

Universidad Nacional de Ingeniería

PROGRAMA ACADÉMICO ESCUELA DE GRADUADOS



**“ANÁLISIS Y SELECCIÓN DE VARIABLES EN
REGRESIÓN LINEAL”**

T E S I S

Para optar el Grado de Magister en Ciencias, Mención en:

MATEMÁTICAS APLICADAS

Presentada por

CARLOS A. BAZAN CABANILLAS

LIMA • PERU • 1980

I N T R O D U C C I O N

La técnica de la Regresión es, sin lugar a dudas , uno de los puntos de mayor aplicación en Estadística, sobre todo la regresión de tipo lineal, ya que, en problemas lineales es más viable llegar a resultados tangibles con menor esfuerzo. Sin embargo, el problema de la regresión lineal continúa siendo un problema abierto, siendo la selección óptima de las variables de regresión y la multicolinealidad de los datos dos aspectos importantes en actual discusión.

Una razón fundamental para que este problema no haya sido resuelto, es que no ha sido bien definido. Parece que no hay un único problema, sino que cada uso particular que se da a la regresión define un problema, y para cada caso se hace necesaria una respuesta. Por otra parte, razones de orden económico, técnico y estadístico exigen la determinación del "mejor subconjunto de regresión". El estudio de estos aspectos se ha visto impulsado últimamente por el apoyo que brinda la ciencia de la computación.

El propósito del presente trabajo es proporcionar los conceptos fundamentales de la regresión lineal y los criterios para la selección de las variables en dicho modelo. Se estudiará el problema de la multicolinealidad de las variables de predicción y se expondrá la técnica de regresión por aristas, con el objeto de superar las inconveniencias de la multicolinealidad.

El problema de la selección de variables consta de tres pasos: 1) El uso de algoritmos que proporcione información para el análisis, por ejemplo se podría usar regresión

por pasos.

2) El uso de criterios para analizar y seleccionar las variables, por ejemplo, determinar cuando se debe detener la regresión por pasos, y

3) La estimación de los coeficientes de la ecuación de regresión, por ejemplo por eliminación Gaussiana.

En la solución de un problema concreto, la obtención de "datos buenos" presupone la homogeneidad de las varianzas, la ausencia de datos erráticos, la preferencia por la ortogonalidad de las columnas de la matriz de diseño de experimentos, etc. Un problema serio es la ausencia de ortogonalidad (multicolinealidad) lo cual origina problemas de estimación de parámetros, de selección de variables y de interpretación de resultados empíricos. Por lo tanto, se hace necesario la determinación de métodos robustos a la multicolinealidad, en este sentido la regresión por aristas trata de superar este problema.

Los dos únicos tipos de ajuste de curvas que se discuten serán; el de mínimos cuadrados, que se usa en los Capítulos Nos. 1, 2, 3 y 4; y el de aristas que se usa en el Capítulo No.5.

En el Capítulo No.1 se presenta la notación y los conceptos fundamentales; en el Cap.2 se presentan los algoritmos computacionales; el Cap.3 nos proporciona los criterios para la selección de variables; en el Cap.4 se presenta el problema de multicolinealidad; y, en el Cap.5 se ofrece la alternativa de regresión por aristas, con el objeto de superar los inconvenientes de la multicolinealidad.

Carlos Bazán C.

Lima, abril de 1978

C O N T E N I D O

		pp.
Capítulo No. 1	Notación y conceptos básicos.	1
Capítulo No. 2	Técnicas de selección de variables.	14
Capítulo No. 3	Criterios de selección	21
Capítulo No. 4	Multicolinealidad	35
Capítulo No. 5	Regresión por aristas	48
	Resumen	56
	Bibliografía	58

CAPITULO No 1

NOTACION Y CONCEPTOS BASICOS

1.1 Notación e hipótesis.

Asumamos que existe una relación lineal entre t variables: $x' = (x_1, x_2, \dots, x_t)$, una variable aleatoria de perturbación e , y una variable dependiente $y = y(x, e)$.

Si tenemos n ($n \geq t + 1$) observaciones de y , a la vez que de x , podemos escribir

$$y_i \equiv \beta_0 + \sum_{j=1}^t x_{ij} \beta_j + e_i, \quad 1 \leq i \leq n. \quad (1.1)$$

Los coeficientes β y los parámetros de la distribución de e son desconocidos, y el problema consiste en obtener sus estimaciones. En forma matricial se tiene

$$Y = X\beta + e, \quad (1.2)$$

donde :

Y es un n vector columna de observaciones de y ,

X es la $n \times (t + 1)$ matriz de diseño de experimentos definida según (1.1) Los elementos de la primera columna de X son iguales a 1.

β es un $(t + 1)$ vector columna, no conocido a priori; y

e es un n vector columna aleatorio no observado, definido según (1.1).

Asumamos también las hipótesis (1.3)

1.3.a) $E(e) = 0,$

1.3.b) $V(e) = E(e e') = \sigma^2 I_n,$

1.3.c) X es una matriz de rango $t + 1$

1.3.d) En x_1, x_2, \dots, x_t están incluidas todas las variables importantes en la determinación de y mediante (1.2), pero

también pueden estar incluidas otras no importantes, en la práctica esta suposición debe ser tomada con cautela ya que es frecuente que no se disponga de datos para alguna variable importante, este problema se discute en la sección 1.2.

1.3.e En el proceso de selección de variables se eliminan r términos, esto equivale a fijar r coeficientes β con el valor cero.

El término asociado a β_0 no será eliminado. Sin pérdida de generalidad, podemos suponer que se eliminan las r variables x_{t-r+1}, \dots, x_t . Sean $p = t + 1 - r$, los términos que se retienen,

Hagamos dos convenciones:

1- Los parámetros estadísticos y las cantidades relacionadas con los modelos restringidos de p y r términos, serán indicados por los subíndices p y r respectivamente; y los relacionados con el "modelo completo" de $t + 1$ términos no llevarán subíndice. Luego, particionando convenientemente, (1.2) se escribe

$$Y = X_p \beta_p + X_r \beta_r + e. \quad (1.4)$$

2- Debido a que en nuestro medio, la bibliografía disponible está, en su mayoría, en el idioma Inglés, para facilidad de los lectores, cuando se usan siglas éstas se mantendrán en su forma original, por ejemplo, RSS_p corresponde a "Residual sum of squares" (suma de los cuadrados de los residuales) del modelo de p términos.

Definición 1.1. Entre todos los modelos de p términos, se llamará el "mejor" modelo de tamaño p a aquel que tenga mínimo RSS_p . Se enfatiza que el adjetivo "mejor" indica estrictamente

mínimo RSS_p , y que realmente el modelo podría no ser el mejor para otro uso que se le destine. Además, la definición de mejor es aplicada solo a la muestra en uso y esto no implica que la relación se mantiene necesariamente para la población.

1.2 Consecuencias de especificar el modelo con un número de variables diferente al adecuado

Razones de orden económico, técnico e inclusive estadístico (por ejemplo, variabilidad de los parámetros estimados y de la predicción) hacen recomendable la eliminación de algunas variables. En esta sección, bajo la suposición de linealidad del modelo, se hace una revisión de las consecuencias de una especificación incorrecta del modelo, ya sea por la eliminación de variables importante y/o por la retención de variables superfluas.

Estimación de los parámetros β y σ^2

Asumamos que el modelo "completo" de $t + 1$ términos es correcto, y sea

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_p \\ \hat{\beta}_r \end{pmatrix}$$

El estimador mínimo - cuadrático de β para el modelo de $t + 1$ términos. Entonces, el rango de $X = t + 1$, se tiene que

$$\hat{\beta} = (X'X)^{-1}X'Y, \quad [1.5.a]$$

$$E(\hat{\beta}) = \beta \quad [1.5.b]$$

y

$$v(\hat{\beta}) = \sigma^2(X'X)^{-1}. \quad [1.5.c]$$

Se demuestra fácilmente que $\hat{\beta}$ es el estimador lineal inseguro de mínima varianza de β . Sea

$$\hat{Y} = X\hat{\beta}, \quad [1.6.a]$$

la predicción de Y en X , definemos el vector de residuos ξ

$$\xi = Y - \hat{Y}, \quad (1.6.b)$$

luego un estimador muy apropiado para σ^2 es

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\xi' \xi}{\text{grados de libertad de } \xi} \quad (= df) \quad (1.6.c) \\ &= \frac{(Y - \hat{Y})'(Y - \hat{Y})}{df} = \frac{Y'Y - \hat{Y}'Y}{df} \\ &= \frac{Y'(I - X(X'X)^{-1}X')Y}{df} = \frac{e'(I - X(X'X)^{-1}X')e}{df} \end{aligned}$$

Como $M \equiv I - X(X'X)^{-1}X'$ es idempotente, se tiene que su rango es igual a su traza $= n - t - 1 = df$ y según (1.3.a) y (1.3.b)

$$E(\hat{\sigma}^2) = \frac{\text{traza } M}{n - t - 1} = \sigma^2 \quad (1.6.d)$$

por lo tanto, $\hat{\sigma}^2$ es insesgado y si $e \rightarrow N(0, \sigma^2)$, el corolario 4.7.1, Graybill[1961]* nos indica que $\frac{(n - t - 1)\hat{\sigma}^2}{\sigma^2} \rightarrow \chi_{n-t-1}^2$.

En el modelo restringido, si llamamos $\tilde{\beta}_p$ al estimador mínimo-cuadrático de β_p , se tiene

$$\tilde{\beta}_p = (X_p'X_p)^{-1}X_p'Y, \quad (1.7.a)$$

$$E(\tilde{\beta}_p) = \beta_p + A\beta_r. \quad (1.7.b)$$

donde

$$A = (X_p'X_p)^{-1}X_p'X_r \quad (1.8)$$

$$V(\tilde{\beta}_p) = \sigma^2(X_p'X_p)^{-1}. \quad (1.7.c)$$

Análogamente al caso anterior, un estimador de σ^2 es

$$\tilde{\sigma}^2 = \frac{(Y - \tilde{Y})'(Y - \tilde{Y})}{df} = \frac{Y'(I - X_p(X_p'X_p)^{-1}X_p')Y}{df} \quad (1.9.a)$$

* Si Y tiene distribución $N(\mu, \sigma^2 I)$, entonces $\frac{Y'AY}{\sigma^2}$ tiene distribución $\chi^2(K, \lambda)$, donde $\lambda = \frac{\mu' A \mu}{2\sigma^2}$, si y solo si A es una matriz idempotente de rango K . Además $E(\chi^2(K, \lambda)) = K + 2\lambda$.

y, si $e \rightarrow N(0, \sigma^2 I)$, por el corolario 4.7.1 de Graybill(1961) se tiene que

$$\frac{\tilde{\sigma}^2(n-p)}{\sigma^2} \longrightarrow \chi^2(n-p, \lambda) \text{ con } \lambda = \frac{\beta_r' X_r' [I - X_p (X_p' X_p)^{-1} X_p'] X_r \beta_r}{2\sigma^2}$$

Por lo tanto

$$E(\tilde{\sigma}^2) = \sigma^2 + \frac{\beta_r' X_r' [I - X_p (X_p' X_p)^{-1} X_p'] X_r \beta_r}{n-p} \quad (1.9.b)$$

El estadístico que mide la variabilidad de $\tilde{\beta}_p$ con respecto a su valor esperado $E(\tilde{\beta}_p)$ es $V(\tilde{\beta}_p)$ y el estadístico que mide la variabilidad de $\tilde{\beta}_p$ con respecto a su valor real esperado β_p es la media de los cuadrados de los errores (mean square error (MSE($\tilde{\beta}_p$))).

$$\begin{aligned} \text{MSE}(\tilde{\beta}_p) &= E[(\tilde{\beta}_p - \beta_p)(\tilde{\beta}_p - \beta_p)'] = E[(\tilde{\beta}_p - E(\tilde{\beta}_p) + A\beta_r)(\hat{\beta}_p - E(\hat{\beta}_p) + A\beta_r)'] \\ &= E[(\hat{\beta}_p - E(\tilde{\beta}_p))(\tilde{\beta}_p - E(\tilde{\beta}_p))'] + 2E[(\tilde{\beta}_p - E(\tilde{\beta}_p))\beta_r' A'] + A\beta_r \beta_r' A' \\ &= V(\tilde{\beta}_p) + A\beta_r \beta_r' A' \end{aligned} \quad (1.10)$$

