49
25j

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

"MULTICOLINEALIDAD. EL PROBLEMA REVISADO"

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A :

RAUL JUAREZ SANCHEZ

Facultad de Ciencias
UNAM

VNIVERSIDAD NACIONAL
AVFNºMA DE
MEXICO

M. en C. Virginia Abrín Batule
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis:

"MULTICOLINEALIDAD. EL PROBLEMA REVISADO"

realizado por RAUL JUAREZ SANCHEZ

con número de cuenta 8809308-4 , pasante de la carrera de ACTUARIA

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	MAT. MARGARITA ELVIRA CHAVEZ CANO
Propietario	DR. JOSE RODOLFO MENDOZA BLANCO
Propietario	M. en C. BEATRIZ EUGENIA RODRIGUEZ FERNANDEZ
Suplente	MAT. LUIS BRISEÑO AGUIRRE
Suplente	M. en C. MARIANO LOZANO MARTINEZ

Consejo Departamental de Matemáticas

M. en C. ALEJANDRO BRAVO MOJICA

MATEMATICAS

A mi madre, Alberto, Dorothy y Juan Luis.

A la profesora Margarita Chávez Cano, a la profesora Beatriz Rodríguez Fernández, al profesor José Rodolfo Mendoza Blanco- propietarios de esta tesis -. A los profesores Luis Briseño Aguirre y Mariano Lozano García, suplentes, respectivamente que aceptaron amablemente ser sinodales de este trabajo.

Por último a Billy Joe Armstrong y al Carita.

ÍNDICE

INTRODUCCIÓN

En el modelo de análisis de regresión múltiple $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, en ciertas ocasiones se presenta el problema de multicolinealidad, el cual puede describirse someramente como la existencia de dependencias cercanas a la lineal entre un subconjunto de columnas de la matriz $\mathbf{X}$ de datos originales.

Este trabajo presenta una visión retrospectiva de los trabajos más importantes con respecto a la multicolinealidad que se han desarrollado en las últimas décadas; abarcando de manera extensa las proposiciones hechas por teóricos como Judge, Belsley, Kuhmar y Montgomery por citar algunos.

En el capítulo **I** explicamos los recursos o fuentes que pueden ocasionar que la multicolinealidad se presente, para poder saber cuándo ocurre esto último necesitamos previamente una definición clara del concepto de multicolinealidad, ésta definición es introducida también en el capítulo **I**.

En el capítulo **II** se presentan las consecuencias estadísticas de la multicolinealidad, una de las cuales es la inestabilidad de los coeficientes estimados, esto es, cuando al cambiar de una muestra a otra los coeficientes difieren en escaso margen y los estimadores de los coeficientes de regresión varían en una cuantía considerable.

En el capítulo III se introducen los distintos métodos para la detección del problema, entre los cuales podemos mencionar: la descomposición en proporciones de varianza, los factores de inflación de varianza, la matriz de correlación, entre otros. En este capítulo comparamos las virtudes y defectos de cada uno de estos métodos y recomendamos la aplicación de alguno en particular.

En el capítulo IV que creo es, junto con el capítulo III, la parte medular de este trabajo, se abordan con detalle las soluciones propuestas en los últimos años al problema de la multicolinealidad. Aquí hacemos una distinción entre el carácter de cada una de las soluciones las cuales pueden dividirse en dos: soluciones ad-hoc y las soluciones tradicionales. Entre los métodos más importantes de solución se encuentran: regresión ridge, regresión en componentes principales y aquellos métodos que se relacionan de manera directa con los recursos o fuentes que dan origen a la multicolinealidad.

Para finalizar, en el capítulo V realizamos el análisis completo de un modelo sencillo de regresión lineal propuesto por Judge, et. al. Este ejercicio es particularmente ilustrativo, ya que analizamos los métodos de detección y solución que consideramos más adecuados para ser aplicados en este problema en particular.

CAPÍTULO I

DEFINICIÓN Y FUENTES DE LA MULTICOLINEALIDAD

1.1. Definición

La obtención de una definición clara del problema conocido como multicolinealidad, es un trabajo al que se han dedicado varios teóricos, entre los cuales mencionamos a Harvey y Johnston (1963). Ellos la definen como una situación en la cual dos o más variables están altamente correlacionadas.

Otros, como Farrar y Glaubert (1967), la asocian con la "debilidad" o "deficiencia" de los datos. Este tipo de enfoques tienen a su favor su naturaleza meramente intuitiva pero no profundizan en los daños que la multicolinealidad ocasiona y las causas que la originan. Podemos encontrar serias debilidades en las definiciones propuestas por los teóricos arriba mencionados, entre las cuales mencionamos las siguientes:

i) Las variables predictoras no necesitan ser estocásticas

ii) Sugieren que una correlación elevada es inherente a las variables predictoras

iii) El término "deficiente" podría relacionarse con el proceso de recolección de los datos

Una definición más detallada que las anteriores, es la propuesta por Mason, Gunst y Webster (1974) , estos autores definen la multicolinealidad como una dependencia cercana a lo lineal entre las columnas de las variables predictoras. Partimos del modelo:

$$\mathbf{y} = \beta_0\mathbf{X}_0 + \beta_1\mathbf{X}_1 + \ldots + \beta_k\mathbf{X}_k + \varepsilon \qquad (1.1.1)$$

$\mathbf{X}_0 = \mathbf{1}' = (1, 1, \ldots, 1)$

$\mathbf{X} = [X_{ij}]$ es la matriz de datos iniciales

$$\varepsilon_i \sim NID(0, \sigma^2), \qquad i = 1, 2, \ldots, n \qquad (1.1.2)$$

Donde NID, indica que las variables aleatorias son normales idénticamente distribuidas. Estos autores la definen de la forma siguiente:

DEFINICIÓN: Sea un modelo de regresión lineal como en **(1.1.1)**. Si para alguna $\eta \geq 0$ específica, existe un vector $\mathbf{C}' = (C_0, C_1, \ldots, C_k)$, con no todos sus elementos iguales a cero, tal que:

$$\sum_{j=0}^{K} C_j\mathbf{X}_j = \delta, \qquad ||\delta|| \leq \eta\, ||\mathbf{C}'||, \qquad (1.1.3)$$

se dice que existe multicolinealidad entre las variables predictoras.

La importancia de la definición anterior radica en el hecho de hacer explícita la relación entre las variables predictoras, aunque no se haga aún explícito el origen o la causa de la multicolinealidad . La definición anterior puede traer a discusión la polémica permanente en torno a la ortogonalidad de las columnas de la matriz de datos **X**, en el

modelo $\mathbf{y = X\beta + \varepsilon}$. Algunos autores han tratado de medir la fuerza de la multicolinealidad basados en criterios que están relacionados con la no ortogonalidad de las columnas de la matriz.

Belsley, Kuh y Welsch (1980), asocian la multicolinealidad con la matriz de variables predictoras **X** únicamente; para ellos, la multicolinealidad no guarda relación con las propiedades estocásticas de **(1.1.1).**

Smith y Campbell (1980) retoman la definición anterior y aseveran que la no existencia de ortogonalidad entre las columnas no está relacionada directamente o indirectamente con el problema, puesto que las columnas de **X** pueden ser transformadas de tal forma que las nuevas columnas sean ortogonales. Estos autores han sido criticados por el razonamiento arriba mencionado. Los críticos de este trabajo puntualizan que el razonamiento seguido por aquellos ignora la influencia de las variables predictoras originales y no el de las transformadas, en la respuesta.

Estas críticas pueden ser resumidas en las dos siguientes:

i) Las variables transformadas son combinaciones lineales de los datos originales por lo que la definición no es invariante bajo cambio de escala.

ii) En un enfoque Bayesiano, es más fácil obtener información a-priori de los parámetros originales del modelo que en combinaciones lineales arbitrarias.

Gunst (1983), analiza las consecuencias de transformaciones lineales en el modelo **(1.1.1)** y en la definición **(1.1.3)**. Sea la transformación:

$$\mathbf{X}_j \rightarrow \mathbf{Z}_j = a_j \mathbf{X}_0 + b_j \mathbf{X}_j \ , \qquad j= 1, 2,... k$$

Para propósitos de la explicación ulterior, entenderemos:

$$\mathbf{X}_0 = \mathbf{1}' = (1, 1,..., 1)$$

Para detallar el impacto que tiene en nuestra definición el efecto de este tipo de transformaciones, retomamos la definición **(1.1.3)** , sustituyendo las variables transformadas, obteniendo así:

$$\sum_{j=0}^{K} \mathbf{C}_j^* \mathbf{Z}_j = \delta \ , \qquad ||\delta|| \leq \eta^* \ ||\mathbf{C}^*|| , \qquad \textbf{(1.1.4)}$$

con

$$\mathbf{Z}_0 = \mathbf{1}' = (1, 1,..., 1)$$

$$\mathbf{C}_0^* = \mathbf{C}_0 - \sum_{j=0}^{K} \mathbf{C}_j^* a_j \mathbf{C}_j b_j^{-1}, \qquad j = 1, 2,..., k$$

$$\eta^* = \frac{(\eta \ ||\mathbf{C}||)}{||\mathbf{C}^*||}$$

De aquí, la definición de multicolinealidad es invariante bajo cambios de escala y traslación del origen.

