UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

ESTIMACION SESGADA EN EL MODELO LINEAL

T E S I S

Que para obtener el título de:

A C T U A R I O

p r e s e n t a

GUILLERMO RAFAEL BAZ TELLEZ

México, D. F. 1976

A Po

A la memoria de mi madre

A mi padre

A Margui, Tere y Eduardo

A Anni, Daniel y Liesl

A mis amigos

Agradecimientos. Al Dr. Enrique de Alba por la paciencia que tuvo al dirigir esta tesis.
Al Dr. Ignacio Méndez por haber proporcionado material bibliográfico.
Gran parte de esta tesis fue elaborada siendo becario del CONACYT.

I N D I C E

Pág.

INTRODUCCION 1

CAPITULO I

1.1. La Distribución Normal Multivariada. 3
1.2. El Modelo Lineal. Propiedades. 11
1.3. La Matriz de Correlación. 14
1.4. El Problema de Multicolinealidad 19

CAPITULO II

2.1. El Error Cuadrático Medio y la Condición de Admisibilidad. . 25
2.2. Los Estimadores de Hoerl y Kennard 27
2.2.a) La Traza de H.K.. 34
2.2.b) Un ejemplo del método de H.K.. 39
2.3. Otros Métodos de Estimación Sesgada. 43
2.3.a) Propuestas de Estimadores Sesgados con Propiedades Análogas a $\hat{\beta}_k$ 44
2.3.a.1. Estimadores H.K. Generalizados y Estimadores Análo-- gos. 44
2.3.a.2. Estimadores de Marquardt 47
2.3.a.3. Estimadores de Componentes Principales. 53
2.3.a.4. Estimadores contraídos. 56
2.3.b) Obtención del Valor de k 60
2.3.c) Criterios de Estimación Propuestos 61

CAPITULO III

3.1. Inferencia Bayesiana. 64
3.1.a) Estimadores bayesianos 70

3.2. Justificación Bayesiana de los Estimadores H.K. 71
3.2.a) Crítica del Método de la Traza H.K. 77
3.3. Justificación Bayesiana de Otros Estimadores Sesgados . . . 80

CONCLUSIONES 89

APENDICE A 92

APENDICE B 96

APENDICE C 98

BIBLIOGRAFIA 109

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

El objetivo de esta tesis es exponer y discutir un método de estimación para los modelos lineales propuesto y desarrollado por Arthur E. Hoerl y Robert W. Kennard (15) (16), método que ellos llamaron "Ridge Regression" (traducido en el presente trabajo como "Método H.K.").

La tesis está estructurada en tal forma que pueda servir a personas que han llevado un curso de regresión como los que se imparten actualmente en la Facultad de Ciencias de la UNAM. En consecuencia se incluyen varias secciones que facilitan la comprensión de los principales aspectos de este trabajo.- Tales secciones son las: 1,1, 1.3, 1.4 y 3.1. Frecuentemente se hace referencia al apéndice A, en donde incluímos varios resultados de álgebra lineal.

El capítulo I contiene resultados bastante conocidos y es de carácter eminentemente introductorio para la discusión del capítulo II.

En el capítulo II presentamos los resultados expuestos originalmente por Hoerl y Kennard además de las diversas publicaciones relacionadas con el método H.K.

El capítulo III se inicia con una breve introducción a la inferencia bayesiana a partir de la cual damos un enfoque bayesiano al método H.K. En es

te capítulo criticamos dicho método.

Uno de los motivos que nos llevó a desarrollar esta tesis fue la analogía que existe entre el método de marquardt para estimación de parámetros en modelos no lineales y el método de Hoerl y Kennard. Una discusión del método de Marquardt aparece en (4). También nos interesó ampliar la discusión del enfoque bayesiano del método H.K.

C A P I T U L O I

1.1. La distribución normal multivariada

Se dice que las variables aleatorias $X_1,\ldots,X_p$ tienen una distribución normal p-variada si existen p variables aleatorias $Z_1, \ldots, Z_p$, independientes, distribuídas cada una según una N(0,1) y si existen p constantes $\mu_1,\ldots, \mu_p$ y una matriz $A_{pxp} = (a_{ij})$, no singular, tales que

$$\begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & & a_{pp} \end{pmatrix} \begin{pmatrix} Z_1 \\ \vdots \\ Z_p \end{pmatrix} + \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix}$$

o, en notación matricial

$$\underline{X} = A\,\underline{Z} + \underline{\mu}\ .* \qquad (1.1.1)$$

Para derivar la función de densidad conjunta

$$f_{\underline{X}'}(\underline{x}') = f_{X_1,\ldots,X_p}(x_1,\ldots x_p)$$

---

* Esta será la notación que se utilice. Las matrices aparecen con letras mayúsculas, los vectores columna aparecen subrayados.

veamos que la función de densidad conjunta de $Z_1, \ldots, Z_p$ es de la forma

$$f_{\underline{Z}'}(\underline{z}') = f_{Z_1 \ldots Z_p}(z_1, \ldots z_p) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}(z_1^2 + z_2^2 + \ldots z_p^2)}$$

$$= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\underline{z}'\underline{z}} .$$

Recordemos que

$$f_{\underline{X}'}(\underline{x}') = f_{\underline{Z}'}(\underline{z}') \left| \frac{d\,\underline{z}'}{d\,\underline{x}'} \right|$$

donde $\frac{d\,\underline{z}'}{d\,\underline{x}'}$ es el Jacobiano de la transformación.

En este caso

$$\underline{Z} = A^{-1}(\underline{X} - \mu)$$

y si denotamos $A^{-1} = (b_{ij})$ tenemos que

$$\frac{\partial z_i}{\partial x_j} = b_{ij}$$

En consecuencia, el Jacobiano de la transformación es el valor absoluto del determinante de A^{-1}. Por consiguiente

$$f_{\underline{X}'}(\underline{x}') = \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\left[(\underline{x} - \underline{\mu})'(A^{-1})'A^{-1}(\underline{x} - \underline{\mu})\right]} \left|\left| A^{-1} \right|\right|$$

Sea $V = AA'$

de donde

$$V^{-1} = (AA')^{-1} = (A')^{-1} A^{-1} = (A^{-1})' A^{-1}$$

y de A.3 y A.2 en el apéndice A

$$\frac{1}{|V|} = \left|V^{-1}\right| = \left|(A^{-1})'\right|\left|A^{-1}\right| = \left|A^{-1}\right|^2$$

Claramente V es una matriz simétrica positiva definida. Utilizando la nueva notación, la función de densidad queda como

$$f_{\underline{X}'}(\underline{x}') = \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sqrt{|V|}} \quad e^{-\frac{1}{2}(\underline{x}-\underline{\mu}')V^{-1}(\underline{x}-\underline{\mu})} \qquad (1.1.2)$$

En adelante, la notación

$$\underline{X} \sim N_p(\underline{\mu}, V) \qquad (1.1.3)$$

significará que $\underline{X}$ tiene una función de densidad dada por (1.1.2).

Es interesante analizar la definición de la normal p-variada. Cada elemento X_i del vector $\underline{X}$ es una combinación lineal de $Z_1,\ldots,Z_p$, de aquí que en general, X_i y X_j no serán independientes. Se puede verificar que

$$V(\underline{X}) = E((\underline{X}-\underline{\mu})(\underline{X}-\underline{\mu})') = \begin{pmatrix} E((X_1-\mu_1)(X_1-\mu_1)) & \cdots & E((X_1-\mu_1)(X_p-\mu_p)) \\ \vdots & & \vdots \\ E((X_p-\mu_p)(X_1-\mu_1)) & \cdots & E((X_p-\mu_p)(X_p-\mu_p)) \end{pmatrix}$$

$$= (\sigma_{ij}) = V,$$

por lo que a la matriz V se le conoce como matriz de varianza-covarianza.

También se puede demostrar que

$$E(\underline{X}) = \underline{\mu}$$

Algunos teoremas que nos serán de utilidad son los siguientes (extraído de referencias (28), (33) y (34)).

Teorema 1:

Si $\underline{X} \sim N_p(\underline{\mu}, V)$ y B es una matriz mxp de rango $m \leq p$, entonces

$$B\underline{X} \sim N_m(B\underline{\mu},\ BVB') \qquad (1.1.4)$$

En particular, si P es una matriz ortogonal pxp tal que PVP' = Λ, donde Λ es la matriz de eigenvalores de V, entonces

$$P\underline{X} \sim N_p\ (P\underline{\mu},\ \Lambda) \qquad (1.1.5)$$

En este caso las covarianzas son iguales a 0.

Teorema 2:

Si $\underline{X}$ es un vector aleatorio tal que

$$E(\underline{X}) = \underline{\mu},\ V(\underline{X}) = E\ ((\underline{X} - \underline{\mu})(\underline{X} - \underline{\mu})') = V$$

y A es una matriz pxp, entonces

$$E(\underline{X}'\ A\ \underline{X}) = tr\ (AV) + \underline{\mu}'\ A\ \underline{\mu}. \qquad (1.1.6)$$

En particular, si A = I

$$E(\underline{X}'\ \underline{X}) = tr\ V + \underline{\mu}'\ \underline{\mu}$$

y de A.6

$$E(\underline{X}'\underline{X}) = \sum_{i=1}^{p} \lambda_i + \underline{\mu}'\ \underline{\mu} \qquad (1.1.7)$$

donde λ_i, i = 1,...,p son los eigenvalores de V.

En especial, nos interesará el teorema 2 cuando $\underline{X} \sim N_p\ (\underline{\mu},\ V)$.

Teorema 3:

Si $\underline{X} \sim N_p(\underline{\mu},\ V)$ la varianza de $\underline{X}'\ A\ \underline{X}$ está dada por

$$\text{Var}\,(\underline{X}'A\underline{X}) = 2\,\text{tr}\,(AV)^2 + 4\underline{\mu}'\,AVA\,\underline{\mu} \qquad (1.1.8)$$

Si $A = I$

$$\text{Var}\,(\underline{X}'\underline{X}) = 2\,\text{tr}\,V^2 + 4\underline{\mu}'\,V\underline{\mu}$$

y de A.8

$$\text{Var}\,(\underline{X}'\underline{X}) = 2\sum_{i=1}^{p}\lambda_i^2 + 4\underline{\mu}'\,V\underline{\mu} \qquad (1.1.9)$$

Las demostraciones de estas propiedades se pueden encontrar en las referencias dadas.

Pasemos a analizar la estructura geométrica de la normal multivariada. Por simplicidad tomemos $p=2$ y una matriz de varianza-covarianza

$$V = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

y vector de medias

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

De (1.1.2) podemos ver que toda la información acerca de la localización de la distribución se encuentra en el exponente. La función de densidad para el ejemplo está dada por

$$f_{X_1X_2}(x_1, x_2) = \frac{1}{\sqrt{2\pi}\,\sigma_1\sigma_2}\; e^{-\frac{1}{2}\left[(x_1-\mu_1,\, x_2-\mu_2)\begin{pmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}\right]}$$

Igualando $f_{X_1X_2}(x_1, x_2)$ a una constante nos queda después de sencillas operaciones,

$$\frac{1}{\sigma_1^2}(x_1-\mu_1)^2 + \frac{1}{\sigma_2^2}(x_2-\mu_2)^2 = c$$

que es la ecuación de una elipse con centro en (μ_1, μ_2) y con ejes paralelos a las coordenadas. Dependiendo de el valor de σ_1^2 y σ_2^2, las elipses se verán - como a) y b) de la figura 1.1.A. Si introducimos covarianzas, las elipses no serán paralelas a las coordenadas, como se ilustra en la figura 1.1.Ac.

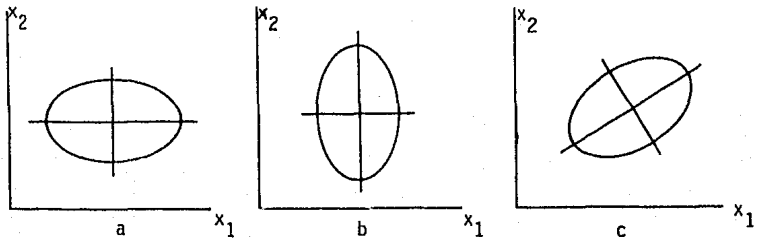


Figura 1.1.A. Elipsoides de concentración: a) sin covarianza, $\sigma_1^2 > \sigma_2^2$; b) sin covarianza, $\sigma_1^2 < \sigma_2^2$; - c) con covarianza.

A estas elipses se les llama elipsoides de concentración en el caso- bivariado e hiperelipsoides de concentración en el caso multivariado. Los -- ejes de las hiperelipsoides están en la dirección de los eigenvectores de la - matriz V. Para ver esto, recordemos que la distancia de $\underline{\mu}$ al hiperelipsoide - es máxima en la dirección del eje mayor. Entonces, si maximizamos

$$(\underline{X} - \underline{\mu})' (\underline{X} - \underline{\mu})$$

sujeto a

$$(\underline{X} - \underline{\mu})' V^{-1} (\underline{X} - \underline{\mu}) = c$$

encontraremos el eje mayor de la hiperelipsoide. Bastará con verificar que el eje mayor está en la dirección del eigenvector correspondiente al mayor eigen- valor y que los demás ejes están también en la dirección de los eigenvectores- restantes. La demostración es bastante sencilla utilizando multiplicadores de Lagrange (28). Sea

$$L(\underline{X}) = (\underline{X} - \underline{\mu})'(\underline{X} - \underline{\mu}) - \lambda((\underline{X} - \underline{\mu})' V^{-1} (\underline{X} - \underline{\mu}) - c)$$

entonces

$$\frac{dL(\underline{X})}{d\underline{X}} = 2(\underline{X} - \underline{\mu}) - 2\lambda V^{-1}(\underline{X} - \underline{\mu})$$

Igualando a $\underline{0}$ llegamos a que

$$(I - \lambda V^{-1})(\underline{X} - \underline{\mu}) = \underline{0} \qquad (1.1.10)$$

o de manera equivalente

$$(V - \lambda I)(\underline{X} - \underline{\mu}) = \underline{0}$$

que tiene solución no trivial si $|V - \lambda I| = 0$. En consecuencia, el eje mayor está en la dirección de los eigenvectores de V. Para ver en dirección de cuál eigenvector se encuentra el eje mayor, premultiplicamos (1.1.10) por $(\underline{X} - \underline{\mu})'$- con lo que

$$(\underline{X} - \underline{\mu})' (I - \lambda V^{-1}) (\underline{X} - \underline{\mu})$$

$$= (\underline{X} - \underline{\mu})' (\underline{X} - \underline{\mu}) - \lambda(\underline{X} - \underline{\mu})' V^{-1}(\underline{X} - \underline{\mu})$$

$$= (\underline{X} - \underline{\mu})' (\underline{X} - \underline{\mu}) - \lambda c = 0$$

tenemos entonces que la distancia

$$(\underline{X} - \underline{\mu})' (\underline{X} - \underline{\mu}) = \lambda c$$

es máxima si λ es máximo. Fácilmente se verifica que los demás ejes también están en dirección de eigenvectores y que las hiperelipsoides están más alargadas en la dirección de los ejes correspondientes a eigenvalores grandes, es decir, los eigenvalores serán factores de elongación de las elipses. En la figura 1.1.B podemos apreciar lo anterior.

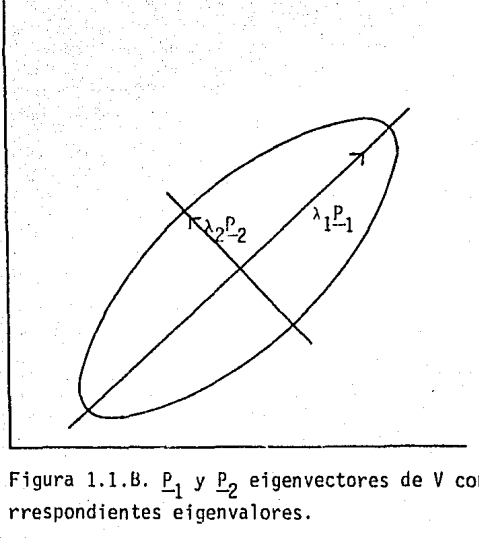


Figura 1.1.B. $\underline{P}_1$ y $\underline{P}_2$ eigenvectores de V con sus correspondientes eigenvalores.

Por último consideramos la caracterización de la distribución normal singular dada por Anderson (Ver Searle (34), p, 67). Sea

$$\underline{Z} \sim N_p\ (\underline{0},\ I)$$

Entonces se dice que

$$\underline{X} = A\underline{Z} + \underline{\mu}$$

se distribuye según una normal singular multivariada

$$\underline{X} \sim NS_j(\mu,\ AA') \qquad (1.1.11)$$

si AA' no es de rango completo sino de rango j. Veremos una aplicación de esta distribución en el capítulo 3.

En la próxima sección veremos que varias propiedades del estimador de mínimos cuadrados son propiedades que posee la normal multivariada. El análisis presentado acerca de la estructura geométrica de la normal multivariada servirá para comprender mejor los efectos del problema de multicolinealidad que se presenta en la sección 1.4 y de los métodos de estimación sesgada, que se presentan en el capítulo 2.

1.2. El modelo lineal; propiedades

Consideremos el modelo lineal

$$Y_i = b_o X_{io} + b_1 X_{i1} + \ldots\ldots + b_p X_{ip} + e_i,\ i = 1, \ldots, n \qquad (1.2.1)$$

con

$$E(e_i) = 0 \text{ para toda } i$$

$$E(e_i e_j) = \begin{cases} 0 & \text{si } i \neq j \\ \sigma^2 & \text{si } i = j \end{cases}$$

que podemos expresar en forma matricial como

$$\underline{Y} = \mathbf{X}\,\underline{b} + \underline{e} \qquad (1.2.2)$$

donde

$\underline{Y}$ es un vector n x 1

$\mathbf{X}$ es una matriz n x (p + 1) de constantes

$\underline{b}$ un vector (p + 1) x 1 de parámetros desconocidos y

$\underline{e}$ un vector aleatorio n x 1 tal que $E(\underline{e})=\underline{0}$, $E(\underline{e}\,\underline{e}') = \sigma^2 I_n$

acerca del cual se supone generalmente que

$$\underline{e} \sim N_n(\underline{0}, \sigma^2 I_n) \qquad (1.2.3)$$

Es frecuente que $X_{io} = 1$ para toda i, de tal manera que (1.2.1) queda de la forma

$$Y_i = b_o + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_p X_{ip} + e_i,\ i = 1, \ldots, n \qquad (1.2.4)$$

Decimos que el modelo lineal es de rango completo si $\mathbf{X}$ es de rango completo, en otro caso diremos que es de rango incompleto. En este trabajo nos interesará el modelo de rango completo. A menos que se especifique lo contrario, nos referiremos a este modelo; si bien trataremos un problema en donde la matriz $\mathbf{X}$ "dista poco" de ser singular.

Algunos de los objetivos del análisis del modelo lineal son:

1) Estimación de los parámetros desconocidos $\underline{b}$ y σ^2

2) Probar hipótesis acerca de los parámetros.

3) Predecir valores de Y para alguna combinación de $X_0, \ldots, X_p$.

Son bastante conocidas las propiedades del estimador de mínimos cuadrados.

$$\hat{\underline{b}} = (X'X)^{-1} X'\underline{Y} \qquad (1.2.5)$$

que es el que minimiza la suma de cuadrados residual (SCR)

$$SCR(\underline{b}) = (\underline{Y} - X\,\underline{b})' (\underline{Y} - X\underline{b}) = \underline{e}'\underline{e} \qquad (1.2.6)$$

así como las propiedades del estimador de σ^2:

$$s^2 = \frac{(\underline{Y} - X\,\underline{b})' (\underline{Y} - X\,\underline{b})}{n - (p + 1)} \qquad (1.2.7)$$

Mencionamos a continuación algunas de estas propiedades (34):

1. Dentro de la clase de estimadores lineales insesgados de $\underline{b}$, $\hat{\underline{b}}$ es el estimador de mínima varianza. Por esta propiedad se dice que $\hat{\underline{b}}$ es el mejor estimador lineal insesgado (MELI) de $\underline{b}$.

2. s^2 es un estimador insesgado de σ^2

Bajo la suposición de normalidad (1.2.3) tenemos además:

3.a) $\hat{\underline{b}} \sim N_p (\underline{b}, \sigma^2(X'X)^{-1})$ (1.2.8)

b) $\frac{(n - p)s^2}{\sigma^2} \sim \chi^2_{(n-p+1)}$

4. $\hat{\underline{b}}$ y s^2 son independientes.

5. $\hat{\underline{b}}$ es el estimador de máxima verosimilitud.

6. $\hat{\underline{b}}$ y s^2 son estimadores suficientes, completos, suficientes minimales, consistentes y eficientes.

7. De (1.1.6) tenemos que $E(\hat{\underline{b}}'\hat{\underline{b}}) = \sigma^2 \text{ tr } (X'X)^{-1} + \underline{b}'\underline{b}$ (1.2.9)

de donde

$$E((\hat{\underline{b}} - \underline{b})'(\hat{\underline{b}} - \underline{b})) = \sigma^2 \text{ tr } (X'X)^{-1} \qquad (1.2.10)$$

8. De (1.1.8) tenemos que

$$\text{Var } (\hat{\underline{b}}'\hat{\underline{b}}) = 2\sigma^4 \text{ tr } (X'X)^{-2} + 4\sigma^2 \underline{b}' (X'X)^{-1} \underline{b}$$

y como $\hat{\underline{b}} - \underline{b} \sim N_p (\underline{0}, \sigma^2 (X'X)^{-1})$, entonces

$$\text{Var } ((\hat{\underline{b}} - \underline{b})'(\hat{\underline{b}} - \underline{b})) = 2\sigma^4 \text{ tr } (X'X)^{-2} \qquad (1.2.11)$$

9. Si $\underline{L}$ es un vector $(p+1) \times 1$ de constantes, $\underline{L}'\hat{\underline{b}}$ tiene mínima varianza en la clase de estimadores lineales insesgados de $\underline{L}'\underline{b}$, además

$$\underline{L}'\hat{\underline{b}} \sim N_p (\underline{L}'\underline{b}, \sigma^2 \underline{L}'(X'X)^{-1}\underline{L}) \qquad (1.2.12)$$

Podemos apreciar que el estimador de mínimos cuadrados satisface varios criterios de optimalidad (eficiencia, suficiencia minimal, máxima verosimilitud, mínimos cuadrados) lo cual lo hace bastante atractivo, sin embargo no temos que varias propiedades dependen de la estructura de la matriz X'X.

Veremos en las próximas secciones que esta estructura es crítica -- cuando se quiere una estimación precisa de $\underline{b}$, cuando se quiere hacer una predicción de Y en una región de interés o cuando se quieren probar hipótesis -- acerca de los elementos de $\underline{b}$. Para analizar esta estructura presentamos en la próxima sección la matriz de correlación; en la sección que le sigue planteamos el problema de multicolinealidad y analizamos los efectos que tiene sobre las propiedades del estimador de mínimos cuadrados.

