Revista de la
Unión Matemática Argentina
Volumen 29, 1979.



# MEJORA FICTICIA DE LA BONDAD DE AJUSTE DEBIDO A QUE SE AJUSTA UNA FAMILIA DE MODELOS

Aldo José Viollaz

*Dedicado al Profesor Luis A. Santaló*

1. INTRODUCCION. Hasta no hace mucho tiempo era muy común que el estadístico enfocara el análisis de un conjunto de datos teniendo en mente un modelo preconcebido, es decir, elegido independientemente de los datos. En la actualidad y debido a la influencia del Análisis de Datos (Data Analysis) esta actitud está cambiando gradualmente y hoy es frecuente comenzar el análisis de un conjunto de datos teniendo en mente una familia de modelos, dentro de la cual, en base a los datos y de acuerdo a cierto criterio prefijado, se debe elegir el modelo más apropiado. Esta elección constituye un problema de decisión múltiple cuya solución frecuentemente implica ajustar cada uno de los modelos de la familia dada y calcular la medida de la concordancia de los datos con el modelo según cierto patrón fijado. El objeto de este trabajo es estudiar la mejora ficticia de la medida de concordancia de los datos con el modelo debido a que se ajusta una familia de modelos a los mismos datos. En este trabajo se considera el caso de variables aleatorias gaussianas que satisfacen a una familia de modelos de regresión con variables independientes no estocásticas ortogonales, y como medida de concordancia entre los datos y el modelo se usa el error cuadrático medio.

2. DEFINICIONES. Sean

$$X_{(n)} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} \tag{2.1}$$

$$\beta_{(n)} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} \tag{2.2}$$

donde las columnas de $X_{(n)}$ forman un conjunto de n vectores ortonormales, esto es $\sum_\sigma x_{\sigma i}\, x_{\sigma j} = \delta_{ij}$, donde $\delta_{ij}$ es la delta de Kronecker.

Sea r entero tal que $0 \leqslant r < n$ y supongamos que el vector aleatorio $Y = (Y_1, Y_2, \ldots, Y_n)'$ satisface el modelo de regresión:

$$(2.3) \qquad Y = X_{(r)}\beta_{(r)} + e \quad , \quad E(e) = 0 \quad , \quad E(ee') = \sigma^2 I$$

donde $X_{(r)}$ es la matriz que se obtiene conservando las primeras r columnas de $X_{(n)}$ y $\beta_{(r)} = (\beta_1, \beta_2, \ldots, \beta_r)'$.

Para cada entero q, $r \leqslant q < n$, sean:

$$P(q) = \{(p_1, p_2, \ldots, p_s)' : p_1 = 1,\ p_2 = 2, \ldots, p_r = r < p_{r+1} < \ldots\ p_s \leqslant q\}$$

$$Q(q) = \{(p_1, p_2, \ldots, p_s)' : p_1 = 1,\ p_2 = 2, \ldots, p_r = r,\ \{p_{r+1}, \ldots, p_s\} \subset$$

$$\subset \{r+1, r+2, \ldots, n\}\ ,\ s \leqslant q\}.$$

Sea $P = (p_1, p_2, \ldots, p_s)'$ tal que $P \in P(q)$ o $P \in Q(q)$. Defina $X_{(P)}$ como la matriz cuya i-ésima columna es la columna $p_i$-ésima de $X_{(n)}$. Analogamente defina $\beta_{(P)}$ como el vector cuya i-ésima componente es la $p_i$-ésima componente de $\beta_{(n)}$. De las definiciones de $P(q)$ y $Q(q)$ y de (2.3) se sigue que el vector $(Y_1, Y_2, \ldots, Y_n)'$, para todo $P \in P(q)$ o $P \in Q(q)$, satisface al modelo

$$(2.4) \qquad Y = X_{(P)}\beta_{(P)} + e \quad , \quad E(e) = 0 \quad , \quad E(ee') = \sigma^2 I.$$