Para analizar más de cerca el efecto que tiene en el problema el hecho de centrar los datos, hacemos lo siguiente:

Sea m_j la media de n valores del vector $\mathbf{X}_j$ y defínase:

$$\mathbf{Z}_j = \mathbf{X}_j - m_j \mathbf{X}_0 , \qquad j = 1, 2, ..., k$$

si

$$\sum_{j=0}^{K} \mathbf{C}_j^* \mathbf{Z}_j = \delta , \quad \text{con} \qquad ||\delta|| \leq \eta ||\mathbf{C}^*|| ,$$

entonces

$$\sum_{j=0}^{K} \mathbf{C}_j \mathbf{X}_j = \sum_{j=0}^{K} \mathbf{C}_j^* \mathbf{Z}_j \qquad (1.2)$$

con

$$\mathbf{C}_0^* = \mathbf{C}_0 + \sum_{j=0}^{K} \mathbf{C}_j m_j \quad \text{y} \quad \mathbf{C}_j^* = \mathbf{C}_j \qquad j = 1, 2, ..., k$$

$\mathbf{Z}_0$ puede no parecer involucrada en el problema , cuando se usan datos centrados, ya que :

i) $\mathbf{Z}_0' \mathbf{Z}_j = 0$ para $j = 1, 2, ..., k$

ii) $\mathbf{C}_0^*$ es aproximadamente cero en **(1. 2)**

Si observamos que

$$\sum_{j=0}^{K} C_j^* X_j = \delta \Rightarrow C_0 + \sum_{j=0}^{K} C_j m_j = \bar{\delta} \Rightarrow C_0 = \bar{\delta}$$

Entonces, en la definición de multicolinealidad al escoger $\eta \, ||C||$ suficientemente pequeño, $C_0^* = \delta$ es aproximadamente cero ya que:

$$||\delta|| = \{ \sum_{i=1}^{n} (\delta_i - \delta) + n \delta^2 \}^{1/2} \leq \eta \, ||C||$$

Según Gunst, sin considerar si la multicolinealidad entre las variables predictoras involucra al vector constante **1'**, la multicolinealidad equivalente entre las variables predictoras centradas puede no aparecer para involucrar el término constante ya que C_0 es aproximadamente cero. Este autor revela también que una multicolinealidad puede surgir simplemente a causa de los orígenes de las variables predictoras no constantes que son mucho más grandes que cero. Si reescribimos la ecuación:

$$\sum_{j=0}^{K} C_j X_j = \sum_{j=0}^{K} C_j^* Z_j$$

como

$$\sum_{j=0}^{K} C_j X_j = (C_0 + (\sum_{j=1}^{K} C_j^* m_j)) X_0 + \sum_{j=1}^{K} C_j^* Z_j$$

Para cualquier Z_j fijo ($j = 1, 2, ..., k$) las C_j pueden elegirse suficientemente pequeñas de tal forma que $\sum_{j=0}^{K} C_j^* Z_j$ sea

arbitrariamente cercano a cero aún si no existe multicolinealidad entre las variables predictoras no constantes.

Marquardt y Snee (1975) recomiendan centrar y estandarizar las variables predictoras para eliminar el mal condicionamiento "no esencial" en la matriz **X**. Centrar las variables es recomendable para eliminar multicolinealidad debida a las fuentes de las variables predictoras .

La estandarización, a su vez, es recomendable, ya que las variables predictoras cuyas escalas son de órdenes de gran magnitud diferentes pueden conducir a una variedad de problemas numéricos asociados con la inversión de **X'X**.

Gunst (1983), recomienda estandarizar las variables de tal forma que se incremente la estabilidad numérica de las variables predictoras y se reduzca la posibilidad de interpretar incorrectamente la multicolinealidad cuando las escalas de las variables difieren en orden de magnitud.

Sigamos analizando las definiciones que han sido propuestas para el concepto de multicolinealidad. Judge, et. al. (1985), definen multicollnealidad como aquel problema que existe cuando hay al menos una dependencia lineal entre las columnas de la matriz de observaciones **X**, éste es el caso extremo, esta dependencia tiene

como consecuencia que la matriz **X** no sea de rango completo. Judge, et. al., parten de este caso extremo para realizar una discusión sobre la multicolinealidad, según ellos, en el caso en el que el rango de **X** sea menor que k, se cumple al menos una relación de la forma:

$$c_1\mathbf{X}_1 + c_2\mathbf{X}_2 + \ldots + c_k\mathbf{X}_k = 0 \qquad (1.1.5)$$

donde las c_i son constantes y no todas iguales a cero y $\mathbf{X}_i$ es la i-ésima columna de **X**. La presencia de relaciones como la **(1.1.5)** implica que ciertas funciones paramétricas reales lineales, tales que:

$$\mathbf{W}'\beta = \mathbf{W}_1\beta_1 + \mathbf{W}_2\beta_2 + \ldots + \mathbf{W}_k\beta_k$$

son estimables. De hecho $\mathbf{W}'\beta$ posee un estimador lineal insesgado si y sólo si $\mathbf{W}'$ puede ser expresado como una combinación de los renglones de **X**.

A continuación, analizaremos los recursos o fuentes de la multicolinealidad. Esto es, revisaremos algunas de las situaciones comunes que provocan posiblemente la aparición del problema.

1.2 Fuentes de la multicolinealidad

Los recursos de la multicolinealidad son aquellas situaciones relacionadas, en algunos casos con la matriz de datos y otras con el procedimiento de muestreo empleado. En este apartado, comentaré brevemente las opiniones al respecto de algunos autores como Montgomery y Peck (1982), Hooking, Pendleton (1983), Judge et. al. (1985), Mason, Gunst y Webster (1975).

Según Montgomery y Peck (1982) existen cuatro recursos de multicolinealidad , estos son:

1.2.1. El método empleado en la recolección de datos

Estos autores aseveran que el problema surge cuando el analista muestrea solamente sobre un subespacio de la región de las regresoras definidas (aproximadamente) por:

$$\sum_{j=0}^{K} t_j X_j = 0 \qquad (1.2.1.1)$$

En general, si existen más de dos regresoras, los datos caerán aproximadamente en un hiperplano definido en **(1.2.1.1).** La multicolinealidad ocasionada por la técnica de muestreo no es inherente al modelo ó a la población que está siendo muestreada.

Mason, Gunst y Webster (1975), afirman a su vez que la multicolinealidad puede ocurrir a causa de deficiencias muestrales, no obstante, al igual que Montgomery, reconocen el verdadero significado de "deficiencia muestral", para ellos, no significa falta de cuidado en el muestreo o en un diseño inadecuado sino que el proceso de recoleción de los datos produce resultados inesperados debido a que estos datos se encuentran en un subespacio de la región p-dimensional de las variables predictoras.

Mason et. al. (1975), señalan un punto importante. La multicolinealidad ocasionada por deficiencias muestrales es quizás, la

más común y la que puede causar más problemas. Debemos tener cuidado cuando tratemos de solucionar el problema y sepamos que este último tiene su origen en el proceso de muestreo.

1.2.2 Restricciones del modelo ó de la población

Según Montgomery y Peck (1982), las restricciones en el modelo ó en la población que está siendo muestrada, pueden ocasionar multicolinealidad. Para ilustrar este punto sugieren el siguiente ejemplo: se desea conocer el efecto del ingreso familiar (X_1) y el tamaño de la habitación (X_2) en el consumo eléctrico residencial. Aquí, es evidente que las restricciones físicas implicarán la existencia de multicolinealidad , la cual existirá sin considerar el método de muestreo empleado.

Gunst et. al. (1975), comentan a su vez, acerca de la multicolinealidad inherente a la población, según ellos, este tipo de multicolinealidad está relacionada a aquella que surge de la especificación del modelo, distinguiéndose de esta última en que se debe a características de la población y no a la definición de variables predictoras . Este tipo de multicolinealidad se debe a algunas características de la población particular o del fenómeno bajo estudio. Se puede esperar que éstas ocurran con cualquier conjunto de datos, sin considerar el método de muestreo utilizado. Estos autores, hacen una distinción relevante, la multicolinealidad debida a restriccion en la población, restringe a su vez el uso de diseños experimentales ya que

esta multicolinealidad no puede ser eliminada con la elección apropiada de unidades experimentales.

1.2.3 Especificación del modelo

Para Montgomery y Peck (1982), uno de los problemas que surgen cuando se ajusta un polinomio en una sola variable, es que al incrementar el orden del polinomio, la matriz $\mathbf{X'X}$ llega a ser mal condicionada ó, si los valores de $\mathbf{X}$ están limitados a un rango poco extenso, puede haber un mal condicionamiento significativo ó multicolinealidad en las columnas de la matriz $\mathbf{X}$. Por ejemplo, si $\mathbf{X}$ varía entre 1 y 2, la variable $\mathbf{X^2}$ varía entre 1 y 4 induciendo así multicolinealidad.