1.3. La matriz de correlación

Para encontrar el estimador de mínimos cuadrados es necesario invertir a la matriz X'X. Recordemos que

$$X'X = \begin{pmatrix} \Sigma X_{io}^2 & \Sigma X_{io}X_{i1} & \cdots & \Sigma X_{io}X_{ip} \\ \Sigma X_{io}X_{i1} & \Sigma X_{i1}^2 & \cdots & \vdots \\ \vdots & \vdots & & \vdots \\ \Sigma X_{io}X_{ip} & \cdots & \cdots & \Sigma X_{ip}^2 \end{pmatrix} \qquad (1.3.1)$$

o bien, si $X_{io} = 1$ para $i = 1,\ldots,n$

$$X'X = \begin{pmatrix} n & \Sigma X_{i1} & \cdots & \Sigma X_{ip} \\ \Sigma X_{i1} & \Sigma X_{i1}^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \Sigma X_{ip} & \cdots & \cdots & \Sigma X_{ip}^2 \end{pmatrix} \qquad (1.3.2)$$

Por varias razones es preferible utilizar la matriz de correlación en vez de X'X. Estas se mencionan a continuación:

1. La inversión de X'X puede llevar a errores de redondeo debido principalmente a que los números que aparecen en los cálculos son de diferente orden, o bien, porque X'X puede estar cerca de ser singular y hacer que el algoritmo de inversión sea ineficiente (6).

2. Al utilizar la matriz de correlación tenemos una medida de la asociación lineal de las variables independientes $X_1,\ldots, X_p$ así como de éstas últimas con la respuesta Y.

Esto aumenta la comprensión de la estructura del espacio de paráme-

tros y también se aplica en diversos métodos utilizados en la selección de variables.

El procedimiento para obtener la matriz de correlación consiste en los siguientes pasos (6):

1. Centrar las observaciones
2. Estandarizar las observaciones.

Primer paso.

Sea el modelo

$$Y_i = b_o + b_1 X_{i1} + \ldots + b_p X_{ip} + e_i,\ i = 1,\ldots,n \quad (1.3.3)$$

Sea $\overline{X}_j = \sum_{i=1}^{n} \frac{X_{ij}}{n}$ $\quad j = 1,\ldots,p$

y $x_{ij} = X_{ij} - \overline{X}_j$ $\quad j = 1,\ldots,p$

Podemos expresar las observaciones en términos de desviaciones con respecto a la media (centrar las observaciones):

$$Y_i = b_o + b_1 \overline{X}_1 + b_2 \overline{X}_2 + \ldots + b_p \overline{X}_p + b_1 (X_{i1} - \overline{X}_1) +$$

$$+ b_2(X_{i2} - \overline{X}_2) + \ldots + b_p(X_{ip} - \overline{X}_p) + e_i$$

$$= b_o' + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip} + e_i,\ i = 1,\ldots,n \quad (1.3.4)$$

en donde $b_o' = b_o + b_1 \overline{X}_1 + \ldots + b_p \overline{X}_p$

En notación matricial,

$$\underline{Y} = \underline{1}\, b_o' + \overline{X}\, \underline{b} + \underline{e} \quad (1.3.5)$$

donde $\underline{1}$ es un vector n x 1 de unos

$$\bar{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \qquad \underline{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$$

Como $\sum_{i=1}^{n} x_{ij} = 0$ para $j = 1,\ldots,p$ tenemos que las ecuaciones normales de (1.3.5) son

$$\begin{pmatrix} n & 0 & \cdots & 0 \\ 0 & \Sigma x_{i1}^2 & \cdots & \Sigma x_{i1}x_{ip} \\ \vdots & \vdots & & \vdots \\ 0 & \Sigma x_{i1}x_{ip} & \cdots & \Sigma x_{ip}^2 \end{pmatrix} \begin{pmatrix} b_o' \\ b_1 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} \Sigma Y_i \\ \Sigma x_{i1}Y_i \\ \vdots \\ \Sigma x_{ip}Y_i \end{pmatrix} \qquad (1.3.6)$$

Por lo tanto, independientemente del valor que tomen $b_1, b_2,\ldots,b_p$, el estimador de b_o' es

$$\hat{b}_o' = \frac{\Sigma Y_i}{n} = \bar{y} \qquad (1.3.7)$$

y en consecuencia reducimos el problema a resolver el sistema de ecuaciones

$$\begin{pmatrix} \Sigma x_{i1}^2 & \cdots & \Sigma x_{i1}x_{ip} \\ \vdots & & \vdots \\ \Sigma x_{i1}x_{ip} & \cdots & \Sigma x_{ip}^2 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix} = \begin{pmatrix} \Sigma x_{i1}Y_i \\ \vdots \\ \Sigma x_{ip}Y_i \end{pmatrix} \qquad (1.3.8)$$

o bien

$$\bar{X}'\bar{X}\,\underline{b} = \bar{X}'\underline{Y}.$$

Segundo paso:

Sea $$s_j = \sqrt{\sum_i \frac{x_{ij}^2}{n}} = \sqrt{\sum_i \frac{(X_{ij} - \bar{X}_j)^2}{n}}$$

y A una matriz diagonal p x p

$$A = \begin{pmatrix} \frac{1}{s_1} & & & \\ & \frac{1}{s_2} & & \\ & & \ddots & \\ & & & \frac{1}{s_p} \end{pmatrix}$$

entonces (1.3.5) es igual a

$$\underline{Y} = \underline{1}\, b_o' + \bar{X}AA^{-1}\underline{\beta} + \underline{e} = \underline{1}\, b_o' + X\,\underline{\beta} + \underline{e} \qquad (1.3.9)$$

con

$$X = \bar{X}A = \begin{pmatrix} \frac{x_{11}}{s_1} & \frac{x_{12}}{s_2} & \cdots & \frac{x_{1p}}{s_p} \\ \frac{x_{21}}{s_1} & \frac{x_{22}}{s_2} & \cdots & \frac{x_{2p}}{s_p} \\ \vdots & & & \vdots \\ \frac{x_{n1}}{s_1} & \frac{x_{n2}}{s_2} & \cdots & \frac{x_{np}}{s_p} \end{pmatrix}$$

$$\underline{\beta} = A^{-1}\underline{\beta} = \begin{pmatrix} s_1 b_1 \\ s_2 b_2 \\ \vdots \\ s_p b_p \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

Luego,

$$X'X = \begin{pmatrix} \frac{\sum x_{i1}^2}{s_1^2} & \frac{\sum x_{i1}x_{i2}}{s_1 s_2} & \cdots & \frac{\sum x_{i1}x_{ip}}{s_1 s_p} \\ \frac{\sum x_{i1}x_{i2}}{s_1 s_2} & \frac{\sum x_{i2}^2}{s_2^2} & \cdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \frac{\sum x_{i1}x_{ip}}{s_1 s_p} & \cdots & \cdots & \frac{\sum x_{ip}^2}{s_p^2} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ r_{1p} & \cdots & \cdots & 1 \end{pmatrix} \qquad (1.3.10)$$

donde r_{ij} = correlación de X_i y X_j. Todos los números que aparecen en la matriz X'X estan entre -1 y 1. Esto evita los errores de redondeo mencionados anteriormente y permite tener una medida de la asociación lineal de las variables, como lo es r_{ij}.

Por otra parte, tomemos el lado derecho de la igualdad (1.3.8) y dado que

$$\sum (X_i - \overline{X})\, Y_i = \sum (X_i - \overline{X})\, (Y_i - \overline{Y})$$

tenemos la igualdad

$$\begin{pmatrix} \sum x_{i1} & Y_i \\ \sum x_{i2} & Y_i \\ \sum x_{ip} & Y_i \end{pmatrix} = \begin{pmatrix} \sum x_{i1}\,(Y_i - \overline{Y}) \\ \sum x_{i2}\,(Y_i - \overline{Y}) \\ \sum x_{ip}\,(Y_i - \overline{Y}) \end{pmatrix} \qquad (1.3.11)$$

Estas ecuaciones son las que se obtienen si se supone el modelo

$$\underline{Y} = \underline{Y} - \underline{1}\,\overline{y} = X\,\underline{\beta} + \underline{e} \qquad (1.3.12)$$

La notación (1.3.12) es solamente convencional (ver apéndice B) puesto que el modelo correcto es (1.3.9). Utilizaremos la notación (1.3.12) cuando el interés primordial se centre en la estimación de $\underline{\beta}$ y no tanto en la estimación de b_0'.

1.4. El problema de multicolinealidad

Hicimos notar en la sección 1.2 que varias de las propiedades del estimador de mínimos cuadrados dependen de X'X (de X'X en aquella sección). Sobre X'X el único supuesto que hemos hecho es que sea de rango completo, es decir, las columnas (o los renglones) son linealmente independientes y por lo tanto forman una base en el espacio p-dimensional. Esta base es ortogonal si -- X'X es una matriz diagonal y, dado que es una matriz de correlación, su determinante vale 1. En caso que X'X no sea diagonal tenemos el resultado

$$0 < \left|X'X\right| < 1 \qquad (1.4.1)$$

al cual se llega utilizando la desigualdad A.4. Si X'X es de rango incompleto tenemos que $\left|X'X\right| = 0$. Entonces, un síntoma que indica que nos acercamos a una matriz de rango incompleto es, que el determinante es cercano a 0. Sean - $\lambda_1, \lambda_2, \ldots \lambda_p$ los eigenvalores de X'X. De A.5 sabemos que

$$\left|X'X\right| = \lambda_1\,\lambda_2 \ldots \lambda_p$$

entonces otro síntoma de que nos hemos alejado bastante de la ortogonalidad, - es el hecho que algunos eigenvalores son cercanos a 0.

Estos 2 síntomas nos sirven para detectar la presencia del problema- que los econometristas han llamado multicolinealidad. Una manera de definir - el problema de muticolinealidad, como lo definen Farrar y Glauber (tomado de - (43)), es en términos de desviaciones de la ortogonalidad. Por ejemplo, Fa- - rrar y Glauber proponen 3 pruebas que "detectan la presencia y el patrón de multicolinealidad" bajo la suposición de que los renglones de X provienen de una- distribución normal multivariada. Varias críticas se han hecho a Farrar y - Glauber dado que una prueba es incorrecta (43) además de que la suposición -- acerca de los renglones de X provenientes de una distribución normal multivá-- riada es muy restrictiva (20) (31). El problema básico presente en el enfoque de Farrar y Glauber es que se avocan a detectar y remediar los síntomas pero - no necesariamente resuelven el problema de multicolinealidad. El defecto de - su enfoque radica en su definición del problema de multicolinealidad. Siguiendo el enfoque de Johnston (19) y Silvey (36) presentamos las consecuencias de- la multicolinealidad para de allí buscar la esencia del problema.

Recordando que

$$\hat{\underline{\beta}} \sim N_p \left(\underline{\beta}, \sigma^2 (X'X)^{-1}\right) \qquad (1.4.2)$$

vemos que la multicolinealidad o no ortogonalidad se refleja en la matriz de - varianza-covarianza de $\hat{\underline{\beta}}$. Tomemos como ejemplo la siguiente matriz de correlación:

$$X'X = \begin{pmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{pmatrix}$$

El determinante

$$D = |X'X| = 1 + 2\, r_{12}\, r_{13}\, r_{23} - r_{12}^2 - r_{13}^2 - r_{23}^2$$

y la inversa es

$$(X'X)^{-1} = \frac{1}{D} \begin{pmatrix} 1 - r_{23}^2 & r_{13}r_{23} - r_{12} & r_{12}r_{23} - r_{13} \\ r_{13}r_{23} - r_{12} & 1 - r_{13}^2 & r_{12}r_{13} - r_{23} \\ r_{12}r_{23} - r_{13} & r_{12}r_{13} - r_{23} & 1 - r_{12}^2 \end{pmatrix} \qquad (1.4.3)$$

Por ejemplo, si $r_{12}=.1$, $r_{13}=.8$, $r_{23}=.6$, entonces $D = .086$

$$(X'X)^{-1} = \begin{pmatrix} 7.442 & 4.419 & -8.605 \\ 4.419 & 4.186 & -6.046 \\ -8.605 & -6.046 & 11.512 \end{pmatrix} \qquad (1.4.4)$$

De (1.4.2) y (1.4.3) vemos que una de las consecuencias de la multicolinealidad es la estimación imprecisa de los parámetros reales . Compárese de (1.4.4), Var $(\hat{\beta}_3) = 11.512\ \sigma^2$ con la varianza de $\hat{\beta}_3$ en un sistema ortogonal:

$$\text{Var}\ (\hat{\beta}_3) = \sigma^2$$

La diferencia es dramática. Además $\hat{\beta}_i$ y $\hat{\beta}_j$, $i \neq j$, están en el ejemplo muy correlacionados, luego entonces, los errores que se cometan al estimar estarán correlacionados. Esto lo podemos ver más claramente transformando la matriz $(X'X)^{-1}$ en una matriz de correlación C. En el ejemplo

$$C = \begin{pmatrix} 1 & .792 & -.929 \\ .792 & 1 & -.871 \\ -.929 & -.781 & 1 \end{pmatrix} \qquad (1.4.5)$$

Si un elemento de $\underline{\hat{\beta}}$ se encuentra alejado del valor real, debido a la

estructura de la matriz de varianza-covarianza, otros elementos de $\hat{\underline{\beta}}$ también - se encontrarán alejados del valor real. En el ejemplo tenemos que la correlación de $\hat{\beta}_1$ y $\hat{\beta}_2$ es .792; si $\hat{\beta}_1$ está sobre estimado, $\hat{\beta}_2$ también tenderá a estar sobreestimado. En este caso $\hat{\beta}_3$ tenderá a estar bajoestimado.

De lo anterior podemos deducir que $\hat{\underline{\beta}}$ puede estar muy alejado de $\underline{\beta}$; - de hecho tenemos de (1.2.10) que

$$E((\hat{\underline{\beta}} - \underline{\beta})'(\hat{\underline{\beta}} - \underline{\beta}) = \sigma^2 \, tr(X'X)^{-1} \qquad (1.4.6)$$

y de A.9 y A 6 resulta

$$\sigma^2 \, tr(X'X)^{-1} = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} \qquad (1.4.7)$$

donde λ_i, $i = 1,\ldots,p$ son los eigenvalores de X'X. Si algún λ_i es cercano a 0, la esperanza (1.4.6) será grande relativa a la esperanza cuando X'X es ortogonal. En el ejemplo $E((\underline{\beta} - \underline{\beta})' (\hat{\underline{\beta}} - \underline{\beta})) = 23.140\sigma^2$ la podemos comparar con $3\sigma^2$ que es la esperanza que se obtendrá en un sistema ortogonal. Por último, de (1.2.11)

$$Var ((\hat{\underline{\beta}} - \underline{\beta})' (\hat{\underline{\beta}} - \underline{\beta}) = 2\sigma^4 \, tr(X'X)^{-2}$$

y de A.8 resulta que

$$2\sigma^4 \, tr(X'X)^{-2} = 2\sigma^4 \sum_{i=1}^{p} \frac{1}{\lambda_i^2} \qquad (1.4.8)$$

Resumiendo lo anterior podemos decir que la multicolinealidad trae - como consecuencia una falta de precisión en la estimación: los parámetros estimados pueden estar muy alejados de los parámetros reales y además algunos o todos los parámetros estimados pueden estar muy correlacionados. Además "resulta extremadamente difícil, si no imposible, desenmascarar las influencias relativas de las diferentes variables" (19)

En la figura 1.4.A se ilustran gráficamente algunas de estas conse-- cuencias

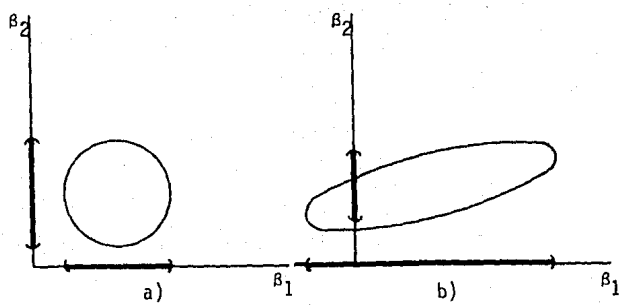


Figura 1.4.A. Intervalos de confianza para β_1 y β_2 a un mismo nivel α para a) X ortogonal b) X no ortogonal.

Otra consecuencia de la multicolinealidad se presenta cuando se quiere utilizar alguno de los métodos de selección de variables. Marquardt (25) anota que las condiciones presentes en la multicolinealidad "desestabilizan todos los criterios que uno puede calcular a partir de la estimación por mínimos cuadrados, llevando a una selección de variables altamente inestable".

Por último, Johnston (19) ve como una consecuencia de la multicolinealidad que el añadir algunas observaciones adicionales al conjunto original de datos produce algunas veces cambios sustanciales en los parámetros estima-- dos originalmente.

Esta última consecuencia nos da la clave para encontrar la esencia - del problema. Cuando X'X es no ortogonal, estamos trabajando con menos información que cuando X'X es ortogonal. Vemos entonces que la esencia del problema- es la falta de información. Es por esto que disentimos del enfoque dado, por-

ejemplo, por Coxe (3), en donde trata al problema de multicolinealidad como un problema numérico. Visto como un problema de falta de información, Silvey (36) estudia cómo se afecta la precisión de las estimaciones al aumentar n, el número de observaciones. Analizando la estructura del espacio de observaciones y viendo en qué direcciones la información es escasa, propone incrementar la precisión de las estimaciones eligiendo optimamente nuevos valores de las variables independientes. Sin embargo, este método no es aplicable cuando por alguna razón no es posible elegir (o diseñar) los nuevos valores de las variables independientes. Además, el problema de multicolinealidad se presenta generalmente en estos casos.

En el próximo capítulo veremos como Hoerl y Kennard buscan también incrementar la precisión de las estimaciones, pero a diferencia de Silvey, que se basa en el estimador de mínimos cuadrados, ellos proponen otro tipo de estimadores.

C A P I T U L O I I

2.1. El error cuadrático medio y la condición de admisibilidad

Ante la imposibilidad de encontrar dentro de la clase de estimadores lineales insesgados de $\underline{\beta}$ un estimador con menor varianza que el estimador de - mínimos cuadrados y con las características poco atractivas que este último -- posee en problemas de multicolinealidad, entonces la proposición de buscar dentro de otras clases de estimadores (por ejemplo estimadores lineales sesgados, lineales con un sesgo fijo, no lineales, etc.) alternativas para resolver este tipo de problemas, es bastante atractiva. Para poder comparar a $\hat{\underline{\beta}}$ con - - otros estimadores necesitamos un criterio más general que el de mínima varianza. El criterio que más se ha utilizado es el de mínimo error cuadrático me-- dio (ECM). Este criterio se puede plantear como sigue (2):

Dada una clase de posibles candidatos a estimadores, encuentre aquel estimador $\hat{\underline{\beta}}^*$ que minimice

$$ECM(\tilde{\underline{\beta}}) = E\left((\tilde{\underline{\beta}} - \underline{\beta})' (\tilde{\underline{\beta}} - \underline{\beta})\right) \qquad (2.1.1)$$

Observemos que si tomamos la clase de estimadores lineales insesgados, el criterio de mínimo ECM coincide con el de mínima varianza.

Si desarrollamos (2.1.1) llegamos a que

$$ECM(\tilde{\underline{\beta}}) = E\left((\tilde{\beta} - E(\tilde{\beta}))' (\underline{\tilde{\beta}} - E(\tilde{\beta}))\right) + (E(\underline{\tilde{\beta}})-\beta)'(E(\underline{\tilde{\beta}})-\underline{\beta}) \qquad (2.1.2)$$

El primer término del lado derecho de la igualdad corresponde a la - varianza del estimador y el segundo es una medida del sesgo del mismo. (Aunque estamos utilizando la notación para el modelo lineal, podemos pensar en otro - tipo de estimadores, (2.1.2) sigue siendo válido.) Surge naturalmente la duda de si tiene sentido perder la propiedad de insesgamiento y tomar en cuenta estimadores sesgados. Veamos la figura 2.1.A.

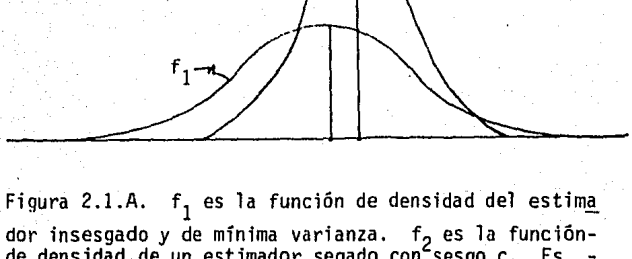


Figura 2.1.A. f_1 es la función de densidad del estimador insesgado y de mínima varianza. f_2 es la función de densidad de un estimador sesgado con sesgo c. Es - claro que preferimos en este caso al estimador con función de densidad f_2.

Mayer y Willke (23) proponen la siguiente condición de admisibilidad: "Una clase E de estimadores es admisible de acuerdo al criterio de ECM, si para cada problema existe $\hat{\underline{\rho}} \in E$ tal que

$$ECM(\hat{\underline{\rho}}) < ECM(\underline{\beta})" \qquad (2.1.3)$$

Esta condición de admisibilidad juega un papel importante en la presente tesis, ya que en parte justifica el uso de los estimadores de Hoerl y -- kennard que se presentan en la siguiente sección, así como otros que se discuten en la sección 2.3.

2.2. Los estimadores de Hoerl y Kennard

Dentro de la clase de estimadores lineales insesgados de $\underline{\beta}$ no es posible encontrar otro estimador que sea superior al de mínimos cuadrados de acuerdo al criterio de ECM. Por lo tanto, la búsqueda de otro estimador que satisfaga la condición de admisibilidad debe buscarse en otra clase de estimadores, por ejemplo, estimadores lineales sesgados. Hoerl y Kennard (15) (16) proponen una clase de estimadores lineales sesgados que satisfacen la condición de admisibilidad (2.1.3). Esta nueva alternativa para resolver el problema de multicolinealidad ha despertado gran interés dado que en la práctica ha dado buenos resultados. Ellos proponen una clase de estimadores de la forma

$$\hat{\underline{\beta}}_k = (X'X + kI)^{-1} X'\underline{Y} = W_k X'\underline{Y} \qquad (2.2.1)$$

$k > 0$, que de aquí en adelante llamaremos estimadores H.K.

Dado que $X'X$ y kI son matrices simétricas positivas definidas, W_k es simétrica positiva definida. Cada elemento de la clase es una transformación lineal de $\hat{\underline{\beta}}$ debido a que

$$\begin{aligned} \hat{\underline{\beta}}_k &= (X'X + kI)^{-1} (X'X) (X'X)^{-1} X'\underline{Y} \\ &= ((X'X)^{-1} (X'X + kI))^{-1} \hat{\underline{\beta}} = (I + k(X'X)^{-1})^{-1} \hat{\underline{\beta}} \\ &= Z_k \hat{\underline{\beta}} \end{aligned} \qquad (2.2.2)$$

De la definición de Z_k podemos ver que también es una matriz simétrica positiva definida.