Predicción y estimación

Sea x un vector fijo de ingreso de $t+1$ componentes, sea la variable dependiente:

$$Y(x) \equiv y \quad (1.11.a)$$

Según el modelo lineal se tiene

$$E(y) \equiv \bar{y} = x'\beta \quad (1.11.b)$$

la estimación \hat{y} de \bar{y} , la esperanza y la varianza de predicción de \hat{y} están dadas por

$$\hat{y} = x'\hat{\beta} \quad (1.11.c)$$

$$E(\hat{y}) = x'\beta \quad (1.11.d)$$

y

$$V(\hat{y}) = x'(X'X)^{-1}x\sigma^2 \quad (1.11.e)$$

Para la predicción simple, se tiene

$$\hat{Y}(x) \equiv \hat{y} \quad (1.12.a)$$

como $y = \bar{y} + e$,

$$\hat{y} = \hat{\bar{y}} + \hat{e} = \hat{\bar{y}} + \hat{e}, \quad (1.12.b)$$

con $\hat{\bar{y}}$ y \hat{e} independientes, luego

$$E(\hat{y}) = x' \beta \quad (1.12.c)$$

y

$$VP(\hat{y}) = \sigma^2(1 + x'(X'X)^{-1}x). \quad (1.12.d)$$

Análogamente, en el modelo restringido, para un x_p de p variables

$$Y_p(x_p) = y_p \quad (1.13.a)$$

y

$$E(y_p) \equiv \bar{y}_p = x_p' \beta_p. \quad (1.13.b)$$

Para una estimación \tilde{y} de \bar{y}_p , se tiene

$$\tilde{y}_p = x_p' \tilde{\beta}_p, \quad (1.13.c)$$

$$E(\tilde{y}) = x_p' \beta_p + x_p' A \beta_r \quad (1.13.d)$$

y

$$V(\tilde{y}) = \sigma^2 x_p' (X_p' X_p)^{-1} x_p. \quad (1.13.e)$$

El estadístico que mide la variabilidad de \tilde{y} con respecto a su media es $V(\tilde{y})$, y el estadístico que mide la variabilidad de \tilde{y} con respecto al valor real esperado $\bar{y} = x' \beta$, es la media del cuadrado del error de estimación.

$$MSE(\tilde{y}) = E(\tilde{y} - \bar{y})^2. \quad (1.13.f)$$

En el Teorema 1.1 se verá que

$$MSE(\tilde{y}) = V(\tilde{y}) - [x_p' A \beta_r - x_p' \beta_r]^2. \quad (1.13.g)$$

Y, para predicción simple de $Y_p(x_p) = \tilde{y}$, se tiene

$$\tilde{y} = \bar{y}_p + e, \quad (1.14.a)$$

$$E(\tilde{y}) = x_p' \beta_p + x_p' A \beta_r, \quad (1.14.b)$$

$$VP(\tilde{y}) = \sigma^2(1 + x_p'(X_p' X_p)^{-1}x_p) \quad (1.14.c)$$

y

$$MSEF(\tilde{y}) = E(\tilde{y} - y)^2. \quad (1.14.d)$$

En el Teorema 1.1 se verá que

$$MSEF(\tilde{y}) = VP(\tilde{y}) - [x_p' A \beta_r - x_p' \beta_r]^2. \quad (1.14.e)$$

Con los conceptos anteriores probaremos seis propiedades muy importantes:

Teorema 1.1.

- 1) $\tilde{\beta}$ es sesgado, excepto cuando $X_p' X_r \beta_r = 0$
 y
 $\tilde{\epsilon}^2$ es sesgado, excepto cuando $\beta_r = 0$.
- 2) $V(\hat{\beta}_p) - V(\tilde{\beta}_p)$ es semidefinida positiva. Esto indica que los estimadores de las componentes de β_p dados por $\hat{\beta}_p$ son generalmente más variables que los dados por $\tilde{\beta}_p$, ya que $V(\hat{\beta}_i) \geq V(\tilde{\beta}_i)$.
- 3) Si la matriz $V(\beta_r) - \beta_r \beta_r'$ es semidefinida positiva, se tiene que
 $V(\hat{\beta}_p) - \text{MSE}(\tilde{\beta}_p)$ es semidefinida positiva, por lo tanto,
 $V(\hat{\beta}_i) - \text{MSE}(\tilde{\beta}_{pi}) \geq 0$.
- 4)
$$\text{MSE}(\tilde{y}) = V(\tilde{y}) - (x_p' A \beta_r - x_r' \beta_r)^2 \quad (1.13.g)$$
 y

$$\text{MSEP}(\tilde{y}) = \text{VP}(\tilde{y}) - (x_p' A \beta_r - x_r' \beta_r)^2 \quad (1.14.e)$$
- 5) $\text{VP}(\hat{y}) \geq \text{VP}(\tilde{y})$.
- 6) Si, $V(\hat{\beta}_r) - \beta_r \beta_r'$ es semidefinida positiva, entonces
 $\text{VP}(\hat{y}) \geq \text{MSEP}(\tilde{y})$.

Prueba. En toda esta prueba asumiremos la notación siguiente:

$$X = (X_p, X_r),$$

entonces,

$$X'X = \begin{bmatrix} X_p' X_p & X_p' X_r \\ X_r' X_p & X_r' X_r \end{bmatrix} \equiv C \equiv \begin{bmatrix} C_{pp} & C_{pr} \\ C_{rp} & C_{rr} \end{bmatrix}. \quad (1.15.a)$$

Sea

$$(X'X)^{-1} \equiv B \equiv \begin{bmatrix} B_{pp} & B_{pr} \\ B_{rp} & B_{rr} \end{bmatrix}. \quad (1.15.b)$$

Las matrices B y C son definidas positivas. Con el objeto de su utilización posterior, hagamos el siguiente cálculo:

$$\begin{aligned}
B_{pr} \cdot B_{rr}^{-1} &= B_{pr}(C_{rr} - C_{rp}C_{pp}^{-1}C_{pr}) \\
&= B_{pr}C_{rr} - B_{pr}C_{rp}C_{pp}^{-1}C_{pr} \\
&= -B_{pp}C_{pr} - (I - B_{pp}C_{pp})C_{pp}^{-1}C_{pr} \\
&= -C_{pp}^{-1}C_{pr} \\
&= -(X_p'X_p)^{-1}X_p'X_r.
\end{aligned}$$

Según (1.8) se tiene

$$B_{pr}B_{rr}^{-1} = -A. \quad (1.15.c)$$

Parte 1. Según (1.7.b), $E(\tilde{\beta}_p) = \beta_p + (X_p'X_p)^{-1}X_p'X_r\beta_r$, luego $\tilde{\beta}_p$ es sesgado, excepto cuando $X_p'X_r\beta_r = 0$.

De (1.15.a) y (1.15.c) se obtiene

$$B_{rr}^{-1} = X_r'(I - X_p(X_p'X_p)^{-1}X_p')X_r,$$

la cual también es definida positiva, aplicando (1.9.b)

$$E(\tilde{\sigma}^2) = \sigma^2 + \frac{\beta_r'B_{rr}^{-1}\beta_r}{n-p},$$

finalmente, se tiene que $\tilde{\sigma}^2$ es sesgada, excepto cuando $\beta_r = 0$.

Parte 2. Según (1.15.a) y (1.15.b) vemos que

$$C_{pp}^{-1} = (X_p'X_p)^{-1} = B_{pp}^{-1} - B_{pr}B_{rr}^{-1}B_{rp}.$$

Como $V(\hat{\beta}) = (X'X)^{-1}\sigma^2$, se tiene que $V(\hat{\beta}_p) = B_{pp}\sigma^2$ y

$V(\tilde{\beta}_p) = (X_p'X_p)^{-1}\sigma^2$, luego:

$$V(\hat{\beta}_p) - V(\tilde{\beta}_p) = B_{pr}B_{rr}^{-1}B_{rp}\sigma^2. \quad (1.15.d)$$

Veamos que esta matriz es semidefinida positiva. Sea Z_p un vector cualquiera de p componentes, entonces

$$Z_p'B_{pr}B_{rr}^{-1}B_{rp}\sigma^2 Z_p = \sigma^2(B_{rp}Z_p)'B_{rr}^{-1}(B_{rp}Z_p) \geq 0.$$

puesto que B_{rr}^{-1} es definida positiva.

Si se toma $Z_p' = (0, \dots, 0, \underset{\substack{\uparrow \\ i}}{1}, 0, \dots, 0)$, se tiene

$$Z_p'(V[\hat{\beta}_p] - V[\tilde{\beta}_p])Z_p = V[\hat{\beta}_i] - V[\tilde{\beta}_i] \geq 0, \\ \text{luego, } V[\hat{\beta}_i] \geq V[\tilde{\beta}_i].$$

Parte 3. Según (1.10) y (1.15.d) se tiene

$$\text{MSE}(\tilde{\beta}_p) = V(\tilde{\beta}_p) + A\beta_r\beta_r'A \quad \text{y} \quad V(\tilde{\beta}_p) = V(\hat{\beta}_p) - \sigma^2 B_{pr}B_{rr}^{-1}B_{rp}, \\ \text{por lo tanto } V(\hat{\beta}_p) - \text{MSE}(\tilde{\beta}_p) = \sigma^2 B_{pr}B_{rr}^{-1}B_{rp} - A\beta_r\beta_r'A.$$

Sea Z_p un vector cualquiera de p componentes, entonces

$$Z_p'(V(\hat{\beta}_p) - \text{MSE}(\tilde{\beta}_p))Z_p = Z_p'(\sigma^2 B_{pr}B_{rr}^{-1}B_{rp} - A\beta_r\beta_r'A)Z_p$$

Según (1.15.c) se tiene

$$\begin{aligned} Z_p'(V(\hat{\beta}_p) - \text{MSE}(\tilde{\beta}_p))Z_p &= Z_p'(\sigma^2 B_{pr}B_{rr}^{-1}B_{rp} - B_{pr}B_{rr}^{-1}\beta_r\beta_r'B_{rp})Z_p \\ &= Z_p'B_{pr}B_{rr}^{-1}(B_{rr}\sigma^2 - \beta_r\beta_r')B_{rr}^{-1}B_{rp}Z_p \\ &= (B_{rr}^{-1}B_{rp}Z_p)'(V(\hat{\beta}_r) - \beta_r\beta_r')(B_{rr}^{-1}B_{rp}Z_p) \\ &\geq 0, \end{aligned}$$

ya que por hipótesis, $V(\hat{\beta}_r) - \beta_r\beta_r'$ es semidefinida positiva.

Si se toma $Z_p' = [0, \dots, 0, \underset{\substack{\uparrow \\ i}}{1}, 0, \dots, 0]$, se tiene

$$Z_p'(V(\hat{\beta}_p) - \text{MSE}(\tilde{\beta}_p))Z_p = V[\hat{\beta}_i] - \text{MSE}[\tilde{\beta}_i] \geq 0, \text{ luego} \\ V[\hat{\beta}_i] \geq \text{MSE}[\tilde{\beta}_i].$$

Parte 4. Probarémos solamente (1.14.e), la prueba de

(1.13.g) es similar.

Calculemos directamente MSEF.

$$\begin{aligned} \text{MSEF}(\hat{y}) &= E(y - \hat{y})^2 = E((y - \hat{y})(y - \hat{y})') \\ &= E((x_p'\beta_p + x_r'\beta_r + e - x_p'\tilde{\beta}_p)(x_p'\beta_p + x_r'\beta_r + e - x_p'\tilde{\beta}_p)') \\ &= E(x_p'(\beta_p - \tilde{\beta}_p)(\beta_p - \tilde{\beta}_p)'x_p + 2x_p'(\beta_p - \tilde{\beta}_p)(x_r'\beta_r + e)' + \\ &\quad + (x_r'\beta_r + e)(x_r'\beta_r + e)') \\ &= x_p'E[(\beta_p - \tilde{\beta}_p)(\beta_p - \tilde{\beta}_p)']x_p + 2x_p'E[(\beta_p - \tilde{\beta}_p)(x_r'\beta_r + e)'] \\ &\quad + 2x_p'E[(\beta_p - \tilde{\beta}_p)e'] + x_r'\beta_r\beta_r'x_r + 2x_r'\beta_rE(e) + E(ee') \end{aligned}$$

Prueba de la parte 5.

Sea $x' = (x_p', x_r')$ un vector de ingreso, entonces,
 $\hat{y} = x' \hat{\beta}$ y $\tilde{y} = x_p' \tilde{\beta}_p$, de donde

$$\begin{aligned} VP(\hat{y}) - VP(\tilde{y}) &= \sigma^2 (x' (X'X)^{-1} x - x_p' (X_p' X_p)^{-1} x_p) \\ &= \sigma^2 (x_p' B_{pr} B_{rr}^{-1} B_{rp} x_p + x_r' B_{rr} x_r + 2x_p' B_{pr} x_r) \quad (1.15.d) \end{aligned}$$

Definamos $Z' \equiv (x_p' B_{pr}; x_r')$, Z tiene $2r$ componentes; reemplazando en la expresión anterior, se tiene

$$VP(\hat{y}) - VP(\tilde{y}) = Z' \begin{pmatrix} B_{rr}^{-1} & I \\ I & B_{rr} \end{pmatrix} Z,$$

Aplicando el Lema anterior se tiene $VP(\hat{y}) \geq VP(\tilde{y})$.

Parte 6. Probemos antes un Lema.

Lema 1.2. Sea M una matriz $k \times k$ semidefinida positiva entonces

$$\bar{M} = \begin{pmatrix} M & M \\ M & M \end{pmatrix}$$

es una matriz $2k \times 2k$ semidefinida positiva.

Prueba. Sea P una matriz ortogonal que diagonaliza a M , luego

$$P' M P = D = \begin{pmatrix} d_1 & & 0 \\ & d_2 & \\ 0 & & \dots & d_k \end{pmatrix},$$

con $d_i \geq 0$ ya que M es semidefinida positiva. Sea

$$\bar{P} = \begin{pmatrix} P & 0 \\ 0 & P \end{pmatrix},$$

\bar{P} es ortogonal ya que P lo es, luego

$$\bar{P}' \bar{M} \bar{P} = \begin{pmatrix} D & D \\ D & D \end{pmatrix} = \begin{pmatrix} d_1 & & 0 & d_1 & & 0 \\ & \dots & & & \dots & \\ 0 & & d_k & 0 & & d_k \\ d_1 & & 0 & d_1 & & 0 \\ & \dots & & & \dots & \\ 0 & & d_k & 0 & & d_k \end{pmatrix},$$

esta matriz tiene los determinantes de los k primeros menores

principales mayores o iguales a cero, y los k siguientes iguales a cero, por lo tanto $\bar{P}'\bar{M}\bar{P}$ es semidefinida positiva.