Gunst et. al. (1975), son más explícitos al comentar sobre esta causa de la multicolinealidad. Estos autores analizan el efecto de introducir multicolinealidad exacta a partir de la definición de las variables predictoras. En particular, Gunst et. al., comentan el siguiente experimento, un experimento mezcla (Cornell, 1981), en él, las variables predictoras se restringen de tal forma que:

$$\sum_{j=0}^{K} \mathbf{X}_{ij} = 1, \qquad i = 1,2,\ldots,k \qquad \textbf{(1.2.3.1)}$$

En dichos modelos, las columnas de $\mathbf{X}$ son linealmente dependientes si se incluye el término constante. De esta forma,

según ellos, si un subconjunto de k variables predictoras representa variables indicadoras para los k niveles de la variable categórica, ocasiona una restricción similar a **(1.2.3.1)**, esto es una dependencia lineal con el término constante.

Cuando se desconoce la forma correcta del modelo a utilizar, Gunst, et. al. (1975) reconocen que la posibilidad de introducir multicolinealidad debida a variables predictoras relacionadas exactamente ó de manera cercana a la lineal aún es mayor. Es el caso en el que, al introducir las variables cualitativas ó algunas variables que actúen como tales puede surgir multicolinealidad "espuria" entre dos ó más variables predictoras, ya que las variables están influídas simultáneamente por alguna otra variable ó fenómeno. Dentro del mismo artículo, estos autores reconocen que es de antemano esperada la multicolinealidad debida a la especificación del modelo, ya que si la multicolinealidad es inherente al modelo la persona que esté realizando la regresión debe saber que debe ser eliminada una de las variables. A pesar de ellos, estos autores no recomiendan la eliminación arbitraria de variables hasta que haya fuerte evidencia de que la fuente de la multicolinealidad sea inherente a la forma en que se especificó el modelo utilizado.

1.2.4 Sobredefinición del modelo

Para Montgomery y Peck (1982), definen un modelo sobredefinido como aquel que tiene más variables regresoras que observaciones.

Este tipo de casos es frecuente en las investigaciones médicas, en las cuales existe de antemano un número reducido de observaciones, y la información recolectada es insuficiente en comparación con el número elevado de variables regresoras involucradas. Montgomery aconseja en estos casos eliminar algunas de las variables regresoras en consideración.

Gunst (1983) coincide al respecto, al afirmar que tamaños de muestra insuficientes (i.e., $n < p+1$), donde p es el número de variables predictoras, producen multicolinealidad exacta entre las variables predictoras. Este autor comprende en este sentido que un modelo sobredefinido dé origen a la multicolinealidad ya que se tienen más variables ó parámetros en el modelo que observaciones con las cuales puedan estimarse las primeras. La multicolinealidad debida a este recurso en particular, no aporta en lo absoluto una indicación de que exista redundancia entre las variables predictoras. Este autor señala a su vez, que esta multicolinealidad puede desaparecer fácilmente al introducir un tamaño de muestra adecuado.

Estas son las cuatro principales fuentes ó recursos de la multicolinealidad según Montgomery y Peck (1982). Gunst (1983) concuerda con ellos, pero introduce una quinta fuente, la multicolinealidad asociada a los outliers.

1.2.5 Multicolinealidad asociada a outliers.

Según Gunst (1983), la multicolinealidad asociada a los outliers es similar a la ocasionada por deficiencias muestrales, ya que, para él, ambas comparten una naturaleza artificial. Este autor entiende por "outliers" la variación inusual que tienen una o más observaciones dentro de un grupo de datos. En el caso múltiple, un outlier puede tener valores muy grandes o pequeños en cada una de sus variables o en un subconjunto de las variables. Mencionamos anteriormente la naturaleza artificial de la multicolinealidad inducida por el outlier. Gunst (1983) señala que la multicolinealidad puede desaparecer rápidamente si eliminamos el outlier del conjunto de datos e ilustra este último comentario con el siguiente ejemplo: en el modelo **(1.1.1)**, con únicamente **p=2** variables predictoras y con X_{11} y X_{12} teniendo valores inusualmente grandes. Para un $k>0$ fijo, háganse $X_{11}=\varphi$ y $X_{21}=k\varphi$. Cuando $\varphi \to \infty$ se induce multicolinealidad entre $\mathbf{X_1}$ y $\mathbf{X_2}$ como lo demuestra al hacer uso de la definición **(1.1.3)**.

Este autor normaliza las columnas no constantes de **X** de tal manera que las dos tengan longitud unitaria:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{11} \\ 1 & X_{21} & X_{22} \\ & . & . \\ & . & . \\ 1 & X_{n1} & X_{n2} \end{bmatrix} = \begin{bmatrix} 1 & \varphi/\lambda_1 & k\varphi/\lambda_2 \\ 1 & X_{21}^*/\lambda_1 & X_{22}^*/\lambda_2 \\ & . & . \\ & . & . \\ 1 & X_{n1}^*/\lambda_1 & X_{n2}^*/\lambda_2 \end{bmatrix}$$

donde X_{i1}^*, X_{i2}^* representan los valores de las variables predictoras para i=1,2,...,n y

$$\lambda_1^2 = \varphi^2 + \sum_{i=1}^{n} X_{i1}^{*2} \quad , \quad \lambda_2^2 = k\,\varphi^2 + \sum_{i=2}^{n} X_{i2}^{*2}$$

Hace la observación de que cuando $\varphi \to \infty$, $X_{1j} \to 1$ y $X_{ij} \to 0$ para i=2,3,...,n. De aquí, para cada $\eta > 0$, existe un valor de φ, digamos φ^*, tal que para todo $\varphi \geq \varphi^*$, $X_1 - X_2 = \delta$ con $||\delta|| \leq \sqrt{2}\ \eta$ y de acuerdo a la definición, existe multicolinealidad .

En el presente capítulo hicimos una breve recapitulación de las definiciones más importantes que se han dado a la multicolinealidad, detallamos a su vez, el impacto que tienen las transformaciones lineales de las variables predictoras en la definición. Mostramos que la definición sigue cumpliéndose después de las transformaciones, posteriormente comentamos las fuentes o recursos de la multicolinealidad que son más ampliamente aceptados como tales.

CAPÍTULO II

CONSECUENCIAS ESTADÍSTICAS DE LA MULTICOLINEALIDAD

En el presente capítulo analizaremos algunos efectos de la multicolinealidad sobre los estimados de los parámetros, en particular comentaremos el caso en el que p=2, el número de los parámetros; este caso es particularmente interesante, por su simplicidad y por lo explícito que aparecen los efectos de la multicolinealidad.

2.1 Efectos de la multicolinealidad

Hocking y Pendleton (1983) proponen como recurso para visualizar la multicolinealidad su "valla de alambre", con lo cual, muestran un conjunto típico de datos con dos variables regresoras que son colineales. Dado que el ajuste de un plano regresor para estos datos es análogo al hecho de balancear el plano sobre los "puntos", el plano será inestable, esto es inestimable y las pendientes fuertemente influenciadas por un "alambre". La predicción estará sujeta a grandes variaciones si los datos son ligeramente modificados. Los autores citados, ilustran lo anterior analizando la situación siguiente: se desea desarrollar un modelo para predecir el promedio de puntos basado en las calificaciones obtenidas en exámenes de ciencia (X_1) y

matemáticas (X_2). Dado que X_1 y X_2 estarán fuertemente correlacionados, la situación anterior aparecerá como en la figura 1.

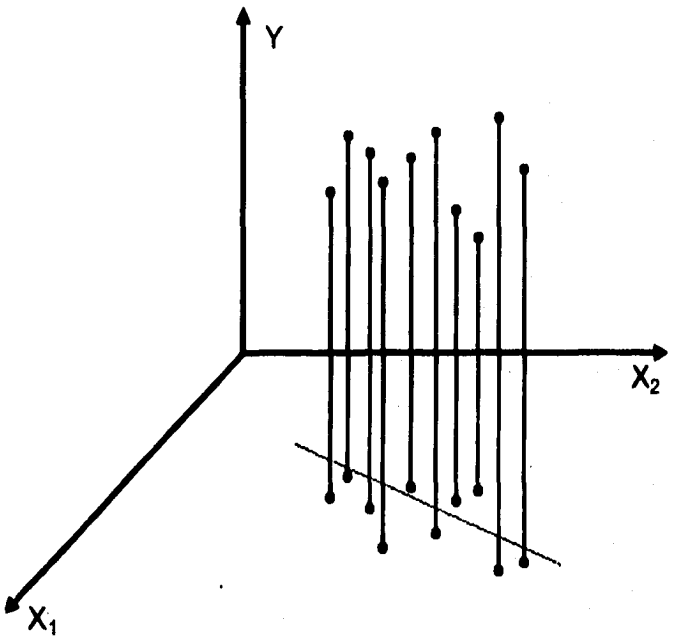

Figura 1. Las variables X_1 y X_2 están fuertemente correlacionadas.