De (2.2.2) resulta que $E(\hat{\underline{\beta}}_k) = Z_k \underline{\beta}$ y por lo tanto $\hat{\underline{\beta}}_k$ es un estimador lineal sesgado de $\underline{\beta}$. De (1.1.4) sabemos que

$$\hat{\underline{\beta}}_k = Z_k \hat{\underline{\beta}} \sim N_p(Z_k\underline{\beta}, Z_k(X'X)^{-1} Z_k') \qquad (2.2.3)$$

En la discusión acerca de estos estimadores utilizaremos las siguientes propiedades de las matrices Z_k y W_k.

Propiedad 1

$$Z_k = I - k\, W_k \qquad (2.2.4)$$

Demostración

Si escribimos Z_k en la forma

$$Z_k = (X'X + kI)^{-1} X'X \qquad (2.2.5)$$

y premultiplicamos (2.2.5) por W_k^{-1} obtenemos que

$$W_k^{-1} Z_k = (X'X + kI)(X'X + kI)^{-1} X'X = X'X$$

Por otro lado, premultiplicando el lado derecho de (2.2.4) por W_k^{-1} resulta que

$$W_k^{-1} (I - k\, W_k) = (X'X + kI)\,(I - k\, W_k) = X'X + kI - kI = X'X$$

por lo tanto, se cumple (2.2.4).

Propiedad 2

Sean $\lambda_1, \lambda_2, \ldots, \lambda_p$ los eigenvalores de $X'X$ y denotemos por $\lambda_i(W_k)$ y $\lambda_i(Z_k)$, $i = 1\ldots, p$ los eigenvalores de W_k y Z_k respectivamente. Entonces

$$\lambda_i(W_k) = \frac{1}{(\lambda_i + k)} \qquad (2.2.6)$$

$$\lambda_i(Z_k) = \frac{\lambda_i}{(\lambda_i + k)} \qquad (2.2.6')$$

Demostración:

Tomemos la ecuación característica de W_k^{-1}:

$$0 = \left| W_k^{-1} - \xi I \right| = \left| X'X + kI - \xi I \right| = \left| X'X - (\xi - k)I \right|$$

que es la ecuación característica de X'X. Por lo tanto se da la igualdad

$$\lambda_i = \xi_i - k \qquad i = 1,\ldots,p$$

de donde

$$\lambda_i(W_k) = \frac{1}{\xi_i} = \frac{1}{\lambda_i + k} \qquad i = 1,\ldots,p$$

y entonces (2.2.6) queda demostrado. El resultado (2.2.6') se demuestra en -- forma análoga.

Como primera aplicación de estas propiedades tenemos el siguiente - teorema.

Teorema 1

$$\hat{\underline{\beta}}_k' \hat{\underline{\beta}}_k < \hat{\underline{\beta}}' \hat{\underline{\beta}} \qquad (2.2.7)$$

y $\hat{\underline{\beta}}_k' \hat{\underline{\beta}}_k$ es una función decreciente en k.

Demostración

$$\hat{\underline{\beta}}_k' \hat{\underline{\beta}}_k = \hat{\underline{\beta}}' Z_k' Z_k \hat{\underline{\beta}} = \hat{\underline{\beta}}' (Z_k)^2 \hat{\underline{\beta}}$$

De A.8 sabemos que los eigenvalores de Z_k^2 son $(\frac{\lambda_i}{\lambda_i+k})^2 = \delta_i < 1$. Entonces tomando la descomposición espectral de Z_k^2 (ver A.7)

$$Z_k^2 = \delta_1 \underline{P}_1 \underline{P}_1' + \delta_2 \underline{P}_2 \underline{P}_2' + \ldots + \delta_p \underline{P}_p \underline{P}_p'$$

$$I = \underline{P}_1 \underline{P}_1' + \underline{P}_2 \underline{P}_2' + \ldots + \underline{P}_p \underline{P}_p'$$

tenemos que

$$\hat{\underline{\beta}}' Z_k^2 \hat{\underline{\beta}} = \delta_1 \hat{\underline{\beta}}' \underline{P}_1 \underline{P}_1' \hat{\underline{\beta}} + \delta_2 \hat{\underline{\beta}}' \underline{P}_2 \underline{P}_2' \hat{\underline{\beta}} + \ldots + \delta_p \hat{\underline{\beta}}' \underline{P}_p \underline{P}_p \leq \delta_{max} \hat{\underline{\beta}}' \hat{\underline{\beta}}$$

donde $\delta_{max} = \max_i \delta_i$ y como $\delta_{max} < 1$ resulta que

$$\hat{\beta}_k' \hat{\beta}_k < \hat{\beta}' \hat{\beta}$$

que es lo que queríamos demostrar.

Vemos entonces que una de las propiedades de estos estimadores es -- que se encuentran más cerca del origen que el estimador de mínimos cuadrados.

A continuación presentamos los teoremas que demuestran que los estimadores H.K. forman una clase admisible. Para ello, veamos que

$$\begin{aligned} ECM(\hat{\beta}_k) &= E\left((\hat{\beta}_k - E(\hat{\beta}_k))'(\hat{\beta}_k - E(\hat{\beta}_k))\right) + (E(\hat{\beta}_k) - \beta)' (E(\hat{\beta}_k) - \beta) \\ &= E\left((\hat{\beta} - \beta)' Z_k' Z_k (\hat{\beta} - \beta)\right) + (Z_k \beta - \beta)' (Z_k \beta - \beta) \\ &= \sigma^2 tr Z_k' (X'X)^{-1} Z_k + \beta'(Z_k - I)'(Z_k - I)\beta \end{aligned} \qquad (2.2.8)$$

Como $Z_k = I - kW$ y $Z_k' (X'X)^{-1} = W_k$

tenemos que

$$\begin{aligned} ECM(\hat{\beta}_k) &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \beta' (X'X + kI)^{-2} \beta \\ &= \gamma_1(k) + \gamma_2(k) \end{aligned}$$

Teorema 2. $\gamma_1(k)$ es una función continua, monótona decreciente en k.

Corolario 2.1. $\lim_{k \to 0^+} \left(\frac{d\gamma_1}{dk}\right) = -2\sigma^2 \sum_{i=1}^{p} \left(\frac{1}{\lambda_i^2}\right)$

Las demostraciones son triviales.

Teorema 3. $\gamma_2(k)$ es una función continua, monótona creciente.

Demostración

Sea Λ la matriz de eigenvalores de X'X y sea P una matriz ortogo-- nal tal que

$$X'X = P'\Lambda P \quad y \quad PX'XP' = \Lambda$$

(ver A.7.)

Si $\underline{\alpha} = P\underline{\beta}$ $\beta = P'\underline{\alpha}$ entonces

$$\gamma_2(k) = k^2\, \underline{\beta}'\,(X'X + kI)^{-2}\, \underline{\beta} = k^2 \underline{\alpha}' P (X'X + kI)^{-2}\, P' \underline{\alpha}$$

De A.9 y A.8 resulta que

$$\gamma_2(k) = k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \qquad (2.2.9)$$

Como $\lambda_i + k$ es mayor que 0 para toda i y toda $k \geq 0$, claramente es - una función continua. Notemos que $\gamma_2(0) = 0$ y que para $k > 0$ podemos escri-- bir a (2.2.9) como

$$\gamma_2(k) = \sum_{i=1}^{p} \frac{\alpha_i^2}{(1 + \frac{\lambda_i}{k})^2}$$

Para toda i y toda $k>0$, $\frac{\lambda_i}{k}$ es una función monótona decreciente de - donde $\frac{\alpha_i}{(1+\frac{\lambda_i}{k})^2}$ es monótona creciente.

Por lo tanto $\gamma_2(k)$ es monótona creciente.

Corolario 3.1. $\gamma_2(k) \to \underline{\beta}'\underline{\beta}$ cuando $k \to \infty$

Corolario 3.2. $\lim_{k\to 0^+}(\frac{d\gamma_2}{dk}) = 0$

El siguiente teorema demuestra que la clase de estimadores de Hoerl-y Kennard es una clase admisible.

Teorema 4. Existe $k > 0$ tal que $ECM(\hat{\underline{\beta}}_k) < ECM(\hat{\underline{\beta}})$

Demostración:

$$\frac{dECM(\hat{\underline{\beta}}_k)}{d_k} = \frac{d_{\gamma_1}(k)}{d_k} + \frac{d_{\gamma_2}(k)}{d_k}$$

$$= -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} + 2k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i+k)^3} \qquad (2.2.10)$$

De los corolarios 1.1 y 2.2

$$\lim_{k\to 0^+} \left(\frac{d\ ECM(\hat{\underline{\beta}}_k)}{d_k} \right) = -2\sigma^2 \sum_{i=1}^{p} \left(\frac{1}{\lambda_i^2}\right)$$

Ahora, si encontramos $k > 0$ tal que

$$\frac{d\ ECM(\hat{\underline{\beta}}_k)}{d_k} < 0$$

para esos valores de k el $ECM(\beta_k)$ será una función monótona decreciente. De - (2.2.10):

$$-2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} + 2_k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i+k)^3} < 0$$

si y solo si $\sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^3} > \frac{k}{\sigma^2} \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i+k)^3}$

que se cumple si

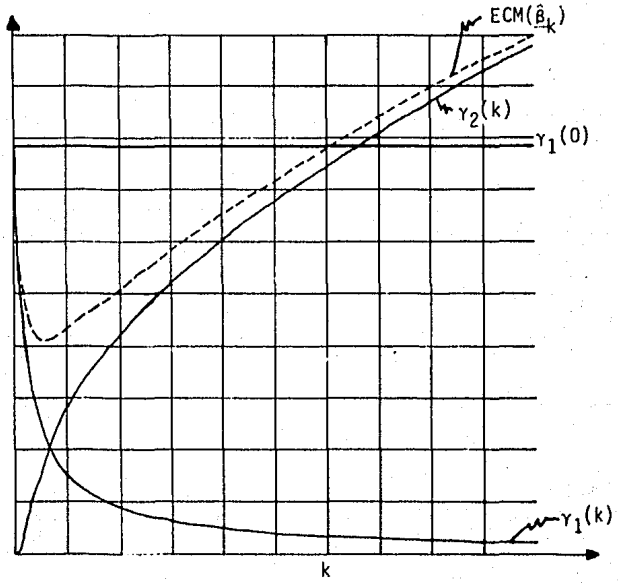
$$\frac{k}{\sigma^2} \alpha_i^2 < 1 \quad \text{para toda } i.$$

Por lo tanto, si tomamos

$$k < \frac{\sigma^2}{\alpha^2_{max}} \quad \text{donde} \quad \alpha^2_{max} = \max_i (\alpha_i^2) \qquad (2.2.11)$$

se satisface la condición de admisibilidad.

En la gráfica 2.2.A se ilustra este teorema.



Gráfica 2.2.A. Tomado de Hoerl y Kennard (15)

Este es el resultado fundamental que presentan Hoerl y Kennard, ya - que nos garantiza la existencia de un valor de k con ECM menor que el de mínimos cuadrados. El problema que se presenta es que (2.2.11) depende tanto de - $\underline{\beta}'\underline{\beta}$ como de σ^2 que son desconocidos. Hasta ahora no existe método para encontrar valores de k que nos garantice una reducción en ECM, a pesar de que se-

ha hecho un esfuerzo considerable en esa dirección como veremos en la sección (2.4).

Hoerl y Kennard propusieron un método al que nosotros llamamos la traza de H.K. (ridge trace) y que presentamos a continuación.

2.2.a). La traza de H.K.

Sabemos que $\hat{\beta}$ minimiza la suma de cuadrados residual, SCR,

$$\phi(\underline{B}) = SCR\ (\underline{B}) = (Y - X\underline{B})'(Y - X\underline{B})$$

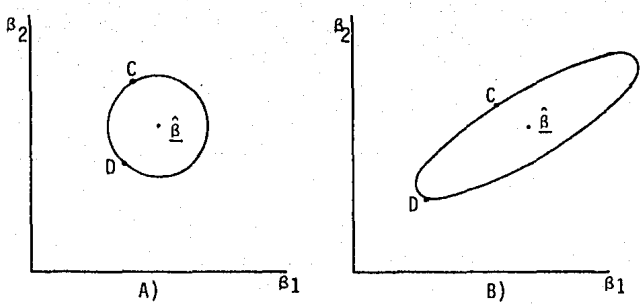
Para cualquier vector $\underline{B}$ que estime a β, podemos escribir la SCR en términos de $\hat{\beta}$ de la siguiente manera

$$SCR\ (\underline{B}) = (\underline{Y} - X\hat{\beta})'(\underline{Y} - X\hat{\beta}) + (\underline{B} - \hat{\beta})'X'X(\underline{B} - \hat{\beta})$$

$$= \min_{\underline{B}^*} \phi\ (\underline{B}^*) + (\underline{B} - \hat{\beta})'X'X\ (\underline{B} - \hat{\beta}) \qquad (2.2.12)$$

Como la forma cuadrática

$(\underline{B} - \hat{\beta})'X'X\ (\underline{B} - \hat{\beta})$ es siempre ≥ 0, al introducir un vector $\underline{B}$ que estime a β nos alejamos del óptimo obtenido por el criterio de mínimos cuadrados. Los vectores $\underline{B}$ tales que $(\underline{B} - \hat{\beta})\ X'X\ (\underline{B} - \hat{\beta}) = c$ forman la superficie de un hiperelipsoide centrado en $\hat{\beta}$. Notemos que si la matriz X'X tiene algunos eigenvalores muy pequeños, nos podemos alejar del vector de mínimos cuadrados sin un notable incremento en la suma de cuadrados residual. Esto se ilustra en la gráfica 2.2.B.



Gráfica 2.2.B. C y D producen el mismo incremento en la suma de cuadrados residual en ambas figuras. En A X'X es una matriz ortogonal. En B es una matriz no ortogonal. En B nos podemos alejar más de $\underline{\hat{\beta}}$ con el mismo incremento de la SCR.

Hoerl y Kennard (15) anotan que el método de estimación de mínimos - cuadrados "no refleja la sensibilidad de la solución al criterio de optimización".

Una caracterización de la clase de estimadores H.K. es la siguiente:

Dado $(\underline{B} - \underline{\hat{\beta}})'X'X\ (\underline{B} - \underline{\hat{\beta}}) = c$ (2.2.13)

existe un valor k tal que

$$\underline{\hat{\beta}}_k'\ \underline{\hat{\beta}}_k = \min_{\underline{B}}\ \underline{B}'\underline{B} \qquad (2.2.14)$$

Esto significa que, de todos los vectores que satisfacen (2.2.13), $\underline{\hat{\beta}}_k$ es el - que tiene longitud mínima. Podemos ver que (2.2.14) es cierta resolviendo

$$\underset{\underline{B}}{\text{Min}}\ \underline{B}'\underline{B}$$

sujeto a $(\underline{B} - \hat{\underline{\beta}})'X'X\,(\underline{B} - \hat{\underline{\beta}}) = c$

Solución

Utilizando multiplicadores de Lagrange, sea el problema

$$\text{Min } F = \underline{B}'\underline{B} + \left(\frac{1}{k}\right)\,(\underline{B} - \hat{\underline{\beta}})'\,X'X\,(\underline{B} - \hat{\underline{\beta}}) - c$$

donde $\frac{1}{k}$ es el multiplicador;

$$\frac{dF}{d\underline{B}} = 2\underline{B} + \left(\frac{1}{k}\right)\left(2(X'X)\,\underline{B} - 2(X'X)\hat{\underline{\beta}}\right) = 0$$

de donde

$$\underline{B} = (X'X + kI)^{-1}\quad X'\underline{Y} = \hat{\underline{\beta}}_k$$

Ya que $\hat{\underline{\beta}}$ es estimador máximo verosímil, puede verse que cualquier otro vector que estime a $\underline{\beta}$ decrementará la función de verosimilitud. Ahora -- bien, ¿por qué escoger el vector que minimiza $\underline{B}'\underline{B}$? Hoerl y Kennard argumentan que, si bien vectores con mayor longitud nos dan el mismo incremento en la SCR o, equivalentemente, el mismo decremento en la función de verosimilitud, "no siempre tendrán el mismo significado físico. Está implicada una restricción en los valores posibles de $\underline{\beta}$, que no se hace explícita en la formulación del modelo lineal general".

Como veremos en el capítulo 3, en términos bayesianos, lo anterior equivale a introducir información a priori acerca de los parámetros. La figura 2.2.C muestra el estimador $\hat{\underline{\beta}}_k$ para diferentes valores de la constante c dada en (2.2.13).

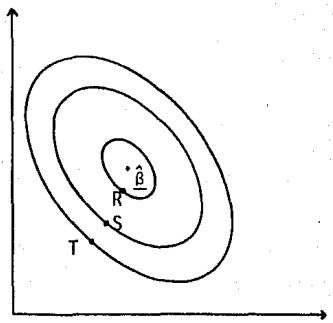


Figura 2.2.C. Los puntos R,S,T representan a los vectores $\underline{\hat{\beta}}_k$ para diferentes valores de $c=(\underline{\hat{\beta}}_k-\underline{\hat{\beta}})'X'X(\underline{\hat{\beta}}_k-\underline{\hat{\beta}})$. A medida que c crece la longitud de $\hat{\beta}_k$ decrece.

En lugar de fijar c y de allí seleccionar el valor de k, es preferible fijar k y a partir de esto ver cuál fue el incremento en la SCR. El método de H.K. consiste en graficar los coeficientes de $\underline{\hat{\beta}}_k$ y la SCR para diferentes valores de k entre 0 y 1, tal como se ilustra en la figura 2.2.D. Tomada de un ejemplo de Hoerl y Kennard (16).

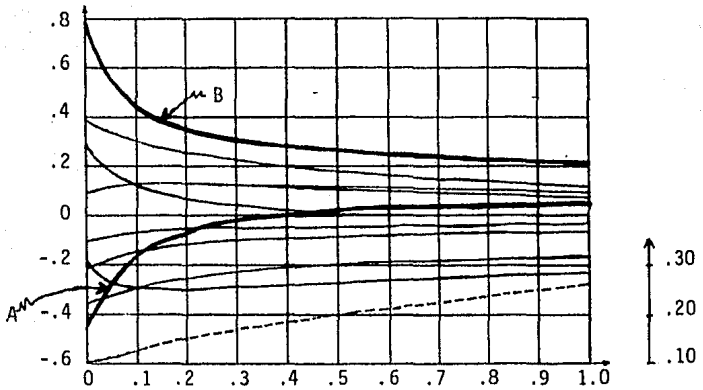


Figura 2.2.D. La traza de H.K. La escala que aparece a la derecha corresponde a la SCR. Las líneas continuas son los coeficientes de $\underline{\hat{\beta}}_k$.

Las recomendaciones que pueden llevar a seleccionar un valor adecuado de k son las siguientes:

1. Estabilidad de los coeficientes de $\underline{\hat{\beta}}_k$. En la figura 2.2.D algunos coeficientes cambian drásticamente de valor para k entre 0 y .2. Hoerl y Kennard dicen que estos coeficientes no "retienen su poder predictivo". - - Véanse A y B en la misma figura.

2. Cambios en los signos de los coeficientes. Puede haber signos - que no vayan de acuerdo con la naturáleza de los datos. Esto se verá con claridad en el ejemplo en 2.2.b).

3. La SCR da una indicación de cuanto se ha afectado el criterio de optimización. Se busca que ésta no sea grande respecto a la SCR ($\underline{\hat{\beta}}$), o bien,- "grande relativa a lo que sería una varianza razonable para el proceso que genera los datos".

4. "Los valores absolutos de los coeficientes no serán grandes en - relación a las variables en donde ellos representan tasas de cambio" (16)

2.2.b). Un ejemplo del método H.K.

Presentamos ahora un ejemplo que sirva para aclarar los puntos anteriores. Este ejemplo proviene de una simulación que describimos en el apéndice C. Pretendemos que puede servir para una mejor comprensión del método de - Hoerl y Kennard. Además, nos concentramos solamente en el problema de estimación, que es una parte del análisis de un conjunto de datos.

Supongamos que tenemos el problema de estimación de parámetros bajo- la suposición de que el modelo está dado por:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 + e \qquad (2.2.15)$$

Para obtener la matriz de correlación transformamos (2.2.15) en

$$Y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e \qquad (2.2.16)$$

La matriz de correlación y su inversa son

$$X'X = \begin{pmatrix} 1 & .7684 & -.3607 \\ .7684 & 1 & 0 \\ -.3607 & 0 & 1 \end{pmatrix} \quad (X'X)^{-1} = \begin{pmatrix} 3.5784 & -2.7496 & 1.2907 \\ -2.7496 & 3.1128 & -.9918 \\ 1.2907 & -.9918 & 1.4656 \end{pmatrix}$$

Los eigenvalores de X'X son

$$\lambda_1 = 1.8489 \qquad \lambda_2 = 1 \qquad \lambda_3 = .1511$$

El determinante de X'X es .2794. Vamos a suponer que conocemos de -

la "naturaleza del fenómeno" que $\beta_i > 0$ para $i=1,\ldots,3$.

En la tabla siguiente aparecen los estimadores para las unidades -- transformadas y para las unidades originales

k	$\hat{\beta}_{k1}$	$\hat{\beta}_{k2}$	$\hat{\beta}_{k3}$	$\hat{\sigma}_{ko}$	$\hat{\sigma}_{k1}$	$\hat{\sigma}_{k2}$	$\hat{\sigma}_{k3}$
0	1.0594	-0.0669	0.4010	0.0461	2.9652	-0.2182	1.2282
.025	0.9623	0.0075	0.3570	0.0461	2.6934	0.0246	1.0935
.05	0.8881	0.0617	0.3231	0.0461	2.4857	0.2011	0.9894
0.75	0.8293	0.1023	0.2958	0.0461	2.3211	0.3337	0.9060
.1	0.7812	0.1335	0.2733	0.0461	2.1866	0.4355	0.8371
.2	0.6515	0.2055	0.2115	0.0461	1.8234	0.6703	0.6479
.3	0.5725	0.2363	0.1734	0.0461	1.6025	0.7709	0.5210
.4	0.5175	0.2497	0.1468	0.0461	1.4483	0.8144	0.4496
.5	0.4758	0.2544	0.1270	0.0461	1.3317	0.8298	0.3889
.6	0.4425	0.2544	0.1115	0.0461	1.2387	0.8300	0.3417
.7	0.4150	0.2519	0.0991	0.0461	1.1616	0.8218	0.3037
.8	0.3916	0.2479	0.0890	0.0461	1.0962	0.8087	0.2724
.9	0.3714	0.2431	0.0804	0.0461	1.0394	0.7928	0.2463
1.0	0.3535	0.2378	0.0732	0.0461	0.9895	0.7756	0.2241

Tabla 2.2.E. En los últimos cuatro renglones aparecen los estimadores en las unidades originales.

Con ayuda para seleccionar el valor de k aparece la traza de H.K. en la figura 2.2.F.