Veamos que \bar{M} es semidefinida positiva. Sea Z un vector cualquiera de $2k$ componentes, Sea $V = \bar{P}'Z$, luego $Z'\bar{M}Z = V'(\bar{P}'\bar{M}\bar{P})V \geq 0$, por lo tanto \bar{M} es semidefinida positiva.

Prueba de la parte 6. Sea $x' = [x_p', x_r']$ un vector de ingreso, entonces $\hat{y} = x'\hat{\beta}$ y $\tilde{y} = x'\tilde{\beta}_p$. Según (1.15.d), (1.14.c) y (1.15.c) se tiene

$$VP(\hat{y}) - VP(\tilde{y}) = \sigma^2 [x_p' B_{pr} B_{rr}^{-1} B_{rp} x_p + x_r' B_{rr} x_r + 2x_p' B_{pr} x_r] ,$$

$$MSEP(\tilde{y}) = VP(\tilde{y}) + (x_p' A \beta_r - x_r' \beta_r)^2$$

y

$$A = -B_{pr} B_{rr}^{-1}$$

con lo cual se consigue

$$\begin{aligned} VP(\hat{y}) - MSEP(\tilde{y}) &= \sigma^2 [x_p' B_{pr} B_{rr}^{-1} B_{rp} x_p + x_r' B_{rr} x_r + 2x_p' B_{pr} x_r] \\ &\quad - (x_p' A \beta_r - x_r' \beta_r)^2 \qquad (1.15.c) \\ &= \sigma^2 [x_p' B_{pr} B_{rr}^{-1} B_{rp} x_p + x_r' B_{rr} x_r + 2x_p' B_{pr} x_r] \\ &\quad - x_p' B_{pr} B_{rr}^{-1} \beta_r \beta_r' B_{rr}^{-1} B_{rp} x_p - x_r' \beta_r \beta_r' x_r \\ &\quad - 2x_p' B_{pr} B_{rr}^{-1} \beta_r \beta_r' x_r . \end{aligned}$$

Definamos $Z' \equiv [x_p' B_{pr} B_{rr}^{-1}, x_r']$, Z tiene $2r$ componentes; reemplazando en la expresión anterior se tiene

$$VP(\hat{y}) - MSEP(\tilde{y}) = Z' \begin{bmatrix} \sigma^2 B_{rr} - \beta_r \beta_r' & \sigma^2 B_{rr} - \beta_r \beta_r' \\ \sigma^2 B_{rr} - \beta_r \beta_r' & \sigma^2 B_{rr} - \beta_r \beta_r' \end{bmatrix} Z ,$$

como $V(\hat{\beta}_r) = \sigma^2 B_{rr}$ y por hipótesis $V(\hat{\beta}_r) - \beta_r \beta_r'$ es semidefinida positiva; aplicando el lema anterior se tiene $VP(\hat{y}) \geq MSEP(\tilde{y})$, que es lo que queríamos probar.

OBSERVACIONES

- Las propiedades (2) y (5) recomiendan la eliminación de variables.
- Las propiedades (3) y (6) dan una condición bajo la cual la precisión del modelo restringido es mejor que la del modelo completo no obstante el sesgo.
- Si las variables x_r son extrañas, esto es $\beta_r = 0$, las propiedades (2) y (5) indican una pérdida de precisión en la estimación y predicción de las variables incluidas.

CAPITULO No 2*

TECNICAS DE SELECCION DE VARIABLES

En este trabajo consideraremos el problema general que se presenta al tratar de determinar las relaciones entre las variables de predicción x , y sus roles, tomadas separadas o en conjunto al describir la respuesta y . Nuestro objetivo es la selección de un subconjunto de variables de predicción para la ecuación final.

Para proporcionar información, se hace necesario considerar ajustes con diferentes combinaciones de variables de predicción. Si el número de variables de predicción t , es pequeño, se pueden considerar las 2^t combinaciones; pero si t es grande, tal posibilidad resulta antieconómica.

Describiremos los rasgos principales de algunos métodos computacionales, preferentemente de mínimos cuadrados, aunque también mencionaremos el método de regresión por aristas (ridge regression). En el capítulo No 3 se presenta una discusión de la forma adecuada de interpretar los resultados.

2.1.-Todas las regresiones posibles

Si el número t de variables de predicción no es muy grande, se puede considerar todas las 2^t regresiones. En la actualidad hay un buen número de algoritmos que tratan este problema (ver Hoeking 1976 pg. 71). Por ejemplo Morgan y Tatar trabajan con más de diez variables. La idea básica es la de realizar el cómputo en tal forma que dos conjuntos consecutivos difieran en una sola variable. Otra idea, un tanto distinta, desarrollada por Newton y Spurrell (1967) es la que

*Una idea básica de las técnicas de selección de variables se da en el Cap. No 6 de Draper (1966).

considera que la suma de los cuadrados debido a la regresión : $(\sum(\hat{Y}_i - Y_i)^2)$ es la suma de "elementos" básicos, y desarrollan un esquema para calcular estos elementos básicos, sin necesidad de evaluar todos los conjuntos.

La comparación de la eficiencia de estos algoritmos no es sencilla y las referencias no proporcionan información suficiente. Sin embargo, para el caso en que la selección se haga en base a la suma de los cuadrados de los residuales, se informa que el segundo algoritmo de Furnival parece eficiente.

2.2.- Métodos de regresión por pasos

Debido al gran volumen de cálculos en la evaluación de todas las 2^k regresiones, se han propuesto otros métodos en los que se evalúa un número menor de alternativas, ya sea por incremento o por eliminación de una variable en cada paso. A estos métodos se los llama de regresión por pasos y son la variación de dos ideas básicas llamadas selección hacia adelante (Forward selection: FS) y eliminación hacia atrás (Backward elimination: BE). Draper (1966) proporciona una exposición detallada. Veamos los lineamientos generales de estos métodos:

Selección hacia adelante: Este algoritmo empieza sin ninguna variable, y va adicionando una a una las variables, hasta que todas se incluyen o hasta que se satisface un criterio de finalización. Entre las variables hábiles para admisión, se elige la que tiene el mayor coeficiente de correlación parcial con Y dado que han sido admitidas las otras variables. Esto es, la variable i es adicionada a la ecuación de p términos si:

$$F_i = \max_j F_j = \max \left[\frac{RSS_p - RSS_{p+j}}{S_{p+j}^2} \right] > F_{in} \quad (2.1)$$

donde el subíndice $p + j$ se refiere a las cantidades calculadas cuando la variable j es adicionada al grupo de p variables. F_i sigue $\text{dist. } F(1, n - p - 1)$. La especificación del valor F_{in} determina una regla de finalización del algoritmo.

En la sección 3.4 se da un breve resumen de las reglas de finalización y los resultados de un estudio de simulación.

Eliminación hacia atrás : En este caso se empieza con la ecuación que incluye a todas las variables y luego se las va eliminando una a una. La variable que se elimina es la que tiene el menor coeficiente de correlación parcial, dado que se mantienen las otras variables. Esto es, la variable i es eliminada de la ecuación de p términos si:

$$F_i = \min_j F_j = \min_j \left[\frac{RSS_{p-j} - RSS_p}{\sigma_p^2} \right] < F_{out} \quad (2.2)$$

donde el subíndice $p-j$ se refiere a la RSS calculada, cuando la variable j es eliminada del grupo de p variables. F_i sigue $\text{dist. } F(1, n - p)$. La especificación del valor F_{out} determina una regla de finalización del algoritmo.

A partir de estas dos ideas se desarrollan otras combinaciones, siendo la más aceptada, la descrita por Efroymsen - (ES) la cual es esencialmente el proceso FS en el que se incluye la posibilidad de eliminar una variable en cada paso. Comúnmente, el proceso ES se le llama regresión por pasos.

Los procesos de regresión por pasos son criticados por no cumplir en general diversas condiciones recomendables, entre ellas:

- 1.- No se garantiza en forma absoluta la obtención del "mejor" subconjunto de p variables.
- 2.- A menudo, el orden de inclusión y eliminación de variables no es el más adecuado.

3.-Las subrutinas típicas de cómputo proporcionan solamente la regresión de un solo conjunto de p variables, el cual no es necesariamente el mejor. Además, como observara Gorman y Toman (1959) el mejor subconjunto de p variables no siempre es único. Por lo tanto, es deseable obtener información de las mejores combinaciones de p variables para tomar una decisión final. Esta última condición es importante.

La idea intuitiva de estos procesos es que, para problemas moderadamente bien definidos, el subconjunto de p variables elegido puede coincidir con el que se elige para todas las regresiones, y el volumen de cálculos es relativamente pequeño. Un proceso ideal sería aquel que para cada subconjunto de p variables de la mejor combinación y, además, otras combinaciones cercanas a la ideal, con un volumen de cálculos no muy elevado. En la siguiente sección se dan herramientas sobre este proceso.

2.3. Subconjuntos óptimos

Varios autores han desarrollado métodos que permiten determinar la bondad de las combinaciones de p variables. La idea básica es la desarrollada por Hockin y Lealie (1967), que consiste en lo siguiente: Supongamos que nuestro objetivo es obtener el mejor subconjunto de regresión que elimina a K variables; para ello, ordenemos las variables de menor a mayor - según el coeficiente de regresión parcial al eliminar una variable dado que se mantienen todas las otras. Denotemos por $Q(i, j, \dots)$ a la suma de los cuadrados de los residuales cuando las variables x_i, x_j, \dots son eliminadas, luego $Q(i) \leq Q(i, \dots)$. Supongamos que se consigue $Q(1, 2, \dots, k) \leq Q(k+1)$, por la desigualdad anterior se tiene que $Q(1, 2, \dots, K) \leq Q(K+1, \dots)$.

Por lo tanto, las k variables que nos conviene eliminar son : X_1, X_2, \dots, X_k . Si $Q(1, 2, \dots, k) > Q(k + 1)$ se evalúan otros conjuntos. La extensión de esta idea es desarrollada e incorporada en un programa llamado SELECT, cuya eficiencia se mide por el número de evaluaciones que se requiere para obtener un subconjunto óptimo que elimina a k variables, hay que considerar que este número de evaluaciones depende también de los datos particulares que se usan; una idea de la eficacia nos la da, el número de evaluaciones que se requiere para determinar la mejor combinación de variables para cada tamaño; para dos conjuntos particulares de datos de 15 y 26 variables, SELECT requirió de la evaluación de 1,465 y 3,546 subconjuntos de un total de $2^{15} = 32,768$ y $2^{26} = 678,108,864$ subconjuntos posibles respectivamente.

Otras alternativas en las que se considera la obtención de los mejores subconjuntos para cada tamaño, y otras modalidades de evaluación se mencionan en Hocking(1976).

2.4.-Métodos subóptimos

Debido a la inseguridad de los métodos de regresión por pesos y el alto volumen de cálculos que requieren los métodos de selección óptima, se desarrolla alternativas intermedias, entre ellas Barr y Goodnight(1971) en el Statistical Analysis System (SAS) Regression Program, proponen un esquema basado en el máximo crecimiento del estadístico R^2 . Este proceso es esencialmente la aplicación del proceso de Efroymsen (ES) que se desarrolla de la forma siguiente: Si se tiene la mejor combinación de $m-1$ variables se incrementa la variable que da el máximo crecimiento de R^2 , y se realiza el intercambio de una variable incluida con una excluida,

si es que esto aumenta R^2 . Este proceso continúa hasta que no haya cambio y, finalmente, se habrá obtenido la mejor combinación de m variables. Pero recordemos que la definición "mejor" se aplicaba en el contexto de mínimo RSS y no en el de máximo R^2 . Por lo tanto, la mejor combinación de SAS puede ser inferior a la mejor (en el sentido original) combinación determinada por SELECT.

No he encontrado un trabajo en el que se cotejen estos métodos; pero en Hocking (1976) se menciona que varios usuarios han reportado diferencias.

2.5. - Regresión por aristas (Ridge Regression)

Para problemas que involucren estadísticos de predicción ^{no} ortogonales, Hoerl y Kennard (1970) sugieren el estimador sesgado de "Ridge"

$$B = B(K) = [X'X + K I]^{-1} X'Y, \quad (2.3)$$

donde X ha sido estandarizado. La constante K se determina por inspección de la "traza" de arista, o sea el gráfico de $B(K)$ versus K . En el Capítulo 5, desarrollaremos con mayor detalle los ideas del método de regresión por aristas, que es uno de los temas más importantes de este trabajo.

En el contexto de esta sección, nótese que aunque la regresión por aristas no está diseñada explícitamente para la eliminación de variables, hay inherente una eliminación de las variables cuyos coeficientes en (2.3) tienden rápidamente a cero al crecer K . Hoerl y Kennard (1970) sugieren que tales variables "no pueden soportar su potencia de predicción" y deberían ser eliminadas. Con respecto a consideraciones de cómputo, la regresión por aristas es bastante eficiente ya que se puede obtener razonablemente buenos gráficos de la traza usando solo pocos

valores de K . La parte difícil, que es discutido en el Cap. 5 es la determinación de K , y cuán pequeña debe ser $\theta_1(K)$ para justificar la eliminación de x_1 . Marquandt(1974) sugiere que la eliminación de variables no es una cuestión de coeficientes iguales a cero, sino más bien que todas las variables debieran ser retenidas con influencia menor si $\theta_1(K)$ es pequeña. Esta idea, por supuesto, ignora los motivos económicos y prácticos para la eliminación de variables.