Hocking y Pendleton (1983) señalan también, que la multicolinealidad puede causar problemas en el proceso de selección de variables. Un proceso de selección de variables intentará seleccionar un subconjunto de las variables originales que disminuya el error cuadrático medio (**ECM**) de la predicción o de la estimación. Ahora, dado que las variables predictoras presentan correlación, ya sea en la forma de correlación simple o como estos autores señalan, en

forma de relaciones lineales más complejas, los métodos de selección de variables pueden no coincidir. En otras palabras, si borramos alguna o algunas de las variables del conjunto original, el error cuadrático medio no disminuye.

El enfoque seguido por Belsley, Kuh y Welsch (1980), es distinto, según estos autores la multicolinealidad atribuye sus características a aquéllas de la matriz de datos **X** y no a los aspectos estadísticos del modelo $\mathbf{y} = \mathbf{X}\beta + \varepsilon$. Según esta última afirmación, la multicolinealidad es un problema de los datos y no un problema estadístico. No obstante, ellos comparten la preocupación por los daños que puede ocasionar la multicolinealidad, a la eficiencia de la estimación por mínimos cuadrados. En cualquier caso una exposición gráfica del problema es más intuitiva, de esta forma estos autores, detallan la naturaleza de la multicolinealidad geométricamente en la figura 2.

El modelo es, en esta ocasión:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad i=1,2,...,n$$

caso en el cual se tienen tres parámetros a estimar. Dado que es difícil separar la influencia de las variables explicativas en la variable de respuesta y debemos tener cuidado al eliminarlas. En la figura 2 se muestran la forma en que se encuentran dispersos los datos.

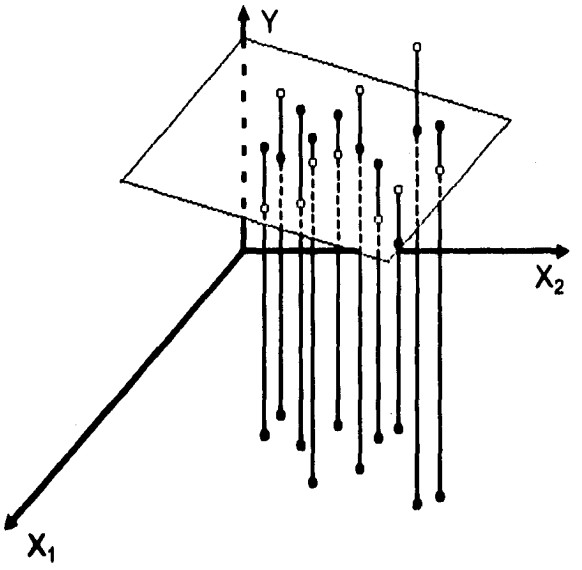

Figura 2. No existe colinealidad entre X_1 y X_2 - todos los coeficientes de regresión están bien determinados. Una variación pequeña en un parámetro del plano de regresión causará un cambio relativamente grande en la suma de cuadrados de los residuales.

En el plano **X1, X2** están los puntos (X_1, X_2), (los puntos se denotan por ♦), en la parte de arriba se muestra la "nube" de datos que resulta cuando es incluida la dimensión Y (los puntos se denotan por ■). Esta nube de datos provee un plano de mínimos cuadrados bien definido. En este caso, dado que el plano está bien definido, los parámetros son estimados con precisión. La figura 3, esboza el caso en el que existe fuerte colinealidad, el caso en el que existe colinealidad casi perfecta puede verse en la figura 1. En la figura 3, el plano de mínimos cuadrados está mal definido en el sentido de que

apuntándolo a lo largo del eje mayor de la nube resulta en una pequeña variación de la suma de cuadrados de los errores. Belsley et. al., anticipan en este sentido una de las consecuencias más importantes: la elevada varianza que puede deberse al hecho de que los estimadores de mínimos cuadrados sean imprecisos, esto es, estén asociados a elevadas variaciones cuando se toman muestras ligeramente distintas.

Para Belsley et al (1980)., es importante la habilidad personal para diagnosticar multicolinealidad en la regresión lineal. Regresaremos a este apartado en el capítulo 4 cuando tratemos los diagnósticos de la multicolinealidad, que son de cierta manera intuitivos y se basan en ocasiones en la aplicación de procedimientos de inspección gráficos.

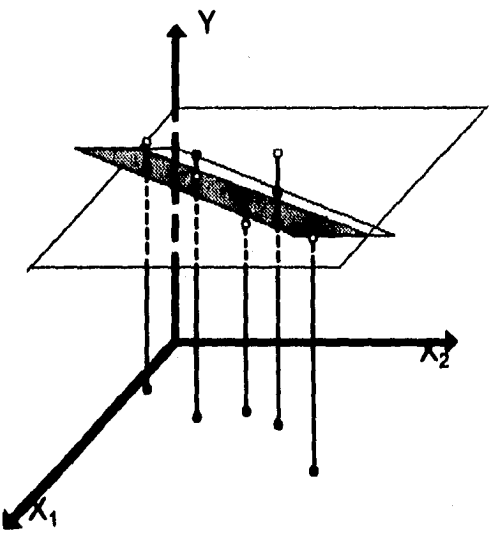

Figura 3. Colinealidad severa. Todos los coeficientes de regresión están mal determinados. Un cambio simultáneo en todos los parámetros del plano de regresión puede causar pequeños cambios en la suma de cuadrados de los residuales.

Montgomery y Peck (1982) han revisado la situación de los efectos potenciales en los estimadores de mínimos cuadrados. En su trabajo, se restringen a analizar el modelo siguiente, suponiendo que $\mathbf{X_1}$, $\mathbf{X_2}$ y $\mathbf{y}$ han sido cambiadas de escala a longitud unitaria

$$\mathbf{y} = \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \varepsilon$$

Como sabemos, las ecuaciones normales son

$$(\mathbf{X'X})\hat{\beta} = \mathbf{X'y}$$

En este caso y bajo la suposición de que $\mathbf{X_1}$, $\mathbf{X_2}$ y $\mathbf{y}$ han sido cambiadas de escala a longitud unitaria

$$\mathbf{X'X} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}$$

por lo que

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

donde r_{12} es la correlación simple entre $\mathbf{X_1}$ y $\mathbf{X_2}$ y r_{jy} es la correlación simple entre $\mathbf{X_j}$ y $\mathbf{y}$, j=1,2. La inversa de $\mathbf{X'X}$ es

$$\mathbf{C} = (\mathbf{X'X}) = \begin{bmatrix} 1/(1-r_{12}^2) & -r_{12}/(1-r_{12}^2) \\ -r_{12}/(1-r_{12}^2) & 1/(1-r_{12}^2) \end{bmatrix} \quad (2.1.1)$$

y los estimadores de los coeficientes de regresión son:

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12} r_{2y}}{(1 - r_{12}^2)} \quad , \quad \hat{\beta}_2 = \frac{r_{2y} - r_{12} r_{1y}}{(1 - r_{12}^2)} \qquad (2.1.2)$$

En el caso en el que existe una fuerte multicolinealidad entre X_1 y X_2, el coeficiente de correlación r_{12} será grande. En la expresión **(2.1.1)**, cuando $|r_{12}| \to 1$, y dado que

$\text{Var}(\hat{\beta}_j) = C_{jj}\sigma^2 = 1/(1-r_{12}^2)\,\sigma^2 \to \infty$ y con $(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2$ dependiendo de si $r_{12} \to 1$. La multicolinealidad severa tiene como consecuencia grandes varianzas y covarianzas para los estimadores de mínimos cuadrados, si bien, como Montgomery señala ésta no es la única causa de ello.

Consideremos, ahora en un modelo más general, en este nuevo modelo con más variables, la multicolinealidad produce efectos similares. Si las X_i ($i = 1,2,...,p$) y **Y** han sido cambiadas de escala y tienen longitud unitaria, los elementos de la matriz $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ son

$$C_{jj} = \frac{1}{(1 - R_j^2)} \quad , \qquad j = 1,2,...,p \qquad (2.1.3)$$

donde R_j^2 es el coeficiente de determinación múltiple de la regresión de X_j, en las restantes (p -1) variables regresoras. De **(2.1.3)** puede verse, que en el caso de existir una fuerte multicolinealidad entre X_j y un subconjunto de las (p -1) variables regresoras el valor de R_j^2 será cercano a la unidad.

Además, dado que $Var(\hat{\beta}_j) = C_{jj}\sigma^2 = 1/(1-R_j^2)\sigma^2$ una multicolinealidad severa provoca que la varianza del estimador de mínimos cuadrados del coeficiente β_j sea muy grande.

