USANDO MODELOS DE REGRESIÓN DE DOS NIVELES

Por: MARIO MIGUEL OJEDA RAMIREZ
ROBERTO BEHAR GUTIERREZ

RESUMEN:
En este artículo se presenta una breve revision de los aspectos metodológicos y computacionales clave para las aplicaciones de los modelos de regresión en dos niveles. Se da una serie de recomendaciones para implementar una modelación jerárquica y se presenta un ejemplo ilustrativo. Se apuntan algunos temas que requieren investigación.

I. INTRODUCCION

La regresión es una de las técnicas más utilizadas por los investigadores de las ciencias sociales y de la conducta. Con frecuencia se pueden encontrar reportes de investigación y artículos que presentan resultados de análisis de regresión sobre datos de estudios educativos (Goldstein, 1987) y pocas veces los investigadores resisten la tentación de correr una regresión sobre los datos de una encuesta levantada con un diseño en varias etapas (conglomerados). A esta realidad contribuyen varios factores, pero se destaca el hecho de que en los estudios educativos se desea estudiar relaciones causa-efecto sobre unidades primarias (estudiantes), pero también se desea estudiarlas sobre grupos o unidades agregadas (salones o escuelas). En el caso de las muestras complejas más que interesar la inferencia descriptiva (estimar medias, totales, razones) con frecuencia interesa la inferencia analítica (el estudio de una relación causa-efecto). En ambos casos un modelo estadístico de regresión es requerido y se plantea como un modelo superpoblacional. En tal sentido la población de referencia u objetivo (tarjet population) es más general que la población finita o de muestreo (sampled population).

El advenimiento de los modelos superpobla cionales (Cassel et al., 1977) ha permitido que la modelación se utilice sobre datos de encuestas complejas, y en este sentido se ha abierto una serie de posibilidades de explotación de la información de encuestas (Skinner et al., 1989). Indirectamente con estos desarrollos se han beneficiado áreas que producen datos de estudios observacionales, como la epidemiología, la antropología social y la investigación clínica. En todos estos campos las muestras observan estructuras de anidamiento (individuos en grupos,

grupos dentro de grupos más generales, etc.) y por tanto la aplicación directa de los modelos de regresión estándar, cuando estos se justifican, puede producir resultados incompatibles con la realidad conocida. En general, se sabe que los errores estándar de las estimaciones se sobreestiman y por tanto efectos significativos pueden no estarse identificando en tales análisis. Esto es debido a que se ignora la estructura de la muestra, o lo que se conoce en muestreo como el efecto de diseño (Potthoff et al., 1992). En resumen se puede decir que la correlación intraclase en las muestras anidadas o con estructura jerárquica produce efectos graves en la inferencia bajo el esquema de los modelos de regresión estándar (Goldstein, 1987; Bryk y Raudenbush, 1992; Longford, 1993). Además de que con frecuencia el objetivo es estudiar la variabilidad de las ecuaciones sobre los grupos; es decir, la unidad de estudio es un grupo de individuos (en estudios comparativos de rendimientos escolares por escuela, por ejemplo).

Los modelos lineales jerárquicos tienen una tradición bastante reciente, pero en la actualidad se encuentran ya disponibles con desarrollos a nivel teórico, metodológico y computacional, lo que permite usarlos en aplicaciones. Este artículo presenta una revisión de los principios, procedimientos y estrategias para realizar aplicaciones de los modelos de regresión en dos niveles. Está orientado a los científicos y profesionales que cotidianamente se enfrentan a datos con dos niveles de jerarquía, susceptibles de modelarse a través de la regresión. Se incluye, en la sección final, una aplicación concreta para ilustrar el procedimiento general propuesto.

II. MODELO GENERAL DE REGRESION EN DOS NIVELES

Los modelos lineales jerárquicos (Bryk y Raudenbush, 1992) constituyen una formulación general que permite la consideración de datos con estructura jerárquica. Este tipo de datos son comunes en estudios educativos, estudios de medidas repetidas, estudios de muestreo bietápico, entre otros. La idea de la estructura jerárquica es que las unidades aparecen agrupadas, y que varios grupos forman grupos en otro nivel y así sucesivamente. Por eso se les llama también modelos lineales multinivel (Goldstein, 1987).

El ejemplo típico de una situación de modelación lineal jerárquica lo constituye el del estudio del rendimiento escolar (Y) a partir de rendimiento previo, condiciones socioeconómicas, anteceden tes educativos del padre, etc.$(x_1, x_2, \ldots, x_p)$. Si consideramos que los estudiantes están agrupados por escuela, podemos proponer un modelo como el siguiente:

$$y_{ij} = \beta_{oi} + \beta_{1i} x_{1ij} + \cdots + \beta_{pi} x_{pij} + e_{ij} \qquad \text{(II.1)}$$

$$j = 1,2,\cdots, n_i \; ; i = 1,2,\cdots, G$$

Nótese que para cada una de las escuelas se está proponiendo una ecuación de regresión, que describe la relación causa efecto sobre los n_i estudiantes. Podemos suponer que son tales que $E(e_{ij})=0$ y $Var(e_{ij}) = \sigma_e^2$ para todo j y todo i. Si tenemos información de cada escuela, por ejemplo una calificación para la calidad del claustro docente (W), podríamos modelar la variabilidad en las ecuaciones de regresión; es decir, modelar los coeficientes de las regresiones. El modelo propuesto sería:

$$\begin{aligned} \beta_{0i} &= \gamma_{00} + \gamma_{01} w_i + u_{0i} \\ &\vdots \qquad\qquad\qquad \vdots \\ \beta_{pi} &= \gamma_{p0} + \gamma_{p1} w_i + u_{pi} \end{aligned} \qquad \text{(II.2)}$$

En general, podríamos postular un modelo que considere varias variables , y no necesariamente usar las mismas para modelarlas .

Nótese que las ecuaciones en (II.2) constituyen un sistema de (p+1) ecuaciones; es decir, son una regresión (p+1)-variada, donde "las res-

puestas" son los coeficientes de regresión en la ecuación (II.1), y la o las variables explicatorias son las W. Por eso estos modelos también son referidos como modelos de regresión con coeficientes aleatorios (Longford, 1995). En este sentido, asociadas a las ecuaciones en (II.2), tenemos suposiciones para un error aleatorio p-variado , $\mathbf{u}_i = (u_{oi}, u_{1i}, \cdots u_{pi})'$que se postulan independientes e identicamente distribuidos, y se establece:

a) $E\{\mathbf{u}_i\} = 0$

b) $Var\{\mathbf{u}_i\} = \boldsymbol{\Omega} = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01} & \cdots & \sigma_{0p} \\ \sigma_{10} & \sigma_{11}^2 & \cdots & \sigma_{1p} \\ \vdots & & & \\ \sigma_{pp} & \sigma_{p2} & \cdots & \sigma_{pp}^2 \end{bmatrix}$

donde $\boldsymbol{\Omega}$ es la matriz de varianzas σ_{ii}^2 y covarianzas $(\sigma_{ij}; i \neq j)$ de los errores $\mathbf{u}_i$;1,2,...,G. Esta presentación nos muestra que el modelo lineal jerárquico permite simultáneamente hacer un estudio de individuos y un estudio de grupos, en el sentido de que se consideran tanto variables explicatorias para los individuos (X_1,...,X_p) como para los grupos (W). El estudio de la variabilidad de las ecuaciones de regresión sobre la muestra de grupos es una posibilidad que con ninguna otra técnica puede lograrse, y éste se hace con base en la estimación e inferencia sobre las varianzas y covarianzas; por tal motivo a estos modelos se les llama también de componentes de varianza y covarianza (Laird y Ware, 1982).