CAPITULO No 3

CRITERIOS DE SELECCION

3.1. Funciones de criterio

Como se verá en la siguiente sección, la selección del subconjunto o subconjuntos de variables de regresión apropiados depende del uso a que se destine la regresión. La evaluación de la información, en cada caso, será llevada a cabo mediante las funciones de criterio. Este término engloba a todos aquellos estadígrafos que se usan para decidir si una selección de variables es apropiada. Muchas de estas funciones son simples funciones de RSS. En esta sección se da una visión global de estos criterios.

1.-Media de los cuadrados de los residuales (Residual mean square),

$$RMS_p = \frac{RSS_p}{n-p} .$$

2.-Cuadrado del coeficiente de correlación múltiple (Squared multiple correlation coefficient),

$$R_p^2 = 1 - \frac{RSS_p}{TSS} = \frac{\tilde{\beta}'_p X'_p Y - n\bar{y}^2}{Y'Y - n\bar{y}^2}$$

donde $TSS = \sum_{i=1}^n (y_i - \bar{y})^2 .$

3.- R^2 ajustado (adjusted R^2),

$$\bar{R}_p^2 = 1 - \frac{n-1}{n-p} [1 - R_p^2] .$$

4.-Promedio de la varianza de predicción (average prediction variance),

$$J_p = \left(\frac{n+p}{n} \right) RMS_p .$$

5.-Total estandarizado de la media de cuadrados de los errores de estimación (standardized total mean squared error of estimation),

$$C_p = \frac{RSS_p}{\sigma^2} + 2p - n.$$

6.-Promedio del la media del cuadrado del error de predicción
(average prediction mean squared error),

$$S_p = \frac{RMS_p}{n-p-1}.$$

7.-Suma estandarizada de los cuadrados de los residuales (stan-
darized residual sum of squares),

$$RSS_p^* = e_p' D_p^{-1} e_p,$$

donde: $e_p = y - \tilde{y}_p$ y $D_p = \text{diag}(I - X_p(X_p'X_p)^{-1}X_p')$.

8.-Suma de los cuadrados de predicción (prediction sum of squa-
res),

$$PRESS_p = e_p' D_p^{-2} e_p.$$

Apreciamos que los seis primeros criterios son fun-
ciones simples de RSS_p . Se puede argumentar que ellos son e-
quivalentes, lo cual es cierto; pero cada criterio ofrece una
perspectiva de interpretación distinta, la que debe ser consi-
derada en forma heurística, ya que no se han determinado las
propiedades exactas de estos criterios.

3.2.-Usos de regresión

Los criterios que se usan para seleccionar el sub-
conjunto o subconjuntos de variables de regresión apropiados
dependen del uso al que se destine la regresión. Mallows (1973)
proporciona la siguiente lista de usos potenciales de la ecu-
ción de regresión:

- a) Descripción pura,
- b) Predicción simple,
- c) Extrapolación,
- d) Estimación de parámetros,

a) Control, y

f) Construcción de modelos

Los aspectos fundamentales que se consideran para cada uso son:

a) Minimizar RSS (suma de los cuadrados de residuos)

a la vez que se mantiene la mayor cantidad de variables en el modelo. Es costumbre muy generalizada en este aspecto usar R^2 , el coeficiente de correlación múltiple, en vez de RSS.

b) Minimizar MSEF (media del cuadrado del error de la predicción) según las pautas fijadas en la sección 1.2. Teorema 1.1, parte (4).

c) Minimizar MSEF con garantía de ausencia de multicolinealidad. Cuando hay multicolinealidad entre las variables de predicción, la predicción que corresponde a datos fuera del rango de ajuste suele ser ineficaz.

d) Estimadores insesgados y de mínima varianza. Sin embargo, en algunos casos se recomienda el uso de estadísticos sesgados, siempre y cuando, ellos mejoren la predicción, en especial si es una extrapolación lo que se busca.

e) Al igual que en d) se busca estimadores insesga- dos de mínima varianza. En gran número de casos, hay que considerar el costo experimental.

f) El objetivo consiste en determinar la forma o mo- delo de relación entre variables. Se busca la evaluación rápi- da de diversas ecuaciones de regresión para una elección poste- rior.

3.3 Evaluación de subconjuntos de regresión

Esta sección será dividida en cuatro partes. Las tres primeras son enfoques complementarios de la evaluación de los subconjuntos de regresión y la cuarta parte resume las recomendaciones de las tres primeras.

3.3.1. Aspectos básicos

En el capítulo No 1, Sección 2, Teorema 1.1, hemos visto que si $V(\hat{\beta}_r) - \beta_r \beta_r'$ es semidefinida positiva, es posible estimar parámetros y respuestas con menor variabilidad utilizando el subconjunto de regresión. Escribamos $V(\hat{\beta}_r) = B_{rr} \sigma^2$ donde B_{rr} es la submatriz adecuada de $B = (X'X)^{-1}$. Una forma más operativa de verificar si $V(\hat{\beta}_r) - \beta_r \beta_r'$ es semidefinida positiva (adfp) nos la da el siguiente Lema y Teorema.

Lema 3.1. Sea M una matriz $k \times k$ definida positiva, sean x e y dos vectores cualesquiera de k componentes cada uno, entonces

$$x' M x y' M y \geq x' M y y' M x$$

Prueba. Sea p una matriz ortogonal que diagonaliza a M , luego

$$P' M P = D = \begin{bmatrix} d_1^2 & & & \\ & d_2^2 & & \\ & & \ddots & \\ & & & d_k^2 \end{bmatrix},$$

con $d_i^2 > 0$, ya que M es definida positiva.

Sea $u = P'x$, $v = P'y$; por lo tanto

$$\begin{aligned} x' M x y' M y - x' M y y' M x &= u' P' M P u v' P' M P v - u' P' M P v v' P' M P u \\ &= u' D u v' D v - u' D v v' D u \\ &= \left(\sum u_i^2 d_i^2 \right) \left(\sum v_i^2 d_i^2 \right) - \left(\sum d_i^2 v_i u_i \right)^2 \end{aligned}$$

utilizando la desigualdad de Cauchy schwartz se tiene

$$x' M x y' M y \geq x' M y y' M x$$

Teorema 3.1. $V(\hat{\beta}_r) - \beta_r \beta_r'$ es semidefinida positiva si y solo si :

$$\frac{\beta_r' B_{rr}^{-1} \beta_r}{\sigma^2} \leq 1 \quad (3.1)$$

Prueba. Condición suficiente: sea x un vector cualquiera de r componentes. Definamos $Z = B_{rr} x$. Luego;

$$x'(V(\hat{\beta}_r) - \beta_r \beta_r')x = z'a_{rr}^{-1}z\sigma^2 - z'a_{rr}^{-1}\beta_r\beta_r'a_{rr}^{-1}z,$$

como B_{rr}^{-1} es definida positiva, por el lema anterior se tiene

$$\begin{aligned} x'(V(\hat{\beta}_r) - \beta_r \beta_r')x &\geq z'a_{rr}^{-1}z\sigma^2 - z'a_{rr}^{-1}\beta_r\beta_r'a_{rr}^{-1}\beta_r \\ &= z'a_{rr}^{-1}z\sigma^2\left(1 - \frac{\beta_r'a_{rr}^{-1}\beta_r}{\sigma^2}\right) \end{aligned}$$

utilizando la hipótesis y el hecho de que B_{rr}^{-1} es definida positiva

$$x'(V(\hat{\beta}_r) - \beta_r \beta_r')x \geq z'a_{rr}^{-1}z\sigma^2 \geq 0.$$

Condición necesaria: Supongamos que $V(\hat{\beta}_r) - \beta_r \beta_r'$ es semidefinida positiva, luego

$$\beta_r'a_{rr}^{-1}(a_{rr}\sigma^2 - \beta_r\beta_r')B_{rr}^{-1}\beta_r \geq 0,$$

por lo tanto, $\beta_r'a_{rr}^{-1}\beta_r\sigma^2 - (\beta_r'a_{rr}^{-1}\beta_r)^2 \geq 0$, de donde $\frac{\beta_r'a_{rr}^{-1}\beta_r}{\sigma^2} \leq 1$.

Por su puesto (3.1) no se puede evaluar ya que se desconoce β y σ^2 ; pero (3.1) se puede aproximar en términos del estadístico $F(r, n-t-1)$ asociado a la hipótesis $\beta_r = 0$, ver Draper (1960). Se espera que

$$F = \frac{\hat{\beta}_r'a_{rr}^{-1}\hat{\beta}_r}{\hat{\sigma}^2} \leq \frac{1}{r}. \quad (3.2)$$

luego, en la suposición de que el modelo es lineal y que se cumple (3.2), es razonable eliminar a x_r y obtener mejores estimados de los coeficientes β_p y de la respuesta y . Además, como estos resultados son válidos para cualquier x de ingreso, estos resultados se pueden usar para extrapolación, teniendo en cuenta las hipótesis del Capítulo 1 para el modelo.

Un criterio menos restrictivo que (3.2) es el que se sugiera en Johnston (1972) de eliminar a las variables cuyo estadístico t asociado es menor que 1. Esta condición conviene más para la predicción simple.

La condición (3.2) parece adecuada para la extrapolación moderada, pero restrictiva para la predicción. Una condi-

ción más razonable para la predicción es que $VP(\hat{y}_i) - \text{MSEP}(\tilde{y}_{pi})$ sea no negativa, promediando para los datos en uso se tiene:

Teorema 3.2.

$$(1) \quad \frac{1}{n} \sum_{i=1}^n VP(\hat{y}_i) = \frac{\sigma^2}{n} (n+t+1)$$

$$(2) \quad \frac{1}{n} \sum_{i=1}^n [VP(\hat{y}_i) - \text{MSEP}(\tilde{y}_{pi})] = \frac{r\sigma^2}{n} [1 - \beta_r' B_{rr}^{-1} \beta_r / r\sigma^2] \geq 0$$

Demostración. si se toma $X = (X_p, X_r)$, se tiene

$$(X'X) = \begin{pmatrix} X_p'X_p & X_p'X_r \\ X_r'X_p & X_r'X_r \end{pmatrix} \equiv C = \begin{pmatrix} C_{pp} & C_{pr} \\ C_{rp} & C_{rr} \end{pmatrix}$$

$$(X'X)^{-1} \equiv B = \begin{pmatrix} B_{pp} & B_{pr} \\ B_{pr} & B_{rr} \end{pmatrix}$$

Parte (1).

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n VP(\hat{y}_i) &= \frac{1}{n} \sum_{i=1}^n \sigma^2 [1 + x_i' (X'X)^{-1} x_i] \\ &= \frac{\sigma^2}{n} (n + \text{tr}z[X(X'X)^{-1}X']) \\ &= \frac{\sigma^2}{n} (n + \text{tr}z[X'X(X'X)^{-1}]) \\ &= \frac{\sigma^2}{n} (n + \text{tr}z[I_{t+1}]) \\ &= \frac{\sigma^2}{n} (n+t+1). \end{aligned}$$

Parte (2). Aplicando (1.15.e) se consigue

$$\begin{aligned} \sum_{i=1}^n [VP(\hat{y}_i) - \text{MSEP}(\tilde{y}_{pi})] &= \sum_{i=1}^n [\sigma^2 (x_{ip}' B_{pr} B_{rr}^{-1} B_{rp} x_{pi} + x_{ir}' B_{rr} x_{ri} \\ &\quad + 2x_{ip}' B_{pr} x_{ri}) - \sum_{i=1}^n (x_{ip}' \beta_r - x_{ir}' \beta_r)^2] \\ &= W - \theta. \end{aligned}$$

Como $B_{pr} B_{rr}^{-1} B_{rp} = B_{pp} - C_{pp}^{-1}$, desarrollando por separado estas sumas se tiene.

$$W = \sigma^2 \text{tr}z \left[X \begin{pmatrix} B_{pp} - C_{pp}^{-1} & B_{pr} \\ B_{rp} & B_{rr} \end{pmatrix} X' \right]$$

$$= \sigma^2 \text{tr}z \left(X'X \begin{pmatrix} B_{pp} - C_{pp}^{-1} & B_{pr} \\ B_{rp} & B_{rr} \end{pmatrix} \right)$$

$$= \sigma^2 \text{tr}z \left(X'XB - X'X \begin{pmatrix} C_{pp}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \right)$$

como $C_{pp} = X'_p X_p$, se tiene

$$W = \sigma^2 [(t+1) - \text{tr}z(I_p)]$$

$$= \sigma^2 (t+1-p) = \sigma^2 r.$$

Calculemos ahora θ .