Montgomery señala el efecto que produce la multicolinealidad en el valor absoluto de los estimadores de mínimos cuadrados. Ejemplifican lo anterior, considerando la distancia al cuadrado de $\hat{\beta}$ al vector real de parámetros β, esto es,

$$L_1^2 = (\beta - \hat{\beta})'(\beta - \hat{\beta}) \qquad (2.1.4)$$

Posteriormente calculamos la esperanza de la distancia cuadrada L_1^2

$$\begin{aligned} E(L_1^2) &= E((\hat{\beta} - \beta)'(\hat{\beta} - \beta)) \\ &= \sum_{j=1}^{p} E(\hat{\beta}_j - \beta_j)^2 \\ &= \sum_{j=1}^{p} Var(\hat{\beta}_j) \\ &= \sigma^2 Tr(X'X)^{-1} \end{aligned} \qquad (2.1.5)$$

Además, con la presencia de multicolinealidad y al analizar la matriz $X'X$ desde la perspectiva del análisis numérico y no estadístico se concluye que la matriz $(X'X)^{-1}$ está mal condicionada, esto es, algunos de los valores propios de la matriz $X'X$ son pequeños. Montgomery utiliza el hecho conocido de que la traza de una matriz simétrica es igual a la suma de sus valores propios, por lo que (2.1.5) toma la forma:

$$E(L_1^2) = \sigma^2 \sum_{j=1}^{p} (1/\lambda_j) \qquad \lambda_j > 0, \qquad j=1,2,...,p$$

donde las λ_j son los valores propios de $\mathbf{X'X}$.

El mal condicionamiento de la matriz $\mathbf{X'X}$, implica que la distancia del estimador de mínimos cuadrados $\hat{\beta}$ al parámetro real β puede ser grande. Más adelante discutiremos con mayor detalle el significado de esto.

Dentro de un enfoque Bayesiano, la expresión L_1^2 correspondería a la función de pérdida:

$$L_1^2 = (\hat{\beta} - \beta)' (\hat{\beta} - \beta)$$

con una correspondiente función de riesgo, esto es la pérdida esperada al utilizar la función de pérdida L_1^2

$$\rho(\beta - \hat{\beta}) = E[(\hat{\beta} - \beta)' (\hat{\beta} - \beta)]$$

que corresponde a $E(L_1^2)$ en **(2.1.5)**.

Es interesante detenernos en este punto, para comentar que sucederá, si en el modelo

$$\mathbf{y = X\beta + \varepsilon}$$

se realizan transformaciones lineales particulares. En el capítulo II, veíamos, que las transformaciones lineales no producían efecto en la multicolinealidad del problema. En la sección siguiente introducimos la transformación que conduce al modelo ortonormal, modelo que volverá

a ser tomado cuando comentemos acerca de la regresión en componentes principales.

2.2 El modelo ortonormal

En nuestro modelo $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, suponemos que $\mathbf{X}$ es una matriz de diseño (nxk), de rango k y de este modo $\mathbf{X}'\mathbf{X}$ es una matriz (kxk) simétrica definida positiva. Utilizando **A.22** del apéndice y dado que $\mathbf{A}$ es una matriz cuadrada, real, simétrica existe una matriz $\mathbf{S}^{1/2}$, tal que

$$\mathbf{S}^{-1/2}\mathbf{X}'\mathbf{X}\,\mathbf{S}^{-1/2} = \mathbf{I}_n$$

donde,

$$\mathbf{S}^{1/2}\mathbf{S}^{1/2} = \mathbf{S} = \mathbf{X}'\mathbf{X} = \mathbf{C}'\Lambda\mathbf{C}$$

$$\Lambda = \mathbf{CSC}', \text{ y } \mathbf{S}^{-1/2} = \mathbf{C}'\Lambda^{-1/2}\mathbf{C}$$

de aquí, $\mathbf{C}$ es una matriz ortogonal de orden (nxn) y Λ es una matriz diagonal cuyos elementos son las raíces características de $\mathbf{X}'\mathbf{X}$. Por lo tanto, podemos escribir el modelo

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

en forma equivalente como

$$\mathbf{y} = \mathbf{X}\,\mathbf{S}^{-1/2}\mathbf{S}^{1/2}\beta + \varepsilon = \mathbf{Z}\theta + \varepsilon \qquad (2.2.1)$$

donde

$$\theta = \mathbf{S}^{1/2}\beta,\ \mathbf{Z} = \mathbf{X}\mathbf{S}^{-1/2}$$

y de aquí

$$Z'Z = I_k$$

La media y la varianza de los vectores aleatorios ε y $\mathbf{y}$ permanecen sin cambios. Esta transformación se conoce como la reducción canónica, y en este caso se ha efectuado una transformación lineal del espacio β al espacio θ que, reparametriza el problema.

En el modelo **(2.2.1)**, el estimador de mínimos cuadrados de $\theta = S^{1/2}\beta$ es

$$\hat{\theta} = S^{1/2}\hat{\beta} = (S^{-1/2}X'XS^{-1/2})^{-1}S^{-1/2}X'y = (Z'Z)^{-1}Z'y = Z'y \quad (2.2.2)$$

con matriz de varianzas y covarianzas

$$E[S^{1/2}(\hat{\beta}-\beta)(\hat{\beta}-\beta)' S^{1/2}] = \sigma^2 S^{1/2}(X'X)^{-1} S^{1/2} = \sigma^2 I_n \quad (2.2.4)$$

donde supondremos que ε se distribuye normal en el caso en el que los elementos del vector de coeficientes θ, sean normales, independientes e idénticamente distribuidos. Todas las propiedades muestrales que se conocen para el modelo simple de mínimos cuadrados se cumplen para este modelo reparametrizado.

Dado que $Z'Z = I_n$, **(2.2.1)** puede escribirse como

$$Z'y = \theta + Z'\varepsilon$$

ó

$$Z = \theta + w \quad (2.2.4)$$

donde Z tiene una distribución normal multivariada con vector de medias θ y matriz de varianzas y covarianzas no singular I_n.

CAPÍTULO III

DETECCIÓN DE MULTICOLINEALIDAD

3.1 Introducción

La determinación de la severidad y la forma de las dependencias cercanas a la lineal entre las columnas de la matriz **X** es un paso inicial necesario antes de intentar remediar la situación para el caso de diseños muestrales pobres. En el presente capítulo, haremos una revisión histórica de los métodos usados sin dedicarnos específicamente a la defensa de alguno en particular. Entre otros, mencionaremos los propuestos por Farrar y Glauber (1967), Kuhmar (1975), Marquardt y Snee (1975), Belsley, Kuh y Welsch (1980), Gunst (1983), entre otros.

Durante este capítulo, haremos uso de algunos resultados acerca de matrices y vectores propios, así como de valores propios. Cada uno de estos resultados se encuentra en el Apéndice.

3.2 Identificación de la multicolinealidad

Retomando el modelo estandarizado y redefiniendo el modelo de regresión múltiple de la forma siguiente:

$$y = \alpha_0 1' + \alpha_1 Z_1 + \ldots + \alpha_k Z_k + \varepsilon \qquad (3.2.1)$$

donde

$$\alpha_0 = \beta_0 + \sum_{j=1}^{k} \beta_j m_j , \quad 1' = (1, 1, 1, \ldots, 1)$$

$$\alpha_j = \beta_j d_j,$$

$$Z_j = (x_j - m_j 1') / d_j,$$

y

$$d_j^2 = (x_j - m_j 1)' (x_j - m_j 1)$$

Si

$$Z = (1, Z_1, Z_2, \ldots, Z_k),$$

entonces tenemos que:

$$1'(Z_1, Z_2, \ldots, Z_k) = 0',$$

Gunst (1983) propone modificar la definición (**1.1.3**), a fin de obtener otra que especifique más claramente la multicolinealidad entre las variables predictoras. La definición propuesta por él es la siguiente:

DEFINICIÓN: Sea un modelo de regresión lineal múltiple definido como en (**3.2.1**). Si para alguna $\eta \geq 0$ especificada, existe un vector $C' = (C_1, C_2, \ldots, C_k)$, con no todos sus elementos iguales a cero, tales que

$$\sum_{j=1}^{k} C_j Z_j = \delta \quad \text{con} \; ||\delta|| \leq \eta \, ||C|| \qquad (3.2.2)$$

entonces se dice que existe multicolinealidad entre las variables predictoras no-constantes. Menciona entre algunos de los métodos más eficaces para detectar multicolinealidad, los siguientes: coeficientes de correlación por parejas, factores de inflación de varianza y las medidas basadas en las raíces características (valores propios) y vectores característicos (vectores propios) de la matriz $\mathbf{Z'Z}$. A continuación veremos el primero de ellos, el análisis de la matriz de correlación.

3.2.1 Análisis de la matriz de correlación

Gunst (1983) subraya que este método ha sido uno de los más utilizados para detectar multicolinealidad . Si r_{jk} es el coeficiente de correlación entre $\mathbf{Z}_j$ y $\mathbf{Z}_k$, en el modelo (3.2.1). Cuando $|r_{jk}| \rightarrow 1$ los vectores $\mathbf{Z}_j$ y $\mathbf{Z}_k$ se aproximan a la dependencia lineal. Veamos esto último con más detalle.

De la definición (3.2.2), si $|r_{jk}| \geq 1 - \eta^2$, existe multicolinealidad entre $\mathbf{Z}_j$ y $\mathbf{Z}_k$. El valor exacto de η a ser utilizado en (3.2.2) puede determinarse por las condiciones que relacionan las metas del investigador o considerando la naturaleza de las variables predictoras, y puede ser función de n y k.