El estimador de Gauss-Markoff o estimador mínimo cuadrático de $\beta_{(P)}$ que denotamos por $\hat{\beta}_{(P)}$ satisface las ecuaciones normales

$$X'_{(P)}\, X_{(P)}\, \hat{\beta}_{(P)} = X'_{(P)}\, Y$$

Debido a la ortogonalidad de la matriz $X_{(P)}$ se sigue que

$$(2.5) \qquad \hat{\beta}_{p_i} = \sum_{j=1}^{n} x_{jp_i}\, Y_j$$

El correspondiente estimador insesgado de $\sigma^2$ está dado por

$$(2.6) \qquad \hat{\sigma}^2_{(P)} = \frac{1}{n-s} \sum_{k \in \overline{P}} Z_k$$

donde $Z_k = \left(\sum_{j=1}^{n} x_{jk}\, Y_j\right)^2$ y $\overline{P}$ es el complemento del conjunto cuyos elementos son las componentes del vector P, respecto del conjunto $\{1, 2, \ldots, n\}$.

A cada vector de índices P corresponde una ecuación de regresión y viceversa. Con P denotaremos tanto al conjunto de índices como su correspondiente ecuación de regresión.

Consideremos las familias de ecuaciones de regresión $P(q)$ y $Q(q)$ y con

sideremos el procedimiento que selecciona dentro de la familia $\mathcal{P}(q)$ (o $\mathcal{Q}(q)$) la ecuación (o ecuaciones) de regresión que minimiza $\hat{\sigma}^2_{(P)}$. Dicha ecuación (o ecuaciones) recibe el nombre de "mejor ecuación de regresión" o "ecuación de regresión óptima" y será denotada por $P_{opt}$. Entonces

$$\min_{P} \hat{\sigma}^2_{(P)} = \hat{\sigma}^2_{(P_{opt})} \tag{2.7}$$

donde P varía en $\mathcal{P}(q)$ o en $\mathcal{Q}(q)$ según sea el caso en consideración. Para cada ecuación $P \in \mathcal{P}(q)$ (o $\mathcal{Q}(q)$) $E(\hat{\sigma}^2_{(P)}) = \sigma^2$. Sin embargo $P_{opt}$ es un vector de índices aleatorios y en general

$$E(\hat{\sigma}^2_{(P_{opt})}) \neq \sigma^2 . \tag{2.8}$$

Para cada vector observado y de Y, corresponde una ecuación óptima $P_{opt(y)}$ con error cuadrático medio $\sigma^2$ el cual se estima mediante

$$\hat{\sigma}^2_{(P_{opt(y)})} .$$

Por lo tanto resulta natural definir la mejora ficticia de la bondad de ajuste inducida por el método de selección, como la razón:

$$\lambda = \frac{E(\hat{\sigma}^2_{(P_{opt})})}{\sigma^2} . \tag{2.9}$$

Sean

$$U = \min_{P \in \mathcal{P}(q)} \hat{\sigma}^2_{(P)} \tag{2.10}$$

$$V = \min_{P \in \mathcal{Q}(q)} \hat{\sigma}^2_{(P)} \tag{2.11}$$

En este trabajo consideramos solamente la mejora ficticia de la bondad de ajuste inducida por procedimientos de selección que consisten en minimizar el error cuadrático medio $\hat{\sigma}^2_{(P)}$ sobre ciertas familias de ecuaciones de regresión. En la Sección 3 se prueba que la variable V puede escribirse como la media aritmética de las primeras (n-q) componentes de cierto estadístico de orden (Teorema 3.1) y se encuentra una expresión asintótica para la mejora ficticia de la bondad de ajuste $\lambda$ para el caso en el cual P varía en $\mathcal{Q}(q)$. En la Sección 4 se encuentra una representación para U en función de un estadístico de orden (Teorema 4.1) y se encuentra una aproximación por exceso para $\lambda$ para el caso en el cual P varía en $\mathcal{P}(q)$. Además, se realiza un estudio exploratorio por el método de Montecarlo que parece indicar que la aproximación es muy buena. En la Sección 5 se aplican los resultados de las secciones 3 y 4 al estudio de la mejora ficticia de la bondad de ajuste para el caso en el

cual P varía sobre familias $P'(q)$ y $Q'(q)$ que son de uso frecuente en la aplicación práctica de procedimientos de selección de la ecuación de regresión óptima.