$$\theta = \sum_{i=1}^n [x_{ip} A \beta_r \beta_r' A' x_{pi} + x_{ir} \beta_r \beta_r' x_{ri} - 2x_{ip} A \beta_r \beta_r' x_{ri}]$$

$$= \sum_{i=1}^n x_i \begin{bmatrix} A \beta_r \beta_r' A' & -A \beta_r \beta_r' \\ -\beta_r \beta_r' A' & \beta_r \beta_r' \end{bmatrix} x_i$$

como $A = -B_{pr} B_{rr}^{-1}$ se tiene

$$\theta = \text{tr}z \left(X \begin{bmatrix} B_{pr} B_{rr}^{-1} \beta_r \beta_r' B_{rr}^{-1} B_{rp} & B_{pr} B_{rr}^{-1} \beta_r \beta_r' \\ \beta_r \beta_r' B_{rr}^{-1} B_{rp} & \beta_r \beta_r' \end{bmatrix} X' \right)$$

$$= \text{tr}z \left(X'X \begin{bmatrix} B_{pr} B_{rr}^{-1} \beta_r \beta_r' B_{rr}^{-1} B_{rp} & 0 \\ 0 & \beta_r \beta_r' \end{bmatrix} \right)$$

$$= \text{tr}z \left(\begin{array}{cc} C_{pp} B_{pr} B_{rr}^{-1} \beta_r \beta_r' B_{rr}^{-1} B_{rp} + C_{pr} \beta_r \beta_r' B_{rr}^{-1} B_{rp} & 0 \\ 0 & C_{rp} B_{pr} B_{rr}^{-1} \beta_r \beta_r' + C_{rr} \beta_r \beta_r' \end{array} \right),$$

como $C_{pp} B_{pr} = -C_{pr} B_{rr}$ y $C_{rp} B_{pr} = I_r - C_{rr} B_{rr}$, se tiene

$$\theta = \text{tr}z \left(\begin{array}{cc} 0 & 0 \\ 0 & B_{rr}^{-1} \beta_r \beta_r' \end{array} \right)$$

$$= \text{tr}z(\beta_r' B_{rr}^{-1} \beta_r)$$

$$= \beta_r' B_{rr}^{-1} \beta_r.$$

Finalmente se consigue

$$\frac{1}{n} \sum_{i=1}^n (VP(\hat{y}_i) - \text{MSEP}(\hat{y}_{pi})) = \frac{\sigma^2}{n} \left(1 - \frac{\beta_r' B_{rr}^{-1} \beta_r}{r \sigma^2} \right).$$

Nótese que:

$$\sum_{i=1}^n (x'_{ip} \beta_p - x'_{ir} \beta_r)^2 = \beta_r' B_{rr}^{-1} \beta_r. \quad (3.3.a)$$

Tomando valores estimados a (3.3) se tiene:

$$F = \frac{\hat{\beta}_r' B_{rr}^{-1} \hat{\beta}_r}{r \hat{\sigma}^2} \leq 1. \quad (3.4)$$

Por lo tanto, se puede resumir que si se cumple (3.2) y el objetivo es la extrapolación o la estimación de los coeficientes β_r , se puede eliminar X_r . Por otra parte, si se cumple (3.4) y el objetivo es la predicción, se elimina X_r .

Un indicador del cambio de variabilidad de los parámetros estimados debido a la eliminación de variables es la ganancia relativa por predicción (Relative gain for prediction RGP)

$$RGP(\tilde{y}) = \frac{VP(\hat{y}) - MSE_P(\tilde{y}_p)}{VP(\hat{y})} \quad (3.5)$$

que, promediando según los datos en uso, ver el Teorema 3.2, se puede escribir

$$RGP = \frac{r(1 - \beta_r' B_{rr}^{-1} \beta_r / r \sigma^2)}{n + t + 1} \quad (3.6)$$

Por lo tanto si se tiene como objetivo la predicción de y , y el estimado de (3.6) RGP es ≥ 0 , ó en casos convenientes no es muy inferior a cero se puede eliminar X_r .

Resultados idénticos se obtienen para la respuesta media (estimación de y) con el único cambio del denominador de (3.6) por $t + 1$. Esta diferencia refleja que la variabilidad inherente del sistema supera a la variabilidad debida a la estimación de los parámetros.

3.3.2. Interpretación de C_p

Con el objeto de ilustrar los conceptos de la sección 3.2 utilizaremos el estadístico C_p , ver Mallows [1973].

Como se mencionó al final de la sección 1, otros estadísticos también pueden ilustrar la sección 3.2 en forma equivalente a C_p . Se define a C_p como el estimador de $T_p =$ total estandarizado de la media de los cuadrados de los errores de estimación de los datos X_p en uso (Standardized total mean square error of estimation for the current data X_p)

$$T_p = \frac{1}{\sigma^2} \sum_{i=1}^n \text{MSE}[\tilde{y}_i]. \text{ ver (1.13.e)}$$

Teorema 3.3.

$$T_p = \frac{E(\text{RSS}_p)}{\sigma^2} + 2p - n \quad (3.7)$$

Demostración. Teniendo en cuenta (1.13.g), (1.13.e) y (3.3.a) se consigue

$$\begin{aligned} T_p &= \frac{1}{\sigma^2} \sum_{i=1}^n [\sigma^2 x'_{ip} (X'_p X_p)^{-1} x_{pi} + [x'_{ip} A \beta_T - x'_{ir} \beta_T]^2] \\ &= \sum_{i=1}^n [x'_{ip} (X'_p X_p)^{-1} x_{pi}] + \frac{\beta'_r B_{rr}^{-1} \beta_r}{\sigma^2} \end{aligned}$$

como:

$$B_{rr}^{-1} = C_{rr} - C_{rp} C_{pp}^{-1} C_{pr}, \text{ se tiene que}$$

$$\begin{aligned} T_p &= \text{Trz}(X'_p (X'_p X_p)^{-1} X_p) + \frac{\beta'_r (C_{rr} - C_{rp} C_{pp}^{-1} C_{pr}) \beta_r}{\sigma^2} \\ &= \text{Trz}(X'_p X_p (X'_p X_p)^{-1}) + \frac{\beta'_r X'_r (I - X_p (X'_p X_p)^{-1} X'_p) X_r \beta_r}{\sigma^2} \\ &\quad + (n-p) \frac{\sigma^2}{\sigma^2} - (n-p) \end{aligned}$$

recordando (1.9.b), se consigue

$$\begin{aligned} T_p &= \text{Trz}(I_p) + \frac{(n-p) E(\text{RSS}_p)}{\sigma^2} - (n-p) \\ &= p + \frac{E(\text{RSS}_p)}{\sigma^2} + p - n \\ &= \frac{E(\text{RSS}_p)}{\sigma^2} + 2p - n \end{aligned}$$

[Cuando se considere la predicción, se usa MSEP y el total anterior suma $T_p + n$], luego C_p está dado por

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - n. \quad (3.8)$$

Mallows(1973) proporciona indicaciones para la calibración de C_p , y recomienda que se minimice a C_p y que la condición $C_p \approx p$ indica sesgo pequeño. Como

$$F = \frac{RSS_p - RSS}{r \hat{\sigma}^2}$$

se tiene que

$$\begin{aligned} F-1 &= \frac{1}{r} \left(\frac{RSS_p}{\hat{\sigma}^2} - \frac{RSS}{\hat{\sigma}^2} - r \right) \\ &= \frac{1}{n} \left(\frac{RSS_p}{\hat{\sigma}^2} - n + t + 1 - r \right) \\ &= \frac{1}{r} \left(\frac{RSS_p}{\hat{\sigma}^2} - n + p \right) \\ &= \frac{1}{r} (C_p - p) \end{aligned}$$

por lo tanto

$$C_p - p = r(F - 1) \quad (3.9)$$

Teniendo en cuenta (3.8) y (3.9) las condiciones (3.2) y (3.4) son equivalentes a :

$$p - r = 2p - t - 1 \leq C_p \leq 2p - t, \quad (3.10)$$

y

$$p - r = 2p - t - 1 \leq C_p \leq p. \quad (3.11)$$

Cuando el objetivo es la extrapolación o la estimación de parámetros y se cumple(3.10), se recomienda eliminar X_r . Para el caso de predicción se elimina X_r si se cumple (3.11). La expresión (3.11) es consistente con la recomendación

de que C_p sea pequeño o cercano a p .

La relación de C_p con RGP la obtenemos por medio de (3.5)

$$\widehat{RGP} = \frac{p - C_p}{n + t + 1} \quad (3.12)$$

Para el caso en que un C_p sea mayor que p , al eliminar X_r se produce una pérdida de exactitud. En este punto, si el objetivo es la predicción, se recomienda buscar la combinación de p elementos que minimice C_p y se elimina su X_r asociado. Para el caso de estimación de respuesta media, la expresión RGP se obtiene al eliminar n en el denominador de (3.12).

En forma análoga se deducen otras aplicaciones de C_p equivalentes a las usadas con otros estadísticos. Por ejemplo, la condición $F \leq 2$ es equivalente a $C_p \leq t + 1$.

3.3.3. Otras funciones de criterio.

\underline{RMS}_p ; Esta función de criterio, e inclusive su gráfico, son analizados muy frecuentemente. La selección de p se hace considerando tres aspectos (i) mínimo RMS_p , (ii) el valor de p tal que $RMS_p = RMS$ o (iii) el valor de p tal que el lugar geométrico del mínimo RMS_p crece bruscamente cuando p decrece. De (3.8) se tiene

$$C_p = (n-p) \left(\frac{RMS_p}{\hat{\sigma}^2} - 1 \right) + p, \quad RMS = \hat{\sigma}^2 \quad (3.13)$$

Esta fórmula indica una relación directa entre C_p y RMS_p . La relación (3.10), que indica eliminación de X_r cuando el objetivo es extrapolación y estimación de parámetros, se expresa como

$$\frac{n-t-1}{n-p} RMS \leq RMS_p \leq \frac{n-t}{n-p} RMS \quad (3.14)$$

Las condiciones (ii) y (iii) se basan en que -----

$RMS_p / RMS \approx 1$. En este caso $C_p \approx p$, pero hay que tener cuidado en que C_p magnifica la diferencia del cociente con la unidad por el factor $(n-p)$ lo cual hace muy sensible a C_p a los cambios de RMS_p / RMS . Las condiciones (ii) y (iii) se recomiendan para la eliminación de X_p cuando el objetivo es la predicción, teniendo siempre en cuenta la discusión de (3.12).

R_p^2 : este criterio probablemente es uno de los más utilizados y su gráfico es empleado para la determinación de p , a medida que p decrece el lugar geométrico del máximo R_p^2 se mantiene constante hasta un p en el que decrece bruscamente, esto indica el p adecuado. La relación entre R_p^2 y C_p es

$$C_p = (n - t - 1) \frac{1 - R_p^2}{1 - R^2} + 2p - n, \quad (3.15)$$

esto indica que pequeñas variaciones de R_p^2 son más sensibles por C_p debido al factor de amplificación $(n-t-1)$. Por ello, el criterio R_p^2 indica la eliminación de más variables que C_p . Se critica al estadístico R_p^2 porque hay casos en los que se eliminan variables importantes. También se presentan dificultades para extraer conclusiones a partir de los gráficos y, como alternativa, se sugiere a R_p^2 ajustado: \bar{R}_p^2 . \bar{R}_p^2 sugiere elegir el p para el que \bar{R}_p^2 es máximo. Su relación con RMS_p es

$$\bar{R}_p^2 = 1 - (n-1) \frac{RMS_p^2}{TSS}, \quad (TSS = Y'Y - n\bar{Y}^2) \quad (3.16)$$

y es exactamente equivalente a buscar el mínimo de RMS_p .

J_p : se obtiene al computer el total de la varianza de predicción en el conjunto de datos X_p en uso y de la estimación de σ^2 por RMS_p . Este estadístico tiene la desventaja de ignorar el sesgo (ver 1.14.c).

S_p : tienen una definición similar a C_p . S_p se cri-

gine a partir del promedio de la media del cuadrado del error de predicción de los datos X en uso, donde X y Y se distribuyen según una normal multivariante, $\text{var}[1.14.d]$, en este caso se tiene que

$$E(\text{MSEP}) = \frac{(n+1)(n-1)}{n} \frac{E(\text{RMS}_p)}{n-p-1} \quad (3.17)$$

El estadístico S_p se obtiene de eliminar el factor $\frac{(n+1)(n-1)}{n}$ y estimar $E(\text{RMS}_p)$ por RMS_p . Aceptando la hipótesis de normalidad multivariante, en predicción se sugiere utilizar la ecuación que tenga mínimo S_p y, como el promedio se toma sobre todos los datos, se puede usar el conjunto con S_p mínima para extrapolación moderada.

Para el problema de predicción y control, Lindley [1968] desarrolla una aproximación bayesiana al problema de selección de variables, incluyendo en su análisis el costo de observación de las variables X_p . En este estudio se recomienda minimizar la cantidad

$$\frac{\hat{\beta}'_r B_{rr}^{-1} \hat{\beta}_r}{n} + U_p, \quad (3.18)$$

donde U_p es el costo de observación. Anteriormente ya se ha recomendado el subconjunto que minimiza el primer sumando.

Para el problema de control de la salida alrededor de un punto y_0 , Lindley recomienda seleccionar el subconjunto de p variables que minimiza

$$\frac{\hat{\beta}'_r B_{rr}^{-1} \hat{\beta}_r}{n} + U_p + \sigma^2 \left(\frac{r}{n} + \frac{y_0^2}{\sigma^2 + \hat{\beta}'_p (X'_p X_p) \hat{\beta}_p} \right). \quad (3.19)$$

3.3.4.-Recomendaciones para la selección de variables

A continuación se da un resumen de recomendaciones para la selección de variables, según los usos de la ecuación

que se mencionan en la segunda Sección de este Capítulo.