Además, el grado en el que $|r_{jk}|$ excede a $1 - \eta^2$ puede utilizarse para indicar la fuerza de la multicolinealidad. Este autor no fué el único en comentar acerca de este método, Montgomery, et. al. (1982),

también indicaron este método y propusieron el análisis de la matriz de correlación como un método sencillo de detección, basta -según ellos- inspeccionar los elementos arriba de la diagonal, los r_{ij} en **X'X**, notemos que aquí se supone que la matriz **X** ha sido centrada y cambiada de escala, suposición que también hace Gunst (1983). Si las regresoras X_i y X_j están involucradas en una dependencia cercana a la lineal, entonces $|r_{ij}|$ será cercana a la unidad.

Para ilustrar lo anterior, en la tabla 3.1 se presentan datos relacionados al porcentaje de conversión del n-heptano o acetileno y tres variables explicativas (Marquardt y Snee, (1975)). Como una superficie en las tres regresoras se considera adecuado para ajustar los datos del proceso químico, se propone el modelo:

$$P = \delta_0 + \delta_1 T + \delta_2 H + \delta_3 C + \delta_{12} TH + \delta_{13} TC + \delta_{23} HC + \delta_{11} T2 + \delta_{22} H^2 + \delta_{33} C^2 + \varepsilon$$

donde

P = Porcentaje de conversión

$$T = \frac{\text{Temperatura} - 1212.50}{80.623}$$

$$H = \frac{H_2 / (\text{n-heptano}) - 12.44}{5.662}$$

$$C = \frac{\text{Tiempo de contacto} - 0.0403}{0.0318}$$

Cada una de las regresoras ha sido centrada y cambiada de escala. Además, ya que tiempo de contacto y temperatura del reactor están altamente relacionados, existen problemas potenciales de multicolinealidad.

Tabla 3.1 Datos del acetileno

Observación i	Conversión n- heptano a acetileno	Temperatura del reactor (^{0}C)	Razón de H2 al n-heptano (Razón Molar)	Tiempo de contacto (Segundos)
1	49.0	1300	7.50	0.0120
2	50.2	1300	9.00	0.0120
3	50.5	1300	11.00	0.0115
4	48.5	1300	13.50	0.0130
5	47.5	1300	17.00	0.0135
6	44.5	1300	23.00	0.0120
7	28.0	1300	5.30	0.0400
8	31.5	1200	7.50	0.0380
9	34.5	1200	11.00	0.0320
10	35.0	1200	13.50	0.0260
11	38.0	1200	17.00	0.0340
12	38.5	1200	23.00	0.0410
13	15.0	1100	5.30	0.0840
14	17.0	1100	7.50	0.0980
15	20.5	1100	11.00	0.0920
16	29.5	1100	17.00	0.0860

De esta forma la matriz X'X en la forma de correlación para los datos anteriores es:

$$\mathbf{X'X} = \begin{bmatrix} 1 & 0.224 & -0.958 & -0.132 & 0.443 & 0.205 & -0.271 & 0.031 & -0.577 \\ & 1 & -0.24 & 0.039 & 0.192 & -0.023 & -0.148 & 0.498 & -0.224 \\ & & 1 & 0.194 & -0.661 & -0.274 & 0.501 & -0.018 & 0.765 \\ & & & 1 & -0.265 & -0.975 & 0.246 & 0.398 & 0.274 \\ & & & & 1 & 0.323 & -0.972 & 0.126 & -0.972 \\ & & & & & 1 & -0.279 & -0.374 & 0.358 \\ & & & & & & 1 & -0.124 & 0.874 \\ & & & & & & & 1 & -0.158 \\ & & & & & & & & 1 \end{bmatrix}$$

La matriz **X'X** revela la elevada correlación entre la temperatura del reactor (X_1) y el tiempo de contacto (X_3), lo cual ya se intuía desde el principio del análisis, ya que $r_{13} = -0.958$. Además, existen otros coeficientes de correlación elevados entre X_1X_2 y X_2X_3, X_1X_3 y X_1^2, X_2^2 y X_3^2. Concluyen que esto no es sorprendente ya que estas variables se generaron de los términos lineales e involucran a regresoras altamente correlacionadas X_1 y X_3.

Según Belsley, et. al.(1980), en el análisis de la matriz de correlación, de la inversa de ésta se debe ser cuidadoso al hacer uso de este método, ya que mientras un coeficiente de correlación elevado entre dos variables explicativas puede señalar un posible problema de multicolinealidad. La ausencia de correlaciónes altas no puede tomarse como evidencia de la ausencia del problema. Esto último es ilustrado por Montgomery, et. al (1982), que consideran los datos de la tabla 3.2. Estos datos fueron generados artificialmente por Webster, Gunst y Mason (1974). Se requería que:

$\sum_{j=1}^{4} X_{ij} = 10$ para las observaciones 2 a 12, mientras que,

$\sum_{j=1}^{4} X_{1j} = 11$ para la observación 1

Las respuestas fueron generadas de la relación

$$y_i = 10 + 2.0\,X_{i1} + 1.0\,X_{i2} + .02\,X_{i3} - 2.0\,X_{i4} + 3.0\,X_{i5} + 10.0\,X_{i6} + \varepsilon_i$$

donde $\varepsilon_i \sim N(0,1)$

Tabla 3.2

Observación	y_i	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	X_{i6}
1	10.006	8.000	1.000	1.000	1.000	0.541	-0.099
2	9.737	8.000	1.000	1.000	0.000	0.130	0.070
3	15.087	8.000	1.000	1.000	0.000	2.116	0.115
4	8.422	0.000	0.000	9.000	1.000	-2.397	0.252
5	8.625	0.000	0.000	9.000	1.000	-0.046	0.017
6	18.289	0.000	0.000	9.000	1.000	0.365	1.504
7	5.958	2.000	7.000	0.000	1.000	1.996	-0.885
8	9.313	2.000	7.000	0.000	1.000	0.228	-0.055
9	12.960	2.000	7.000	0.000	1.000	1.380	0.502
10	5.541	0.000	0.000	0.000	10.000	-0.798	-0.399
11	8.756	0.000	0.000	0.000	10.000	0.257	0.101
12	10.937	0.000	0.000	0.000	10.000	0.440	0.432

Para estos datos, la matriz **X'X** en la forma de correlación es

$$\mathbf{X'X} = \begin{bmatrix} 1 & 0.052 & -0.343 & -0.498 & 0.417 & -0.192 \\ & 1 & -0.4232 & -0.371 & 0.485 & -0.317 \\ & & 1 & -0.355 & -0.505 & 0.494 \\ & & & 1 & -0.215 & -0.087 \\ & & & & 1 & -0.123 \\ & & & & & 1 \end{bmatrix}$$

De la matriz pueden verse las correlaciones por parejas r_{ij}, ninguna de ellas es elevada y no puede concluirse en lo absoluto que existan dependencias basándonos básicamente en la matriz **X'X**. Es por esta razón que Belsley, et. al., dicen que es claramente posible para tres o más variables ser colineales aún cuando en parejas no estén altamente correlacionadas. Por ello, la matriz de correlación es totalmente incapaz de diagnosticar tal situación.

A continuación comentaremos un método que está relacionado con el análisis de la matriz de correlación, el análisis de los llamados factores de inflación de varianza.

3.2.2 Factores de inflación de varianza

En los análisis anteriores, hemos supuesto que la matriz **X** de datos había sido centrada y cambiada de escala a longitud unitaria, por lo cual $\mathbf{R}^{-1} = (\mathbf{X'X})^{-1}$, donde por **R** entenderemos la matriz de correlación de las variables explicativas.

Los elementos de la diagonal de $\mathbf{R}^{-1}$, los r^{jj} se conocen como factores de inflación de varianza, los $\mathbf{VIF}_j$ [Chatterjee y Price (1977)] y su valor en el diagnóstico se sigue de la relación

$$\mathbf{VIF}_j = \frac{1}{(1 - R_j^2)}$$

donde R_j^2 es el coeficiente de correlación múltiple de $\mathbf{X}_j$ que surgen al hacer una regresión de $\mathbf{X}_j$ en las variables explicativas restantes.

Para Belsley, et. al., un alto **VIF** indica un R_j^2 cercano a uno y de aquí multicolinealidad. Una ventaja de este método, es su uso como indicación global de multicolinealidad . Su debilidad, como la del análisis de la matriz de correlación, es su incapacidad para distinguir entre varias dependencias coexistentes.

Montgomery, et. al. (1982), sugieren incluso valores derivados de la experiencia práctica, según ellos si alguno de los VIF excede 5 ó 10,

esto es una indicación de que los coeficientes de regresión están siendo estimados pobremente a causa de la multicolinealidad. Estos autores, sugieren una interpretación interesante para los VIF. Dado que la longitud del intervalo de confianza (teoría normal) del j-ésimo coeficiente de regresión puede escribirse como

$$L_j = 2 (C_{jj} \hat{\sigma}^2)^{1/2} t_{\alpha/2,\, n-p-1}$$

y la longitud del intervalo correspondiente basado en un diseño de referencia ortogonal con el mismo tamaño de muestra y valores de la raíz media cuadrada (rms) [i.e. $rms = \sum_{i=1}^{n} (X_{ij} - \overline{X_j})^2/n$ es una medida de la desviación del regresor X_j] como el diseño original es

$$L^* = 2 \hat{\sigma}\, t_{\alpha/2,n-p-1}$$

La razón de estos dos intervalos de confianza es $L_j / L^* = C_{jj}^{1/2}$. Donde $C = [X'X]^{-1} = [C_{ij}]$. Así la raíz cuadrada del j-ésimo VIF indica la medida en la que será más grande el intervalo de confianza debido a la multicolinealidad.