3. LA VARIABLE V. En esta sección supondremos que el vector de errores e definido en (2.3) tiene distribución normal con $E(e) = 0$, $E(ee') = \sigma^2 I$. Bajo este supuesto las variables aleatorias $Z_k/\sigma^2$ forman un conjunto de variables aleatorias independientes cada una con distribución $\chi^2_{(1)}$.

TEOREMA 3.1. *La variable aleatoria* V *definida por* (2.11) *puede expresarse como*

$$V = \frac{1}{n-q} \sum_{k=1}^{n-q} Z_{(k)} \tag{3.1}$$

*donde* $Z_{(1)}, Z_{(2)}, \ldots, Z_{(n-r)}$ *es el estadístico de orden correspondiente a* $Z_{r+1}, Z_{r+2}, \ldots, Z_n$.

*Demostración.* Sea $P = (p_1, p_2, \ldots, p_s)' \in Q(q)$. Entonces se tiene

$$\frac{1}{n-s} \sum_{k \in P} Z_k \geqslant \frac{1}{n-s} \sum_{k=1}^{n-s} Z_{(k)}$$

Por otra parte puesto que $Z_{(1)} \leqslant Z_{(2)} \leqslant \ldots \leqslant Z_{(n-q)}$ se tiene $\frac{1}{n-s} \sum_{k=1}^{n-s} Z_{(k)} \geqslant \frac{1}{n-q} \sum_{k=1}^{n-q} Z_{(k)}$.

Por lo tanto $\frac{1}{n-s} \sum_{k \in P} Z_k \geqslant \frac{1}{n-q} \sum_{k=1}^{n-q} Z_{(k)}$ y entonces $V \geqslant \frac{1}{n-q} \sum_{k=1}^{n-q} Z_{(k)}$.

Por otra parte, puesto que $(n-q)^{-1} \sum_{k=1}^{n-q} Z_{(k)}$ es una de las variables de la familia sobre la cual se toma el mínimo, vale la desigualdad en el otro sentido. Esto completa la demostración.

La siguiente proposición la cual es un caso particular del Corolario 4.4 de Bickel (1967) es útil pues nos da una aproximación conveniente para $E(V)$.

PROPOSICION 3.1. *Sean* $X_1, X_2, \ldots, X_m$ *variables aleatorias independientes e identicamente distribuídas con función de distribución* G *y función de densidad* g. *Supongamos que* $E(X_1^2) < \infty$ *y que existe* A *tal que* g *es monótona para* $|x| > A$. *Definamos* $V_m$ *mediante*

$$V_m = \frac{1}{m-p} \sum_{k=1}^{m-p} X_{(k)}$$

*donde* $X_{(1)}, X_{(2)}, \ldots, X_{(m)}$ *denota el estadístico de orden. Entonces si* $0 < \alpha < 1$

$$\lim_{m\to\infty,\,(m-p)/m\to\alpha} \sqrt{m}\,[E(V_m) - \frac{1}{\alpha}\int_0^{\alpha} G^{-1}(t)\,dt] = 0. \tag{3.3}$$

Sean F y f las funciones de distribución y de densidad de probabilidad de las variables $Z_k$, respectivamente. Bajo el supuesto de distribución gaussiana del vector de errores e se satisfacen todas las hipótesis de la Proposición 3.1. Identificando $(Z_{r-1}, Z_{r-2}, \ldots, Z_n)'$ con $(X_1, X_2, \ldots, X_m)$ se deduce que E(V) puede aproximarse por

$$E(V) \cong \frac{1}{\alpha}\int_0^{\alpha} F^{-1}(u)\,du \quad \text{donde} \quad \alpha = (n-q)/(n-r). \tag{3.4}$$

Puesto que

$$\sigma^2 = E(Z_k) = \int_0^1 F^{-1}(u)\,du \tag{3.5}$$

la función $\tilde{\lambda} = \tilde{\lambda}(\alpha)$ definida por

$$\tilde{\lambda}(\alpha) = \frac{1}{\alpha}\,\frac{\int_0^{\alpha} F^{-1}(u)\,du}{\int_0^1 F^{-1}(u)\,du} \tag{3.6}$$

puede tomarse como una aproximación para $\lambda$.