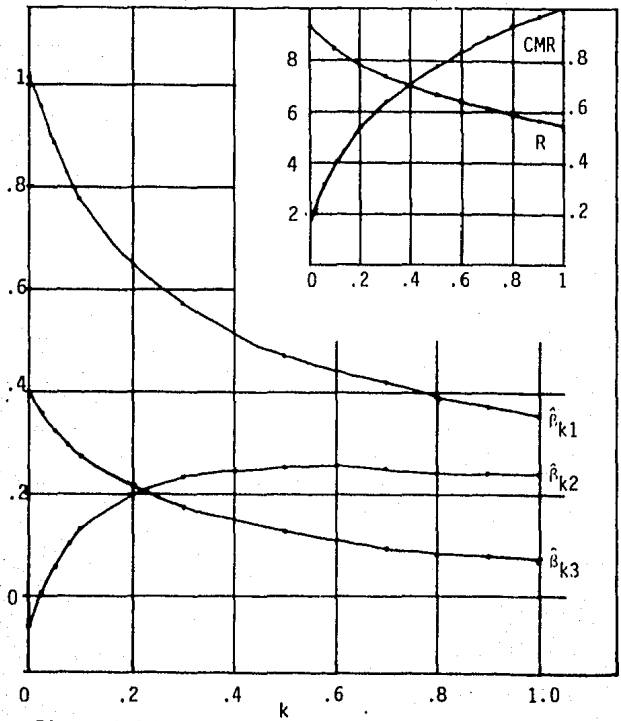


Figura 2.2.F. La traza de H.K. Arriba, a la derecha aparece graficado R, el coeficiente de correlación múltiple así como el cuadrado medio residual (CMR).

Siguiendo los lineamientos de Hoerl y Kennard el análisis de la traza de H.K. es el siguiente:

1. Los estimadores $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ no son estables. Un incremento pequeño en el valor de k trae como consecuencia un cambio brusco en los valores de los estimadores.

2. Dada la naturaleza del fenómeno, $\hat{\beta}_2$ no tiene sentido ya que es -

menor que 0. El incrementar el valor de k de 0 a 0.025 lleva a $\hat{\beta}_{k2}$ a 0 y en adelante se vuelve positivo. Comparemos lo que pasa a $\hat{\beta}_{k2}$ con lo que sucede a $\hat{\beta}_{k1}$. Las variables 1 y 2 tienen un coeficiente de correlación simple de .7684. Por otro lado, la covarianza de $\hat{\beta}_1$ y $\hat{\beta}_2$ es proporcional a -2.7496. Esta covarianza negativa lleva a que estos estimadores tiendan a estar muy alejados entre sí. Como consecuencia de esto, $\hat{\beta}_2$ es negativo y posiblemente $\hat{\beta}_1$ este sobreestimado. La covarianza negativa de $\hat{\beta}_2$ y $\hat{\beta}_3$ contribuye también a que $\hat{\beta}_2$ sea negativo y quizás subestimado.

3. Dada la covarianza positiva de $\hat{\beta}_1$ y $\hat{\beta}_3$ y la negativa de $\hat{\beta}_2$ y $\hat{\beta}_3$ existe cierta tendencia a sobreestimar $\hat{\beta}_3$.

4. Entre k=0 y k=.1 ocurre un cambio brusco en el valor que toman los estimadores. En el intervalo (.2, .4) el sistema se estabiliza. Sin embargo, como el CMR crece considerablemente, es aconsejable elegir un valor de k del intervalo (.1, .2). En este intervalo los signos van de acuerdo con la naturaleza de los datos ya que todos son positivos. Además se reduce tanto la sobreestimación de $\hat{\beta}_{k1}$ como la subestimación de $\hat{\beta}_{k2}$.

Para comparar los estimadores tomamos como criterios

1) $$S_{\underline{b}^*} = \sqrt{\sum_{j=o}^{4} \frac{(b_j^* - b_j)^2}{4}}$$

donde b_j^* es un estimador de b_j (b_j son conocidos por tratarse de una simulación).

2) $$S_{\underline{y}^*} = \sqrt{\sum_{i=1}^{n} \frac{y_i^* - E(y_i)^2}{n}}$$

donde y_i^* es un estimador de $E(Y_i)$. Para k=.1 y k=.2 se presentan estos valores comparados con el de mínimos cuadrados.

	S_{b^*}	S_{y^*}	R
Mínimos Cuadrados	.8216	1.1868	.9339
Hoerl y Kennard (k=.1)	.4684	.9107	.8421
Hoerl y Kennard (k=.2)	.3378	.7965	.7859

Para estos 2 criterios los estimadores H.K. superan al estimador de-mínimos cuadrados a pesar de que el coeficiente de correlación múltiple disminuye. Conviene entonces utilizar los parámetros estimados por el método H.K.- que los estimados por mínimos cuadrados.

2.3. Otros métodos de estimación sesgada

Como consecuencia de las publicaciones de Hoerl y Kennrad han aparecido una gran cantidad de artículos. Es nuestro interés presentar un panorama general de lo que se discute en ellos y para esto hemos intentado hacer una -- clasificación de los mismos. La clasificación la hicimos de acuerdo al enfo-- que particular de cada artículo, esto es, algunos proponen estimadores con propiedades análogas a los de Hoerl y Kennard, otros se orientan a la selección- adecuada de un valor de k, otros dan un enfoque bayesiano, etc. La clasificación hecha fue la siguiente (algunos artículos aparecen en más de un criterio):

a) Propuesta de estimadores con propiedades análogas a $\hat{\beta}_k$, véanse - las referencias (1), (9), (10), (11), (23), (24), (25), (26), (42).

Mc Callum (26) presenta los métodos de estimación por componentes principales- que, si bien no aparecieron como consecuencia de los artículos de Hoerl y - - Kennard, resulta conveniente incluirlos en esta sección para relacionarlos con el método de Marquardt (24).

b) Búsqueda de una manera explícita o heurística de obtener el valor de k, véanse referencias (9), (13), (27), (30).

c) Propuesta de criterios de estimación, véanse las referencias (11) (23), (30), (39).

d) Un enfoque bayesiano de los estimadores sesgados, véanse las referencias (5), (8), (22), (37), (41).

En este capítulo presentaremos los aspectos que consideramos más importantes o representativos de las investigaciones hechas hasta ahora de los - primeros 3 incisos y en el siguiente capítulo presentaremos el inciso d).

Parte de la discusión la aplazamos hasta el capítulo 3.

2.3.a) Propuestas de estimadores sesgados con propiedades análogas a $\hat{\beta}_k$.

2.3.a.1. Estimadores H.K. generalizados y estimadores análogos

Una primera propuesta la hacen Hoerl y Kennard en su artículo original (15), si bien no intentan una aplicación del método. Una simulación en -- donde se aplica este método se encuentra en (9). Sea P una matriz ortogonal tal que reduce a X'X a una matriz diagonal; i.e.

$$P(X'X)\,P' = \Lambda$$

y

$$P'P = PP' = I$$

donde Λ es la matriz diagonal de eigenvalores de X'X. Sea

$$W = XP$$

y $$\underline{\alpha} = P'\underline{\beta}$$

Podemos escribir entonces al modelo

$$\underline{Y} = X\underline{\beta} + \underline{e}$$

como

$$\underline{Y} = X\, PP'\underline{\beta} + \underline{e} \quad = W\underline{\alpha} + \underline{e} \tag{2.3.1}$$

de donde

$$W'W = P'X'XP = \Lambda$$

Definimos entonces el estimador H.K. generalizado como

$$\hat{\underline{\alpha}}^* = (W'W + K)^{-1} W'\underline{Y} \tag{2.3.2}$$

donde K es una matriz diagonal con elementos k_i, i=1,...,p mayores o iguales a cero. Se puede demostrar que esta clase de estimadores satisface la condición de admisibilidad (2.1.3) i.e. $\exists\ \hat{\underline{\alpha}}^*$ tal que

$$ECM(\underline{\alpha}^*) = E(\hat{\underline{\alpha}}^* - \underline{\alpha})'(\hat{\underline{\alpha}}^* - \underline{\alpha}) < E((\hat{\underline{\alpha}} - \underline{\alpha})'(\hat{\underline{\alpha}} - \alpha))$$

con $$\hat{\underline{\alpha}} = (W'W)^{-1} W'\underline{Y}$$

La demostración es muy parecida a la que se hizo para $\hat{\underline{\beta}}_k$ obteniendo, después de un poco de álgebra,

$$ECM(\hat{\underline{\alpha}}^*) = \sum_{i=1}^{p} (\sigma^2\lambda_i + \alpha_i^2 k_i^2)/(\lambda_i + k_i)^2 \tag{2.3.3}$$

Derivando a (2.3.3) con respecto a k_i, i=1,...,p y bajo las restricciones $k_i \geq 0$, i=1,...p se llega a que la condición de admisibilidad se satisface si

$$k_i < \sigma^2/\alpha_i^2 \qquad i = 1,\ldots,p \tag{2.3.4}$$

Otra vez se presenta el problema de que se desconocen σ^2 y $\underline{\alpha}$; por lo tanto Hoerl y Kennard proponen un procedimiento iterativo para obtener valores de k_i. Utilizando la notación de Hemmerle (13), el método que proponen es

$$k_{i(j)} = \hat{\sigma}^2/(\hat{\alpha}^*_{i(j)})^2, \quad i=1,\ldots,p \tag{2.3.5}$$

donde j indica la iteración j-ésima. Se inicia el procedimiento con el estimador de mínimos cuadrados, i.e.

$$k_{i(o)} = \hat{\sigma}^2/(\hat{\alpha}_{i(o)}{}^*)^2 = \hat{\sigma}^2/\hat{\alpha}_i^2$$

La interación continúa hasta que

$$(\hat{\underline{\alpha}}^*{}_{(j+1)})'(\hat{\underline{\alpha}}^*{}_{(j+1)}) \approx (\hat{\underline{\alpha}}^*{}_{(j)})'(\hat{\underline{\alpha}}^*{}_{(j)}) \tag{2.3.5'}$$

De (2.3.5') podemos ver que de hecho el método se basa en la estabilidad de $(\hat{\underline{\alpha}}^{*\prime})(\hat{\underline{\alpha}}^*)$. Hoerl y Kennard argumentan que ya que hay una tendencia a sobreestimar $\underline{\alpha}'\underline{\alpha}$, al introducir $k_i \geq 0$, la longitud del vector disminuye, siendo la nueva longitud más congruente con la realidad. Nosotros criticaremos en el capítulo siguiente el uso de esta noción de estabilidad, en particular argumentaremos que el que $(\hat{\underline{\alpha}}^*)'(\hat{\underline{\alpha}}^*)$ se haya estabilizado no garantiza que las particulares k_i resulten en un estimador con menor ECM que el de mínimos cuadrados.

Dentro del enfoque de Hoerl y Kennard es posible hacer ciertas modificaciones al estimador H.K. generalizado. Por ejemplo, para no modificar tanto $\hat{\underline{\alpha}}$ puede resultar conveniente asignar un valor $k_i=0$ para aquellos coeficientes cuyos correspondientes eigenvalores sean mayores que 1. Tomando en cuenta que $\hat{\alpha}_i \sim N(\alpha_i, \frac{\sigma^2}{\lambda_i})$ tenemos que para $\lambda_i >> 1$ la $V(\hat{\alpha}_i) = \frac{\sigma^2}{\lambda_i}$ será pequeña com-

parada con σ^2, o sea, α_i se estima en una forma relativamente precisa. Esta -- idea de no modificar aquellos coeficientes $\hat{\alpha}_i$ que estén relativamente bien estimados es la que lleva a Guilkey y Murphy (9) a proponer otro tipo de estimadores análogos al estimador H.K. generalizado. Ellos definen que un eigenva-- lor es pequeño si

$$\lambda_i < 10^{-c} \lambda_{max}$$

con $\lambda_{max} = \max_i \lambda_i$ y c una constante arbitraria, típicamente utilizan c=1,2,3. Así, si $\lambda_i \geq 10^{-c} \lambda_{max}$ se asigna $k_i=0$. Resulta claro que estos métodos tampoco garantizan una reducción del ECM; lo único que garantiza (2.3.4) es que - existe un estimador con menor ECM. Lo que sí se puede verificar es que el sesgo introducido es menor que con el estimador H.K. generalizado. En un experimento de simulación que hacen Guilkey y Murphy reportan que tanto el estimador H.K. generalizado como la modificación propuesta dan mejores resultados que el estimador de mínimos cuadrados. El criterio que se utilizó para comparar este estimador con el de mínimos cuadrados fue la razón

$$\frac{(\hat{\underline{\beta}}^* - \underline{\beta})'(\hat{\underline{\beta}}^* - \underline{\beta})}{(\hat{\underline{\beta}} - \underline{\beta})'(\hat{\underline{\beta}} - \underline{\beta})}$$

donde $\hat{\underline{\beta}}^* = P\hat{\underline{\alpha}}^*$

La modificación hecha produjo resultados ligeramente superiores a -- las del estimador H.K. generalizado.

2.3.a.2. Estimadores de Marquardt

Otra clase importante de estimadores sesgados es la propuesta por - Marquardt (24). El propuso un método de estimación utilizando inversas generalizadas y encontró que esta clase de estimadores tiene propiedades muy simila-

res a las de los H.K. Como es sabido, se pueden utilizar inversas generalizadas en modelos lineales de rango incompleto. Ahora bien, cuando se tiene el problema de multicolinealidad, se trabaja con matrices que distan poco de ser singulares. Marquardt dice que los eigenvalores de X'X se pueden agrupar cualitativamente dentro de 3 clases: mayores que cero, ligeramente mayores que 0 y precisamente 0. Argumenta que cuando X'X es casi singular, resulta natural tratar de invertirla utilizando inversas generalizadas. Cuando algunos eigenvalores son pequeños, dice Marquardt, no existe un rango claramente asignable a la matriz. Marquardt generaliza el concepto de rango y permite que el rango ρ sea una variable continua en $0 \leq \rho \leq p$. Para aclarar lo anterior, supongamos que X'X es de rango $r \leq p$, entonces si $\lambda_1 \geq \ldots \geq \lambda_r$ y $\underline{P}_1, \underline{P}_2, \ldots \underline{P}_r$ son los eigenvalores y eigenvectores correspondientes de X'X, de A.7:

$$X'X = \sum_{i=1}^{r} \lambda_i \, \underline{P}_i \underline{P}_i' \qquad (2.3.6)$$

y de A.13

$$(X'X)^{+} = \sum_{i=1}^{r} \frac{1}{\lambda_i} \, \underline{P}_i \underline{P}_i' \qquad (2.3.7)$$

es una inversa generalizada de X'X (ver. A.11). De hecho, se puede comprobar que (2.3.7) es la inversa generalizada de Moore-Penrose (ver A.12).

Sea ρ^* la parte entera de ρ y sea $d\rho = \rho - \rho^*$ Definimos

$$(X'X)\rho^{+} = \sum_{j=1}^{\rho^*} \frac{1}{\lambda_j} \underline{P}_j \underline{P}_j' + \frac{d\rho}{\lambda_{\rho^*+1}} \underline{P}_{\rho^*+1} \underline{P}_{\rho^*+1}' \qquad (2.3.8)$$

y definimos la clase de estimadores de inversas generalizadas (o de Marquardt) como:

$$\hat{\underline{\beta}}_{\rho}^{+} = (X'X)_{\rho}^{+} \quad X'\underline{Y} \qquad (2.3.9)$$

El nombre de inversas generalizadas se debe a que si ρ fuera un entero y X'X de rango ρ, $(X'X)^{+}_{\rho}$ sería una inversa generalizada de X'X. Algunos de los teoremas que presenta Marquardt son los siguientes:

Teorema 1

Sea $\hat{\underline{\beta}}^{+}_{r}$ la solución a las ecuaciones normales suponiendo que X'X es de rango r. $\hat{\underline{\beta}}^{+}_{r}$ minimiza

$$\phi\ (\underline{B}) = (\underline{Y} - X\ \underline{B})'\ (\underline{Y} - X\ \underline{B})$$

dentro del espacio r-dimensional generado por los eigenvectores $\underline{P}_1, \underline{P}_2, \ldots \underline{P}_r$.- La demostración es inmediata usando la suposición de que X'X es de rango r.

Teorema 2

$\left|\left|\hat{\underline{\beta}}^{+}_{\rho}\right|\right|^2 = (\hat{\underline{\beta}}^{+}_{\rho})'(\hat{\underline{\beta}}^{+}_{\rho})$ es una función creciente en ρ, continua en pedazos. Este teorema es análogo a (2.2.7) donde $(\hat{\underline{\beta}}_{k}')(\hat{\underline{\beta}}_{k})$ era una función decreciente en k.

Teorema 3

$\hat{\underline{\beta}}^{+}_{\rho}$ es una transformación lineal de $\hat{\underline{\beta}}$ y depende solo de X y ρ, i.e.

$$\hat{\underline{\beta}}^{+}_{\rho} = Z_{\rho}\hat{\underline{\beta}} \qquad (2.3.10)$$

(Análogo a (2.2.2)

Corolario 3.1.

$$\hat{\underline{\beta}}^{+}_{\rho} \sim N(Z_{\rho}\underline{\beta},\ \sigma^2 Z_{\rho}\ (X'X)^{-1}\ Z_{\rho}') \qquad (2.3.11)$$

$$\text{y además } \sigma^2\ Z_{\rho}\ (X'X)^{-1}\ Z_{\rho}' = \sigma^2\ (X'X)^{+}_{\rho}$$

(Análogo a (2.2.3)

Teorema 4

El error cuadrático medio de $\hat{\underline{\beta}}_{\rho}^{+}$ es

$$ECM(\hat{\underline{\beta}}_{\rho}^{+}) = tr(V(\hat{\underline{\beta}}_{\rho}^{+})) + \underline{\beta}'(Z_{\rho} - I)'(Z_{\rho} - I)\underline{\beta}$$

Corolario 4.1.

$tr(V(\hat{\underline{\beta}}_{\rho}^{+}))$ es una función creciente en ρ.

(Análogo al teorema 2, sección 2.2)

Corolario 4.2.

$\underline{\beta}'(Z_{\rho} - I)'(Z_{\rho} - I)\underline{\beta}$ es una función monótona decreciente en ρ.

(Análogo al teorema 3, sección 2.2)

Teorema 5.

Una condición suficiente para que se satisfaga la condición de admisibilidad (2.1.3) es que para alguna $r < p$.

$$\sum_{j=r+1}^{p} \frac{1}{\lambda_j} > \frac{1}{\sigma^2}(\underline{\beta}'\underline{\beta}) \qquad (2.3.12)$$

(Análogo al teorema 4, sección 2.2)

Las demostraciones de estos teoremas se pueden encontrar en (24).

Otra vez nos encontramos con el problema de que no conocemos $\underline{\beta}'\underline{\beta}$ y σ^2. Para ilustrar geométricamente estos resultados tomemos el ejemplo que utiliza Marquardt (24).

Sea $\lambda_1 = 1.98$, $\lambda_2 = .02$

$$\underline{P}_1 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{pmatrix} \qquad \underline{P}_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{pmatrix} \qquad X'\underline{Y} = \begin{pmatrix} \frac{26}{5}\sqrt{\frac{1}{2}} \\ 5\sqrt{\frac{1}{2}} \end{pmatrix}$$

$$\underline{P}_1\underline{P}_1' = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} \qquad \underline{P}_2\underline{P}_2' = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

Así, si $\rho \geq 1$

$$\hat{\underline{\beta}}_\rho^+ = \left[\frac{1}{1.98} \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix} + \frac{d\rho}{.02} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{pmatrix} \quad X'\underline{Y} \right]$$

$$= \begin{pmatrix} (\frac{1}{3.96} + \frac{d\rho}{.04}) \frac{26}{5}\sqrt{\frac{1}{2}} + (\frac{1}{3.96} + \frac{d\rho}{.04}) 5\sqrt{\frac{1}{2}} \\ (\frac{1}{3.96} - \frac{d\rho}{.04}) \frac{26}{5}\sqrt{\frac{1}{2}} + (\frac{1}{3.96} - \frac{d\rho}{.04}) 5\sqrt{\frac{1}{2}} \end{pmatrix}$$

Cuando $\rho = 2$ tenemos la solución de mínimos cuadrados:

$$\hat{\underline{\beta}}_2^+ = \hat{\underline{\beta}} = \begin{pmatrix} 5.3569 \\ -1.7142 \end{pmatrix}$$

Si $\rho = 1$

$$\hat{\underline{\beta}}_1^+ = \begin{pmatrix} \frac{1}{3.96} \frac{26}{5}\sqrt{\frac{1}{2}} + \frac{1}{3.96} 5\sqrt{\frac{1}{2}} \\ \frac{1}{3.96} \frac{26}{5}\sqrt{\frac{1}{2}} + \frac{1}{3.96} 5\sqrt{\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} 1.8213 \\ 1.8213 \end{pmatrix}$$

Notemos que $\underline{P}_1$ y $\underline{P}_2$ siguen siendo los mismos para diferentes valores de ρ, mientras que con $\hat{\underline{\beta}}_k$ se generan nuevos eigenvectores para cada k. Del corolario 3.1. los elipsoides de concentración para una α dada, se ven como en -

la figura 2.3.A.

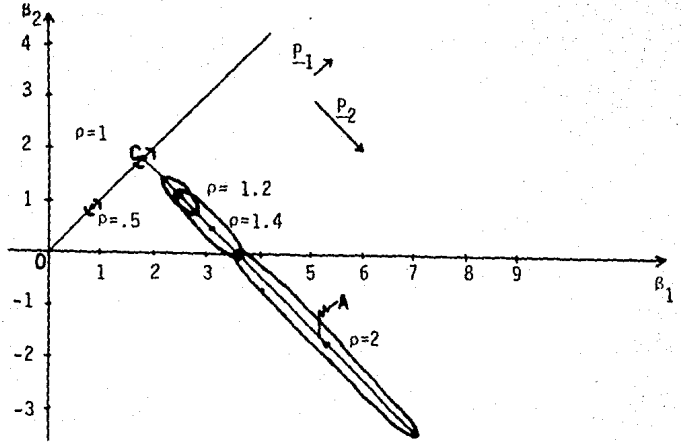


Figura 2.3.A. Elipsoides de concentración de $\underline{\hat{\beta}}_p^+$ para ρ = .5, 1, 1.2, 1.4 y 2 a un mismo nivel α . En la recta AC están todas las soluciones para ρ si $1 \leq \rho \leq 2$. En la recta OC están todas las soluciones para ρ si $-0 \leq \rho \leq 1$. Si $\rho \leq 1$ tenemos elipsoides degeneradas, estoes, las soluciones están en un subespacio lineal (una recta que pasa por el origen en este caso).

Los resultados para varios valores de son los siguientes:

ρ	β_1^+	β_2^+	$(\underline{\hat{\beta}}_p^+)'(\underline{\hat{\beta}}_p^+)$
2	5.3569	-1.7142	5.62
1.6	3.9427	-.300	3.95
1.4	3.2356	.4071	3.26
1.2	2.5284	1.1142	2.76
1	1.8213	1.8213	2.58
.5	.9107	.9107	1.29

Como veremos a continuación el método de Marquardt es una generalización de los métodos de estimación sesgada que utilizan componentes principales.