3.3 La técnica de Farrar y Glauber

Farrar y Glauber (1967) critican algunos de los métodos comúnmente empleados en la detección y solución del problema de la multicolinealidad, proponiendo además un método basado en el empleo de información de R y R^{-1}. Como ellos mismos comentan, el análisis del determinante de la matriz $X'X$ no es suficiente para proporcionar información sobre la interacción que ejercen las distintas variables

entre sí. Para ellos, la relación entre ortogonalidad y el valor del determinante de la matriz **X'X**

$$0 \leq |\mathbf{X'X}| \leq 1$$

ha sido discutida con un enfoque eminentemente numérico. De aquí, debería ser posible intuir propiedades de distribución bajo la hipótesis de ortogonalidad y el valor del determinante $|\mathbf{X'X}|$, ó por medio de una transformación conveniente de $|\mathbf{X'X}|$ para obtener una estadística que provea una medida útil de la presencia y la severidad de la multicolinealidad.

Estos autores suponen que **X** tiene distribución normal multivariada. Por un resultado obtenido por Wishart, las varianzas muestrales y la covarianzas, se distribuyen conjuntas de acuerdo a la función de densidad que lleva su nombre. Utilizan además la hipótesis adicional de ortogonalidad y obtienen los momentos y la distribución para las matrices de correlación de la muestra. Se demuestra que el r-ésimo momento de $|\mathbf{X'X}|$ es

$$M_r(|\mathbf{X'X}|) = \frac{[\,\Gamma((n-1)/2)\,]^{p-1} \prod_{i=2}^{p} \Gamma((n-i)/2 + r)\,]}{[\,\Gamma((n-1)/2 + r)\,]^{p-1} \prod_{i=2}^{p} \Gamma((n-i)/2)\,]} \qquad (3.3.1)$$

donde n es el tamaño de la muestra y p=k+1 el número de variables.

Teóricamente, podría obtenerse la función de densidad para $|\mathbf{X'X}|$ de **(3.3.1)** pero para valores p>2, las soluciones explícitas no son fácilmente obtenibles.

Bartlett, al comparar los momentos de menor orden de **(3.3.1)** con los de la distribución χ^2, obtuvo una transformación de $|\mathbf{X'X}|$

$$\chi^2_{|\mathbf{X'X}|}{}^{(v)} = -[\, n - 1 - 1/6\,(2p + 5)\,] \log|\mathbf{X'X}|\,]$$

que se distribuye aproximadamente como χ-cuadrada con $v=\frac{1}{2}p(p-1)$ grados de libertad.

Como O'Hagan y McCabe (1974), lo señalan, esta estadística nos permite obtener enunciados acerca de la presencia de multicolinealidad en la población, si uno define multicolinealidad en términos de desviaciones de la ortogonalidad. Estos autores puntualizan que esta estadística nos dice únicamente el nivel de probabilidad al cual la hipótesis nula de ortogonalidad en la población sería rechazada.

Farrar y Glauber (1967) son explícitos al respecto, al señalar que los niveles de probabilidad proveen una medida cardinal de la extensión en la cual las **X** son interdependientes. Belsley, et. al. (1980), señalan que Farrar y Glauber propusieron el uso de la medida

$$r_{ij} = \frac{-r^{ij}}{(\sqrt{r^{ii}})(\sqrt{r^{jj}})}$$

esto es, la correlación parcial entre $\mathbf{X}_i$ y $\mathbf{X}_j$ ajustada para todas las otras variables **X**, para investigar las tendencias de interdependencia en

mayor detalle; critican a su vez, que así como **R**, det(**R**) no puede diagnosticar la presencia de varias dependencias cercanas simultáneas. Algunos otros investigadores, como Kumar (1975) han criticado esta medida, basada en desviaciones de la ortogonalidad de las columnas de **X**, esto a causa de que, la medida propuesta por Farrar y Glauber (1967) frecuentemente indica multicolinealidad, cuando de hecho, ésta no existe.

Otros, como Kumar (1975), puntualizan que la hipótesis supuesta por Farrar y Glauber (1967) de que la matriz de datos **X** sea estocástica, no tiene relevancia en el modelo de regresión lineal estándar en el cual **X** se supone fija.

Welsch, et. al., señalan que el uso de los elementos $r_{ij} = - r^{ij} / (\sqrt{r^{ii}})(\sqrt{r^{jj}})$ por Farrar y Glauber (1967), podría indicar correlaciones elevadas aún cuando las variables X_i, X_j no estuvieran involucradas en ninguna relación colineal. Es importante señalar, que si bien la multicolinealidad causa problemas y reduce la precisión de los estimadores, la multicolinealidad no es en si misma un fenómeno sujeto a prueba estadística.

En las siguientes páginas revisaremos el método que cuenta con mayor aceptación y que se basa en el análisis de valores propios de la matriz **X'X**, el cual estará basado en su totalidad en el trabajo de Kuh, Welsch y Belsley (1980).

3.4 Análisis de los valores propios de la matriz X'X

El análisis de los valores propios y vectores propios de la matriz **X'X** o la matriz de correlación **R** ha sido empleado por muchos años en el tratamiento de la multicolinealidad , con buenos resultados. Smith (1974) y Silvey (1969) sugirieron el uso de los valores propios de **X'X** como una clave de la presencia de multicolinealidad , como ellos lo sugirieron, la multicolinealidad está indicada por la presencia de un valor propio "pequeño". Surgen las primeras dificultades si nos preguntamos sobre el significado de la palabra "pequeño" en este punto, algunos correrían el riesgo de asociar "pequeño" con un valor cercano al cero; otros, como Thisted (1978a) tienden a asociar la multicolinealidad con la existencia de un valor propio pequeño en relación a los demás. De aquí en adelante, resaltaremos la importancia de las características numéricas y de estabilidad de la matriz **X'X** en nuestro análisis.

Dado que ninguno de los métodos anteriormente descritos provee un método completamente exitoso de detección, se vuelve necesario voltear hacia otros campos, en los cuales podamos encontrar la base de un diagnóstico exitoso. Belsley, Kuh y Welsch (1980), retoman en su trabajo resultados provenientes del Análisis Numérico, algunos de los cuales no involucran en absoluto problemas como el que afrontamos, pero que se relacionan en gran medida con nuestro problema.

El esfuerzo en la detección correcta de la multicolinealidad debe dirigirse primordialmente al análisis y a la manipulación de la matriz **X'X**, así que revisaremos algunos resultados útiles. Los analistas numéricos se preocupan, entre otras cosas, de las propiedades (condicionamiento) de una matriz **A** de un sistema de ecuaciones lineales **AZ = C**, que permitan obtener una solución **Z** estable.

En nuestro caso, las ecuaciones normales son $(X'X)\hat{\beta} = X'y$ con matriz de varianza-covarianza $\sigma^2(X'X)^{-1}$. La multicolinealidad, ocasiona entonces que la matriz $A=(X'X)^{-1}$ sea mal condicionada y que la solución $\hat{\beta}$ y su matriz de varianza-covarianza respectiva sean numéricamente inestables. Es saludable entonces, que dada su estrecha relación, las técnicas del análisis numérico hayan sido utilizadas por los estadísticos y los econometristas en un grado creciente.

Tenemos que subrayar las diferencias entre el enfoque que el analista numérico recomendaría en una situación particular (cuando se tenga que eliminar una o varias columnas de una matriz de datos) y el que el econometrista sugeriría en tal situación. No obstante, como Belsey, Kuh y Welsch (1980) lo señalan, las técnicas de los analistas numéricos tienen mucho que ofrecernos a los usuarios de mínimos cuadrados para diagnosticar multicolinealidad.

Durante su exposición. Belsley, Kuh y Welsch (1980) retoman y amplían las sugerencias dadas por Silvey (1967), una de las cuales ya se abordó al principio de esta sección. El uso de la noción del número de condición - propiedad del análisis numérico - exhibe en toda su extensión la fuerza de la sugerencia hecha por Silvey (1967), si bien él no se percató de ello, al examinar la descomposicón de la varianza estimada de cada coeficiente de regresión en una forma que exhibe la degradación de cada coeficiente causada por las relaciones colineales.

Estos autores explotan la sugerencia de Silvey (1967), utilizando técnicas del análisis numérico para proveer un conjunto de índices (índices de condición) que señalan la presencia de una o más dependencias cercanas entre las columnas de **X** y adaptan la descomposición de varianza de Silvey (1967) en una forma que puede combinarse con los índices de condición para descubrir aquellas variables que están involucradas en alguna relación cercana a la lineal y probar el grado en el que los coeficientes estimados están siendo degradados.

