en el 70 % de los casos. Los resultados se encuentran en las Tablas 1 y 2 del Anexo.

ANEXO

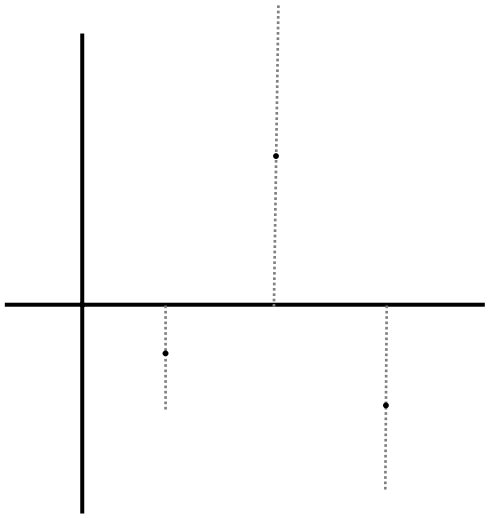

Gráfico No.1

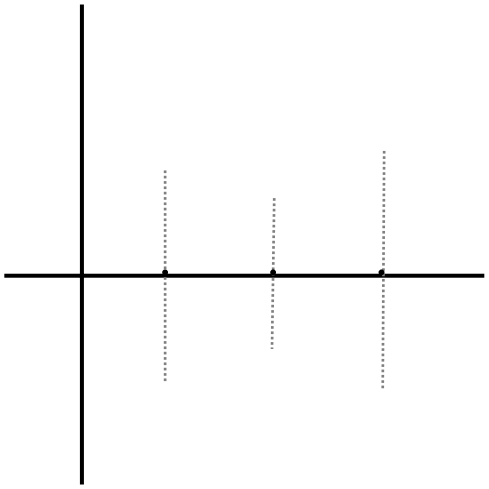

Gráfico No.2

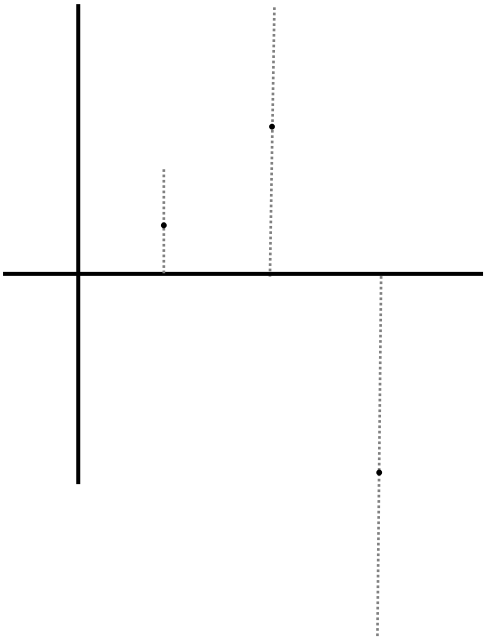

Gráfico No.3

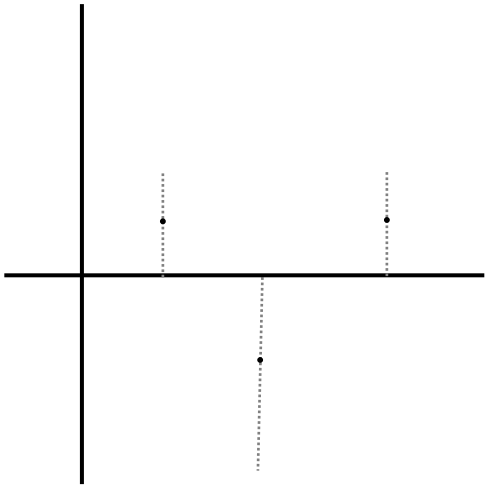

Gráfico No.4

Tabla 1. Modelo $\rightarrow E(Y_{i_j i'_{j'}}) = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + \lambda\alpha_{i_j}\beta_{i'j'}$

$\widehat{CMEP}$	Seleccionado el modelo optimal	$\widehat{CMEP}$ del modelo optimal
2.40293	x	
2.30035	x	
4.32473	no	3.99876
1.23825	x	
3.25822	x	
2.28502	no	2.19432
3.55866	x	
1.65434	x	
5.52838	no	4.31666
2.91463	x	

Tabla 2. Modelo $\rightarrow E(Y_{i_j i'_{j'}}) = \mu + \alpha_{i_j} + \beta_{i'_{j'}} + C_i\beta_{i'_{j'}}$

$\widehat{CMEP}$	Seleccionado el modelo optimal	$\widehat{CMEP}$ del modelo optimal
15.63442	x	
7.25386	x	
6.05034	x	
10.26811	x	
23.46212	no	15.16177
21.30004	no	19.66664
10.11103	no	9.86332
4.35899	x	
9.86870	x	
35.44286	x	

REFERENCIAS

1. BUNKE, O. and E. CASTELL (1998): "Regression and contrast estimates based on adaptive regressograms depending on qualitative explanatory variables", Discussion Paper 20. Sonderforschunsbereich 373. Universität Humboldt. Berlín.

2. DARROCH, J.N. and T.P. SPEED (1983): "Additive and multiplicative models and interaction", **Ann. Stat**. 11(3), 724-738.

3. HUET, S. (1991): "Asymptotic nonlinear regression methods for interactions", INRA, Jouy-en-Josas, Francia.

4. HUMAK, K.M.S. (1983): **Statistiche Methoden der Modellbildung II**. Akademie Verlag, Berlín, RDA.

5. JOHNSON, D.E. and F. GRAYBILL (1972): "An analysis of two-way model with interaction and no replication", JASA 67, 862-868.

6. KOVACS, L.B. (1980): "Combinatorial Methods of Discrete Programming". **Mathematical Methods of Operations Research**, 2.

7. MANDEL, J. (1961): "Non additivity in two-way analysis of variance", **Ann. Stat. Assoc**. 56, 878-888.

8. _____ (1971): "A new analysis of variance models for non-additive data", **Technometrics** 13, 1-18.

9. MILLIKEN, G.A. and F.A. GRAYBILL (1970): "Extensions of the general linear hypothesis model", **Am. Stat. Assoc**. 65, 797-807.

10. SCHEFFÉ, H. (1982): "The Analysis of Variance", Segunda edición. Wiley, New York.

11. WILHELM, U. (1985): "Zum problem der whechselvierkung in der variance analysi", Diplomarbeit. Universität Humboldt. Berlín.

12. ZWANZIG, S. (1979): "The choice of approximative models in nonlinear regression", **Mathematische Operationsforschung and Statistik**, 10.