Definiendo matricialmente la ecuación en (II.1) tendríamos:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{e}_i \quad i = 1,2,\cdots, G \quad \textbf{(II.3)}$$

con $E(\mathbf{e}_{ij}) = \mathbf{0}$ y $Var(\mathbf{e}_{ij}) = \sigma_e^2\mathbf{I}_{ni}$ Si ahora definimos

$$\boldsymbol{\beta}_i = \mathbf{W}_i\boldsymbol{\Gamma} + \mathbf{u}_i \; ; \; i = 1,2,\cdots,G \qquad \textbf{(II.4)}$$

donde es la matriz de coeficientes fijos γ en las ecuaciones (II.2).

Si asumimos que $E(\mathbf{u}_i) = \mathbf{0}$ y $Var\{\mathbf{u}_i\} = \boldsymbol{\Omega}$, entonces tenemos las ecuaciones generales, donde $\mathbf{W}_i = \mathbf{I}_{p+1} \otimes \mathbf{w}_i$, con $\otimes$, con indicando producto Kronecker, y $\mathbf{W}_i$ es el vector de valores de W_1, W_2, W_q para el grupo i-ésimo. Si escribimos la ecuación (II.4) en la (II.3) obtenemos el modelo general lineal jerárquico en dos niveles.

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{W}_i\boldsymbol{\Gamma} + \mathbf{X}\mathbf{u}_i + \mathbf{e}_i \quad i = 1,2,...,G \text{ (II.5)}$$

La ecuación (II.5) se puede escribir de una forma matricial más compacta definiendo las siguientes matrices y vectores.

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_G \end{bmatrix}, \quad \mathbf{X} = \text{diag}(\mathbf{X}_i), \quad \mathbf{W} = \text{diag}(\mathbf{W}_i)$$

$$\mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_G \end{bmatrix} \quad y \quad \mathbf{e} = \begin{bmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_G \end{bmatrix}$$

donde $(\mathbf{A}_i)$es la matriz diagonal por bloques $\mathbf{A}_i$. Así, obtenemos la ecuación que se corresponde con el modelo lineal general mixto,

$$\mathbf{y} = \mathbf{X}\mathbf{W}\boldsymbol{\Gamma} + \mathbf{X}\mathbf{u} + \mathbf{e} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{X}\mathbf{u} + \mathbf{e} \quad \text{(II.6)}$$

cuya estructura de covarianzas se expresa como:

$$\mathbf{V} = \mathbf{X}\left(diag(\boldsymbol{\Omega})\right)\mathbf{X}' + diag\left(\sigma^2\mathbf{I}_N\right) \; :$$

Donde . $N=n_1+...+n_G$ Los parámetros que hay que estimar en este modelo son los elementos de (llamados efectos fijos), y los de $\boldsymbol{\Omega}$ y σ_e^2 , incluidos en **V** (llamados componentes de varianza y covarianza). Este modelo en (II.6) se puede llevar a la forma general

$$\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \mathbf{e}^* \quad \text{(II.7)}$$

con $E(\mathbf{e}^*) = \mathbf{0}$ y Var $(\mathbf{e}^*) = \mathbf{V}$

en donde $\mathbf{X}^*$, $\boldsymbol{\beta}^*$, y $\mathbf{e}^*$ son una matriz de diseño general, un vector de parámetros $\boldsymbol{\beta}^*$, que incluye únicamente a los efectos fijos, y e^* es un vector que incluye a los errores aleatorios.

Si se asume $\mathbf{V}$ conocida se sabe que el estimador de mínimos cuadrados generalizados es:

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{X}^{*t}\mathbf{V}^{-1}\mathbf{X}^*)^{-1}\mathbf{X}^*\mathbf{V}^{-1}\mathbf{y}$$

y también que la matriz de varianzas y covarianzas de este estimador, sería:

$$Var(\hat{\beta}^*) = (\mathbf{X}^{*t}\mathbf{V}^{-1}\mathbf{X}^*)^{-1}$$

Sin embargo, el problema de estimación involucra así mismo a los parámetros desconocidos en $\mathbf{V}$.

III. ESTIMACION Y PRUEBA DE HIPOTESIS EN EL MODELO LINEAL JERARQUICO

El problema de estimación e inferencia en el modelo presentado en la sección anterior es harto complejo, puesto que V, la matriz de varianzas y covarianzas asociada, es asimismo un objetivo de estimación. Este problema tiene una larga tradición de estudio desde que fue formulado como tal (Lindley y Smith, 1971). La solución se ha abordado desde diferentes principios: máxima verosimilitud restringida (Longford, 1987), bayesiano empírico (Bryk y Raudenbush, 1992) y mínimos cuadrados generalizados (Goldstein, 1987). Se han propuesto algoritmos computacionales para la estimación, puesto que la solución requiere de métodos numéricos iterativos. Entre estos algoritmos se encuentran la adaptación particular del algoritmo EM (Dempster, Laird y Rubin, 1977; Bryk y Raudenbush, 1992), el algoritmo de Fisher-Scoring (Longford, 1987) y el algoritmo de Mínimos Cuadrados Generalizados Reponderados Interativamente (MCGRI), (Goldstein, 1987; Goldstein y Rasbash, 1992). Estos algoritmos se han implementado en paquetes computacionales que actualmente se distribuyen comercialmente: el ML3 (y versiones sucesivas, Prosser et al., 1990) el HLM (Bryk y Raudenbush, 1992) el VARCL (Longford, 1993), y versiones especiales de programas en BMDP y SAS, que se obtienen como productos aparte de los sistemas mencionados. Una discusión amplia de los méritos de los sistemas, emanada de una evaluación comparativa, se presenta en el artículo de Kreft et al. (1994); también en Vander Leeden et al. (1996) se puede encontrar información valiosa sobre los principios y procedimientos de inferencia en este contexto de modelación y una evaluación comparativa del software existente.

Invariablemente, la idea de estos procedimientos y algoritmos es realizar una estimación en dos fases: obtener una estimación inicial de la matriz $\mathbf{V}$ y sustituirla en las ecuaciones que la requieren para estimar $\boldsymbol{\beta}^*$; en una segunda fase se mejora la estimación de $\mathbf{V}$ y se repite el proceso hasta que se logra cumplir con un criterio de convergencia. Obviamente las soluciones al problema de estimación e inferencia no son cerradas y éstas dependen de la eficiencia de los algoritmos numéricos, y además están fuertemente afectadas por la cantidad de datos, tanto en la que se refiere al número de grupos como al número de individuos por grupo (Goldstein, 1995).

A manera de ilustración general y para establecer la lógica del funcionamiento del algoritmo MCGRI, a continuación damos una idea superficial de como trabaja.