Mediante el cambio de variable $u = F(t)$, $\tilde{\lambda}$ puede escribirse como

$$\tilde{\lambda}(\alpha) = \frac{1}{\alpha}\,\frac{\int_0^{x} f(t)\,t\,dt}{\int_0^{\infty} f(t)\,t\,dt}, \quad x = F^{-1}(\alpha). \tag{3.7}$$

Sin pérdida de generalidad se puede suponer $\sigma^2 = 1$ de modo que

$$f(t) = (2\pi)^{-1/2}\,t^{-1/2}\exp(-t/2), \quad t \geqslant 0 \tag{3.8}$$

$$F(t) = 2\phi(t^{1/2}) - 1, \quad t \geqslant 0.$$

donde $\phi(t) = (2\pi)^{-1/2}\int_{-\infty}^{t} \exp(-u^2/2)\,du$.

Reemplazando (3.8) en (3.7) y resolviendo las integrales se obtiene

$$\tilde{\lambda}(\alpha) = \frac{1}{\alpha}\,[2\phi(\sqrt{x}) - 1 - \sqrt{\frac{2}{\pi}}\,\sqrt{x}\,\exp(-x/2)] \quad \text{donde} \quad x = F^{-1}(\alpha).$$

En la tabla 3.1 se dan los valores de $\tilde{\lambda}(\alpha)$ para $\alpha$ = 0,25 ; 0,50 ; 0,75; 0,90 ; 0,99.

| $\alpha$ | 0 | 0,25 | 0,50 | 0,75 | 0,90 | 0,99 | 1,00 |
|---|---|---|---|---|---|---|---|
| $\tilde{\lambda}(\alpha)$ | 0,000 | 0,034 | 0,103 | 0,368 | 0,624 | 0,925 | 1,00 |

Tabla 3.1

La tabla 3.1 muestra que la magnitud de la mejora ficticia de la bondad del ajuste es importante aún para valores pequeños de la razón $\beta = 1-\alpha = (q-r)/(n-r)$. Todo el análisis precedente está basado en la hipótesis de que el vector de errores e tiene distribución gaussiana. Es de esperar que resultados análogos valgan si se cambia la hipótesis de normalidad por una hipótesis más general. Sin embargo el análisis de este caso es más difícil pues las variables $Z_k$ no son ya independientes e idénticamente distribuídas, según se deduce de una conocida caracterización de la distribución normal (ver Feller (1966), pág. 77).

4. LA VARIABLE U. En esta sección estudiamos $\lambda$ para el caso de la familia $P(q)$. Supondremos que el vector de errores e tiene distribución normal con parámetros $E(e) = 0$, $E(ee') = \sigma^2 I$.

TEOREMA 4.1. *La variable* U *definida por* (2.10) *puede representarse como:*

$$U = \min_{0\le s\le q-r} \frac{1}{n-s}\Big[\sum_{k=q+1}^{n} Z_k + \sum_{k=1}^{q-s} Z_{(k)}\Big] \tag{4.1}$$

*donde* $Z_{(1)}, Z_{(2)}, \ldots, Z_{(q-r)}$ *es el estadístico de orden correspondiente a* $Z_{r+1}, Z_{r+2}, \ldots, Z_q$.