2.3.a.3. Estimadores de Componentes Principales

Los métodos de estimación usando componentes principales son anteriores a los otros métodos de estimación sesgada, sin embargo, resulta pertinente incluirlos aquí para relacionarlos con el de Marquardt. Sea

$$\underline{Y} = X\underline{\beta} + e$$

y sea

$$W = XP$$

con $PP' = P'P = I$,

$$\underline{\alpha} = P'\underline{\beta} \qquad (2.3.13)$$

y P tal que $X'X = P\Lambda P'$

Se estima entonces el vector de parámetros $\underline{\alpha}$ por mínimos cuadrados.

$$\hat{\underline{\alpha}} = (W'W)^{-1} W'\underline{Y} \qquad (2.3.14)$$
$$= \Lambda^{-1} W'\underline{Y}$$

Como Λ es una matriz diagonal, la covarianza entre dos componentes del vector $\hat{\underline{\alpha}}$ es 0.

Existen varios métodos para obtener estimadores de $\underline{\beta}$ (26) (35). El primer método consiste en asignar un valor de 0 o aquellos coeficientes del vector $\hat{\underline{\alpha}}$ que corresponden a eigenvalores "pequeños". Para aclarar el significado de "pequeños" recordemos que

$$\hat{\underline{\alpha}} \sim N_p(\underline{\alpha}, \sigma^2\Lambda^{-1})$$

o bien, $\hat{\alpha}_i \sim N(\alpha_i, \frac{\sigma^2}{\lambda_i})$, $i=1,..,p$ (2.3.15)

Entonces, si λ_i tiende a 0, la varianza de $\hat{\alpha}_i$ tenderá a ∞. Una vez que se han sustituido por 0 a d coeficientes, $0 \leq d < p$, se tiene el estimador

$$\hat{\underline{\alpha}}(d) \tag{2.3.16}$$

y se transforma a unidades originales mediante la fórmula

$$\hat{\underline{\beta}}(d) = P\hat{\underline{\alpha}}(d) \tag{2.3.17}$$

Un segundo método es probar las hipótesis

$$\alpha_i = 0 \qquad i = 1,\ldots,p$$

que pueden hacerse fácilmente con una t de Student. Cuando no se rechace la hipótesis para alguna i, se sustituye $\hat{\alpha}_i$ por 0. Suponiendo que no se recha-- zan d hipótesis nos queda

$$\hat{\underline{\alpha}}(d) \tag{2.3.18}$$

donde d coeficientes, $0 < d < p$ son iguales a 0. Nótese que (2.3.16) y (2.3.18) no tienen por qué ser los mismos, si bien es de esperarse que algunos coefi- - cientes igualados a 0 coincidan. La transformación a unidades originales es - otra vez

$$\hat{\underline{\beta}}\ (d) = P\ \hat{\underline{\alpha}}$$

Vamos ahora a ver como se relaciona el primer método de componentes- con el de Marquardt cuando ρ es un entero. Si ρ es un entero

$$(X'X)_\rho = \sum_{j=1}^{\rho} \lambda_j \underline{P}_j \underline{P}_j'$$

$$(X'X)_\rho^+ = \sum_{j=1}^{\rho} \frac{1}{\lambda_j} \underline{P}_j \underline{P}_j' \tag{2.3.19}$$

(2.3.19) equivale a tener

$$X'X = P\Lambda P'$$

y asignar un valor 0 a $p - \rho = d$ elementos de la matriz Λ o sea,

$$\Lambda = \begin{pmatrix} \lambda_1 & & & & & \\ & \lambda_2 & & & & \\ & & \ddots & & & \\ & & & \lambda_\rho & & \\ & & & & \lambda_{\rho+1} & \\ & & & & & \ddots \\ & & & & & & \lambda_p \end{pmatrix}$$

y

$$\Lambda_\rho = \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_\rho & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

Por otro lado, hacer 0 a d coeficientes α_i equivale a

$$\Lambda_\rho \, \hat{\underline{\alpha}}(d) = W'\underline{Y}$$

y una inversa generalizada de Λ_ρ es

$$\Lambda_\rho^- = \begin{pmatrix} \frac{1}{\lambda_1} & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \ddots & & & & & & \vdots \\ \vdots & & \ddots & & & & & \vdots \\ \vdots & & & \frac{1}{\lambda_\rho} & & & & \vdots \\ \vdots & & & & 0 & & & \vdots \\ \vdots & & & & & \ddots & & \vdots \\ \vdots & & & & & & 0 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0 \end{pmatrix}$$

de donde se obtiene

$$\hat{\underline{\alpha}}(d) = \Lambda_\rho^- \, W'\underline{Y} = \Lambda_\rho^- \, P'X'\underline{Y}$$

por lo tanto

$$\hat{\underline{\beta}}(d) = P\hat{\underline{\alpha}}(d) = P\Lambda_\rho^- \, P'X'\underline{Y} = (X'X)_p^+ X'\underline{Y} = \hat{\underline{\beta}}_p^+$$

Mc Callum (26) sugiere que es deseable hacer un coeficiente igual a 0 si y solo si se reduce el ECM($\underline{\hat{\beta}}(d)$). Podríamos agregar que es deseable hacer un coeficiente igual a 0 si y solo si se cumple la condición de admisibilidad (2.1.3). Debido a que $\underline{\hat{\beta}}(d)$ es equivalente a $\underline{\hat{\beta}}(\rho)$, ρ entero, vemos que una condición suficiente para el criterio (2.1.3) depende tanto de σ^2 y $\underline{\beta}'\underline{\beta}$. Para el segundo método de estimación por componentes principales es de esperarse que se encuentre una condición semejante a (2.3.12). Las referencias acerca del desempeño que han tenido en la práctica los estimadores de componentes principales son (3) y (26); para los estimadores de Marquardt se puede consultar (24) y (25).

Otro método que podemos enmarcar dentro de los métodos de componentes principales es el método de Gunst, Webster y Mason (10),(42).

Ellos presentan un método muy ingenioso para decidir qué eigenvalores eliminar tomando en cuenta también los eigenvectores correspondientes. Los resultados de las simulaciones son ventajosos a su método aunque si la longitud de $\underline{\beta}$ se incrementa demasiado el método de mínimos cuadrados llega a superar este método.

2.3.a.4. Estimadores contraídos.

Mayer y Willke (23) presentan varias clases de estimadores que satisfacen la condición de admisibilidad y se caracterizan por un parámetro c, $0 < c < 1$, tal que cada elemento de alguna de estas clases es de la forma

$$\underline{\hat{\beta}}(c) = c(X'X)^{-1} X'\underline{Y} = c \ \underline{\hat{\beta}} \qquad (2.3.20)$$

Mayer y Willke los llaman estimadores contraídos (Shrunken estimators) y a c el factor de contracción. Si c es un número real fijo los llaman estimadores -

contraídos determinísticos; si c es función de $\hat{\beta}'\hat{\beta}$, $c = f(\hat{\beta}'\hat{\beta})$, los llaman - estimadores contraídos estocásticos. Presentan el siguiente:

Teorema

La clase de estimadores contraídos determinísticos es admisible.

Demostración

$$ECM(\underline{\hat{\beta}}(c)) = E((c\underline{\hat{\beta}} - E(c\underline{\hat{\beta}}))'(c\underline{\hat{\beta}} - E(c\underline{\hat{\beta}})) + E((c\underline{\hat{\beta}} - \underline{\beta})'(c\underline{\hat{\beta}} - \underline{\beta}))$$

$$= c^2\, ECM(\underline{\hat{\beta}}) + (1 - c)^2\, \underline{\beta}'\underline{\beta}$$

Por lo tanto

$$ECM(\underline{\hat{\beta}}(c) < ECM\ (\underline{\hat{\beta}})$$

se cumple si y solo si

$$c > \frac{\underline{\beta}'\underline{\beta} - ECM(\underline{\hat{\beta}})}{\underline{\beta}\ \underline{\beta} + ECM(\underline{\hat{\beta}})} \qquad (2.3.21)$$

Este teorema justifica parcialmente el uso de esta clase de estimadores.

Puede resultar sorprendente que esta clase de estimadores con formatan sencilla sea admisible (ver figura 2.3.B).

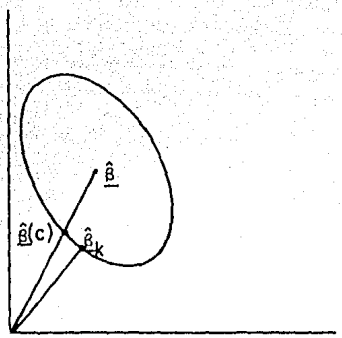


Figura 2.3.B. $\hat{\beta}_k$ y $\hat{\beta}(c)$ para un mismo incremento en la SCR, es decir $(\beta - \hat{\beta})'X'X(\beta - \hat{\beta})$= constante (Tomado de Mayer y Willke (23)).

Una pregunta surge naturalmente: ¿ es la condición de admisibilidad un criterio adecuado para escoger una clase de estimadores? Más adelante trataremos de responderla.

Para atacar el problema de encontrar un valor de c adecuado, Mayer y Willke proponen una clase de estimadores estocásticos contraídos en donde cada miembro de la clase depende tanto de $\hat{\beta}'\hat{\beta}$ como de un parámetro $\delta \geq 0$, es decir, proponen c_δ tal que

$$c_\delta = f(\hat{\beta}'\hat{\beta}, \delta) = \delta\left(\hat{\beta}'\hat{\beta} + (1 + \delta\,\hat{\beta}'\hat{\beta})^{-1}\,\delta(\hat{\beta}'\hat{\beta})^2\right) \qquad (2.3.22)$$

y

$$\hat{\beta}(c_\delta) = c_\delta\,\hat{\beta}$$

La única propiedad conocida de esta clase de estimadores es que de todas las transformaciones lineales que dan un mismo incremento en la SCR existe una $\delta>0$ tal que la varianza del estimador es mínima. A la expresión (2.3.22) se llegó pidiendo precisamente esta propiedad. Mayer y Willke proponen utilizar el criterio de estabilización propuesto por Hoerl y Kennard para esco

ger un valor de δ, el cual no se puede utilizar en los estimadores contraídos-determinísticos (Nótese que c_δ no crece linealmente conδ).

Por último, Mayer y Willke presentan los estimadores de James y Stein:

$$\underline{\hat{\beta}}(c_\xi) = c_\xi \hat{\beta} \qquad (2.3.23)$$

donde

$$c_\xi = \left(1 - \xi s^2 (\underline{\hat{\beta}}' \underline{\hat{\beta}})^{-1}\right)$$

y $s^2 = \underline{Y}'\underline{Y} - \underline{\hat{\beta}}(X'X)^{-1}\underline{\hat{\beta}}$

con $0 < \xi \leq 2(p-2)/(n-p+2) \qquad p \geq 3.$

Sclove (23), (35) demostró que estos estimadores son superiores al - de mínimos cuadrados tomando como criterio el error cuadrático medio ponderado (ECMP) definido como

$$ECMP(\underline{B}) = E((\underline{B} - \underline{\beta})'X'X(\underline{B} - \underline{\beta})) \qquad (2.3.24)$$

o sea, demostró que

$$ECMP\ (\underline{\hat{\beta}}(c_\xi)) < ECMP(\underline{\hat{\beta}})$$

y además demostró que $ECMP(\hat{\beta}(c_\xi))$ es mínimo si $\xi = (p-2)/(n-p+2)$. Entonces, a diferencia de los estimadores sesgados presentados hasta ahora, en esta clase de estimadores sabemos exactamente cuáles estimadores son superiores al de mínimos cuadrados según un criterio, en este caso el ECMP.

En la siguiente sección se presentan métodos de obtención del valor de k, con lo cual de hecho se están generando nuevos estimadores, si bien no - difieren mucho de los presentados en esta sección.

2.3.b) Obtención del valor de k

El obstáculo principal para encontrar un valor de k que nos garantice que el estimador seleccionado tiene menor ECM que los de mínimos cuadrados- es que k depende tanto de $\underline{\beta}$ como de σ^2 que son desconocidos. Una parte de la- investigación en torno a estos estimadores sesgados ha girado en torno a este- problema (9), (13), (27), (30). Inclusive autores que han trabajado con otros estimadores sesgados se han preocupado porque la condición de admisibilidad de- pende de parámetros desconocidos como es el caso de Marquardt (24) y de Mayer- y Willke (23).

El trabajo de Mc Donald y Galarneau (27) es un ejemplo típico de la- preocupación por encontrar un valor adecuado de k. A lo que llegan ellos es a unas reglas empíricas que dan una guía para la selección de k. Un ejemplo de- las reglas que utilizan es la siguiente:

Sea

$$Q = \hat{\underline{\beta}}'\hat{\underline{\beta}} - \hat{\sigma}^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

Regla. Escoja k tal que $\hat{\underline{\beta}}_k' \hat{\underline{\beta}}_k = Q$ si $Q>0$. Si $Q\leq 0$ escoja $k=k'$ don- de k' es un valor prefijado. Este valor de k' puede ser por ejemplo el selec- cionado por la traza H.K. Ellos toman k=0 y $k=\infty$ que correspondería a mínimos- cuadrados el primero y $\underline{0}$ en el segundo caso. Una justificación de estas re- - glas es que Q estima a $\underline{\beta}'\underline{\beta}$ que es una cantidad positiva. Entonces tratamos -- que $\hat{\underline{\beta}}_k' \hat{\underline{\beta}}_k$ también estime a $\underline{\beta}'\underline{\beta}$ igualándola a Q.

Mc Donald y Galarneau concluyen que en realidad no existe una regla - para escoger un valor de k óptima y suponen que sus reglas junto con la traza-

H.K. pueden servir para hacer una selección adecuada.

Guilkey y Murphy (9) hacen una modificación del método H.K. generalizado que ya expusimos en la sección 2.3.

La selección de los valores de k_i, tanto en su método como en el H.-K. generalizado se basaba en la desigualdad (2.3.4)

$$k_i < \frac{\sigma^2}{\alpha_i^2}$$

A partir de los estimadores iniciales $\hat{\sigma}^2$ y $\hat{\alpha}_i^2$ se iniciaba un proceso iterativo (2.3.5) hasta que (2.3.5') ocurre, o bien, hasta que

$$k_{i(j)} \tilde{=} k_{i(j+1)} \qquad (2.3.25)$$

de acuerdo a la notación (2.3.5). Hemmerle (13) encuentra condiciones bajo las cuales existe

$$\lim_{j \to \infty} \hat{\alpha}_i^*(j) = \hat{\alpha}^*$$

y encuentra la forma de evaluar $\hat{\alpha}_i^*$ explícitamente. Supone con esto que los valores que él encuentra son "óptimos".

Estos artículos ilustran el enfoque de las investigaciones. En el último capítulo haremos una crítica de dichos artículos.

2.3.c) Criterios de estimación propuestos

Theobald (39) propuso como criterio de estimación una generalización del ECM que es el ECM generalizado definido como

$$ECMG(\underline{B}) = E((\underline{B} - \underline{\beta})(\underline{B} - \underline{\beta})') \qquad (2.3.26)$$

donde $\underline{B}$ es un estimador de $\underline{\beta}$ y demostró que existe un estimador $\hat{\underline{\beta}}_k$ que supera a $\hat{\underline{\beta}}$ de acuerdo a este criterio si

$$k < \frac{2\sigma^2}{\underline{\beta}'\underline{\beta}} \qquad (2.3.27)$$

Además demostró que si un estimador es superior a otro de acuerdo a (2.3.-26) entonces también lo es de acuerdo al criterio

$$ECM_A(\underline{B}) = E((\underline{B} - \underline{\beta})'A(\underline{B} - \underline{\beta}) \qquad (2.3.28)$$

con A una matriz no negativa definida. En particular, si A=I tenemos el ECM y - si A=X'X tenemos el ECMP que usamos en (2.3.24). En consecuencia, existe $\hat{\underline{\beta}}_k$ con ECMP menor que $\hat{\underline{\beta}}$.

Farebrother (11) exhibió el "estimador" de Theil

$$\hat{\underline{\beta}}^* = \underline{\beta}'\underline{\beta}\, X'(X\underline{\beta}\underline{\beta}'X' + \sigma^2 I_n)^{-1}\, \underline{Y} \qquad (2.3.28')$$

que es el estimador con menor ECM dentro de la clase de estimadores lineales de $\underline{\beta}$. Farebrother reconoce que el estimador no es operacional y después de demostrar que (2.3.28') es igual a

$$\hat{\underline{\beta}}^* = \frac{\underline{\beta}'X'\underline{Y}}{\sigma^2 + \underline{\beta}'X'X\underline{\beta}}\ \underline{\beta}$$

propone como estimador

$$\hat{\underline{\beta}}^{**} = \frac{\hat{\underline{\beta}}'X'\underline{Y}}{s^2 + \hat{\underline{\beta}}'X'X\hat{\underline{\beta}}}\ \hat{\underline{\beta}} \qquad (2.3.29)$$

y sugiere que este último puede reemplazar al estimador que se obtenga utilizando el método subjetivo de Hoerl y Kennard, ya que $\hat{\underline{\beta}}^{**}$ es una aproximación al estimador con menor ECM.

Mayer y Willke (23) presentan también ejemplos en donde dentro de --

cierta clase de estimadores se busca uno con propiedades óptimas, por ejemplo,- mínima varianza.

El problema que se presenta con este tipo de enfoque lo podemos ver - en (2.3.27) y (2.3.28').

Los estimadores dependen de parámetros desconocidos, exceptuando (2.- 3.23), en donde se exhibe un estimador con menor ECMP que $\underline{\beta}$. Respecto a (2.3.- 29) se conocen muy pocas propiedades.

C A P I T U L O III

3.1. Inferencia bayesiana.

La aplicación del teorema de Bayes a problemas de inferencia estadística ha sido y es aún muy discutida. Un punto que se discute es, que mientras en la inferencia estadística clásica los parámetros son desconocidos pero fijos, en la inferencia bayesiana se supone que los parámetros tienen una distribución probabilística y que esta distribución refleja el grado de conocimiento del investigador con respecto a los parámetros, es decir, es una distribución subjetiva. Dentro de la corriente bayesiana se argumenta que es bastante frecuente tener algún conocimiento a priori acerca de los parámetros, o bien, que existe una ignorancia relativa y no absoluta de los mismos (2) (22). Este conocimiento a priori puede ayudar a mejorar las estimaciones, sobre todo en aquellos casos en que las observaciones no nos dan suficiente información acerca de los parámetros. La manera de incorporar al modelo el conocimiento a priori o la ignorancia relativa es introduciendo una distribución a priori para los parámetros. Esta es la parte crucial bajo discusión, porque la distribución a priori es subjetiva ya que depende del grado de conocimiento o ignorancia del investigador.- Se teme que la prior distorsione la información que nos dan las observaciones acerca de los parámetros. Box y Tiao (2) tratan de mostrar en su libro que utilizando distribuciones a priori adecuadas y escogiendo cuidadosamente la estructura del modelo se puede llegar a un mejor conocimiento del mismo.

Recordemos el teorema de Bayes. Sea A un evento y $\{H_i\}i=1, N$ una partición del espacio muestral Ω. El teorema de Bayes dice que

$$P(H_i|A) = \frac{P(H_i)\ P(A|H_j)}{P(A)} = \frac{P(H_j)\ P(A|H_j)}{\sum_{i=1}^{N} P(H_i)\ P(A|H_i)} \qquad (3.1.1)$$

En el caso continuo tenemos que si X y Y son variables aleatorias, entonces la función de densidad de X dado Y=y está dada por

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{f_X(x)\ f_{Y|X}(y|x)}{f_Y(y)} \qquad (3.1.2)$$

donde

$$f_Y(y) = \int_{-\infty}^{-\infty} f(x,y)dx = \int_{-\infty}^{\infty} f_X(x) f_{Y|X}(y|x)dx$$

es un valor fijo. El análogo multivariado de (3.1.2) es tomar vectores $\underline{X}$ y $\underline{Y}$ y encontrar la función de densidad de $\underline{X}$ dado $\underline{Y} = \underline{y}$:

$$f_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y}) = \frac{f_{\underline{X},\underline{Y}}(\underline{x},\underline{y})}{f(\underline{Y})} = \frac{f_{\underline{X}}(\underline{x})\ f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})}{f_{\underline{Y}}(\underline{y})} \qquad (3.1.3)$$

en este caso

$$f_{\underline{Y}}(\underline{y}) = \int_R f(\underline{x}, \underline{y})d\underline{x} = \int_R f_{\underline{X}}(\underline{x})\ f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})d\underline{x}$$

donde R es la región de integración del vector $\underline{X}$. También aquí $f_{\underline{Y}}(\underline{y})$ tiene un valor fijo en la fórmula (3.1.3). (3.1.2) y (3.1.3) son los equivalentes del teorema de Bayes en el caso continuo.

Pasemos a la aplicación del teorema de Bayes a problemas de inferencia. Sean

$\underline{Y}$ un vector de n variables aleatorias y

$\underline{\theta}$ un vector de p parámetros, tomados como variables aleatorias.

De (3.1.3) tenemos que

$$f_{\underline{\theta}|\underline{Y}}(\underline{\theta}|\underline{y}) = \frac{f_{\underline{\theta}}(\underline{\theta})\, f_{\underline{Y}|\underline{\theta}}(\underline{y}|\underline{\theta})}{f_{\underline{Y}}(\underline{y})} \qquad (3.1.4)$$

Notemos que $f_{\underline{Y}}(\underline{y})$ no depende de $\underline{\theta}$ y que la función de densidad posterior $f_{\underline{\theta}|\underline{Y}}(\underline{\theta}|\underline{y})$ depende de $\underline{Y}$ sólo a través de $f_{\underline{Y}|\underline{\theta}}(\underline{y}|\underline{\theta})$. Como $\underline{y}$ es fijo, $f_{\underline{Y}|\underline{\theta}}(\underline{y}|\underline{\theta})$ es función de θ, de hecho, es la función que Fisher llamó función de verosimilitud. $f_{\underline{\theta}}(\underline{\theta})$ representa el conocimiento a priori que tenemos de $\underline{\theta}$. A través de la función de verosimilitud modificamos el conocimiento a priori con la información proveniente de los datos. En general, buscamos obtener más información a través de la función de verosimilitud de la que nos pueda proporcionar un conocimiento a priori, en este sentido, buscamos que la verosimilitud domine a la prior, lo que se ilustra en la figura 3.1.A.

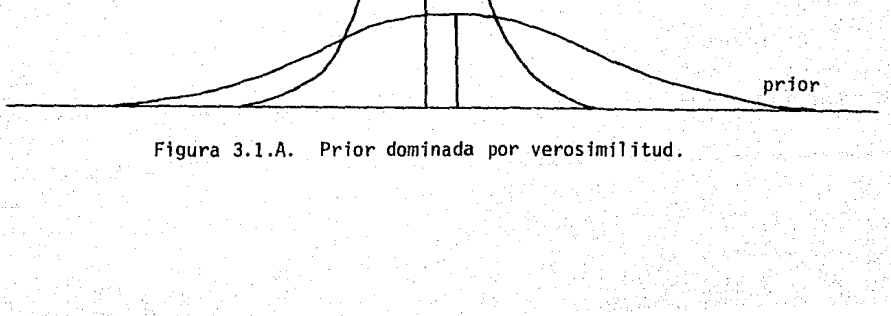


Figura 3.1.A. Prior dominada por verosimilitud.