En la ecuación (II.7) se ajusta el modelo suponiendo que $\mathbf{V} = \mathbf{I}_N\sigma^2$; es decir, se obtiene

$$\hat{\beta}^* = (\mathbf{X}^{*t}\mathbf{X}^*)^{-1}\mathbf{X}^{*t}\mathbf{y}$$

A partir de esta estimación se obtiene

$$\hat{\mathbf{e}}^* = \mathbf{y} - \mathbf{X}^*\hat{\beta}^*$$

con se obtienen las primeras estimaciones para los elementos de **V**, lo que se logra aplicando un algoritmo de mínimos cuadrados a la ecuación de **V**; es decir:

$$\hat{\mathbf{e}}^*\hat{\mathbf{e}}^* = \mathbf{X}^* \left(diag(\Omega)\right) X^t + diag\left(\sigma_e^2 \mathbf{I}_N\right)$$

donde los parámetros desconocidos son Ω y σ_e^2 y el resto de componentes son valores conocidos. Escribir esta ecuación en una forma operable y desarrollar un algoritmo eficiente no fue sencillo. Para detalles a este respecto se recomienda ver Goldstein y Rassbash (1992). La idea es que de este paso se obtiene $\hat{\mathbf{V}}$. Acto seguido se sustituye esta estimación en el estimador de mínimos cuadrados generalizados y se obtiene una segunda estimación de β^*. Esto nos permite producir otros residuos. El proceso se repite hasta que la estimación $\hat{\mathbf{V}}$ de sea estable.

La prueba de hipótesis para el modelo (II.5) se formula tanto sobre los efectos fijos en Γ como sobre los efectos aleatorios o varianzas y covarianzas. En el primer caso, se orienta el interés sobre la tendencia general de los modelos sobre la muestra de grupos, y en el segundo caso interesa establecer juicios concluyentes sobre la dispersión de los coeficientes de los modelos. Para los efectos fijos se usa el enfoque de la hipótesis lineal general refiriendo resultados asintóticos bajo la suposición de normalidad, por lo que contar con un número suficiente de grupos e individuos dentro de grupos es muy importante. Por otro lado, la prueba de hipótesis para las varianzas y covarianzas es generalmente basada en el estadístico de razón de verosimilitudes, también bajo las mismas suposiciones. Buenas revisiones y notas bibliográficas sobre este asunto se pueden encontrar en Longford (1993; 1995) y Goldstein (1987; 1995).

El principal problema en el ajuste y conducción del proceso de inferencia para los modelos lineales jerárquicos es que para un problema particular se requiere implementar una estrategia que permita ir aproximándose a la complejidad máxima del modelo permitida por los datos. En ese sentido, lo que se entiende es que la formulación del modelo no puede ser tal que se plantee un modelo muy complejo, puesto que en tal caso los algoritmos numéricos pueden no funcionar adecuadamente. En la siguiente sección abordamos la problemática de la formulación e implementación de estrategias de modelación lineal jerárquica, siguiendo algunos lineamientos desarrollados en Ojeda (1992, 1993); Behar y Ojeda (1995), y Ojeda et al. (1996).

IV. ESTRATEGIAS DE ANALISIS DE DATOS EN MODELACION LINEAL JERARQUICA

Ante situaciones de análisis de datos en los que la estructura es jerárquica, no es posible proceder de forma directa a la formulación de un modelo complejo. Primero porque es muy difícil lograr un planteamiento realista sin evaluar la importancia de las variables explicativas en una manera exploratoria, y segundo porque un modelo complejo de partida puede no ser soportado por los datos, en el sentido de que el algoritmo numérico no corra por mal condicionamiento. La segunda cuestión es bastante frecuente en problemas reales según la experiencia de los autores.

En lo que sigue presentaremos algunas recomendaciones generales que, de alguna forma, pueden ayudar para hacer aplicaciones concretas de la modelación de regresión lineal jerárquica.

PRIMERO: Haga siempre análisis exploratorios para conocer suficiente sobre las distribuciones marginales y bivariadas de los datos, tanto de manera global como grupo por grupo. Debe haber evidencia suficiente sobre la razonabilidad del modelo de regresión para explicar las relaciones que se están estudiando. Los gráficos de caja, los diagramas de tallo y hoja, los correlogramas, los correlogramas codificados y los diagramas de escalera (*Halft Matrix Plot*) son herramientas que serán de indudable utilidad en esta fase, en la que es posible identificar errores en los datos, datos atípicos y tener una idea del rango de variabilidad en las asociaciones estudiadas.

SEGUNDO: Realice ajustes por separado y de manera global para tener una clara idea de la variación de los coeficientes de regresión estimados. En general esto da una primera buena idea de las variables explicatorias realmente significativas al nivel de individuos. Cabe hacer notar que los coeficientes estimados mediante este procedimiento constituyen en si la primera base para explorar la importancia de las variables explicatorias en el segundo nivel (las). La exploración detallada de las asociaciones entre las W y los β estimados por separado nos puede dar elementos para establecer patrones de comportamiento central y variabilidad. Llamemos a la matriz de coeficientes estimados por separado **B**. Esta es una matriz de datos multivariados de orden G por (p+1). Podemos, usar técnicas multivariadas exploratorias, como Componentes Principales o Análisis de Conglomerados, para explorar patrones de asociación entre variables (los β estimados) e individuos (en este caso los grupos). También podemos hacer un estudio de correlación canónica de **B** con la matriz de datos de las variables .

TERCERO: Una vez agotadas las fases anteriores se debe tener una idea clara de la importancia relativa de las X en la predicción de los valores Y, y se supone que se está en posibilidades de postular el modelo lineal jerárquico. Para este momento tenemos una estimación de , que es la que se obtiene de la regresión global. La primera pregunta fuerte de contestar es saber si hay suficiente variabilidad entre grupos como para justificar el ajuste de un modelo jerárquico; es decir, contestarse si realmente los datos tienen una estructura jerárquica que justifique la complejidad del modelo. Para esto Goldstein (1987) sugiere ajustar un modelo de componentes de varianza simple; es decir, propone ajustar:

$$y_{ij} = \beta_{0i} + e_{ij}$$
$$i = 1,2,\cdots,G\ ;\ j = 1,2,\cdots,n_i \qquad \text{(IV.1)}$$

con modelo en el segundo nivel:

$$\beta_{oi} = \gamma_{oo} + u_i \qquad \text{(IV.2)}$$

Nótese que aquí $\Omega = \sigma_u^2$, al que se le llama componente de varianza entre grupos. Con las estimaciones de σ_e^2 y σ_u^2 se constituye una estimación del coeficiente de correlación intraclase ρ, que sería:

$$\hat{\rho} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_e^2 + \hat{\sigma}_u^2} \qquad \text{(IV.3)}$$

Según Goldstein (1987) basta declarar que se justifica un ajuste de un modelo jerárquico. La experiencia de los autores a este respecto es que en algunos casos valores de hasta 0.08 para $\hat{\rho}$ han dado análisis multinivel interesantes. Por tal motivo recomendamos que el ajuste del modelo de componentes de la varianza se tome con reservas, y que más bien se proceda en correspondencia con la información obtenida en los análisis exploratorios previos.