*Demostración.* Sea $P \in P(q)$, $A(P) = \{r+1, r+2, \ldots, q\} - \{p_{r+1}, p_{r+2}, \ldots, p_s\}$. Entonces

$$\hat{\sigma}^2_{(P)} = \frac{1}{n-s}\Big[\sum_{k=q+1}^{n} Z_k + \sum_{k\in A(P)} Z_k\Big] \le \frac{1}{n-s}\Big[\sum_{k=q+1}^{n} Z_k + \sum_{k=1}^{q-s} Z_{(k)}\Big] \tag{4.2}$$

y por lo tanto

$$U \ge \min_{s} \frac{1}{n-s}\Big[\sum_{k=q+1}^{n} Z_k + \sum_{k=1}^{q-s} Z_{(k)}\Big]. \tag{4.3}$$

Por otra parte el conjunto de vectores P sobre el cual se calcula el mínimo en el segundo miembro de (4.1) es un subconjunto de $P(q)$. Por lo tanto

$$U \le \min_{s} \frac{1}{n-s}\Big[\sum_{k=q+1}^{n} Z_k + \sum_{k=1}^{q-s} Z_{(k)}\Big]. \tag{4.4}$$

De (4.3) y (4.4) se sigue la conclusión del teorema.

COROLARIO 4.1. *Si* $E|X_{(k)}| < \infty$, $k = 1,2,\ldots,n$, *entonces*

$$E(U) \leq \min_{0\leq s\leq q-r} \frac{1}{n-s}\left[(n-q)\sigma^2 + \sum_{k=1}^{q-s} E(Z_{(k)})\right].$$

*Demostración.* La conclusión sigue de inmediato de (4.4).

Supongamos que el vector de errores e tiene distribución normal con parámetros $E(e) = 0$, $E(ee') = \sigma^2 I$. Sin pérdida de generalidad podemos suponer que $\sigma^2 = 1$. Entonces $Z_k$, $k = 1,2,\ldots,n$, tiene distribución $\chi^2_{(1)}$ cuya función de distribución denotamos por F. Una aproximación conveniente para la cota del Corolario 4.1 puede obtenerse reemplazando $\frac{1}{n-s}\sum_{k=1}^{q-s} E(Z_{(k)})$ por la aproximación, proporcionada por el Teorema 3.1, que sigue:

$$\frac{q-r}{n-s}\int_0^{(q-s)/(q-r)} F^{-1}(u)\,du$$

Así se obtiene la siguiente estimación $K(\beta)$ de la cota del Corolario 4.1

$$K(\beta) = \inf_x \frac{(1-\beta) + \beta\int_0^x F^{-1}(u)\,du}{(1-\beta) + x\,\beta}$$

donde $\beta = 1-\alpha = (q-r)/(n-r)$.

La Tabla 4.1 da los valores de $K(\beta)$ para $\beta = 0{,}1;\ 0{,}25;\ 0{,}50;\ 0{,}75;\ 1$.

| $\beta$ | 1,00 | 0,75 | 0,50 | 0,25 | 0,10 |
|---|---|---|---|---|---|
| $K(\beta)$ | 0,00 | 0,46 | 0,69 | 0,86 | 0,95 |

Tabla 4.1

A fin de tener alguna información acerca de la proximidad de la función $K(\beta)$ a la función $\lambda = E(U)/\sigma^2$ se ha estimado $\lambda$ por el método Monte Carlo tomando 100 muestras de tamaño n = 40. Los resultados figuran en la Tabla 4.2.

| $\beta$ | 1,00 | 0,75 | 0,50 | 0,25 | 0,00 |
|---|---|---|---|---|---|
| $\lambda$ | 0,00 | 0,44 | 0,67 | 0,83 | 1,00 |

Tabla 4.2

La comparación de las tablas 4.1 y 4.2 muestra que la función $K(\beta)$ es una buena aproximación para $\lambda$.

Del examen de la Tabla 4.1 se concluye que la magnitud de la mejora ficticia de la bondad de ajuste es importante aunque sensiblemente menor que en el caso discutido en Sección 3.

Es importante hacer notar que las funciones $\lambda(\alpha)$ y $K(\beta)$ dependen de n,

r, q sólo a través de $\beta = (q-r)/(n-r)$.