- a) Para la descripción pura, se recomienda el subconjunto que minimiza RSS_p o que maximiza R_p^2 , eliminándose pocas variables.
- b) Para la predicción se recomienda el subconjunto que cumple con (3.4) o (3.11), o que hace cero (3.6), o que hace cero (3.12), o $RMS_p = RMS$, o que $\min.C_p$, o $\max.R_p^2$, o $\min.S_p$, o $\max.\bar{R}_p^2$.
- Para la estimación de respuesta se recomienda hacer cero la modificación de (3.6) ó (3.12), o $\max.R_p^2$, o $\max.\bar{R}_p^2$.
- c) Para la extrapolación se escoge el subconjunto que cumple (3.2), o (3.10) o (3.14), o que $\min.S_p$.
- d) Para la estimación de parámetros se recomienda que el subconjunto de variables satisfaga (3.2), o (3.10), o (3.14) o que $\max.R_p^2$ o \bar{R}_p^2 .
- e) Para el caso de control se recomienda satisfacer (3.18) o (3.19).
- f) En modelaje se recomienda ayudarse con asesoría técnica y con algoritmos de cómputo eficientes y fijarse alguno de los usos anteriores para la ecuación de regresión.

3.4. Selección de variables según los métodos de regresión por pasos

En esta sección veremos los conceptos de selección para los métodos de selección hacia adelante (FS) y eliminación hacia atrás (BE). Según se vió en el Capítulo No 2, los cálculos son secuenciales. Un punto bastante delicado es la determinación del criterio de finalización, según el uso de la ecuación de regresión mediante determinación de $F_{in} = F(x, 1, n-p-1)$ ó

$F_{out} = F(\alpha, 1, n-p)$. En líneas generales, en BE si F_{in} es pequeño o en FS si F_{out} es pequeño los modelos incluyen numerosas variables; lo contrario sucede si F_{in} ó F_{out} es grande. Se puede recomendar un proceso consistente en procesar BE o FS en su totalidad, obtener un conjunto de cada tamaño y seleccionar al final.

En Hocking(1976) se menciona que en un estudio de simulación, Bendel(1974), compara ocho reglas distintas de finalización para FS. Su estudio incluye a C_p , S_p , F univariada, F falta de ajuste (L_f)

$$L_f = \frac{RSS_p - RSS}{\hat{\sigma}^2(t+1-p)},$$

\bar{R}_p^2 y otros; y se recomienda el uso de F univariada o la prueba $S_p = S_{p+1}$ con nivel $0.1 \leq \alpha \leq 0.4$ para pocos grados de libertad. Para 40 ó mas grados de libertad, C_p y S_p son los mejores, y F univariada es bastante bueno para $.10 \leq \alpha \leq .25$. Se sugiere que el mejor test es F univariada con $\alpha = 0.15$.

Se consigue reducir p si se utilizan métodos de selección mas eficaces. Por ejemplo, si se evalúan todos o al menos los mejores conjuntos.

Para la regresión polinomial se recomienda tener precaución con la regla común de finalización, cuando la siguiente potencia no mejora la ecuación.

Finalmente, se puede obtener información sobre evaluación y verificación de una experiencia en Hocking(1976).

CAPITULO No 4

MULTICOLINEALIDAD

4.1. Aspectos generales

Para resolver el modelo de regresión lineal

$$Y = X\beta + e \quad (7.1)$$

Con $t + 1$ variables x_i independientes y una variable y dependiente por mínimos cuadrados, se requiere que el rango de la matriz de datos X sea $t + 1$, o equivalente, $|X'X| \neq 0$ (en este caso $X'X$ es definida positiva) y la estimación de β es $\hat{\beta} = (X'X)^{-1} X'Y$. Si el modelo es correcto (en el sentido de incluir a todas las variables independientes de importancia) se tiene

$E(\hat{\beta}) = \beta$, $V(\hat{\beta}) = \sigma^2 [X'X]^{-1}$ y para la predicción simple $V(\hat{y}(x)) = \sigma^2 [1 + x'(X'X)^{-1}x]$. Esto nos indica que la variabilidad de $\hat{\beta}$ y de \hat{y} está dada por la magnitud de las componentes de $[X'X]^{-1}$, en sentido más amplio por una norma de $[X'X]^{-1}$. Por lo tanto, es de importancia que las componentes de $[X'X]^{-1}$ tengan una cota razonablemente baja. Esto se consigue cuando $|X'X|$ no es cercana a cero.

Vamos, entonces con más detenimiento el det. de $X'X$. Si $|X'X| = 0$, se tiene que una o más variables independientes se pueden deducir de las otras mediante combinaciones lineales y que no se puede estimar β de (4.1). Vemos, pues, que si algunas variables independientes se pueden deducir linealmente, de las otras, esto imposibilita la deducción de los parámetros β .

Si $|X'X| \approx 0$ se tiene que las variables independientes tienen una relación casi lineal y que $V(\hat{\beta})$ y $V(\hat{y})$ pueden ser muy grandes.

Si $|X'X| \gg 0$, es el caso más favorable. Esta condición se logra cuando $X_i'X_j = 0$, $i \neq j$, o sea cuando las columnas de X son ortogonales.

A los conceptos anteriores se los estudia bajo el nombre de multicolinealidad.

Definición 4.1 : Se dice que las variables independientes X de (7.1) son completamente multicolineales si $|X'X| = 0$, y que hay falta completa de multicolinealidad cuando $X_i'X_j = 0$, $i \neq j$. Entre estos dos casos extremos, tenemos diversos grados de multicolinealidad, la cual no se refiere únicamente a relaciones de tipo lineal entre las variables independientes.

En general, la existencia de multicolinealidad en la matriz X deviene en: (1) la estimación inexacta de los coeficientes de regresión, tanto por la varianza grande como por la inestabilidad numérica de la solución; (2) la especificación incierta del modelo con respecto a la inclusión de variables; y (3) dificultad en determinar en que medida las variables independientes influyen en la variable dependiente y. Del primer punto hemos tenido una idea clara en esta sección y para los otros dos vemos primero los conceptos de regresión ortogonal y de regresión auxiliar.

4.2. Regresión ortogonal

Si las columnas de X son ortogonales dos a dos, se tiene,

$$X_1'X_2 = 0 \quad (4.2)$$

para cualquier partición X_1X_2 de X . Sin pérdida de generalidad se puede asumir que $X = (X_1, X_2)$. Luego $\hat{\beta}$ está dado por

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{bmatrix} [X_1'X_1]^{-1} X_1'Y \\ [X_2'X_2]^{-1} X_2'Y \end{bmatrix} \quad (4.3)$$

Se aprecia una independencia completa de $\hat{\beta}_1$ con $\hat{\beta}_2$ y que $\hat{\beta}_1$ se puede obtener regresionando Y con X_1 y que al agregar X_2 y regresionar Y con $X = [X_1, X_2]$, $\hat{\beta}_1$ no cambia, y lo mismo sucede para $\hat{\beta}_2$.

Cualquier discrepancia de la ortogonalidad dos a dos de las columnas de X , indica la presencia de multicolinealidad, la cual se acentúa a medida que $|X'X|$ tiende a cero.

Téngase presente que el método de mínimos cuadrados no es invalidado por la presencia de multicolinealidad en X . Este método solo falla cuando el rango de X es menor que $t + 1$; lo que sucede es que — los datos no permiten distinguir su influencia sobre la variable dependiente Y .

4.3. Efectos de la multicolinealidad en la especificación del modelo

Supongamos que el modelo de regresión [4.1] de $t + 1$ variables es el adecuado, pero que X tiene un alto grado de multicolinealidad, por lo que los estimados obtenidos son no significativos y se duda de la especificación original. Una alternativa para eliminar la multicolinealidad es la eliminación de las variables que la acentúan. Usemos el esquema de Huang(1970) para ver el efecto de esta restricción.

Supongamos que [4.1] representa el modelo adecuado pero que en realidad la ecuación se ajusta según

$$Y = \bar{X} \bar{\beta} + \bar{e}, \quad (4.4)$$

donde \bar{X} es una submatriz de X , \bar{X} tiene m columnas, $m < t+1$. Por mínimos cuadrados, se tiene que

$$\hat{\bar{\beta}} = (\bar{X}'\bar{X})^{-1} \bar{X}'Y = (\bar{X}'\bar{X})^{-1} \bar{X}'(X\beta + e) \quad (4.5)$$

luego

$$E(\hat{\bar{\beta}}) = Q\beta, \quad (4.6)$$

donde

$$Q = (\bar{X}'\bar{X})^{-1}\bar{X}'X = [(\bar{X}'\bar{X})^{-1}\bar{X}'x_1, \dots, (\bar{X}'\bar{X})^{-1}\bar{X}'x_{t+1}]. \quad (4.7)$$

Apreciamos que una columna j cualquiera de Q está formada por coeficientes de la regresión de x_j como variable dependiente con \bar{x} como variables independientes. Por lo tanto, Q_j indica como influye \bar{x} en la determinación de x_j , según el modelo

$$x_j = q_{1j}\bar{x}_1 + q_{2j}\bar{x}_2 + \dots + q_{mj}\bar{x}_m + U_j \quad (4.8)$$

A esta relación, que asocia la variable "verdadera" x_j con las variables incluidas $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_m$, se le llama regresión auxiliar. Por lo tanto, si se tiene un concepto apriori de lo que es Q , según (4.6) se pueden deducir ciertas características de $\hat{\beta}$.

Para ilustrar estas ideas, veamos un ejemplo concreto. Supongamos que \bar{X} se obtiene a partir de X por la eliminación de una columna, por ejemplo x_{t+1} . Entonces,

$$Q = [I_t, q], \quad q' = [q_1, q_2, \dots, q_t] \quad (4.9)$$

según (4.6) se tiene:

$$\begin{aligned} E[\hat{\beta}_1] &= \beta_1 + q_1\beta_{t+1}, \\ E[\hat{\beta}_2] &= \beta_2 + q_2\beta_{t+1}, \\ &\vdots \\ E[\hat{\beta}_t] &= \beta_t + q_t\beta_{t+1}. \end{aligned} \quad (4.10)$$

El producto $q_i\beta_{t+1}$ de el sesgo de $\hat{\beta}_i$; q_i indica la influencia de \bar{x}_i sobre x_{t+1} , y β_{t+1} es el inf. de x_{t+1} sobre Y , por ejemplo en la determinación de una producción P según las variables independientes: labor L y capital C , se espera la inclusión de la variable dirección gerencial D . Luego, si regresionamos P con L y C únicamente, se espera un sesgo en $\hat{\beta}_L$ y $\hat{\beta}_C$ cuyas direcciones, positiva o negativa, se deducen teniendo en cuenta que la influencia de D (β_{k+1}) sobre

p es positiva y las influencias de L y C (q_L, q_C) sobre D son positivas. Por lo tanto, en la regresión de C y L sobre P se espera signos positivos en β_C y β_L .

4.4. Detección de multicolinealidad *

En la mayoría ^{de} los modelos, los datos presentan cierto grado de multicolinealidad, y el objetivo general no es eliminarla sino reducirla. Una forma de identificar la dependencia entre dos variables es a través de sus coeficientes de correlación, los cuales están dados por los componentes de la matriz $M = X'X$, y una inspección de ella nos permite ver la correlación de las variables dos a dos. El caso de falta completa de multicolinealidad se da para datos ortogonales y se tiene $M = I$ (por lo tanto $|M| = 1$); y el otro extremo se da cuando $|M| = 0$. Para los casos intermedios, en Huang(1970) se sugiere que una condición "tolerable" de multicolinealidad entre x_i y x_j se da cuando

$$|M_{ij}| < R = \sqrt{R^2} \quad (4.11)$$

Pero, es el usuario, el que en definitiva decide $|M_{ij}|$. Nótese que este proceso sirve para analizar grupos de dos variables solamente, y que su aplicación por transitividad, para más de dos variables, ya no es confiable, en este caso se recomienda el coeficiente de correlación múltiple R_i^2 o el estadístico F_i de cada variable x_i con las otras t restantes. Un elevado valor de R_i^2 o de F_i indica un alto grado de relación de x_i con las otras variables.

* Por facilidad, en esta sección se trabajará con los datos tomados como desviaciones de su valor medio y divididos por su desviación estándar, A este proceso se le llama estandarización de los datos.

Algunos conceptos muy importantes se dan en Gunst (1977), al detectar la multicolinealidad con ayuda de valores y vectores propios. Sean λ_i y Z_i los valores y vectores unitarios propios de $M = X'X$. Los λ_i cumplen $\sum \lambda_i = t + 1$. Y, para el caso de ortogonalidad, se tiene $\lambda_i = 1$ para todo i . Formemos la matriz $Z = [Z_1, Z_2, \dots, Z_{t+1}]$. Luego, a partir de $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$, se obtiene

$$V(\hat{\beta}_j) = \sigma^2 \left(\frac{Z_{j1}^2}{\lambda_1} + \frac{Z_{j2}^2}{\lambda_2} + \dots + \frac{Z_{j,t+1}^2}{\lambda_{t+1}} \right), \quad 1 \leq j \leq t + 1, \quad (4.12)$$

lo cual indica la importancia de que λ_j sea grande. Supongamos que $\lambda_j \approx 0$, luego $X'XZ_j = \lambda_j Z_j \approx 0$, se puede esperar que $XZ_j \approx 0$, o sea

$$\sum_{i=1}^{t+1} x_i Z_{ij} \approx 0. \quad (4.13)$$

En esta combinación lineal, los coeficientes Z_{ij} con mayor valor absoluto indican que las variables correspondientes x_i están influyendo más fuertemente en la multicolinealidad.