Dado que $\underline{y}$ es fijo, podemos escribir a (3.1.4) como

$$f_{\underline{\theta}|\underline{Y}}(\underline{\theta}|\underline{y}) = c\, f_{\underline{\theta}}(\underline{\theta}) L(\underline{\theta}|\underline{Y}) \qquad (3.1.5)$$

donde c es una constante y L la función de verosimilitud.

Para ilustrar la forma de obtener la función de densidad posterior - se presenta un ejemplo. Supongamos que tenemos $Y_1,\ldots,Y_n$ variables aleatorias independientes distribuídas según una $N(\mu, \sigma^2)$ con σ^2 conocida y queremos obtener la función de densidad posterior para μ.

Si suponemos que a priori

$$\mu \sim N(m, \sigma^2) \qquad (3.1.6)$$

estamos asignando un mayor peso a un rango de valores cercanos a m.

La función de verosimilitud para este ejemplo está dada por

$$L(\mu|\sigma^2, \underline{Y}) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\left(\sum_1^n (y_i - \mu)^2\right)}$$

$$= \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum y_i^2} \; e^{-\frac{1}{2\mu^2}(n\mu^2 - 2\mu\sum y_i)}$$

y combinándola según (3.1.5) con la distribución a priori (3.1.6) obtenemos

$$f_{\mu|\sigma^2,\underline{Y}}(\mu|\sigma^2, \underline{y}) = c\, f_{\mu}(\mu)\, L(\mu|\sigma^2, \underline{Y})$$

$$\propto e^{-\frac{1}{2\sigma_1^2}(\mu^2 - 2m\mu)} \; e^{-\frac{1}{2\sigma^2}(n\mu^2 - 2\mu\sum y_i)} \qquad (3.1.7)$$

$\propto$ indica proporcionalidad; hemos dejado únicamente los términos que dependen - de μ ya que los demás son constantes. La idea ahora es identificar la función-

la función de densidad posterior a partir de (3.1.7). Pero el lado derecho en (3.1.7) rearreglando términos es igual a

$$e^{-\frac{1}{2}\left[\left(\frac{1}{\sigma_1^2}+\frac{1}{\frac{\sigma^2}{n}}\right)\mu^2 \quad -2\left(\frac{n\bar{y}}{\sigma^2}+\frac{m}{\sigma_1^2}\right)\mu\right]}$$

$$= e^{-\frac{1}{2}\left(\frac{1}{\sigma_1^2}+\frac{1}{\frac{\sigma^2}{n}}\right)\left[\mu^2 - 2\left(\frac{\frac{n\bar{y}}{\sigma^2}+\frac{m}{\sigma_1^2}}{\frac{1}{\sigma_1^2}+\frac{1}{\frac{\sigma^2}{n}}}\right)\mu\right]} \qquad (3.1.8)$$

Completando cuadrados en (3.1.8) llegamos a que μ se distribuye a posteriori según una

$$N\left(\frac{\frac{1}{\sigma_1^2}m+\frac{n}{\sigma^2}\bar{y}}{\frac{1}{\sigma_1^2}+\frac{n}{\sigma^2}}, \quad \frac{1}{\frac{1}{\sigma_1^2}+\frac{n}{\sigma^2}}\right) \qquad (3.1.9)$$

En consecuencia, la media posterior es una ponderación de la media prior y el estimador máximo verosímil con pesos inversamente proporcionales a las varianzas.

El resultado es bastante interesante ya que a medida que aumentamos el tamaño de muestra, el peso asignado a la verosimilitud será más grande. A la larga el peso asignado a la prior será despreciable. La media de la prior viene a ser una especie de "observación hipotética" a la cual secuencialmente le restamos valor conforme vamos teniendo más información experimental. Las observaciones que hagamos tenderán a que reafirmemos o restemos validez a nuestro conocimiento a priori. En los casos en que las observaciones sean insuficientes, una distribución prior adecuada puede ser de mucha utilidad como veremos

en la sección 3.2.

Ahora veamos que sucede si quisiéramos representar con una distribución a priori que el conocimiento previo acerca del parámetro es nulo o prácticamente nulo relativo a la información proveniente de los datos. Entonces, (3.1.9) sería aproximadamente una

$$N(\bar{y}, \frac{\sigma^2}{n}) \qquad (3.1.10)$$

debido a que σ_1^2 sería muy grande. De hecho, la media en (3.1.10) es el estimador de máxima verosimilitud, es decir, toda la información proviene de las observaciones. Ahora, para formar la distribución posterior (3.1.10) es necesario suponer una prior que asigne el mismo peso a todos los intervalos de la misma longitud. Tal distribución tendría función de densidad

$$f_\mu(\mu) = c \qquad -\infty < \mu < \infty , \qquad (3.1.11)$$

o sea, uniforme en la recta. Pero (3.1.11) es una función de densidad impropia ya que

$$\int_{-\infty}^{\infty} f_\mu(\mu)\, d\mu = c\int_{-\infty}^{\infty} d\mu$$

diverge para cualquier valor de c. Sin embargo, la función de densidad posterior sí es propia aunque la prior sea impropia. Como puede verse, el uso de este tipo de distribuciones no informativas puede generar bastante discusión que preferimos omitir puesto que nos desviaría bastente del objetivo de esta tesis. Para una discusión al respecto se puede consultar (2), en donde se da un tratamiento especial a las distribuciones a priori no informativas.

Una vez identificada la función de densidad posterior, las inferencias acerca de los parámetros las podemos hacer utilizando dicha función. En algunos problemas nos puede interesar tener un estimador de punto para ciertos pará

metros. Veamos de manera muy general como se pueden elegir dentro de este enfo que estimadores.

3.1.a. Estimadores bayesianos

Asociado a la acción de escoger un estimador $\hat{\underline{\theta}}$ se define una función- $\ell(\hat{\underline{\theta}}, \underline{\theta})$ que nos indica la pérdida incurrida por escoger $\underline{\theta}$ como estumador cuando θ es el verdadero valor. La función $\ell(\hat{\underline{\theta}}, \underline{\theta})$ es aleatoria ya que $\underline{\theta}$ es aleatorio. El procedimiento más utilizado para generar estimadores de punto es encontrar - aquel $\hat{\underline{\theta}}$ que minimiza la esperanza de la función de pérdida, o sea,

$$\min_{\hat{\underline{\theta}}} E(\ell(\hat{\underline{\theta}}, \underline{\theta})) = \min_{\hat{\underline{\theta}}} \int_R \ell(\hat{\underline{\theta}}, \underline{\theta}) \, f_{\underline{\theta}|\underline{Y}}(\underline{\theta}|\underline{y}) d\underline{\theta} \qquad (3.1.12)$$

Como ejemplo de función de pérdida tomemos la conocida función de pér dida cuadrática

$$\ell(\hat{\underline{\theta}}, \underline{\theta}) = (\underline{\theta} - \hat{\underline{\theta}})' C (\underline{\theta} - \hat{\underline{\theta}})$$

con C una matriz de constantes, positiva definida. Entonces, tenemos que

$$E(\ell(\hat{\underline{\theta}}, \underline{\theta})) = E((\underline{\theta} - \hat{\underline{\theta}})' C (\underline{\theta} - \hat{\underline{\theta}}))$$

$$= E\left((\underline{\theta} - E(\underline{\theta}))' C (\underline{\theta} - E(\underline{\theta}))\right) + (\hat{\underline{\theta}} - E(\underline{\theta}))' C (\hat{\underline{\theta}} - E(\underline{\theta})) \qquad (3.1.13)$$

El segundo término de la expresión anterior es el único que depende - de $\hat{\underline{\theta}}$ y dado que C es positiva definida la esperanza será mínima cuando $\hat{\underline{\theta}} = E(\underline{\theta})$, por lo tanto, la media posterior es el estimador óptimo de acuerdo a (3.1.12) - utilizando la función de pérdida cuadrática. Veremos en la próxima sección como el estimador H.K. corresponde a la media posterior dada una determinada prior.-

Recomendamos la lectura de Box y Tiao ((2), apéndice A.5.6.) para una discusión acerca del uso de estimadores de punto.

En las siguientes secciones veremos como el estimador H.K. y otros estimadores sesgados corresponden a estimadores bayesianos utilizando distribuciones a priori y muy particulares.

3.2. Justificación bayesiana de los estimadores H.K.

Antes de proceder a la justificación bayesiana, veamos que sucedería si se aumentara el modelo

$$\underline{Y} = X\,\underline{\beta} + \underline{e}$$

de la siguiente manera

$$\begin{pmatrix} \underline{Y} \\ \underline{0} \end{pmatrix} = \begin{pmatrix} X \\ \sqrt{k}\ \ I \end{pmatrix} \underline{\beta} + \begin{pmatrix} \underline{e} \\ \underline{e}_1 \end{pmatrix} \qquad (3.2.1)$$

donde $\underline{0}$ es un vector pxl de ceros

$\underline{e}_1$ es un vector pxl

I es una matriz idéntica pxp

De (3.2.1) obtenemos las ecuaciones normales

$$(X'X + kI)\,\hat{\underline{\beta}}_k = X'\underline{Y} + \sqrt{k}\ \ I\ \underline{0}$$

de donde obtenemos el estimador H.K.:

$$\hat{\underline{\beta}}_k = (X'X + kI)^{-1} X'\underline{Y}$$

De 3.2.1 resulta que $\hat{\underline{\beta}}_k$ es una ponderación de valores observados con otros valores no observados; en otras palabras, al modelo le hemos incorporado-información adicional. De aquí surgen naturalmente varias preguntas: ¿qué tipo de información estamos incorporando?; ¿refleja esta información, un conoci-- miento que tenemos del fenómeno observado?; ¿es adecuada esta información?; -- ¿estamos solucionando realmente el problema de multicolinealidad? Trataremos-de contestar estas y otras preguntas en esta sección.

Empecemos viendo como podemos derivar $\hat{\underline{\beta}}_k$ combinando la función de ve rosimilitud con una distribución a priori adecuada. En el modelo lineal, la - función de verosimilitud, suponiendo σ^2 conocida, está dada por

$$L(\underline{\beta}|\sigma^2, \underline{Y}) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^n \quad e^{-1/2\sigma^2 (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})} \qquad (3.2.3)$$

Ahora bien, podemos expresar L en término del estimador máximo vero símil de $\underline{\beta}$ utilizando la igualdad

$$(\underline{Y} - X\underline{\beta})'(\underline{Y} - X\underline{\beta}) = (\underline{Y} - X\hat{\underline{\beta}})'(\underline{Y} - X\hat{\underline{\beta}}) + (\underline{\beta} - \hat{\underline{\beta}})'X'X(\underline{\beta} - \hat{\underline{\beta}})$$

obteniendo

$$L(\underline{\beta}|\sigma^2, \underline{Y}) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \quad e^{-\frac{1}{2\sigma^2}(\underline{y} - X\hat{\underline{\beta}})'(\underline{y} - X\hat{\underline{\beta}}) - \frac{1}{2\sigma^2}(\underline{\beta} - \hat{\underline{\beta}})'X'X(\underline{\beta} - \hat{\underline{\beta}})} \qquad (3.2.4)$$

L es función de $\underline{\beta}$ solo a través del último factor de (3.2.4); los - otros factores son constantes. Por lo tanto

$$L(\underline{\beta}|\sigma^2, \underline{Y}) = c \quad e^{-\frac{1}{2\sigma^2}(\underline{\beta} - \hat{\underline{\beta}})'X'X(\underline{\beta} - \hat{\underline{\beta}})} \qquad (3.2.5)$$

donde c es una constante. (3.2.5) tiene la forma de una distribución normal-multivariada. De (3.1.4) tenemos que

$$f_{\underline{\beta}|\sigma,\underline{Y}}(\underline{\beta}|\sigma, \underline{y}) \propto f_{\underline{\beta}}(\underline{\beta})\, e^{-\frac{1}{2\sigma^2}(\underline{\beta} - \hat{\underline{\beta}})'X'X(\beta - \hat{\underline{\beta}})} \qquad (3.2.6)$$

donde $\propto$ indica proporcionalidad.

De (3.2.6) vemos que la distribución posterior va a ser una ponderación de la prior y la verosimilitud. Box y Tiao sugieren una prior impropia (no informativa)

$$f_{\underline{\beta}}(\underline{\beta}) \propto \text{constante}$$

de tal manera que

$$f_{\underline{\beta}}(\underline{\beta}|\sigma,\underline{y}) \propto e^{-\frac{1}{2\sigma^2}(\underline{\beta} - \hat{\underline{\beta}})'X'X\,(\beta - \hat{\beta})}$$

Introduciendo una constante adecuada para que $f_{\underline{\beta}|\sigma,\underline{Y}}(\underline{\beta}|\sigma,\underline{y})$ integre a 1 llegamos al resultado:

$$f(\underline{\beta}|\sigma, \underline{y}) = \frac{|X'X|^{1/2}}{(\sqrt{2\pi}\,\sigma)^p}\, e^{-\frac{1}{2\sigma^2}(\underline{\beta} - \hat{\underline{\beta}})X'X(\underline{\beta} - \hat{\underline{\beta}})} \qquad (3.2.7)$$

Por lo tanto $\underline{\beta}$ se distribuye a posteriori según una $N_p(\hat{\beta}, \sigma^2(X'X)^{-1})$. Vemos así que con una prior no informativa, seguimos teniendo el problema de no ortogonalidad. Esto es debido a que no hemos incorporado información adicional al modelo que permita resolver el problema. Una posible solución es tratar de introducir una distribución a priori para $\underline{\beta}$, que quizás con un mínimo de información, logre mejorar sensiblemente nuestras estimaciones. Una distribución tal puede ser, por ejemplo

$$\underline{\beta} \sim N_p(\underline{0}, \sigma^2 I) \qquad (3.2.8)$$

Bajo esta suposición la función de densidad posterior es

$$f(\underline{\beta}|\sigma, \underline{y}) \propto e^{-\frac{1}{2\sigma_1^2} - \frac{1}{2\sigma^2}(\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})} \qquad (3.2.9)$$

Ahora el problema consiste en identificar a (3.2.9); para ello nos fijamos en el exponente:

$$\frac{1}{\sigma_1^2}\underline{\beta}'\underline{\beta} + \frac{1}{\sigma^2}\underline{Y}'\underline{Y} - 2\frac{1}{\sigma^2}\underline{Y}'X\underline{\beta} + \frac{1}{\sigma^2}\underline{\beta}'X'X\underline{\beta}$$

$$= \underline{\beta}'(\frac{1}{\sigma^2}X'X + \sigma_1^2 I)\underline{\beta} - 2\frac{1}{\sigma^2}\underline{Y}'X\underline{\beta} + \frac{1}{\sigma^2}\underline{Y}'\underline{Y} \qquad (3.2.10)$$

El último término de (3.2.10) no depende de $\underline{\beta}$ y es fijo, por lo tanto, nos fijamos solamente en los primeros 2 términos y completamos cuadrados:

$$\underline{\beta}'(\frac{1}{\sigma^2}X'X + \frac{1}{\sigma_1^2}I)\underline{\beta} - 2\frac{1}{\sigma^2}\underline{Y}'X\underline{\beta} + \frac{1}{\sigma^2}(X'\underline{Y})'(\frac{1}{\sigma^2}X'X + \frac{1}{\sigma_1^2}I)\frac{1}{\sigma^2}X'\underline{Y}$$

$$= (\underline{\beta} - (\frac{1}{\sigma^2}X'X + \frac{1}{\sigma_1^2}I)^{-1}X'\underline{Y})'(\frac{1}{\sigma^2}X'X + \frac{1}{\sigma_1^2}I)(\underline{\beta} - (\frac{1}{\sigma^2}X'X + \frac{1}{\sigma_1^2}I)^{-1}X'\underline{Y})$$

$$= (\underline{\beta} - (X'X + \frac{\sigma^2}{\sigma_1^2}I)^{-1}X'\underline{Y})'\frac{1}{\sigma^2}(X'X + \frac{\sigma^2}{\sigma_1^2}I)(\underline{\beta} - (X'X + \frac{\sigma^2}{\sigma_1^2}I)^{-1}X'\underline{Y}) = Q$$

Por lo tanto

$$f(\underline{\beta}|\sigma^2, \underline{y}) \propto e^{-\frac{1}{2}Q}$$

Tiene la forma de la distribución normal multivariada

$$N_p((X'X + \frac{\sigma^2}{\sigma_1^2}I)^{-1}X'\underline{Y}, \sigma^2(X'X + \frac{\sigma^2}{\sigma_1^2}I)^{-1}) \qquad (3.2.11)$$

Si definimos

$$k = \frac{\sigma^2}{\sigma_1^2}$$

tenemos que la media posterior es el estimador propuesto por Hoerl y Kennard. - En este caso resulta que la prior es una distribución informativa y en consecuencia existe la necesidad de justificar el uso de esta distribución a priori. Para ello necesitamos saber qué clase de información o conocimiento a priori estamos agregando al modelo y que nos puede llevar al uso de esta distribución a -- priori. Con la distribución (3.2.8) estamos suponiendo que los parámetros se - encuentran cercanos a 0. A medida que k es más pequeño estamos dando menos confianza a la información a priori y más a la proveniente de los datos. Además,- estamos suponiendo misma media y varianza a todos los parámetros, lo cual refleja que nuestro conocimiento o ignorancia relativa es igual para todos los parámetros. Como hacen notar Lindley y Smith (22), para que sea razonable esta - suposición "puede ser necesario cambiar la escala de las variables", por ejem-- plo, escribiendo el modelo de tal manera que X'X resulte en una matriz de correlación.

Por otro lado, al no introducir covarianzas en la distribución a priori estamos suponiendo que no conocemos ninguna relación que guarden los parámetros entre sí, por ejemplo, podemos saber que si β_1 es grande, también será β_3 o si β_2 es grande, β_4 será pequeño. Este tipo de conocimiento previo se podría- incorporar introduciendo covarianzas en la distribución a priori.

La distribución (3.2.8) en donde las variables tienen exactamente la- misma distribución y son independientes entre sí es un caso particular de dis-- tribución a priori que Finetti (22) llamó intercambiable.

Podemos apreciar que se trata de una clase muy especial de distribu-- ciones a priori. Uno de los comentarios de Hoerl y Kennard podría justificar -

el uso de esta distribución a priori en ciertas ocasiones:

"Cualquiera que haya abordado problemas no ortogonales reales ha ob-- servado coeficientes $\hat{\beta}_i$ que son muy grandes en valor absoluto" (15). El hecho- que para ellos sean los coeficientes "muy grandes en valor absoluto" está eviden ciando que tienen ya un cierto conocimiento del fenómeno.

Marquardt y Snee (25) también hacen este tipo de comentarios. Todos- ellos han obtenido buenos resultados en la práctica, lo cual lleva a pensar que- esta clase de distribuciones a priori son adecuadas en determinados problemas.- Ahora bien, Hoerl y Kennard obtuvieron buenos resultados dejando la derivación- bayesiana a nivel de mera interpretación. Por lo tanto, cabe hacer la pregunta: ¿qué desventajas presenta el enfoque de Hoerl y Kennard al tratar de solucionar el problema de multicolinealidad con respecto al enfoque bayesiano?

Aquí hay que insistir en la esencia del problema de multicolinealidad que no es que X'X sea casi singular ni como detectar la multicolinealidad sino que es un problema de falta de información proveniente de los datos. El error- que cometen Hoerl y Kennard es proponer una solución bayesiana al problema de - multicolinealidad y aferrarse a esquemas no bayesianos. ¿Por qué creemos que - esto es un error? Primero, porque el ECM no es un criterio operativo, es decir, a pesar de que podemos encontrar clases de estimadores con miembros de la clase con menor ECM que $\hat{\beta}$, no podemos identificarlos debido a que las condiciones -- siempre dependen de parámetros desconocidos. Segundo, porque al presentar la - solución bayesiana sólo como una mera interpretación no analizan por ejemplo, - el significado que tiene la traza H.K. ¿Qué significa, por ejemplo, que la tra za se haya estabilizado? ¿Qué quiere decir que algunos coeficientes de $\hat{\beta}$ cam--

bien bruscamente de valor al incrementar el valor de k, o como dicen ellos, -- "que no retengan su poder predictivo"?

Por otra parte, el enfoque dado por ellos ha llevado a varios investigadores a adoptar el método de la traza suponiendo la validez del mismo, olvidando la interpretación bayesiana. (ver por ejemplo Mayer y Willke (23)). Por -- otro lado vamos a argumentar que no es posible encontrar un valor óptimo de k y que en todo caso un valor que asignamos a k no debe basarse en la estabilidad - de la traza.

3.2.a. Crítica del método de la traza H.K.

Empecemos por analizar el significado de $k = \frac{\sigma^2}{\sigma_1^2}$. Es un cociente - que lo podemos interpretar como la confianza que tenemos en la información contenida en la verosimilitud sobre la información contenida en la prior.

A medida que σ_1^2 crece le damos más peso a la información proveniente de los datos. Dada la no ortogonalidad de X'X y la ortogonalidad de $\sigma_1^2 I$, en - algunas direcciones la información a priori puede rápidamente dominar a la verosimilitud mientras que en otras tardará más en dominarla. Entonces, si suponemos una distribución a priori para varios parámetros, intercambiable, es decir, con misma información para todos los parámetros, resulta que esta información - será más relevante para los parámetros peor estimados, dependiendo de la orientación relativa de los ejes de la función de verosimilitud. (Figura 3.2.B).

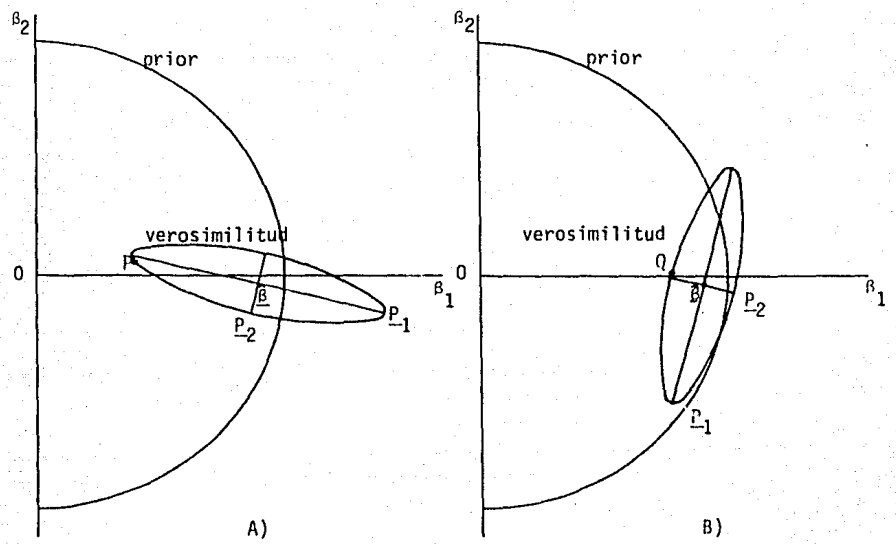


Figura 3.2.B. P y Q son las medidas posteriores. Notemos como la información a priori es más relevante en A que en B. En la dirección del eje mayor $\underline{P}_1$ se tiene menos información que en la dirección del eje menor $\underline{P}_2$. En A) se tiene poca información de las observaciones acerca de β_1 relativa a la que se tiene de β_2. En B) se tiene poca información de β_2 relativa a la que se tiene de β_1. Sin embargo se modifica más $\hat{\beta}_1$ que $\hat{\beta}_2$.