## 5. APLICACION AL ANALISIS DEL PROBLEMA DE SELECCION DE LA MEJOR ECUACION DE REGRESION.

El problema conocido con el nombre de "Selección de la mejor ecuación de regresión" es un problema que admite más de una solución. El concepto: *mejor ecuación de regresión* tiene en general significados diferentes, dependiendo del problema en consideración. Por otra parte la introducción de dicho concepto constituye en el mejor de los casos una simplificación del problema a fin de facilitar su solución. El juicio personal aplicado al análisis del problema en consideración es un ingrediente importante en la solución de todo problema de selección de la mejor ecuación de regresión. En consecuencia no existe un único procedimiento estadístico para resolver el problema. Entre los procedimientos empleados podemos citar:

(1) Elegir la ecuación de regresión que minimiza el estadístico $C_p$ de Mallows (ver Daniel C. (1971), Cap. 6),

(2) Elegir la ecuación de regresión que minimiza el error cuadrático medio estimado,

(3) Procedimientos de selección de las variables regresoras basados en tests F individuales,

(4) Selección de las variables regresoras en base al coeficiente de correlación múltiple $R^2$.

En la solución de un problema dado, además de la aplicación de alguno de los procedimientos arriba descriptos o de algún otro, se tienen en cuenta otros factores, pero no es el momento de considerarlos aquí.

Los procedimientos conocidos con los nombres de: Eliminación progresiva, Ampliación progresiva y Regresión en etapas, pertenecen a los procedimientos definidos por (3).

El coeficiente de correlación múltiple $R^2$ debe emplearse con mucha precaución por cuanto $R^2$ no tiene en cuenta el incremento del error cuadrático medio estimado, por pérdida de grados de libertad (ver Draper and Smith (1966), pág. 63 y 118). El método (2) constituye una variante de (4) sin ese defecto. En efecto, si definimos el coeficiente de correlación múltiple corregido por grados de libertad por

$$R_c^2 = 1 - \frac{\hat{\sigma}^2_{(P)}}{\frac{1}{n-1}\sum_{j=1}^{n}(Y_j-\bar{Y})^2} \tag{5.1}$$

entonces el procedimiento (2) consiste en maximizar $R_c^2$.

Los resultados de las secciones 3 y 4 se refieren a procedimientos del tipo (2) los cuales, en esta sección serán extendidos a familias P'(q)

y $Q'(q)$ que son de uso frecuente en la aplicación de métodos de selección de la mejor ecuación de regresión.

En el caso del procedimiento (1) como la mejor ecuación de regresión se obtiene minimizando $C_p$, obviamente se produce también una mejora aparente de la bondad de ajuste de la ecuación óptima. El autor tiene en progreso un trabajo al respecto.

Un análisis cuidadoso del problema específico a resolver por regla general conducirá a la individualización de un conjunto de variables $X_1, X_2, \ldots, X_q$ que contiene todas las variables relevantes. De manera que la hipótesis de que existe un conjunto de r variables que satisfacen el modelo de regresión (2.3) es plausible. Más aún, esta hipótesis es con frecuencia asumida explicitamente en el uso de los métodos de selección de la mejor ecuación de regresión. (ver Daniel C. (1971), pág. 83).

Las familias de ecuaciones de regresión definidas por $P(q)$ y $Q(q)$, dentro de las cuales se busca la ecuación de regresión óptima, no son familias empleadas en la aplicación práctica de los métodos de selección de la mejor ecuación de regresión y su elección ha sido una cuestión de conveniencia. Sin embargo, los resultados obtenidos en las secciones 3 y 4 son esencialmente válidos para familias empleadas en la práctica y que definimos en lo que sigue.

La situación que se presenta con mayor frecuencia es aquélla en la cual después de un análisis cuidadoso del problema se ha determinado un conjunto de q variables potencialmente importantes: $X_1, X_2, \ldots, X_q$, y la familia en consideración consiste de todas las ecuaciones de regresión que se pueden construir usando un número cualquiera de estas variables, con la condición de que todas contengan un término constante, que escribiremos en la forma $\beta_1 X_1$ con $X_1 = 1$. Esta familia será denotada por $P'(q)$. Luego,

$$(5.2) \qquad P'(q) = \{(p_1, p_2, \ldots, p_s)' : 1 = p_1 < p_2 < \ldots < p_s \leq q\}.$$