El modelo de regresión con coeficientes aleatorios es el siguiente paso, para lo cual se considera recomendable sólo incluir la variable X que mayor evidencia de explicabilidad sobre Y haya dado. El modelo a ajustar sería:

$$y_{ij} = \beta_{oi} + \beta_{1j}\beta_{1i}x_{1ij} + e_{ij} \qquad \text{(IV.4)}$$

El modelo en el segundo nivel sería

$$\beta_{oi} = \gamma_{oo} + u_{oi}$$
$$\beta_{1i} = \gamma_{10} + u_{1i} \qquad \text{(IV.5)}$$

La matriz de varianzas y covarianzas asociada a $\upsilon_i = (u_{oi}, u_{1i})^t$ sería

$$\Omega = \begin{bmatrix} \sigma_{00}^2 & \sigma_{01} \\ & \sigma 2_{11} \end{bmatrix}$$

Para cada estimación de las entradas de esta matriz se obtiene un error estándar asociado, con lo que es posible decidir si la variabilidad del coeficiente es significativa. También se obtiene una estimación de σ_e^2, la que se espera se reduzca en un porcentaje importante respecto de las estimaciones obtenidas del ajuste global y del modelo de componentes de la varianza.

Para evaluar la mejora de la ecuación cuando se introducen más X se recomienda hacer esto de una en una. Nótese que el número de parámetros a estimar en crece conforme más variables X tenga la ecuación en (IV.4). Así por ejemplo, si tenemos tres variables explicatorias, tendríamos:

$$y_{ij} = \beta_{0i} + \beta_{1i}x_{1ij} + \beta_{2i}x_{2ij} + \beta_{3i}x_{3ij} + e_{ij}$$

y entonces

$$\Omega = \begin{bmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} & \sigma_{03} \\ & \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ & & \sigma_2^2 & \sigma_{23} \\ & & & \sigma_3^2 \end{bmatrix}$$

Un modelo sobreparametrizado en este sentido

requiere muchos grupos, y en la práctica el número de grupos está siempre restringido. Se recomienda para una X y una W tener al menos 20 grupos y 5 individuos por grupo.

En lo que se refiere a la inclusión de las variables W, ésta se debe hacer de manera secuencial una más en cada corrida. Esto implicaría que las ecuaciones en (IV.5) sean cambiadas por

$$\beta_{0i} = \gamma_{00} + \gamma_{01} w_i + u_{0i}$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11} w_i + u_{1i} \qquad \text{(IV.6)}$$

No hemos encontrado aplicaciones en las que se requiera ajustar más de una W, aunque sólo hemos tenido problemas reales con tres W candidatas. Lo que pensamos es que usualmente no se diseña la obtención de información para grupos. Goldstein (1987) recomienda usar medias, desviaciones estándar o coeficientes de variación muestrales en cada grupo como variables W. Lo difícil en estos casos es dar una buena interpretación a los resultados de las corridas.

CUARTO: Dadas las estimaciones finales de los efectos fijos y los aleatorios es posible obtener estimaciones de los errores en ambos niveles; es decir, podemos obtener los residuos, $\hat{e}$ y $\hat{u}$ Con estos residuos se debe hacer diagnósticos. En el caso de los residuos en el primer nivel los podemos tratar a partir de las técnicas tradicionales de diagnóstico (Goldstein, 1987; Bryk y Raudenbush, 1992). El problema más complejo se presenta con los residuos en el segundo nivel, que son residuos (p+1) variados. Sea $\mathbf{U}$ la matriz que tiene por renglones a los vectores $\hat{u}_i^t$; $i = 1,2,\cdots,G$. Esta matriz puede ser sometida a técnicas de análisis multivariado como Componentes Principales. Obteniendo la matriz $\mathbf{C}_{(2)}$, de los puntajes *(scores)* de los primeros dos componentes principales, es posible identificar grupos atípicos o incluso ordenarlos. También se pueden usar los últimos componentes (Barnet y Lewis, 1994) para identificar outliers. Ojeda y Juárez-Cerrillo (1996a; 1996b) exploraron la viabilidad del uso de la técnica del Biplot (Gabriel, 1971) en la identificación de grupos atípicos. En este sentido hay poco estudio reportado, y principalmente pocas referencias sobre experiencias prácticas, y esto se abre como una potencial línea de investigación (Seltzer, 1993).

Muchos de los procedimientos aquí mencionados es posible implementarlos con el paquete ML3 y con el apoyo de algún otro paquete estadístico. En este sentido se debe entender que las estrategias de modelación jerárquica son iterativas y requieren una compenetración con el problema bajo estudio. En Ojeda (1992, 1993), Ojeda y Juárez-Cerrillo (1996a; 1996b), Behar y Ojeda (1995) y Ojeda et al. (1996) aparecen algunos ejemplos y recomendaciones prácticas sobre el proceso de modelación jerárquica. Para efectos de puntualizar detalles en la siguiente sección incluimos un ejemplo que tiene como objetivo ilustrar los aspectos mencionados.

EJEMPLO

Este estudio tiene el propósito de modelar la relación que existe entre las habilidades de lectoescritura y el rendimiento escolar en estudiantes de bachillerato, para una población de escuelas, de las cuales se obtuvo una muestra irrestricta aleatoria de 35. En cada escuela seleccionada se obtuvo una muestra de entre 8 y 15 estudiantes, dependiendo del tamaño de la escuela. A cada estudiante seleccionado, al inicio del peiodo escolar se le aplicaron dos test, uno que mide la habilidad de lectura y el otro que mide la de escritura. Los puntajes obtenidos se denominaron LECTURA y ESCRITURA, respectivamente. Al final del periodo escolar se re-

gistró el promedio de calificaciones del estudiante en el periodo regular, al que se le denominó RENDIMIENTO. Para explicar el efecto contextual se evaluó, en cada una de las escuelas en la muestra, la implementación real de una serie de protocolos y procedimientos recomendados por la oficina de educación para mejorar el rendimiento escolar. Con estos registros se confeccionó un indice denominado CALIDAD, los datos fueron transformados linealmente a una escala de 0 a 100. La Figura 1 muestra las distribuciones de las variables en estudio en una forma comparativa por escuela.

En la Figura 2 se muestra la gráfica de dispersión de las variables en estudio, incluyendo un un ajuste de línea recta, con el propósito de identificar las asociaciones más importantes. Al explorar la asociación lineal entre LECTURA y ESCRITURA como explicatorias del RENDIMIENTO, en cada una de las escuelas por separado, se encontró suficiente evidencia que justifica el ajuste de un modelo de regresión lineal. El modelo propuesto es:

$$(RENDIMIENTO)_{ij} = \beta_0 + \beta_1 (LECTURA)_{ij} + \beta_2 (ESCRITURA) + e_{ij}$$

donde $i = 1, 2, \ldots, 35$ y $j = 1, 2, \ldots, n;$

Este modelo se ajustó para cada escuela por separado y de manera global. El cuadro 1, presenta los resultados importantes. En la Figura 3 se presentan los diagramas de dispersión de los coeficientes estimados separadamente con la variable CALIDAD, también incluyendo las distribuciones univariadas representadas por histogramas. Aquí se puede observar que la variable CALIDAD tiene una distribución bastante inadecuada; sin embargo en ausencia de otra información se decidió utilizarla.

Como primera fase en el análisis jerárquico se ajustó un modelo de componentes de la varianza y se obtuvo que el coeficiente de correlación intraclase estimado que resultó $\hat{\rho} = (29.42 / 106.21)$, lo que justifica plenamente un análisis mayor. Sobre todo porque las estimaciones de los componentes de la varianza tuvieron errores estándar estimados pequeños (8.742 y 5.772 respectivamente).