En consecuencia un cambio brusco de valor de los coeficientes de $\underline{\hat{\beta}}$ significa que la información a priori ha sido relevante para esos coeficientes. La distribución a priori cobra importancia, porque si esta información no es adecuada, en vez de remediar el problema lo podemos agravar. La distribución a priori de Hoerl y Kennard puede estar muy alejada de la realidad y por lo tanto sesgar demasiado los resultados, inclusive antes de que la traza se haya estabilizado. Es indudable que en ciertos problemas puede ser una distribución

a priori adecuada, los resultados obtenidos en la práctica corroboran esto, pero no pensamos que el método de Hoerl y Kennard deba ser un procedimiento a ser utilizado siempre que se tenga el problema de multicolinealidad como proponen ellos y Marquardt (24). Creemos que el problema debe enfocarse a conseguir más información e incorporarse ésta al modelo.

Por otro lado, ¿qué significa que la traza se estabilice? Básicamente, que la información a priori relevante ha sido incorporada al modelo y que el seguir incrementando nuestra confianza en la prior ya no nos reporta un cambio sensible en la información ya incluída. Para nosotros la traza H.K. juega entonces un doble papel: por un lado refleja la sensibilidad de $\hat{\beta}$ al conjunto particular de datos; por otro lado refleja el compromiso que se contrae al utilizar esa distribución a priori. Si hay un cambio brusco al pasar de $\hat{\beta}$ a $\hat{\beta}_k$ debemos ser muy críticos con las suposiciones que justifican el uso de esta distribución a priori.

Creemos que con la argumentación anterior la búsqueda de un valor óptimo de k basándose en la noción de estabilidad de la traza pierde sentido, ya que la estabilidad de la misma no nos garantiza que hayamos encontrado un estimador adecuado; únicamente el análisis de las suposiciones a priori puede darnos una relativa seguridad en el estimador resultante. Un ejemplo del peligro de no hacer un análisis de las suposiciones a priori nos lo da Johnson (18). Johnson hace caso omiso de la interpretación bayesiana del método de H.K. Efectúa una pequeña simulación y da los siguientes datos:

$$X'X = \begin{pmatrix} 1 & .9 \\ .9 & 1 \end{pmatrix}$$

Quiere estimar 2 parámetros, β_1 y β_2, en donde el cociente $\beta_1/\beta_2=10$. Obtiene los siguientes resultados:

k	$\hat{\beta}_{1k}$	$\hat{\beta}_{2k}$	$\hat{\beta}_{1k}/\hat{\beta}_{2k}$
0	10	1	10
.1	7.745	2.975	2.51
.2	6.476	3.476	.1.86
.5	5.104	3.604	1.42
1.0	4.013	3.194	1.26

Concluye que "esto debería difícilmente ser tomado como indicativo - de un acercamiento a la realidad". Sin embargo, aparte de la evidencia de que el método en ocasiones no funciona, no aporta ninguna crítica de por qué no - sirve. Es evidente que la suposición de una prior intercambiable no es adecuada. Lo que sí evidencía el ejemplo es que el método de H.K. por sí no resuelve el problema de multicolinealidad sino lo que ayudará a resolverlo será el - análisis de lo adecuado de las suposiciones a priori. Dado que las suposiciones a priori implícitas en el método de Hoerl y Kennard responden a un conjunto de características muy específicas, derivamos en la siguiente sección la -- distribución a posteriori que se obtiene utilizando una priori más general tomando como referencia a Lindley y Smith (22) y Zellner (45).

3.3. Justificación bayesiana de otros estimadores sesgados

En la sección pasada supusimos que $\underline{\beta}$ se distribuía a priori según - una $N_p(\underline{0}, \sigma_1^2 I)$. Vamos ahora a suponer que $\underline{\beta}$ se distribuye a priori como

$$\underline{\beta} \sim N_p(\underline{\mu}\ ,\ V) \tag{3.3.1}$$

de donde en particular resulta la anterior distribución cuando $\underline{\mu} = \underline{0}$ y $V=\sigma_1^2 I$. Análogamente a (3.2.9) tenemos que

$$f(\underline{\beta}|\sigma,\ \underline{y}) \propto e^{-\frac{1}{2}(\underline{\beta} - \underline{\mu})'V^{-1}(\underline{\beta} - \underline{\mu})} \cdot e^{-\frac{1}{2\sigma}(\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})} \tag{3.3.2.}$$

y fijándonos en el exponente para tratar de identificar a la distribución posterior

$$(\underline{\beta} - \underline{\mu})'V^{-1}(\underline{\beta} - \underline{\mu}) + (\underline{Y} - X\underline{\beta})'\frac{1}{\sigma}(\underline{Y} - X\underline{\beta})$$
$$= \underline{\beta}'(\frac{1}{\sigma}X'X + V^{-1})\underline{\beta} - 2(\frac{1}{\sigma}X'\underline{Y} + V^{-1}\underline{\mu})'\underline{\beta} + \{\text{Términos que no dependen de } \underline{\beta}\}$$

de donde completando cuadrados llegamos a la expresión

$$(\underline{\beta} - (\frac{1}{\sigma}X'X + V^{-1})^{-1}(\frac{1}{\sigma}X'\underline{Y} + V^{-1}\underline{\mu}))'$$
$$(\frac{1}{\sigma}X'X + V^{-1})(\underline{\beta} - (\frac{1}{\sigma}X'X + V^{-1})^{-1}(\frac{1}{\sigma}X'\underline{Y} + V^{-1}\underline{\mu}))$$

Por lo tanto $\underline{\beta}$ se distribuye a posteriori como

$$N_p\Big((\frac{1}{\sigma}X'X + V^{-1})^{-1}(\frac{1}{\sigma}X'\underline{Y} + V^{-1}\underline{\mu}),\ (\frac{1}{\sigma}X'X + V^{-1})^{-1}\Big) \tag{3.3.3}$$

Con la distribución a priori (3.3.1), más general que la propuesta por Hoerl y Kennard, podemos identificar qué tipo de información a priori es la que se ha incorporado implícitamente en los estimadores sesgados que se han venido publicando. Empecemos con el caso más sencillo. Banerjee y Carr (1) propusieron -

dentro del enfoque dado por H.K. el estimador

$$\hat{\underline{\beta}}_K = (X'X + K)^{-1} X'\underline{Y} \qquad (3.3.4)$$

con K una matriz con elementos k_i en la diagonal y 0 fuera de la diagonal. Fácilmente podemos comprobar que si suponemos a priori que

$$\underline{\beta} \sim N_p(\underline{0}, \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \ddots & & \\ \vdots & & \ddots & \\ 0 & \cdots & & \sigma_p^2 \end{pmatrix}) \qquad (3.3.5)$$

obtenemos que la media posterior coincide con $\hat{\underline{\beta}}_K$, en donde

$$k_i = \frac{\sigma^2}{\sigma_i^2} \qquad i=1,\ldots,p.$$

En este caso resulta que la prior refleja que el conocimiento a priori que tenemos acerca de los parámetros es que están cercanos a 0, difiriendo de la prior (3.2.8) en que la confianza en este conocimiento a priori es diferente para cada parámetro. Una generalización directa de (3.3.4) que prononemos es

$$\underline{\mu}\,\hat{\underline{\beta}}_K = (X'X + K)^{-1} (X'\underline{Y} + K\underline{\mu}) \qquad (3.3.6)$$

o sea, suponer a priori que

$$\underline{\beta} \sim N_p(\underline{\mu}, \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \ddots & & \\ \vdots & & \ddots & \\ 0 & \cdots & & \sigma_p^2 \end{pmatrix})$$

En el caso de que $\underline{\mu}=\underline{0}$,(3.3.7) coincide con (3.3.6); si $\sigma_1^2=\sigma_2^2=\ldots\sigma_p^2$ (3.3.6) coincide con el estimador propuesto por Swindel (37) que ya utiliza un enfoque bayesiano.

Tomemos ahora los estimadores de Mayer y Willke:

$$\hat{\underline{\beta}}(c) = c\ (X'X)^{-1}\ X'\underline{Y} \tag{3.3.8}$$

Si suponemos que a priori

$$\underline{\beta} \sim N(\underline{0},\ \sigma_1^2(X'X)^{-1}) \tag{3.3.9}$$

tenemos que la media posterior es

$$(\frac{1}{\sigma^2}X'X + \frac{1}{\sigma_1^2}X'X)^{-1}\ (\frac{1}{\sigma^2}X'\underline{Y}) = ((\frac{1}{\sigma^2} + \frac{1}{\sigma_1^2})(X'X))^{-1}(\frac{1}{\sigma^2}X'\underline{Y})$$

$$= (\frac{\sigma^2\sigma_1^2}{\sigma_1^2+\sigma^2})(X'X)^{-1}\ (\frac{1}{\sigma^2}X'\underline{Y}) = \frac{\sigma_1^2}{\sigma_1^2 + \sigma^2}(X'X)^{-1}\ X'\underline{Y}$$

de donde $c = \dfrac{\sigma_1^2}{\sigma_1^2 + \sigma^2}$ y en consecuencia

$$0 \leq c < 1$$

Nosotros ponemos en duda que un conocimiento a priori pueda llevar - al uso de (3.3.9). Creemos que es más factible utilizar cualquiera de las distribuciones a priori anteriores a ésta.

Recordemos ahora al estimador H.K. generalizado (2.3.2):

$$\hat{\underline{\alpha}}^* = (W'W + K)^{-1}\quad W'\ \underline{Y} \tag{3.3.10}$$

De (3.3.4) resulta evidente que la prior que da origen a este estimador es

$$\underline{\alpha} \sim N_p(\underline{0}, \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \ddots & & \\ \vdots & & \ddots & \\ 0 & & & \sigma_p^2 \end{pmatrix}) \tag{3.3.11}$$

ya que el único cambio que ha habido es que ahora se trabaja con $\underline{\alpha} = P\underline{\beta}$. Dado que los elementos de $\underline{\alpha}$ son combinaciones lineales de los elementos de $\underline{\beta}$ suponemos que puede existir en algunas ocasiones algún conocimiento a priori que nos lleve al uso de (3.3.11) o una generalización de la misma que sería suponer $\underline{\mu} \neq \underline{0}$. Hay que tomar muy en cuenta que al utilizar el método tal como lo plantean (9), (13) y (15) estamos incorporando información a priori en aquellas direcciones donde menos información existe proveniente de los datos. Lo que creemos que es interesante es analizar el significado de la prior en las unidades originales; tomando en cuenta que $P\underline{\alpha} = \underline{\beta}$ tenemos que

$$P\underline{\alpha} = \underline{\beta} \sim N_p(\underline{0},\ P\begin{pmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \ddots & & \\ \vdots & & \ddots & \\ 0 & & & \sigma_p^2\end{pmatrix}P') \qquad (3.3.12)$$

lo que significa que al suponer una prior para $\underline{\alpha}$ con media $\underline{0}$, estamos suponiendo una prior con media $\underline{0}$ para $\underline{\beta}$. La objeción que puede ponérsele a la prior (3.3.12) es que la matriz de varianza-covarianza depende de X.

Por último falta analizar el significado que tienen los estimadores de Marquardt y de componentes principales. Para ello veamos como podemos incorporar información a priori no directamente de los parámetros sino de algunas combinaciones lineales de los mismos. Supongamos que a priori no podemos afirmar que existen las restricciones lineales

$$\underline{r} = R\underline{\beta} \qquad (3.3.13)$$

donde $\underline{r}$ es un vector constante jx1 y R una matriz jxp de constantes, j≤p, pero relajando estas restricciones podemos suponer que

$$R\underline{\beta} \sim N_j\ (\underline{r},\ \sigma_1^2 I) \qquad (\text{R de rango j})$$

o bien,

$$R\underline{\beta} = \underline{r} + \underline{n} \qquad (3.3.14)$$

donde $\underline{n} \sim N_j(\underline{0}, \sigma_1^2 I)$

Lo importante ahora es suponer que $\underline{\beta}$ es aleatorio, lo cual modifica el enfoque dado por Theil (38) y Ward (41) en donde toman $\underline{\beta}$ fijo pero $\underline{r}$ aleatorio. Entonces, bajo la suposición de que $\underline{\beta}$ es aleatorio, nos interesa encontrar su distribución. Una manera de encontrarla es tomando una inversa generalizada de R y premultiplicando (3.3.14) por ella. Por conveniencia tomemos como inversa generalizada de R a

$$R^- = (R'R)^+ R'$$

donde $(R'R)^+$ es la inversa generalizada de Moore-Penrose (ver A.12) de R'R. Entonces, de (3.3.14) tenemos que

$$\underline{\beta} = (R'R)^+ R'\underline{r} + (R'R)^+ R' \underline{n}$$

En consecuencia $\underline{\beta}$ se distribuye a priori según una normal singular - ver (1.1.11)

$$NS_j((R'R)^+ R'\underline{r}, \sigma_1^2(R'R)^+ R'R (R'R)+) \qquad (3.3.15)$$

De A.17 tenemos que (3.3.15) es una

$$NS_j((R'R)^+R'\underline{r}, \sigma_1^2(R'R)^+) \qquad (3.3.16)$$

Basándonos en (32) y (41) llegamos a que la media posterior es

$$\hat{\underline{\beta}}_{\underline{r}} = (X'X + \frac{\sigma^2}{\sigma_1^2} R'R)^{-1} (X'\underline{Y} + (R'R) (R'R)^+(R'R)\underline{r})$$

$$= (X'X + \frac{\sigma^2}{\sigma_1^2} R'R)^{-1} (X'\underline{Y} + (R'R)\underline{r}) \qquad (3.3.17)$$

con lo cual generalizamos el resultado de Ward en donde él supone a priori que

$$\underline{\beta} \sim N_p (\underline{\mu}, \sigma_1^2 W) \qquad (3.3.18)$$

con W no singular.

Creemos que este resultado puede servir cuando se tiene algún conocimiento previo de pocas ($j < p$) combinaciones lineales de $\underline{\beta}$. Si el rango de R es igual a p cubrimos el caso considerado por Ward y además hay mas congruencia en la formulación que en Theil (38), donde él supone que la información a priori es

$$\underline{r} = R\,\underline{\beta} + \underline{n}$$

con $\underline{n}$ aleatorio pero $\underline{\beta}$ fijo. Aquí suponemos que $\underline{\beta}$ es aleatorio y que dado que tenemos conocimiento previo de combinaciones lineales de $\underline{\beta}$, indirectamente tenemos información acerca de $\underline{\beta}$.

Veamos la relación que existe de lo anterior con el método de Marquardt. Sea $\underline{P}_p$ el eigenvector correspondiente a λ_p y supongamos que a priori

$$\underline{P}_p'\ \underline{\beta} \sim N(0, \sigma_1^2) \qquad (3.3.18')$$

y de acuerdo con (3.3.17), la media posterior es

$$\hat{\underline{\beta}}_0 = (X'X + \frac{\sigma^2}{\sigma_1^2}\ \underline{P}_p\,\underline{P}_p')\ X'\underline{Y}$$

Recordemos que

$$X'X = \lambda_1 \underline{P}_1 \underline{P}_1' + \lambda_2 \underline{P}_2 \underline{P}_2' + \ldots + \lambda_p \underline{P}_p \underline{P}_p'$$

de donde

$$(X'X + \frac{\sigma^2}{\sigma_1^2} \underline{P}_p\underline{P}_p') = \lambda_1\underline{P}_1\underline{P}_1' + \lambda_2\underline{P}_2\underline{P}_2' + \ldots + (\lambda_p + \frac{\sigma^2}{\sigma_1^2})\underline{P}_p\underline{P}_p' \qquad (3.3.19)$$

Ahora bien, dado que (3.3.19) es de rango completo, podemos tomar su inversa que es igual a

$$\frac{1}{\lambda_1}\underline{P}_1\underline{P}_p' + \frac{1}{\lambda_2}\underline{P}_2\underline{P}_2' + \ldots + \frac{\sigma_1^2}{\sigma_1^2\lambda_p + \sigma^2}\underline{P}_p\underline{P}_p' \qquad (3.3.20)$$

Fijémonos en

$$\frac{\sigma_1^2}{\sigma_1^2\lambda_p + \sigma^2} = \frac{\sigma_1^2}{(\sigma_1^2 + \frac{\sigma^2}{\lambda_p})} \cdot \frac{1}{\lambda_p} = \frac{1}{(1 + \frac{\sigma^2}{\sigma_1^2\lambda_p})} \cdot \frac{1}{\lambda_p}$$

Tenemos que $\frac{\sigma_1^2}{\sigma_1^2 + \frac{\sigma^2}{\lambda_p}}$ está jugando el papel de d_p en (2.3.8). Además,

$$\lim_{\sigma_1^2 \to 0} \frac{\sigma_1^2}{\sigma_1^2 + \frac{\sigma^2}{\lambda_p}} = 0 \qquad (3.3.21)$$

Esto quiere decir que si d_p es pequeña, corresponde a una prior que está proveyendo mucha información en la dirección del eigenvector $\underline{P}_p$. Cuando se alcanza el límite (3.3.21) resulta que (3.3.20) sería una inversa generalizada si X'X fuera de rango p-1. Es claro que si $\sigma_1^2 = 0$, (3.3.19) no tiene sentido. En este caso llegamos a la restricción:

$$\underline{P}_p' \, \underline{\beta} = 0$$

que también la podemos considerar como información a priori ya que una restricción en realidad equivale a conocer algo más acerca del fenómeno estudiado.

Lo importante de la discusión anterior es que con estos métodos se está suponiendo a priori una normal singular con media $\underline{U}$ y en consecuencia son de tomarse en cuenta las implicaciones que estas suposiciones tienen.

C O N C L U S I O N E S

1. El problema de multicolinealidad debe enfocarse de acuerdo a como lo han tratado Silvey (36) y Johnston (19), es decir, como un problema de falta de información y como tal debe tratársele. Se deben descartar aquellos enfoques que lo tratan como un problema numérico exclusivamente como lo hace por ejemplo Coxe (3).

2. Se ha insistido que las técnicas de selección de variables sirven para atacar el problema de multicolinealidad. Sin embargo varias gentes han hecho notar las deficiencias de estas técnicas en presencia de la multicolinealidad y las consecuencias de las mismas (15), (25) (44).

3. Pensamos que la única manera razonable de resolver el problema de multicolinealidad es consiguiendo más información, ya sea a partir de un conocimiento previo del fenómeno, o bien, obteniendo más datos.

4. La condición de admisibilidad que tanto se ha utilizado para derivar nuevas clases de estimadores no parece ser un criterio adecuado a utilizar con los estimadores sesgados puesto que para garantizar que se cumpla con un particular estimador, es necesario conocer los parámetros desconocidos.

5. Con los criterios de estimación propuestos de la sección 2.3.c., no se ha logrado resolver ningún problema pero quizás si crear algo de confu-- sión.

6. Los métodos de estimación sesgada tienen implícitas ciertas suposiciones a priori y debe uno estar consciente de ellas pues si éstas estan muy alejadas de la realidad, en vez de remediar el problema, éste puede agravarse. Por esto la justificación bayesiana de estos estimadores no debe quedarse a nivel de mera interpretación.

7. Una vez que se reconoce que con estos métodos se está introduciendo información a priori cabe hacerse la pregunta ¿es esta distribución a priori adecuada? No comprendemos por qué Marquardt propone que el método de Hoerl y - Kennard es un método a utilizarse siempre que se tenga el problema de no ortogonalidad ya que estamos seguros que en muchos casos la distribución a priori de-ellos puede estar muy alejada de la realidad.

8. Los métodos de estimación H.K. generalizados están incorporando-información precisamente en las direcciones en donde es más imprecisa la estimación, es decir, donde menos información se tiene.

9. Los métodos de Marquardt y de componentes principales están im-- plicando ciertas restricciones que a la vez implican una distribución a priori con media $\underline{0}$.

10. Dada la interpretación bayesiana puede verse que la búsqueda -- del valor de k, ó k_i, óptimo no tiene sentido. Cualquier búsqueda dentro - de los marcos clásicos podemos augurar que será infructuosa. Lo que nos debe- preocupar es saber la importancia de la prior y de la verosimilitud en la -- formación de la posterior.

APENDICE A

Presentamos en este apéndice algunos resultados de matrices. Las demostraciones se encuentran en los siguientes libros:

Johnston (19), Rao (33) y Searle (34)

A.1. Sea C una matriz cuadrada, simétrica y no singular. Entonces

$$(C')^{-1} = (C^{-1})'$$

A.2. Sean C y D dos matrices cuadradas tales que el producto de ellas tiene sentido. Entonces

$$|C\ D| = |C|\ |D|$$

A.3. Sea C es una matriz cuadrada, no singular. Entonces

$$|C^{-1}| = \frac{1}{|C|}$$

A.4. Sea $C = (c_{ij})$ una matriz p x p positiva semidefinida. Entonces

$$|C| \leq c_{11} \cdot \ldots \cdot c_{pp}$$

Sea $C = (c_{ij})$ una matriz p x p. Sean $\lambda_1 \geq \ldots \geq \lambda_p$ los eigenvalores de C y $\underset{\sim}{P}_1, \ldots, \underset{\sim}{P}_p$ los eigenvectores correspondientes. Sea Λ la matriz de eigenvectores. Entonces

A.5 $$|C| = \lambda_1 \cdot \ldots \cdot \lambda_p$$

A.6 $$t_r\, C = \sum_{i=1}^{p} \lambda_i$$

Si además C es simétrica

A.7 La descomposición espectral de C es

$$C = P\Lambda\, P' = \lambda_1 \underline{P}_1 \underline{P}_1' + \ldots + \lambda_p\, \underline{P}_p\, \underline{P}_p'$$

y además $I = P\, P' = \underline{P}_1 \underline{P}_1' + \ldots + \underline{P}_p \underline{P}_p'$

A.8 $C^2 = P\Lambda\, P' P\Lambda\, P' = P\Lambda^2\, P'$;

por lo tanto

$$tr\, C^2 = \sum_{i=1}^{p} \lambda_i 2$$

Si C es simétrica y no singular

A.9. $\frac{1}{\lambda_1}, \ldots \ldots \frac{1}{\lambda_p}$ son los eigenvalores de C^{-1} y

$$C^{-1} = P\Lambda^{-1}\, P' = \frac{1}{\lambda_1}\, \underline{P}_1 \underline{P}_1' + \ldots + \frac{1}{\lambda_p}\, \underline{P}_p \underline{P}_p'$$

A.10 Sean C y D 2 matrices simétricas p x p, no singulares. Sean $\lambda_1, \lambda_2 \ldots, \lambda_p$ los eigenvalores de C y $\underline{P}_1, \ldots, \underline{P}_p$ los eigenvectores correspondientes, sean $\ell_1, \ell_2, \ldots \ell_p$ los eigenvalores de D y $\underline{Q}_1, \ldots, Q_p$ los eigenvectores correspondientes,

Si $\underline{P}_1 = Q_1, \underline{P}_2 = \underline{Q}_2, \ldots, \underline{P}_p = \underline{Q}_p$

entonces

$$C\,D = \ell_1\lambda_1 \underset{\sim}{P}_1 \underset{\sim}{P}_1' + \ell_2\,\lambda_2 \underset{\sim}{P}_2\underset{\sim}{P}_2' + \ldots + \ell_p\,\lambda_p \underset{\sim}{P}_p\underset{\sim}{P}_p'$$

y además

$$\text{tr } C\,D = \sum_{i=1}^{p} \ell_i \lambda_i$$

A.11. Inversas generalizadas.