En Gunst(1977) se presenta un modelo de una variable y , con 24 variables x_i multicolineales, se calculan los tres valores propios más pequeños $\lambda_1 \leq \lambda_2 \leq \lambda_3$ y sus vectores propios asociados Z_1, Z_2 , y Z_3 (ver tabla 4.1). Se aprecia que

$\lambda_1 = 0$, indica completa multicolinealidad entre las variables x_i cuyos $\{Z_{i1}\}$ correspondientes son grandes, en este caso x_9 hasta x_{20} .

$\lambda_2 = .060$, indica multicolinealidad fundamentalmente entre x_4, x_5, x_8 , y x_{24} .

$\lambda_3 = .121$, este λ todavía es pequeño comparado con la unidad (en ortogonalidad $\lambda_i = 1$ para todo i) y puede indicar

multicolinealidad, pero es recomendable reducir por eliminación de variables la multicolinealidad indicada por λ_1 y λ_2 y después en el modelo reducido analizar nuevamente la multicolinealidad.

4.5. Reducción de la multicolinealidad *

Una técnica muy utilizada es la imposición de restricciones lineales en los coeficientes β , tales como la eliminación de coeficientes y otras restricciones.

Supongamos que β cumple

$$P\beta = Q \quad (4.14)$$

donde P es una matriz $(t+1) \times (t+1)$ conocida y Q un $(t+1)$ vector conocido, luego el estimador $\tilde{\beta}$ de β que cumple (4.14) y minimiza los cuadrados de los residuales es

$$\tilde{\beta} = \hat{\beta} + (X'X)^{-1}P'(P(X'X)^{-1}P')^{-1}(Q - P\hat{\beta}), \quad (4.15.a)$$

con

$$E(\tilde{\beta}) = \beta + (X'X)^{-1}P' \cdot [P(X'X)^{-1}P']^{-1}(Q - P\beta) \quad (4.15.b)$$

y

$$V(\tilde{\beta}) = \sigma^2(X'X)^{-1} - \sigma^2(X'X)^{-1}P'[P(X'X)^{-1}P']^{-1}P(X'X)^{-1} \quad (4.15.c)$$

Si $P\beta = Q$, $\tilde{\beta}$ es insesgado y de (4.5) vemos que $V(\tilde{\beta}_1) \leq V(\hat{\beta}_1)$,

cuando la restricción es aproximadamente cierta $\tilde{\beta}$ es sesgado y

la medida de variabilidad que nos interesa es $MSE(\tilde{\beta})$

$$\begin{aligned} MSE(\tilde{\beta}) &= E(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)' \\ &= \sigma^2 [I - (X'X)^{-1}P'[P(X'X)^{-1}P']^{-1}P(X'X)^{-1} \quad (4.16) \\ &+ (X'X)^{-1}P'[P(X'X)^{-1}P']^{-1}(Q - P\beta)(Q - P\beta)'[P(X'X)^{-1}P']^{-1}P(X'X)^{-1} \end{aligned}$$

Si la discrepancia de la restricción $P\beta = Q$ no es significativa, a menudo se tiene que $MSE(\tilde{\beta}_1) \leq V(\hat{\beta}_1)$. Para ilustración consideremos el modelo

$$y = \beta_1 x_1 + \beta_2 x_2 + e,$$

cuyos vectores de observación de x_1 y x_2 son X_1 y X_2 respectivamente.

* En esta sección se trabaja con datos estandarizados.

T A B L A 4.1

Vectores propios de $X'X$ correspondientes a los tres valores propios más pequeños.

Variable	$\lambda_1 = 0$ Z_1	$\lambda_2 = .060$ Z_2	$\lambda_3 = .121$ Z_3
x_1	0.	.003	-.053
x_2	0.	.120	.274
x_3	0.	.105	-.631
x_4	0.	-.337*	.204
x_5	0.	.487*	.115
x_6	0.	.0768	.153
x_7	0.	-.093	-.291
x_8	0.	.425*	.321
x_9	.497	-.005	.091
x_{10}	.300	.022	-.035
x_{11}	.270	-.001	-.025
x_{12}	.348	.015	-.019
x_{13}	.325	-.010	-.057
x_{14}	.139	.022	-.034
x_{15}	.159	.019	.074
x_{16}	.195	-.009	-.107
x_{17}	.139	-.034	.053
x_{18}	.348	.038	.086
x_{19}	.270	-.038	-.077
x_{20}	.270	-.011	-.004
x_{21}	0.	.030	-.156
x_{22}	0.	-.135	.157
x_{23}	0.	.015	-.355
x_{24}	0.	-.631*	.153

Luego:

$$X = (X_1, X_2) ,$$

$$X'X = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

y

$$V \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} ,$$

donde ρ es el coeficiente de correlación entre x_1 y x_2 .

Acumamos la restricción $\beta_2 = 0$ cuya notación matricial es

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \beta = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Luego

$$MS\hat{\beta} = \frac{\sigma^2}{1-\rho^2} \begin{bmatrix} 1-\rho^2 & 0 \\ 0 & 0 \end{bmatrix} + \beta_2^2 \begin{bmatrix} \rho^2 & -\rho \\ -\rho & 1 \end{bmatrix} \quad (4.17.a)$$

Comparamos

$$V(\hat{\beta}_1) = \frac{\sigma^2}{1-\rho^2}$$

con

$$MS(\hat{\beta}_1) = \sigma^2 + \beta_2^2 \rho^2 . \quad (4.17.b)$$

Para algunos valores de ρ y β_2 , la cantidad (4.17.b) es menor que (4.17.a).

La restricción lineal mas utilizada es la restricción cero, o sea que, a algunos β_i se les iguale a cero. Estos β_i son elegidos de acuerdo a que las columnas X_i asociadas tengan alta multicolinealidad con las otras columnas de X y poca relación con el vector y (se analizan los estadísticos F_i , R_i^2 , etc. Una forma muy apropiada de este proceso se presenta en Gunst(1977). La idea es la siguiente :

Se forma la matriz $A = (Y, X)$. Sean $\bar{\lambda}_i$, \bar{z}_i los valores y vectores propios de $A'A$, según la discusión de la sección anterior. Si existe $\bar{\lambda}_j \approx 0$, esto indica multicolinealidad. Las componentes \bar{z}_{ij} del vector propio \bar{z}_j con mayor valor indican que las columnas A_i producen multicolinealidad. Si la columna A_1 de A no está incluida entre las A_i esto nos indica que la

multicolinealidad se da principalmente entre las columnas de X , y que éstas no influyen mayormente en y , y por lo tanto se pueden eliminar. Por ejemplo Gunet(1977) presenta la tabla 4.2 (página siguiente).

$\bar{\lambda}_0 = 0$ indica multicolinealidad completa entre las columnas A_i cuyas $|\bar{z}_{i0}|$ correspondientes son grandes, en este caso x_9 hasta x_{20} y no se influencia a y , ya que $\bar{z}_{00} = 0$.

$\bar{\lambda}_1 = .059$ indica multicolinealidad entre X_4, X_5, X_8 , y X_{24} y no se influencia a y ($\bar{z}_{01} = .059$). Identicamente se puede inspeccionar $\bar{\lambda}_2$, el cual es relativamente mayor que $\bar{\lambda}_0$ y $\bar{\lambda}_1$.

El análisis anterior nos sugiere eliminar variables entre x_9 a x_{20} y entre x_4 ó x_5 ó x_8 ó x_{24} . Otros criterios estadísticos (t, F , etc) nos ayudan a elegir. Repitiendo este proceso dos veces se obtiene la tabla 4.3 con menos datos.

Para este caso $\lambda_0 = 0.100$ puede ser considerado cercano a cero, y las variables que exhiben multicolinealidad fuerte son y, X_5, X_8 , y X_{24} . Como en la multicolinealidad se incluye a y , se recomienda detener la eliminación. En esta forma se ha conseguido un modelo con pocas variables, en el que se reduce la multicolinealidad entre las X_i y, además, se manifiesta una relación fuerte de las X_i retenidas con Y .

T A B L A No 4,2

Tres valores propios de $A'A$ y sus correspondientes vectores propios.

Variable	$\bar{\lambda}_0 = .0$ \bar{z}_0	$\bar{\lambda}_1 = .059$ \bar{z}_1	$\bar{\lambda}_2 = .115$ \bar{z}_2
y	0.	.056	.274
x ₁	0.	-.010	-.127
x ₂	0.	.124	.242
x ₃	0.	.097	-.547
x ₄	0.	-.332*	.215
x ₅	0	.475*	-.050
x ₆	0.	.076	.099
x ₇	0.	-.100	-.269
x ₈	0	.419*	.285
x ₉	.497	-.001	.111
x ₁₀	.300	.019	-.051
x ₁₁	.270	-.004	-.045
x ₁₂	.348	.020	-.009
x ₁₃	.325	-.008	-.041
x ₁₄	.139	.023	-.024
x ₁₅	.139	.024	.097
x ₁₆	.195	-.010	-.098
x ₁₇	.139	-.058	.100
x ₁₈	.348	.034	.034
x ₁₉	.270	-.041	-.058
x ₂₀	.270	-.021	-.045
x ₂₁	0.	.012	-.258
x ₂₂	0.	-.125	.215
x ₂₃	0.	.001	-.412
x ₂₄	0.	-.648*	.035

T A B L A No 4.3

Los valores propios de $A'A$ con sus correspondientes vectores propios para el modelo reducido de 15 variables.

Variable	$\lambda_0 = .100$	$\lambda_1 = .166$
	Z_0	Z_1
y	-.152*	-.648
x_1	.042	.172
x_3	-.033	-.143
x_5	-.239*	-.545
x_6	-.067	.192
x_7	.000	.024
x_8	-.498*	.114
x_{10}	-.012	.118
x_{12}	-.004	.043
x_{17}	.038	-.122
x_{18}	-.015	.221
x_{20}	.070	.105
x_{21}	.001	.273
x_{22}	.033	-.127
x_{23}	.018	.179
x_{24}	.811*	.108

CAPITULO No 5

REGRESION POR ARISTAS *

(Ridge Regression)

5.1. Aspectos generales

La forma usual de estimar los parámetros β , en un proceso de ajuste de curvas, se hace por la técnica de mínimos cuadrados, la cual es buena cuando cuando $X'X$ no discrepa considerablemente de I . En este caso, los valores propios λ_i de $X'X$ no son cercanos a cero; pero, cuando $X'X$ es muy distinta de I , ocurren problemas de inflación o inestabilidad general. Hoerl(1962), sugirió perturbar la diagonal de $X'X$ con el objeto de aproximarla a I y estimar β , según

$$\begin{aligned}\hat{\beta}^* &= (X'X + KI)^{-1} X'Y \\ &= WX'Y ; \quad W = (X'X + KI)^{-1}\end{aligned}\quad (5.1.a)$$

A la estimación y el análisis construidos de acuerdo a (5.1) se le llama "Regresión por aristas", y su relación con la estimación ordinaria está dado por

$$\begin{aligned}\hat{\beta}^* &= [K(X'X)^{-1} + I^{-1}] \hat{\beta} = Z \hat{\beta} , \\ Z &= [K(X'X)^{-1} + I]^{-1} = (X'X + KI)^{-1} X'X .\end{aligned}\quad (5.1,b)$$

Propiedades más importantes de $\hat{\beta}^*$, W y Z

(i) Sean $\xi_i(W)$ y $\xi_i(Z)$ los valores propios de W y Z respectivamente, entonces

$$\xi_i(W) = \frac{1}{\lambda_i + K} \quad (5.2.a)$$

$$y \quad \xi_i(Z) = \frac{\lambda_i}{\lambda_i + K} , \quad (5.2.b)$$

* En este Capítulo se trabaja con datos estandarizados.

donde λ_i son los valores propios de $X'X$. Estas relaciones se obtienen a partir de (5.1.a) y (5.1.b).

$$(ii) \quad Z = I - K(X'X + KI)^{-1} = I - KW, \quad (5.3)$$

Este resultado se obtiene tomando la forma alternativa de $Z = WX'X$.

(iii) Para $K \neq 0$, $\hat{\beta}^*$ es mas pequeño que $\hat{\beta}$; esto es

$$(\hat{\beta}^*)' \hat{\beta}^* < \hat{\beta}' \hat{\beta} \quad (5.4)$$

Prueba. Sea $|\cdot|$ la norma Euclídea en R^n , sea $\|\cdot\|$ la norma espectral, en $M(n \times n)$; en algebra se establece que estas normas son compatibles, esto quiere decir que, para toda matriz M y todo vector X se cumple $|MX| \leq \|M\| \cdot |X|$. Z es definida positiva. Por lo tanto,

$$(\hat{\beta}^*)' \hat{\beta}^* = (Z \hat{\beta})' Z \hat{\beta} = |Z \hat{\beta}|^2 \leq \|Z\|^2 |\hat{\beta}|^2 = \xi_{\max}^2(Z) \hat{\beta}' \hat{\beta}.$$

Como todo $\xi_i(Z)$ es menor que la unidad, (5.4) se verifica fácilmente.