Al ajustar los modelos de regresión con coeficientes aleatorios, tanto para LECTURA como para ESCRITURA por separado, se obtuvo evidencia de que era posible ajustar el modelo.

$$(RENDIMIENTO)_{ij} = \beta_{0i} + \beta_{1i} (LECTURA)_{ij} + \beta_{2i} (ESCRITURA)_{ij} + e_{ij}$$

con modelos al segundo nivel como:

$$\beta_{0i} = \gamma_{00} + u_{0i}$$
$$\beta_{1i} = \gamma_{10} + u_{1i}$$
$$\beta_{2i} = \gamma_{20} + u_{2i}$$

ESCUELA	n	C-DET-AJU	ECM	BETA-0	BETA-1	BETA-2	CALIDAD
1	8	0.963	6.13	49.093	0.144	0.424	100.00
2	15	0.96	3.414	47.119	0.026	0.265	100.00
3	11	0.91	7.177	41.526	0.081	0.279	100.00
4	10	0.921	3.086	53.423	0.104	0.235	100.00
5	11	0.948	3.128	44.524	0.099	0.204	100.00
6	10	0.973	2.154	49.092	0.12	0.267	100.00
7	14	0.877	5.633	45.387	0.16	0.206	100.00
8	10	0.941	7.064	49.125	0.122	0.308	100.00
9	10	0.945	3.914	48.523	0.141	0.248	37.601
10	12	0.974	1.584	44.027	0.11	0.246	37.601
11	9	0.863	4.041	38.718	0.199	0.16	37.601
12	13	0.958	2.238	50.301	.0157	0.257	37.601
13	9	0.975	2.089	48.661	0.129	0.321	37.601
14	15	0.962	2.011	45.416	0.107	0.216	37.601
15	12	0.924	6.614	50.325	0.083	0.342	37.601
16	8	0.978	0.96	44.461	0.155	0.224	37.601
17	13	0.933	3.557	48.072	0.040	0.274	37.601
18	12	0.959	3.863	47.992	0.12	0.270	37.601
19	9	0.985	0.699	43.848	0.126	0.242	37.601
20	12	0.969	2.004	45.023	0.117	0.268	37.601
21	14	0.946	4.144	48.814	0.121	0.273	37.601
22	15	0.95	3.264	47.084	0.138	0.261	37.601
23	9	0.955	2.826	44.361	0.137	0.277	37.601
24	9	0.955	2.826	44.361	0.137	0.277	32.299
25	8	0.977	3.606	44.905	0.168	0.38	32.299
26	12	0.972	4.577	50.051	0.105	0.421	32.299
27	12	0.907	3.66	50.242	0.112	0.221	32.299
28	10	0.953	3.085	44.001	0.067	0.269	32.299
29	15	0.916	6.344	46.375	0.102	0.327	32.299
30	12	0.905	6.015	47.757	0.086	0.233	32.299
31	11	0.948	5.116	45.798	0.122	0.284	32.299
32	10	0.973	2.048	47.965	0.129	0.229	32.299
33	8	0.903	4.582	41.699	0.062	0.264	32.299
34	10	0.951	4.490	46.859	0.199	0.307	32.299
35	8	0.973	2.708	48.771	0.155	0.280	82.767
Global	389	0.712	30.06	46.502	0.115	0.284	

› 1. Resultados relevantes del ajuste de regresión por separado y de manera global. NUMEST úmero de estudiantes, C-DET-AJU es el coeficiente de determinación ajustado, ECM es el uadrático medio, BETA-0, BETA-1 y BETA-2, son los coeficientes de regresión estimados y)AD es el índice de calidad asociado a la escuela.

VARIABLE	ESTIMADORES DE COEFICIENTES γ	ESTIMADORES DE VARIANZA Y COVARIANZAS		
		CONSTANTE	*LECTURA*	*ESCRITURA*
CONSTANTE	46.92 (p < 0.001)	3.44200 (p = 0.0384)		
LECTURA	0.117 (p < 0.001)	0.02690 (p = 0.02690)	0.00104 (p < 0.001)	
ESCRITURA	0.275 (p < 0.001)	0.07450 (p = 0.0050)	0.00012 (p < 0.1)	0.00264 (p < 0.001)

Cuadro 2. Estimadores de los coeficientes fijos en el modelo (γ) así como de los componentes de varianza, mostrando los valores de probabilidad empírica (p-values) usando resultados asintóticas bajo normalidad.

Podemos observar que los coeficientes fijos resultaron altamente significativos. Así mismo, las entradas correspondientes a los componentes de varianza y covarianza resultaron ser seignificativas, excepto $\hat{\sigma}_{12}$.

Al incluir como variable explicatoria a nivel de grupo a CALIDAD; es decir, al plantear los modelos al segundo nivel como:

$$\beta_{0i} = \gamma_{00} + \gamma_{01}(CALIDAD)_i + u_{0i}$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11}(CALIDAD)_i + u_{1i}$$
$$\beta_{2i} = \gamma_{20} + \gamma_{21}(CALIDAD)_i + u_{2i}$$

se obtuvo los resultados que se muestran en el Cuadro 3.

VARIABLE	ESTIMADORES DE LOS COEFICIENTES γ	ESTIMADORES DE LAS VARIANZAS Y COVARIANZAS		
		CONSTANTE	*LECTURA*	*ESCRITURA*
CONSTANTE	43.0200 (p < 0.001)	1.222 (p < 0.1)		
LECTURA	0.0632 (p < 0.001)	-0.0048 (p < 0.1)	0.0006 (p = 0.0449)	
ESCRITURA	0.2080 (p < 0.001)	0.0335 (p = 0.0658)	-0.0005 (p > 0.1)	0.0019 (p < 0.001)
CALIDAD	0.0655 (p < 0.001)			
*CALIDAD*LECTURA*	0.0009 (p < 0.001)			
*CALIDAD*ESCRITURA*	0.0012 (p < 0.001)			

Cuadro 3. Estimadores de los coeficientes fijos en el modelo (γ) incluyendo como variable explicatoria al segundo nivel a CALIDAD, y los respectivos estimadores de los componentes de varianzas y covarianza, mostrando los valores de probabilidad empírica (p - values), usando resultados asintóticos bajo normalidad.

Como puede observarse todos los coeficientes fijos resultaron significativos, lo que indica que el modelo explica sustancialmente a los datos; por otro lado, los componentes de varianza y covarianza indican una variabilidad significativa entre los grupos, excepto en los términos independientes en donde la varianza estimada no resultó ser significativamente distinta de cero. Además, el estimador σ_e^2 resuelto 3.922 con un error estándar estimado de 0.325.

Al hacer un análisis de los residuos de primer nivel, $\hat{e}_{ij}$, se encontró alguna evidencia visual de heterocedasticidad, y una distribución con un sesgo positivo, que aunque pequeño si notorio (ver figuras 4 y 5).

Para evaluar la posibilidad de otro modelo mejor se realizó una transformación logarítmica a la variable respuesta y se ajustó nuevamente el último modelo. Se obtuvieron los resultados que se muestran en el Cuadro 4.