Sea A una matriz cualquiera.

G es una inversa generalizada de A si

$$A\,G\,A = A.$$

G no es única generalmente.

A.12. Inversa de Moore-Penrose.

Sea A una matriz cualquiera.

A^+ es la inversa de Moore-Penrose

si y sólo si

a) $A\,A^+A = A$

b) $A^+AA^+ = A^+$

c) $(A^+A)' = A^+A$

d) $(A\,A^+)' = A\,A^+$

Esta inversa generalizada es única.

A. 13. Sea A una matriz n x m de rango r. Tomemos A'A y sean $\lambda_1, \ldots \lambda_r$ los eigenvalores distintos de 0 de A'A y $\underline{P}_1, \underline{P}_2 \ldots \underline{P}_r$ los eigenvectores correspondientes.

Entonces

$\frac{1}{\lambda_1} \underline{P}_1\underline{P}_1' + \ldots + \frac{1}{\lambda_r} \underline{P}_r\underline{P}_r'$ es una inversa generalizada de A'A y además coincide con la de Moore-Penrose.

A P E N D I C E B

Dado el modelo

$$\underline{Y} = \underline{1}\, b_o' + X\, \underline{\beta} + \underline{e} \qquad (B.1)$$

descrito en (1.3.9), utilizando mínimos cuadrados se llega a que el estimador b_o' es igual a $\bar{y}$ independientemente del valor que tome $\hat{\underline{\beta}}$. Algunos libros (por ejemplo, Draper y Smith (6), Johnston (19)), interesándose sólo en el aspecto-computacional, no hacen ninguna aclaración y restan de $\underline{Y}$ el vector $\underline{1}\,\bar{y}$, -proponiendo el modelo

$$\underline{Y} = \underline{Y} - \underline{1}\,\bar{y} = X\,\underline{\beta} + \underline{\mu} \qquad (B.2)$$

$$\underline{\mu} = \underline{e} - \underline{1}\,\bar{e}$$

donde $\bar{e} = \sum \frac{e_i}{n}$

Es claro que

$$E(\mu_i) = E(e_i - \bar{e}) = 0 \qquad \text{para } i=1,\ldots,n$$

Pero resulta que

$$E(\mu_i\, \mu_j) = \begin{cases} \frac{(n-1)}{n}\,\sigma^2 & \text{si } i=j \\ -\frac{\sigma^2}{n} & \text{si } i\neq j \end{cases}$$

Por lo tanto, resulta que no se cumple que

$$E(\underline{\mu}\,\underline{\mu}') = \sigma^2 I$$

sino que

$$E(\underline{\mu}\,\underline{\mu}') = \sigma^2 \begin{pmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & & \vdots \\ \vdots & & \ddots & \vdots \\ -\frac{1}{n} & \cdots & & \frac{n-1}{n} \end{pmatrix} \qquad (B.3)$$

Podría pensarse en considerar el modelo de Aitken, el cual permite - que las observaciones estén correlacionadas. Pero tenemos que una de las supo siciones básicas es que el determinante (ver Rao (33), p. 221).

$$| \; E(\underline{\mu}\,\underline{\mu}') \; | \neq 0$$

En este caso tenemos que la matriz (B.3) es de rango n-1 y por lo -- tanto su determinante vale 0 (ver Rao (33), p. 67). Entonces, la notación que usemos

$$\underline{Y} = X\underline{\beta} + \underline{e}$$

es solamente convencional. Esto indica que el interés principal radica en $\underline{\beta}$ y no en b_0'. El modelo correcto debe ser B.1 y adoptamos esta convención debido a que el método de Hoerl y Kennard se aplica para $\underline{\beta}$. Existe confusión en li-- bros y artículos respecto a este punto ya que no se aclara debidamente. En el ejemplo que Hoerl y Kennard utilizan, tomado de Gorman y Toman (12), sólo remitiéndonos a esta fuente queda claro que existe un término constante.

APENDICE C

Para llevar a cabo la simulación se modificó el programa listado por Davis (46). Al tomar la decisión de efectuar una simulación se tenían en mente dos objetivos.

1. Proveer de ejemplos para ilustrar el método H.K en esta tesis.
2. Comparar soluciones del método H.K. con soluciones bayesianas y de mínimos cuadrados.

Descripción de la simulación

Basándonos en el ejemplo dado por Marquardt (25) decidimos utilizar el mismo método.

Se generaron 30 conjuntos de datos para cada una de 3 estructuras de la matriz de correlación.

Cada conjunto de datos contenía 8 observaciones generadas por el modelo

$$Y_i = b_o + b_1 X_{i1} + b_2 X_{i2} + b_3 X_{i3} + e_i \qquad i = 1,\ldots, 8$$

en donde se asignaron valores para los parámetros

$$b_o = 0 \qquad b_1 = b_2 = b_3 = 1$$

Los valores de e_i se generaron usando el método polar y se tomó $\sigma = .9$. Los valores X_{ij} se formaron a partir de los siguientes datos:

i	X_{i1}	X_{i2}	X_{i3}
1	-1	-1	-1
2	1	1	-1
3	-1	-1	1
4	1	1	1
5	-1	1	-1
6	1	-1	-1
7	-1	1	1
8	1	-1	1
$\sum X_{ij}$	0	0	0

Fácilmente se puede comprobar que las columnas son ortogonales. Esta estructura correponde a lo que se conoce como un arreglo o diseño factorial 2^3 que aparece en la figura C.1.

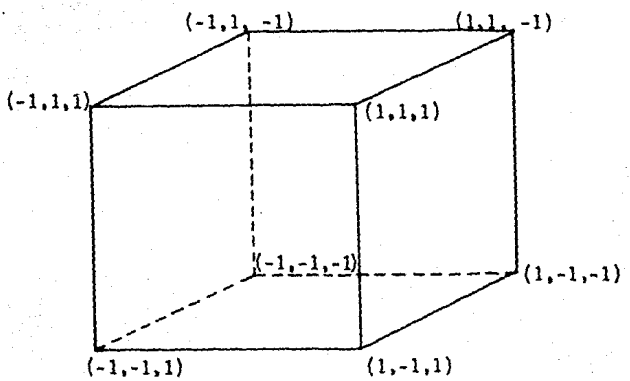


Figura C.1. Diseño factorial 2^3

La ortogonalidad del diseño se pierde modificando alguno o varios de-los puntos. Si modificamos las últimas cuatro observaciones de la siguiente manera:

i	X_{i1}	X_{i2}	X_{i3}
5	-1	$1-2\alpha$	-1
6	1	$-(1-2\alpha)$	-1
7	$-(1-2\alpha)$	1	1
8	$1-2\alpha$	-1	1

$0 \leq \alpha < 1$

podemos verificar fácilmente que la correlación entre $\underline{X}_1$ y $\underline{X}_2$, r_{12}, es $\alpha/(1 - \alpha + \alpha^2)$. En la figura C.2. aparece como se modifica el diseño con $\alpha = .2$

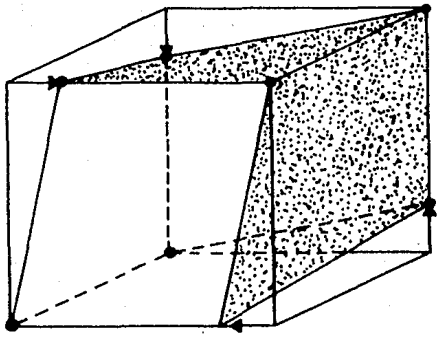


Figura C.2.

Modificamos también los primeros 4 renglones como sigue

i	X_{i1}	X_{i2}	X_{i3}
1	-1	-1	$-(1-2\gamma)$
2	1	$(1-2\gamma)$	-1
3	-1	$-(1-2\gamma)$	1
4	1	1	$1-2\gamma$

logrando con esto que la correlación r_{23} sea negativa. Generamos datos para los siguientes valores de α y γ:

		α .75	.9
γ	.3		X
	.6	X	X

Las matrices de de correlación y sus respectivos eigenvalores son -- los siguientes:

1) $\alpha = .9$

$\gamma = .3$

$$X'X = \begin{pmatrix} 1 & .9397 & -.1769 \\ .9397 & 1 & 0 \\ -.1769 & 0 & 1 \end{pmatrix}$$

$\lambda_1 = 1.9562 \qquad \lambda_2 = 1 \qquad \lambda_3 = .0438$

2) $\alpha = .9$

$\gamma = .6$

$$X'X = \begin{pmatrix} 1 & .7684 & -.3607 \\ .7684 & 1 & 0 \\ -.3607 & 0 & 1 \end{pmatrix}$$

$\lambda_1 = 1.8489 \qquad \lambda_2 = 1 \qquad \lambda_3 = .1511$

3) $\alpha = .75$

$\gamma \quad .6$

$$X'X = \begin{pmatrix} 1 & .6598 & -.3818 \\ .6598 & 1 & 0 \\ -.3818 & 0 & 1 \end{pmatrix}$$

$\lambda_1 = 1.7623 \qquad \lambda_2 = 1 \qquad \lambda_3 = .2377$

Selección del valor de k

La selección del valor de k utilizando el método H.K. resulta poco - práctico en una simulación por el tiempo necesario para analizar las gráficas. Resolvimos tomar una regla de decisión para seleccionar el valor máximo de k "aceptable". Sea

R el coeficiente de correlación múltiple del estimador de mínimos - cuadrados.

R' el coeficiente de correlación múltiple del estimador H.K.

CMR el cuadrado medio residual del estimador de mínimos cuadrados.

CMR' el cuadrado medio residual del estimador H.K.

Decimos que k es un valor aceptable si y sólo si

$$R - R' \leq .05 \qquad (C.1)$$

y si

$$CMR - CMR' \leq .75 \qquad (C.2)$$

Estamos de acuerdo en que tanto .05 como .75 son valores arbitrarios, sin embargo, de acuerdo a la naturaleza de los datos resultan razonables. Los-valores de k que se analizaron fueron

k = .025, .05, .075, .1, .2, .3, .4, .5, .6, 7, .8, .9, 1.0

Para comparar los estimadores se utilizaron los siguientes criterios

1) $$S_{\underline{b}^*} = \sum_{j=0}^{4} \sqrt{\frac{(b_j^* - b_j)^2}{4}}$$

donde b_j^* es un estimador de b_j

2) $$S_{\underline{Y}}^* = \sqrt{\sum_{i=1}^{8} \frac{Y_i^* - E(Y_i)}{8}^2}$$

con Y_i^* estimador de $E(Y_i)$

Para los valores aceptables de k se comparaban los estimadores H.K.- con el estimador de mínimos cuadrados de acuerdo a cada uno de los criterios.- Se concluía que el método H.K. era superior al de mínimos cuadrados de acuerdo al criterio 1(ó 2) si existía un valor aceptable de k tal que

$$S_{\underline{b}k} < S_{\hat{\underline{b}}}$$

(ó $$S_{\hat{\underline{Y}}} < S_{\hat{\underline{Y}}}$$)

Solución bayesiana

Tomamos el estimador propuesto por Swindel (37)

$$\underline{\mu}\ \hat{\underline{\beta}}_k = (X'X + kI)^{-1} (X'\underline{Y} + k\ \underline{\mu})$$

bajo la "suposición a priori" de que $\underline{\mu}$ estaba "cercano" a $\underline{\beta}$ (recuérdese que -- cuando se toma $\underline{\beta}$ se está refiriendo a los parámetros ya transformados según -- (1.3.9). Debido a la estructura del programa, se supuso a priori que

$$\underline{\beta} \sim N(s_y \begin{pmatrix} .35 \\ .35 \\ .35 \end{pmatrix}, \sigma_1^2\ I) \qquad (C.3)$$

$$s_y = \sqrt{\frac{(y_i - \bar{y})^2}{8}}$$

ya que el programa maneja los coeficientes de regresión estandarizados, es decir, encuentra los estimadores

$$\hat{\underline{\beta}} \; / \; s_y = \hat{\underline{a}}$$

y luego transforma a las unidades originales mediante la conocida fórmula

$$\hat{b}_i = \frac{s_y}{s_i} \quad \hat{a}_i$$

Tomamos 3 valores de $k = \frac{\sigma^2}{\sigma_1^2}$ (.1, .2, .3) para cada matriz de correlación (por lo tanto, 10 conjuntos con el mismo valor de k).

Dada la arbitrariedad de las condiciones (C.1) y (C.2) se convino en que se aplicarían también para los estimadores bayesianos. Se compararon los estimadores bayesianos con los otros estimadores de acuerdo a los 2 criterios mencionados anteriormente.

Resultados

En las tablas siguientes aparecen las comparaciones entre los 3 estimadores de acuerdo al lugar en que quedaron (1 si fue el mejor, 2 el segundo, 3 el peor). Dentro de los cuadros aparace el número de veces en que quedaron en cada lugar.

a) Comparación de acuerdo al criterio 1

$\alpha = .9$
$\gamma = .3$

	Estimador		
Lugar	MC	H.K.	Bayes
1	2	6	22
2	5	20	5
3	23	4	3

$\alpha = .9$

$\gamma = .6$

	Estimador		
Lugar	MC	H.K.	Bayes
1	3	2	25
2	5	22	3
3	22	6	2

$\alpha = .75$

$\gamma = .6$

	Estimador		
Lugar	MC	H.K.	Bayes
1	2	4	24
2	9	19	2
3	19	7	4

b) Comparación de acuerdo al criterio 2

$\alpha = .9$

$\gamma = .3$

	Estimador		
Lugar	MC	H.K.	Bayes
1	4	8	28
2	6	14	10
3	20	8	2

$\alpha = .9$

$\gamma = .6$

	Estimador		
Lugar	MC	H.K.	Bayes
1	3	4	23
2	7	18	5
3	20	8	2

$\alpha = .75$

$\gamma = .6$

	Estimadores		
Lugar	MC	H.K.	Bayes
1	3	4	23
2	9	18	3
3	18	8	4

Los resultados dan una clara superioridad del estimador bayesiano sobre el de Hoerl y Kennard y a la vez de este sobre el de mínimos cuadrados. La superioridad del estimador H.K. sobre el de mínimos cuadrados se puede explicar por el hecho de que la priori no se encuentra muy alejada de la realidad.

Otro hubiera sido el caso si se tomara, por ejemplo, $b_1=7$ $b_2=-3$, $b_3=6$. Pero la simulación se trató de que entre el estimador H.K. y el de Bayes, pudiera competir el primero en condiciones mas o menos aceptables. Comparación entre el método H.K. y el estimador de Swindel.

Procedimos a comparar para un mismo valor de k el estimador H.K. con el estimador bayesiano obteniendo los siguientes resultados:

	Estimador	
K	Bayes	H.K.
.1	19	11
.2	22	8
.3	26	4

Lo anterior indica que al incrementar la confianza en la prior el estimador bayesiano nos lleva a mejores resultados que a los que nos lleva el método de Hoerl y Kennard.

Una última comparación que hicimos fue tomar R para el estimador bayesiano, encontrar aproximadamente el mismo valor de R en los estimadores H.K. y comparar el estimador seleccionado con el bayesiano usando los criterios 1 y 2. En todos los casos el estimador bayesiano superó al de Hoerl y Kennard.

Conclusiones

Las conclusiones que se pueden obtener de esta pequeña simulación son limitadas. Sin embargo sí da un indicativo de las ventajas que tiene el dar un enfoque bayesiano a los estimadores H.K. Una vez comprendido el enfoque, si se desea utilizar una distribución a priori, no necesariamente la distribución a priori que se supone en el método H.K. será la que refleje más adecuadamente el conocimiento previo. La simulación indica como un conocimiento previo sólo un poco más cercano a la realidad puede ser más eficiente que el método H.K. aplicado rutinariamente.

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

(1) BANERJEE, K.S., y CARR, R.N. (1971). "A comment on ridge regression. -- Biased estimation for non-orthogonal problems". Technometrics, v. 13, -- 895-898.

(2) BOX, G.E.P. y G.C. TIAO (1972). Bayesian Inference in Statistical Analysis. Addison Wesley.

(3) COXE, K. (1975) "Do principal components solve multicollinearity? The - Longley data Revisited". Presentado en la Reunión Anual de la American-Statistical Association.

(4) DE ALBA GUERRA E. (1970). "Estimación de parámetros en modelos no lineales". Tesis Profesional de Actuario, UNAM.

(5) DEMPSTER, A.P. (1972). "Alternatives to least squares in multiple regression". Multivariate Statistical Inference, Kabe y Gupta.

(6) DRAPER, N., y SMITH, H. (1966). Applied Regression Analysis. New York: Wiley.

(7) DYKSTRA, O. (1971). "The augmentation of experimental data to maximize - X'X". Technometrics, v. 13, 682-688.

(8) GOLDSTEIN, M. (1976). "Bayesian Analysis of Regression Problems". Biometrika 63, 51-58.

(9) GUILKEY, D. Y MURPHY (1975). "Directed Ridge Regression Techniques in - Cases of Multicollinearity", JASA, 769-775.

(10) GUNST, R.F., WEBSTER, J.T. y MASON R.L. (1976). " A Comparison of Least-Squares and Latent Root Regression Analysis". Technometrics 18, 75-84.

(11) FAREBROTHER, R.W. (1975). "The minimun mean square error linear estima-- tor and ridge regression". Technometrics, V. 17, 127-128.

(12) GORMAN, J.W., y TOMAN, R.J. (1966). "Selection of variables for fitting equations to data". Technometrics, v. 8, 27-51.

(13) HEMMERLE, W.J. (1975). "An explicit solution for generalized ridge re-- gression". Technometrics, v. 17, 309-314.

(14) HOCKING, R.R. (1972). "Criteria for selection of a subset regression: -- which one should be used?" Technometrics, v. 14, 967-970.

(15) HOERL, A.E., y KENNARD, R.W. (1970). "Ridge regression: biased estima-- tion for non-orthogonal problems". Technometrics, v. 12, 55-67.

(16) HOERL, A.E., y KENNARD, R.W. (1970). "Ridge regression: applications to nonorthogonal problems". Technometrics, v. 12, 69-82.

(17) HOERL, A.E., y KENNARD, R.W. (1975). "A note on a power generalization- of ridge regression. "Technometrics, v. 17, 269.

(18) JOHNSON, A.F. (1974). En "Letters to the editor", Technometrics, v. 16, 641.

(19) JOHNSTON, J. (1972). Econometric Methods. Tokyo: McGraw-Hill.

(20) KUMAR, K. T. (1975). "Multicollinearity in Regression Analysis". The - Review of Economics and Statistics. 365-366.

(21) LAWSON, C.L., y HANSON, R.J. (1974). Solving Least Squares Problems. En glewood: Prentice-Hall.

(22) LINDLEY, D.V., y SMITH, A.F.M. (1972). "Bayes estimates for the linear- model". JRSS, Series B, v. 34, 1-18.

(23) MAYER, L.S., y WILLKE, T.A. (1973). "On biased estimation in linear models". Technometrics, v. 15, 497-508.

(24) MARQUARDT, D.W. (1970). "Generalized inverses, ridge regression, biased-linear estimation, and nonlinear estimation". Technometrics, v. 12, 591-612.

(25) MARQUARDT, D.W., y SNEE, R.D. (1975). "Ridge regression in practice". The American Statistician, v. 29, n. 1, 3-20.

(26) McCALLUM, B.T. (1970). "Artificial orthogonalization in regression analysis". The Review of Economics and Statistics, V. 52, 110-113.

(27) McDONALD, G.C., y GALARNEAU, D.I. (1975). "A Monte Carlo evaluation of some ridge-type estimators". JASA, v. 70, 407-416.

(28) MORRISON, D.F. (1967). Multivariate Statistical Methods. McGraw-Hill.

(29) OBENCHAIN, R.L. (1975) "Residual Optimality: Ordinary vs. Weighted vs. Biased Least Squares". JASA, 375-379.

(30) OBENCHAIN, R.L. (1975). "Ridge Analysis Following a Preliminary Test of The Shrunken Hypothesis". Technometrics 17, 413-445.

(31) O'HAGAN, J. y McCABE, B. (1975). "Tests for the severity of Multicollinearity in Regression Analysis. A comment". The Review of Economics and Statistics. 368-370.

(32) PRESS, S.J. y ZELLNER, A. (1968). "On Generalized Inverses and Prior Information in Regression Analysis". 1-14.

(33) RAO, C.R. (1973). Linear Statistical Inference and its Applications, second edition. Wiley.

(34) SEARLE, S.R. (1971). Linear Models. Wiley

(35) SIDIK, S. M. (1975) "Comparison of sone Biased Estimation Methods (In-- cluding Ordinary Subset Regression) in the Linear Model". NASA-Langley. 1-45.

(36) SILVEY, S.D. (1969). "Multicollinearity and imprecise estimation". JRSS, Series B, V. 31, 539-552.

(37) SWINDEL, B.F. (1974). "Good ridge estimators based on prior information". Presentado en la reunión anual de la American Statistical Association.

(38) THEIL, H. (1963). "On the Use of Incomplete Prior Information in Regres sion Analysis". JASA. 401-414.

(39) THEOBALD, C.M. (1974). "Generalizations of mean square error applied to ridge regression". JRSS, Series B, V. 36, 103-106.

(40) TUCKER, H.G. (1967). An introduction to Probability and Mathematical -- Statistics. Academic Press.

(41) WARD, J.F. (1974). "Restricted least squares and ridge estimators". Pre sentado en la reunión anual de la American Statistical Association.

(42) WEBSTER, J.T., y GUNST, R.F. y MASON R.L. (1974). "Latent root regression analysis". Technometrics, v. 16, 513-522.

(43) WICHERS, C.R. (1975). "The Detection of Multicollinearity. A comment". - The Review of Economics and Statistics. 366-368.

(44) WONNACOTT, R.J. y WONNACOTT, T.H. (1970). Econometrics. Ed. Wiley

(45) ZELLNER, A. (1971). An Introduction to Bayesian Inference in Econometrics. Ed. Wiley.

(46) DAVIS, J.C. (1973). Statistics and Data Analysis in Geology, Nueva York, Ed. Wiley.