Otra situación que puede presentarse en la práctica es aquélla en la cual se tiene un conjunto de n variables potencialmente importantes $X_1, X_2, \ldots, X_n$, dentro de las cuales se desea elegir la mejor ecuación de regresión con las restricciones de que $X_1$ esté siempre presente y no se usen más de q variables, $q < n$. Denotaremos a esta familia por $Q'(q)$. Luego

$$(5.3) \qquad Q'(q) = \{(p_1, p_2, \ldots, p_s)' : 1 = p_1,\ 2 \leq p_i \leq n,\ i = 2, \ldots, s,\ s \leq q\}.$$

Obviamente se tiene $P(q) \subset P'(q)$ y $Q(q) \subset Q'(q)$. Por lo tanto si $\lambda'$ se define por

$$(5.4) \qquad \lambda' = \frac{1}{\sigma^2} \cdot E\{\min_{P} \hat{\sigma}^2_{(P)}\}$$

donde P varía en $P'(q)$ o en $Q'(q)$, según sea el caso en consideración, se tiene

$$\lambda' \leqslant \lambda.$$

Puesto que $P \in P'(q)-P(q)$ (o $P \in Q'(q)-Q(q)$) implica que se ha omitido por lo menos una variable relevante, $\hat{\sigma}^2_{(P)}$ será una variable (estrictamente) estocásticamente mayor que $\hat{\sigma}^2_{(Q)}$ para todo $Q \in P(q)$ (o $Q \in Q(q)$). De aquí se deduce que en general el número $\lambda$ debe ser una buena aproximación (por exceso) para $\lambda'$. En consecuencia los resultados de las secciones 3 y 4 son válidos para el análisis del procedimiento de elección de la ecuación de regresión óptima sobre las familias $P'(q)$ y $Q'(q)$. La mejora aparente para estos casos es mayor que la que muestran las tablas 3.1 y 4.1. Aún en aquellos casos en que $\lambda$ no sea una buena aproximación para $\lambda'$, los resultados de las secciones 3 y 4 son útiles pues proporcionan cotas para el mínimo de la mejora ficticia, es decir, la mejora ficticia es por lo menos de la magnitud de dichas cotas.

Debe tenerse presente que todo el análisis precedente es válido bajo el supuesto de que el vector $(Y_1,Y_2,\ldots,Y_n)'$ tiene distribución normal y las columnas de $X_{(n)}$ forman un conjunto de n vectores ortonormales. En el caso de la familia $P'(q)$ el supuesto de ortonormalidad es requerido para las primeras q columnas de $X_{(n)}$. Las variables $X_{q+1},\ldots,X_n$ no son regresores sino que representan una partición ortogonal de la suma de cuadrados residual correspondiente al modelo que incluye las variables $X_1,X_2,\ldots,X_q$.

6. AGRADECIMIENTOS. El autor desea expresar su agradecimiento a la Lic. Zulema Cardozo quien ha realizado la simulación cuyo resultado se incluye en la Sección 4.

REFERENCIAS

[1] ANDERSON T. W. (1962), *The choice of the degree of a polynomial regresion as a multiple decision problem*, Ann. Math. Statist., Vol. 33 p. 255-265.

[2] DRAPER N. R. and SMITH H. (1966), *Applied regresion analysis*, John Wiley and Sons.

[3] FELLER W. (1966), *An introduction to Probability Theory and its applications*, Vol. II, John Wiley and Sons.

[4] GORMAN J. W. and TOMAN R. L. (1966), *Selection of variables for fitting equations to data*, Technometrics Vol. 8, p. 27-51.

[5] SCHEFFE H. (1959), *The analysis of variance*, John Wiley and Sons.

[6] BICKEL P. (1967), *Some contributions to the theory of order statistics*, Fifth Berkeley Symposium, p. 575-591.

[7] DANIEL C. and WOOD F. (1971), *Fitting Equations to Data*. Wiley-Interscience.

Instituto de Matemática
Facultad de C. Exactas y Tec.
Universidad Nacional de Tucumán
(4000) San Miguel de Tucumán
Argentina.

Recibido en julio de 1977.
Versión final diciembre de 1977.