VARIABLE	ESTIMADORES DE LOS COEFICIENTES γ	ESTIMADORES DE LAS VARIANZAS Y COVARIANZAS		
		CONSTANTE	*LECTURA*	*ESCRITURA*
CONSTANTE	3.79 (p < 0.001)	0.00155 (p = 0.0076)		
LECTURA	0.0012 (p < 0.001)	0.000009 (p > 0.1)	0.000000 (p > 0.1)	
ESCRITURA	0.0038 (p < 0.001)	0.000000 (p > 0.1)	0.000000 (p > 0.1)	0.000000 (p > 0.1)
CALIDAD	0.00182 (p < 0.001)			
*CALIDAD*LECTURA*	0.000009 (p = 0.0414			
*CALIDAD*ESCRITURA*	0.000005 (p > 0.1)			

Cuadro 4. Estimadores de los coeficientes fijos (γ) incluyendo como variable explicatoria de segundo nivel a CALIDAD cuando la variable RENDIMIENTO fue transformada a la escala logarítmica; también aparecen los estimadores de los componentes de varianza y covarianza, mostrando los valores de probabilidad empírica (p-values), usando resultados asintóticos bajo normalidad.

Como puede observarse el modelo para los efectos fijos sería:

$$LOGE(\hat{RENDIMIENTO})_{ij} = 3.79 + 0.0012(LECTURA)_{ij} + 0.0038(ESCRITURA)_{ij} + 0.000009(CALIDAD)_i * (LECTURA)_{ij}$$

La distribución de los residuos, $\hat{e}_{ij}$, para este modelo muestra una sensible mejora. Así mismo, el efecto de heterocedasticidad se ha atenuado sensiblemente. Esto se puede ver en las figuras 6 y 7.

Para complementar el análisis se exploró los residuos de segundo nivel para el término constante, ya que ninguno de los componentes de varianza y covarianza restantes resultaron ser significativos (ver Cuadro 4). A partir de estos residuos es

posible caracterizar los diferentes tipos de escuelas. En la Figura 8 se presenta el gráfico de los dos primeros componentes principales de los coeficientes de regresión estimados por separado, incluyendo la variable calidad. Se señala el valor de los residuos estandarizados de segundo nivel para el término constante sobre las escuelas que parecen ser atípicas.

REFERENCIAS

BARNETT, V. and LEWIS, T. (1994) Outliers in Statistical Data (Second Edition); New York, Wiley.

BEHAR, R. and OJEDA, M.M. (1995) A two-level regression model for studying achivement engineering students of Universidad del Valle, at Cali, Colombia; por aparecer en Investigación Operacional.

BRYK, A.S. and RAUDENBUSH, S.W. (1992) Hierarchical Linear Models: Applications and Data Analysis Methods; Sage, CA, USA.

CASSEL, C.M., SARNDAL, C.E., and WRETMAN, J.H. (1977) Foundations of Inference in Survey Samplig; Wiley, New York.

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D.B. (1977) Maximun likelihood for incomplete data via the EM algorithm (with discussion) Journal of The Royal Statistical Society, Series B, 39, 1-38.

GABRIEL, K.R. (1971) The biplot display of matrices with applications to principal components analysis; Biometrika, 58, 453-467.

GOLDSTEIN, H. (1987) Multilevel Models in Educational and Social Research; London, Griffin.

GOLDSTEIN, H. (1995) Multilevel Statiscal Models (Second Edition); Wiley New York.

GOLDSTEIN, H. and RASSBASH, J. (1992) Efficient computational procedures for the estimation of parameters in multilevel models based on iterative gneralised least squares; Computational Statistics & Data Analysis, 13, 63-71.

KREFT, I.G.G., DE LEEW, J. and CAR VAR LEEDEN, R. (1994) Review of five Multilevel Analysis Programs: BMDP-5V, GENMOD, HLM, ML3, VARCL; The American Statistician, 48 (4), 324-335.

LINDLEY, D.V. and SMITH, A.F.M. (1972) Bayes estimates for the linear model; Journal of The Royal Statistical Society, Series B, 34, 1-41.

LONGFORD, N. (1993) Random Coefficient Models; Oxford University Press, London.

LONGFORD, N.T. (1995). Random coefficient models. In Handbook of Staristical Models for the Social and Behavioral Sciences, p.p. 519-577, (Eds. Arminger, G., Clogg, C.C. and Sobel, M.E.); Plenum Press, New York.

OJEDA, M.M. (1992) Aspectos Teóricos, Metodológicos y Computacionales en el Análisis de Muestras Complejas; Tesis Doctoral, Universidad de La Habana, Cuba.

OJEDA, M.M. (1993) Multilevel modelling strategies for complex samples; paper presented at 1993 European Meeting of The Psychometric Society, Barcelona, Spain.

OJEDA, M.M. and JUAREZ-CERRILLO S.F. (1996a) Biplot display for diagnostic in a two-level regression model for data analysis of growth curves; Computational Statistics and Data Analysis, 22, 583-597.

OJEDA, M.M. AND JUAREZ-CERRILLO, S.F. (1996b) Identifying multiple outliers with Hadi's method in a 2-level linear model for growth curves; under review process.

OJEDA, M.M., SAHAI, H. AND JUAREZ-CERRILLO, S.F. (1996) Multilevel data analysis with hierarchical linear models; under review process.

POTTHOFF, R.F., WOODBURY, M.A., and MANTON, K.G. (1992) "Equivalent sample size" and "equivalent degrees off reedom" refinements for inference using survey weights under superpopulation models; Journal of The American Statistical Association, 87 (418), 383-396.

PROSER, R., RASSBASH, J. and GOLDSTEIN, H. (1990) ML3: Software for Three-level analysis; Institute of Education, University of London.

SELTZER, M.H. (1993) Sensitivity analysis for fixed effects in the hierarchical model: a Gibbs sampling approach; Journal of Educational Statistics, 18(3), 207-235.

SKINNER, C.J. HOLT,D., and SMITH T.M.F. (eds.) (1989) Analysis of Complex Surveys; Wiley, New York.

VAN DER LEEDEN, R., VRIJBURG, K. AND DE LEEUW, J. (1996). A review of two different approaches for the analysis of growth data using longitudinal mixed linear models; The Statistical Software Newsletter, 583 - 605.