(iv) El estimador de la suma de los cuadrados de los residuales esta dado por

$$\begin{aligned} RSS(\hat{\beta}^*) &= (Y - X \hat{\beta}^*)' (Y - X \hat{\beta}^*) \\ &= Y'Y - (\hat{\beta}^*)' X'Y + (\hat{\beta}^*)' X'X \hat{\beta}^* - (\hat{\beta}^*)' X'Y \\ &= Y'Y - (\hat{\beta}^*)' X'Y - K(\hat{\beta}^*)' \hat{\beta}^* \quad (5.5) \\ &= (\text{suma total de cuadrados} - \text{suma de cuadrados debido a la regresión de } \hat{\beta}^* - \text{cuadrado de longitud de } \hat{\beta}^* \text{ por } K) \end{aligned}$$

Observación: De (5.2.b) y (5.3) se tiene,

$$Z(0) = I$$

y

$$\lim_{K \rightarrow \infty} Z(K) = 0 \quad (5.6)$$

5.2. La traza de aristas

5.2.1 Definición de la traza de aristas

A medida que $X'X$ se desvía de la matriz identidad I , o sea, cuando se tiene valores propios pequeños, la probabilidad de que $\hat{\beta}$ esté cerca a β puede ser muy pequeña ($V(\hat{\beta}) = \sigma^2(X'X)^{-1}$).

La inspección de la correlación entre las variables tomadas dos a dos (elementos de $X'X$) no es suficiente para determinar las relaciones existentes para más de dos factores. Los métodos computacionales actuales no permiten un conocimiento adecuado del espacio de factores y de la sensibilidad de los resultados, en cada caso particular, con excepción de los métodos^{de} componentes principales y regresión por aristas. Se obtiene información más tangible de la traza de las aristas esto es : calcular $\hat{\beta}^*(K)$ y $RSS(\hat{\beta}^*)$ para varios K .

5.2.2 Caracterización de la traza de aristas

Sea B cualquier estimador lineal de β , la suma del cuadrado de los residuales es

$$\begin{aligned} RSS(B) &= (Y - XB)'(Y - XB) \\ &= (Y - X\hat{\beta} - (XB - X\hat{\beta}))'(Y - X\hat{\beta} - (XB - X\hat{\beta})) \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (B - \hat{\beta})'X'X(B - \hat{\beta}) \\ &= RSS_{\min} + \phi(B). \end{aligned}$$

La traza de aristas se define como la B que resuelve el problema siguiente :

$$\text{Minimizar } B'B, \quad (5.8.a)$$

$$\text{Sujeto a } (B - \hat{\beta})'X'X(B - \hat{\beta}) = \phi_0.$$

Utilizando multiplicadores de Lagrange con multiplicador $\frac{1}{K}$,

(5.8.a) es equivalente a

$$\text{Minimizar } F = F(B, K) = B'B + \frac{1}{K}(B - \hat{\beta})'X'X(B - \hat{\beta}) - \theta, \quad (5.8.b)$$

cuya solución es justamente la traza de aristas

$$B = \hat{\beta}^* = (X'X + KI)^{-1}X'Y, \quad (5.8.c)$$

con K elegida de tal forma que se satisface la restricción (5.8.a)

5.3 Propiedades del error cuadrático medio de la regresión por aristas

5.3.1 Varianza y sesgo de un estimador por aristas.

Definamos :

$$L_1^2(K) = \begin{cases} \text{cuadrada de la} \\ \text{distancia de } \hat{\beta}^* \text{ a } \beta, \end{cases}$$

$$L_1^2(K) = (\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta). \quad (5.9.a)$$

Aplicando el operador esperanza se tiene :

$$\begin{aligned} E[L_1^2(K)] &= E[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)] \\ &= E[(\hat{\beta} - \beta)'Z'Z(\hat{\beta} - \beta) + (Z\beta - \beta)'(Z\beta - \beta)] \\ &= \sigma^2 \{ \text{Tr} Z(X'X)^{-1}Z' \} + \beta'(Z-1)'(Z-1)\beta, \end{aligned}$$

utilizando (5.1.b) y (5.1.a)

$$\begin{aligned} E[L_1^2(K)] &= \sigma^2 \{ \text{traza}(X'X + KI)^{-1} - K \text{ traza}(X'X + KI)^{-2} \} \\ &\quad + K^2 \beta'(X'X + KI)^{-2} \beta \end{aligned}$$

aplicando (5.2.a)

$$\begin{aligned} E[L_1^2(K)] &= \sigma^2 \sum \frac{\lambda_i}{(\lambda_i + K)^2} + K^2 \beta'(X'X + KI)^{-2} \beta \\ &= r_1(K) + r_2(K). \end{aligned} \quad (5.9.b)$$

Significado de $r_1(K)$

puesto que $\hat{\beta}^* = Z\hat{\beta} = Z(X'X)^{-1}X'Y$, se sigue que

$$V(\hat{\beta}^*) = \sigma^2 Z(X'X)^{-1}Z'. \quad (5.10)$$

Como $r_1(K) = \text{Traza } \sigma^2 Z(X'X)^{-1}Z = \text{suma de las varianzas de todas las } \hat{\beta}_i^*$, diremos que $r_1(K)$ es la "varianza total" de los estimados de los parámetros.

Significado de $r_2(K)$

Como $r_2(K) = (Z\hat{\beta} - \beta)'(Z\hat{\beta} - \beta)$ y $r_2(0) = 0$, puesto que $Z(0) = I$, $r_2(K)$ puede ser considerado como el cuadrado del sesgo al usar $\hat{\beta}^*$ en vez de $\hat{\beta}$.

5.3.2 Comportamiento comparado de $r_1(K)$ y $r_2(K)$

(i) Según (5.9.b) se tiene

$$r_1(K) = \sigma^2 \sum \frac{\lambda_i}{(\lambda_i + K)^2}$$

luego $r_1(K)$ es una función continua, monótona decreciente en K , y :

$$\lim_{K \rightarrow 0^+} r_1(K) = \sigma^2 \sum \frac{1}{\lambda_i}, \quad \lim_{K \rightarrow \infty} r_1(K) = 0. \quad [5.11.a.]$$

$$(ii) \quad \frac{d}{dK} r_1(K) = -2\sigma^2 \sum \frac{\lambda_i}{(\lambda_i + K)^3}, \quad [5.11.b.]$$

luego :

$$\lim_{K \rightarrow 0^+} \frac{d}{dK} r_1(K) = -2\sigma^2 \sum \frac{1}{\lambda_i^2}$$

(5.11.c)

y si $X'X$ deviene en singular, en este caso algún $\lambda_i \rightarrow 0$, y

$$\lim_{K \rightarrow 0^+} \frac{d}{dK} r_1(K) \rightarrow -\infty$$

(iii) Según (5.9.b), $r_2(K) = K^2 \beta'(X'X + KI)^{-2} \beta$; afirmamos que:

$r_2(K)$ es una función continua monótona creciente.

Prueba. Sea P' la matriz ortogonal que diagonaliza a W^2 ,

esto es

$$PW^2P' = 0 = \begin{bmatrix} d_1 & & & \\ & \ddots & & \\ & & 0 & \\ & & & d_{t+1} \end{bmatrix},$$

las d_i son los valores propios de W^2 , según (5.2.b) se tiene

$$d_i = \frac{1}{(\lambda_i + K)^2}. \text{ Luego,}$$

$$\begin{aligned} r_2(K) &= K^2(F\beta)'PW^2P'F\beta = K^2(F\beta)'DP\beta \\ &= K^2 \sum (F\beta)_i^2 d_i \\ &= K^2 \sum \frac{(\alpha)_i^2}{(\lambda_i + K)^2}, \end{aligned} \quad (5.12.a)$$

donde $\alpha = F\beta$. Cada elemento $\lambda_i + K$ es positivo, por lo tanto r_2 es continua, para $K \neq 0$

$$r_2(K) = \sum \frac{\alpha_i^2}{[1 + (\frac{\lambda_i}{K})]^2}; \quad \lambda_i > 0.$$

Por lo tanto, r_2 es monótona creciente.

Notemos también que

$$\lim_{K \rightarrow 0^+} r_2(K) = 0, \text{ y } \lim_{K \rightarrow \infty} r_2(K) = \alpha'\alpha = \beta'F'F\beta = \beta'\beta. \quad (5.12.b)$$

$$(iv) \quad \frac{d}{dK} r_2(K) = 2K \sum \frac{\lambda_i \alpha_i}{(\lambda_i + K)^3} \quad (5.12.c)$$

luego

$$\lim_{K \rightarrow 0^+} \frac{dr_2}{dK}(K) = 0, \text{ y } \lim_{K \rightarrow \infty} \frac{dr_2}{dK}(K) = 0. \quad (5.12.d)$$

Estos conceptos se visualizan en el gráfico 5.1.

5.3.3 Teorema de existencia de mínimo de $E[L_1^2(K)]$ para $K > 0$.

Hemos visto que $E[L_1^2(K)] = r_1(K) + r_2(K)$ y el gráfico (5.1) nos muestra un mínimo de $L_1^2(K)$.

Como $\lim_{K \rightarrow \infty} r_2(K) = \beta'\beta$, a medida que la magnitud de β crece, se puede pensar que alguna vez $E[L_1^2]$ tome su mínimo en

$\frac{d}{dK} E(L_1^2(K))$ es menor que cero siempre que $K < \frac{\sigma^2}{\alpha_{max}^2}$, luego

$$E(L_1^2(K)) < E(L_1^2(0)).$$

5.4. Recomendaciones para la selección de un $\hat{\beta}^*$ adecuado

Naturalmente el usuario es el que toma la decisión final acerca del sesgo, la varianza y la magnitud del error cuadrático medio. Pero, para su ayuda, se dan los criterios que deben normar la selección de K.

- (i) Para un determinado valor de K el sistema debe mostrarse estable y los valores propios de $X'X + KI$ deben ser mayores que cero.
- (ii) Los coeficientes de $\hat{\beta}_i^*$ deben tener un valor absoluto moderado de acuerdo a la razón de cambio de la variable asociada x_i .
- (iii) Los coeficientes que aparentemente tienen signo incorrecto, deben cambiar de signo al tomar $K = 0$.
- (iv) $RSS(\hat{\beta}^*)$ no debe ser muy grande con respecto a $RSS(\hat{\beta})$ ni con respecto a una varianza razonable acorde con el proceso de generación de datos.

R E S U M E N

Al analizar un proceso que relaciona variables independientes x_i con una variable dependiente y , los conceptos técnicos emergentes del contexto del proceso no son suficientes para explicar la relación que se busca. Es así que se plantean modelos, entre ellos los de tipo lineal, cuya discusión ha sido el objetivo mayor de este trabajo.

Con la hipótesis de que el modelo es lineal, la determinación de la magnitud de la influencia de cada x_i en la respuesta final y , en la mayoría de los casos se ve afectada considerablemente por perturbaciones, las cuales pueden surgir por la inclusión de variables extrañas y/o porque las variables de predicción consideradas guardan relación entre sí (multicolinealidad) y que tomadas en conjunto dificulten la determinación de la influencia de cada una de ellas en la respuesta final. En este sentido las recomendaciones del presente trabajo son:

- i) Determinar con claridad el uso al que se destine la regresión.
- ii) Considerando la variabilidad de los coeficientes β a estimar, la eliminación de variables poco significativas, es ventajosa, no obstante el sesgo que se introduce por eliminación.
- iii) En una primera fase, además de los estadísticos R^2 y cociente F , un criterio bastante elaborado que ayuda en la eliminación de variables es C_p .
- iv) En una segunda fase, el cálculo de los valores propios más cercanos a cero y de sus vectores propios asociados son

de gran ayuda en el análisis y en la eliminación de variables.

v) Finalmente, si la multicolinealidad de los datos persiste y ya no es recomendable la eliminación de variables, la regresión por aristas ayuda a la determinación de una mejor ecuación de ajuste .

B I B L I O G R A F I A

- Averson, J.N. y Mc Cobe, G.P.(1973). Variable selection in a regression analysis. Department of Stat. Univ. of Kentucky.
- Anderson, V.L. y Mc Lean, R.A.(1974). Design of experiments
Marcel Dekker.
- Draper, N.R y Smith, H.(1966). Applied Regression Analysis
Wiley, New York.
- Foreythe, G.E. y Moler, C.B.(1967). Computer solution of linear algebraic systems. Prentice - Hall.
- Graybill, F.A.(1961). An introduction to linear statistical models, Vol.1. Mc Graw - Hill.
- Gunst, R.F. y Mason, R.L.(1977). Advantages of examining multicollinearities in Regression Analysis.
Biometric, Marzo.
- Hocking, R.R.(1976). The analysis and selection of variables in linear regression. Biometrics 32, 1 - 48
- Hoerl, A.E. y Kennard, R.W.(1970). Ridge regression : Biased estimation for nonorthogonal problems.
Technometrics Vol.12, NO 1.
- Huang, D.S.(1970). Regression and econometric methods
Wiley, New York.
- Johnston, J.(1972) Linear Algebra. Addison Wesley.
- Mallows, C.L.(1973). Some comments on C_p . Technometrics. Vol. 15
- Marquardt, P.W. y Snee, R.D.(1973). Ridge regression.
Department of stat. Univ. of Kentucky.