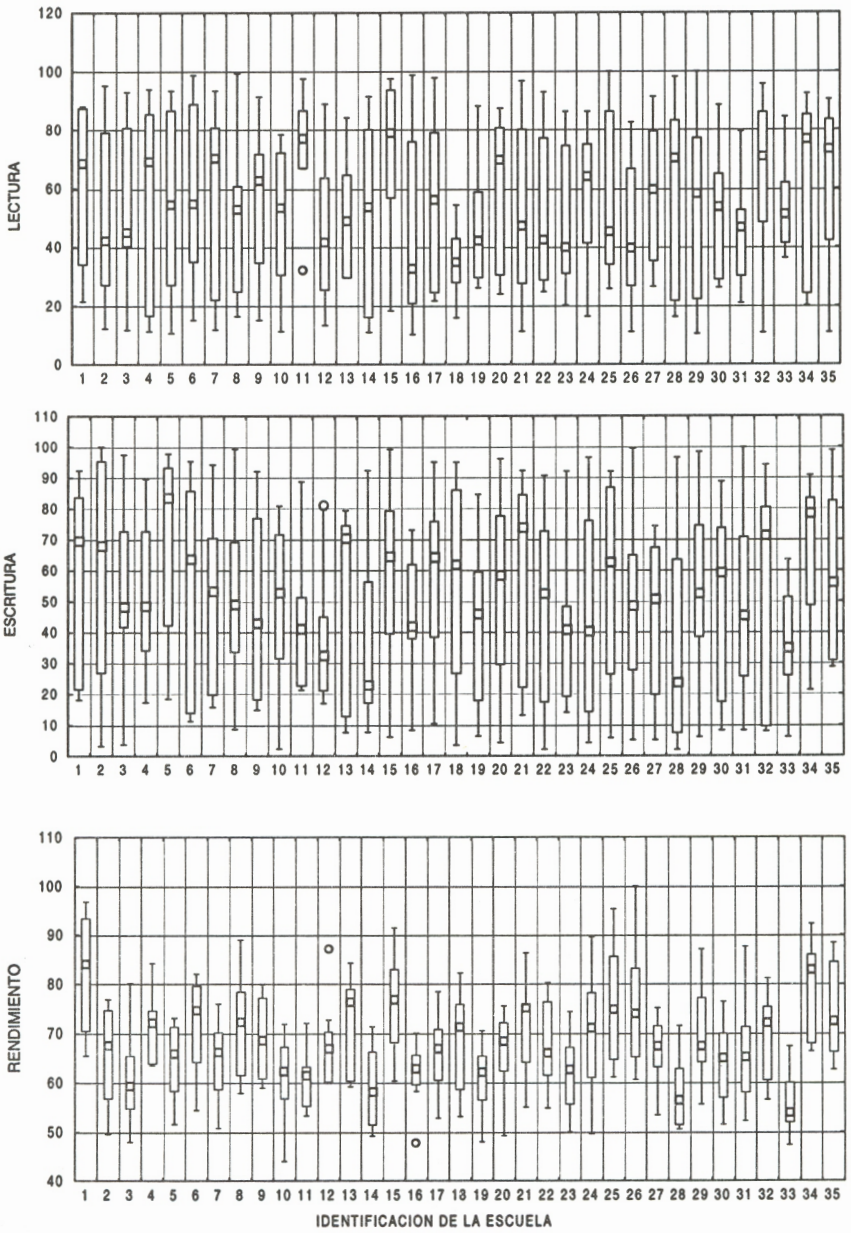


Figura 1. Distribuciones de las variables en estudio mostradas comparativamente por escuela.

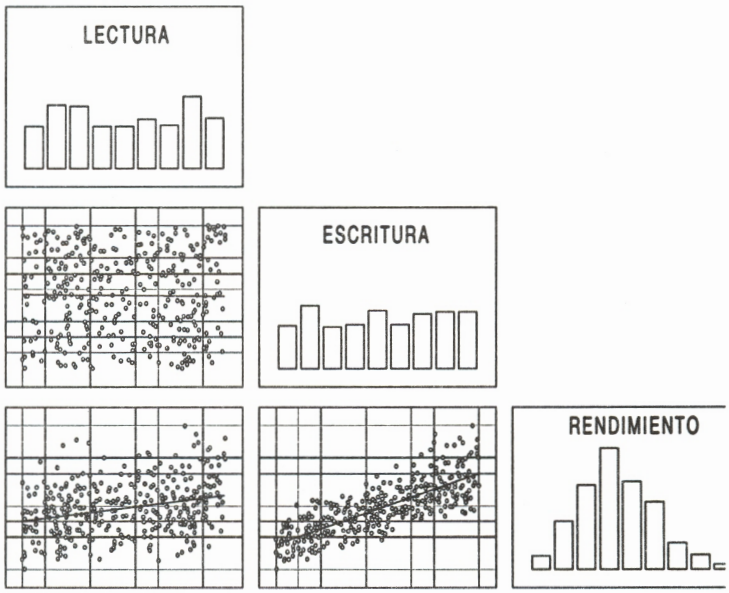


Figura 2. Diagrama de dispersión de las variables en estudio, mostrando histogramas para visualizar las distribuciones univariadas.

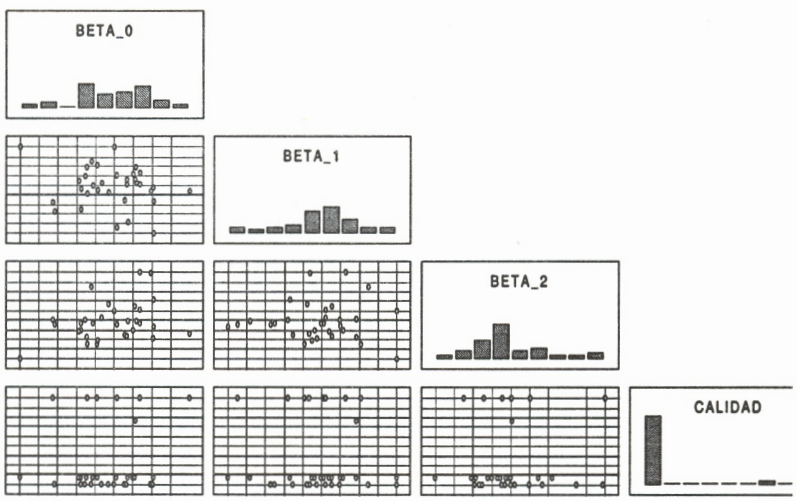
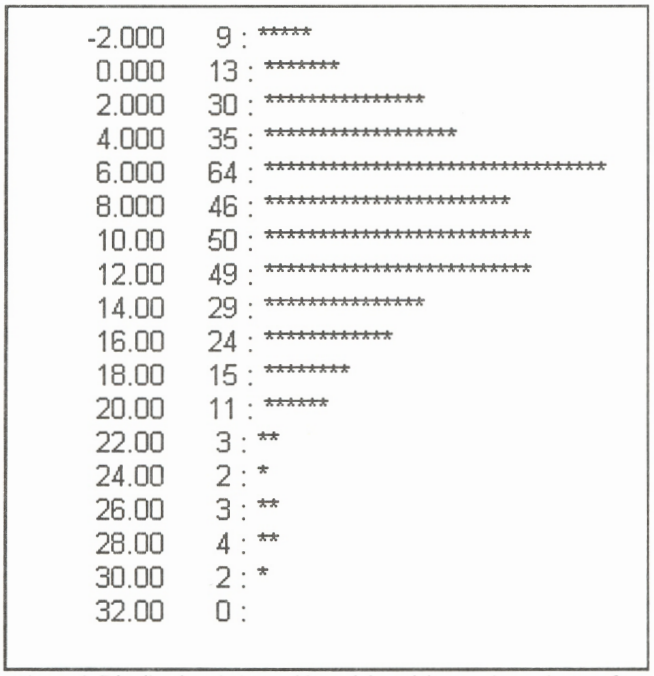


Figura 3. Diagramas de dispersión de los coeficientes estimados por separado y la varial CALIDAD, mostrando histogramas para identificar las distribuciones univariadas.



```
-2.000     9 : *****
 0.000    13 : *******
 2.000    30 : ***************
 4.000    35 : ******************
 6.000    64 : ********************************
 8.000    46 : ***********************
10.00     50 : *************************
12.00     49 : *************************
14.00     29 : ***************
16.00     24 : ************
18.00     15 : ********
20.00     11 : ******
22.00      3 : **
24.00      2 : *
26.00      3 : **
28.00      4 : **
30.00      2 : *
32.00      0 :
```

Figura 4. Distribucion de los residuos del modelo con datos sin transformar.

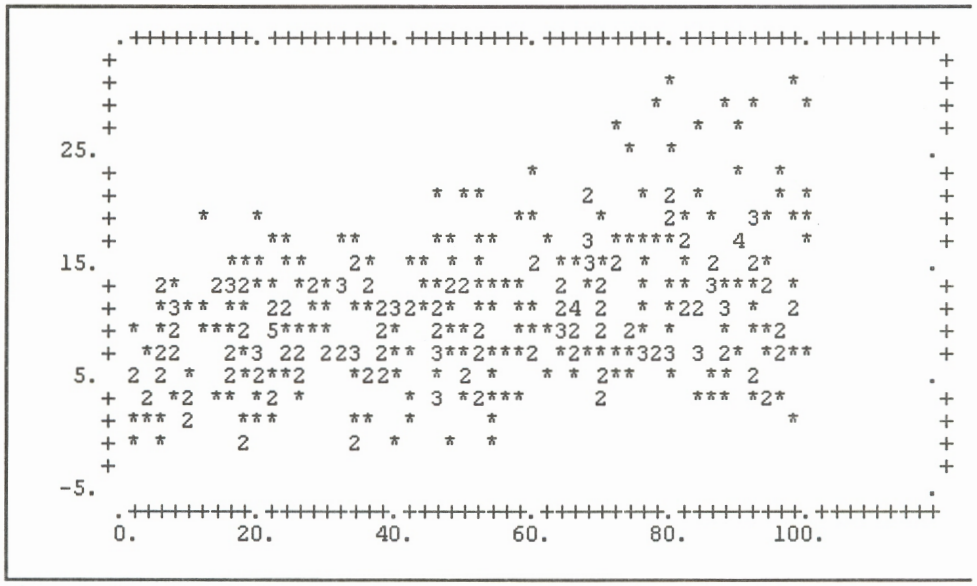


Figura 5. Diagrama de dispersión de residuos contra predichos del modelo con los datos si transformar.

```
      -3.300       1 : *
      -3.000       0 :
      -2.700       2 : **
      -2.400       4 : ****
      -2.100       7 : *******
      -1.800      12 : ************
      -1.500      19 : ********************
      -1.200      26 : **************************
      -0.900      35 : ***********************************
      -0.600      42 : ******************************************
      -0.300      46 : **********************************************
-8.882e-016       46 : **********************************************
       0.300      42 : ******************************************
       0.600      35 : ***********************************
       0.900      27 : ***************************
       1.200      19 : ********************
       1.500      12 : ************
       1.800       7 : *******
       2.100       4 : ****
       2.400       2 : **
       2.700       0 :
       3.000       1 : *
       3.300       0 :
```

Figura 6. Distribución de los residuos estandarizados del modelo usando los d transformados a escala logarítmica.

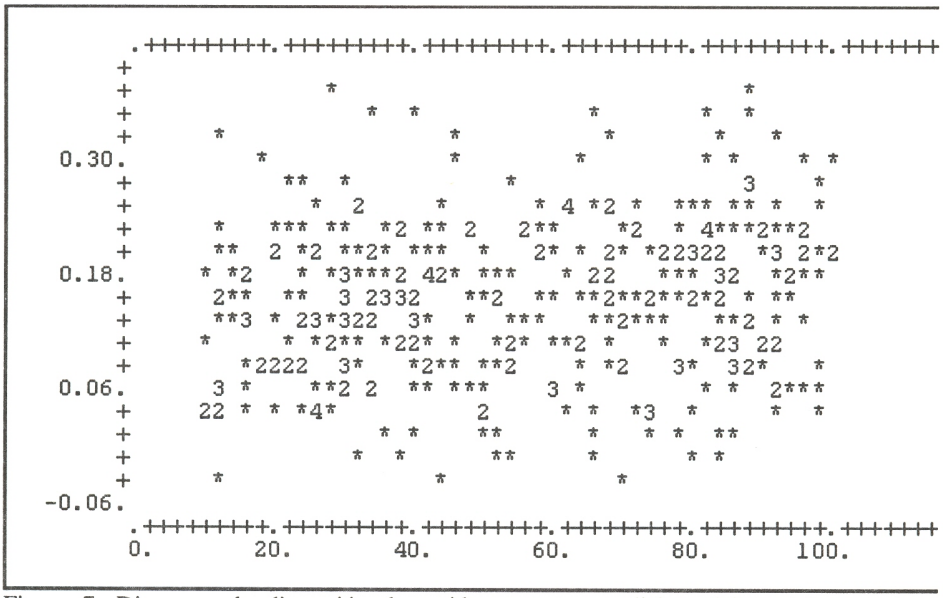


Figura 7. Diagrama de dispersión de residuos contra predichos del modelo con los tranformados a escala logarítmica.

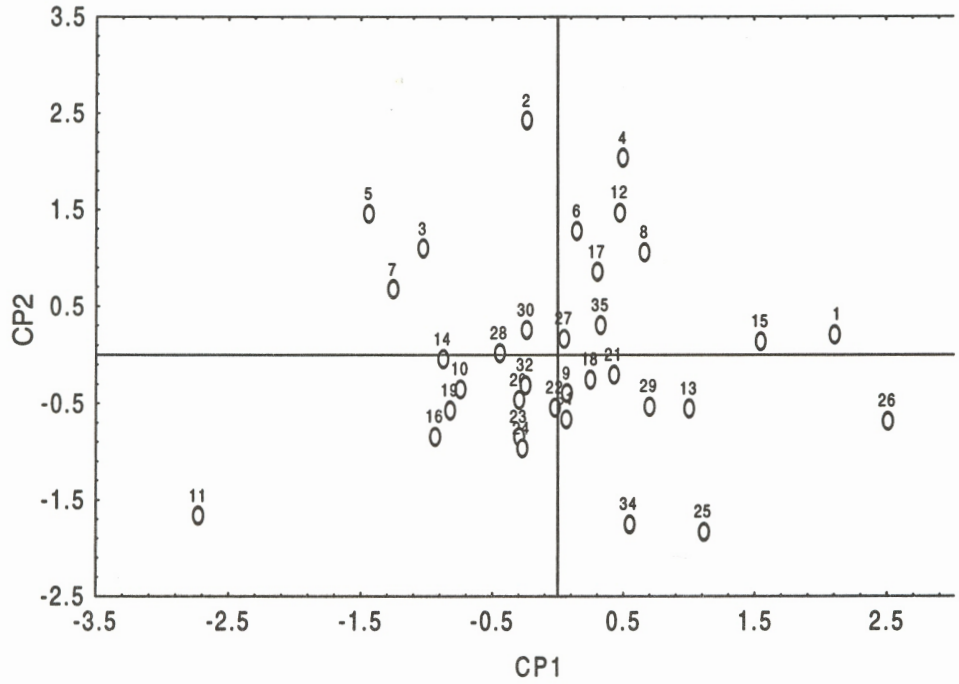


Figura 8. Diagrama de dispersión de los dos primeros componentes principales (que explic 88% de la varianza total) de la matriz de coeficientes de regresión estimados por separ incluyendo la variable CALIDAD.