3.4.1 Descomposición en valores singulares

A fin de resumir de una mejor forma el método propuesto por Belsley, et. al. (1980), tenemos que revisar algunos conceptos y definir, en su caso, otros que sean de relevancia en la exposición.

Cualquier matriz **X**(nxk), que los autores arriba citados consideran una matriz de n-observaciones en k-variables, puede descomponerse (ver el apéndice 1) como:

$$\mathbf{X = UDV'}$$

donde **U** es una matriz (nxk), **V** es una matriz (kxk) y además **U** y **V** satisfacen que

$\mathbf{U'U = V'V = I_k}$ y **D** es diagonal con elementos en la diagonal no negativos μ_i, i=1,2,...,k, estos valores son conocidos como los valores singulares de **X**. Esto es independiente transformaciones lineales sobre la matriz. Belsley, et. al. (1980) sugieren que siempre que esto sea posible la matriz sea centrada y cambiada de escala.

Estos autores señalan que la descomposición en valor singular está relacionada con los conceptos de valores propios y vectores propios, si bien tiene importantes diferencias ya que

$$\mathbf{X'X = (UDV')'(UDV') = (VD'U')(UDV') = VD^2V',}$$

así que **V** es una matriz ortogonal que diagonaliza **X'X** y por ende, los elementos de la diagonal de $\mathbf{D^2}$, los cuadrados de los valores singulares, deben ser los valores propios de la matriz simétrica **X'X**. Además, las columnas ortogonales de **V** deben ser los vectores propios de **X'X**.

La importancia de la descomposición en valores singulares radica en que provee información que acompaña a aquella dada por los

valores y vectores propios de **X'X**. Los autores arriba citados señalan tres razones para preferir el uso de la descomposición en valores singulares:

i) Se aplica directamente a la matriz **X**, que es lo que nos interesa

ii) El número de condición de una matriz se define propiamente en términos de los valores singulares de **X** (norma del espectro)

iii) El concepto de descomposición en valores singulares tiene un uso práctico y analítico en el álgebra de matrices

iv) Existen algoritmos que permiten calcular con mayor facilidad la descomposición en valores singulares de **X** que los que calculan el sistema de valores y vectores propios de **X'X**

Por otro lado, examinan la situación en la que existen dependencias lineales exactas, en este caso, rango **X** = r < k. Como en la DVS de **X**, **U** y **V** son ortogonales (por tanto de rango completo), se debe tener rango **X** = rango **D**. De aquí:

$$\mathbf{X} = \mathbf{UDV'} = \mathbf{U}\begin{bmatrix} \mathbf{D}_{11} & 0 \\ 0 & 0 \end{bmatrix}\mathbf{V'} \qquad (3.4.2.1)$$

donde $\mathbf{D}_{11}$ es rxr y no sigular. Multiplicando por **V** y particionando,

$$\mathbf{XV} = \mathbf{X}[\mathbf{V}_1\ \mathbf{V}_2] = [\mathbf{U}_1\ \mathbf{U}_2]\begin{bmatrix} \mathbf{D}_{11} & 0 \\ 0 & 0 \end{bmatrix} \qquad (3.4.2.2)$$

donde $\mathbf{V_1}$ es kxr, $\mathbf{U_1}$ es nxr, $\mathbf{V_2}$ es kx(k-r) , y $\mathbf{U_2}$ es nx(k-r). De **(3.4.2.2)**, tenemos

$$\mathbf{XV_1 = U_1D_{11}}$$

$$\mathbf{XV_2 = 0}$$

En estas ecuaciones se exhiben las dependencias lineales de **X** y la matriz $\mathbf{V_2}$ proporciona una base ortonormal para el espacio nulo asociado con las columnas de **X**.

Puesto que en la mayoría de las aplicaciones estadísticas, las interrelaciones entre las columnas de **X** no son dependencias exactas, y que las computadoras trabajan con una aritmética finita, no exacta, debemos averiguar en que medida la existencia de un valor singular pequeño μ sugerirá la existencia de relaciones cercanas a la lineal entre las columnas de la matriz **X**. Para coadyuvar en la solución de esta cuestión, desarrollaremos la noción de número de condición de una matriz **X**.

Es preciso definir entonces cuando está mal condicionada una matriz; afortunadamente la descomposición en valor singular nos ayuda de nuevo. Es razonable suponer que una matriz **A** está mal condicionada si el producto de su norma espectral (vea el apéndice 1) con el de $\mathbf{A^{-1}}$ es grande. Esta medida, se conoce como el número de condición de **A**, y provee información resumida del condicionamiento

de **A**; entre más grande es el número de condición, mayor es el mal condicionamiento de la matriz **A**.

El condicionamiento de cualquier matriz cuadrada **A** puede resumirse, entonces, por un número de condición K(**A**) definido como el producto del máximo valor singular de **A** (su norma espectral) y el máximo valor singular de $\mathbf{A}^{-1}$, ó, $K(\mathbf{A}) = ||\mathbf{A}|| / ||\mathbf{A}^{-1}||$, este valor provee una medida de la sensitividad potencial de la solución de un sistema de ecuaciones lineales a cambios en los elementos de **C** y **A** del sistema lineal. De manera general, definimos el número de condición para una matriz **X** (nxk):

$$K(\mathbf{X}) = \frac{\mu_{max}}{\mu_{min}} \geq 1$$

En este momento estamos en condiciones de responder la pregunta sobre el tamaño del valor singular. Anteriormente se dijo que la presencia de un valor propio pequeño sugería la existencia de alguna(s) dependencia(s) entre las columnas de la matriz **X**. Puesto que el grado de mal condicionamiento depende de qué tan pequeño sea el mínimo de los valores singulares en relación con los restantes valores singulares, μ_{max} puede ser de utilidad en la medición de la "pequeñez" de tales valores singulares. Belsley, Kuh y Welsch definen

$$\eta_i = \mu_{max} / \mu_i \quad i=1,2,...,k$$

como el i-ésimo índice de condición de la matriz **X** de datos. Es fácil ver, que $\eta_k \geq 1$ para toda k, es también evidente que el número de condición de la matriz **X** será igual al máximo de los índices de condición.

Estos autores extienden la sugerencia de Kendall-Silvey (1969) como sigue: hay tantas dependencias cercanas entre las columnas de la matriz de datos **X** como índices de condición altos haya (valores singulares pequeños relativos a μ_{max}) además aclaran que el mero hecho de tener valores singulares pequeños no implica el cumplimiento del hecho de tener problemas con el condicionamiento de la matriz, pero que un valor singular sea pequeño en relación a μ_{max} si está relacionado con el problema del mal condicionamiento de **X'X**. Estos autores proponen- enfatizando el cuidado que debe tenerse al tratar este problema -que las dependencias débiles están asociadas con índices de condición entre 5 y 10, mientras que las relaciones fuertes o moderadas están asociadas con números de condición entre 30 y 100.

Como ellos subrayan , el uso de los índices de condición extiende la sugerencia de Kendall-Silvey (1967) en dos formas. La primera, que la experiencia práctica permite responder a la pregunta de cuándo "pequeño " es pequeño (ó cuándo grande es grande), y en segundo lugar la ocurrencia simultánea de varias η grandes indica la presencia simultánea de más de una dependencia cercana.

3.4.2 La descomposición de varianza de los coeficientes de regresión

En la sección anterior, interpretábamos la aparición de un valor singular pequeño en relación a μ_{max} como una indicación de dependencia cercana entre las columnas de **X**. Belsley recoge la sugerencia hecha por Silvey (1967) para mostrar como puede descomponerse la varianza estimada de cada uno de los coeficientes de regresión en una suma de términos, cada uno de los cuales está asociado con un valor singular. De aquí provee un medio para determinar la extensión en la cual las dependencias cercanas, (aquellas con altos índices de condición) degradan (llegan a ser parte dominante de) cada varianza.

Sabemos que la matriz de varianza-covarianza del estimador de mínimos cuadrados $\hat{\beta} = (X'X)^{-1}X'y$ es $\sigma^2(X'X)^{-1}$, donde σ^2 es la varianza de ε en el modelo $y = X\beta + \varepsilon$. Hagamos uso de la descomposición en valor singular, para obtener:

$$X = UDV'$$

$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1} = \sigma^2((UDV')'(UDV)) = \sigma^2VD^{-2}V' \quad (3.4.3.1)$$

y el i-ésimo componente de β

$$Var(\hat{\beta}_k) = \sigma^2 \Sigma v_{kj}^2 / \mu_j^2 \quad (3.4.3.2)$$

donde los μ_j son los valores singulares y $V=[\ V_{ij}\ ]$.

Es importante notar que **(3.4.3.2)** descompone $\mathbf{Var}(\hat{\beta}_k)$ en una suma de componentes, cada una de las cuales está asociada con uno y sólo uno de los k valores singulares μ_j (ó valores propios μ_j^2). Para Belsley, et. al. (1980) esto sugiere que una proporción inusualmente alta de la varianza de dos ó más coeficientes concentrada en los componentes asociados con el mismo valor singular pequeño proporciona evidencia de que la dependencia cercana correspondiente está causando problemas.

Para abundar más al respecto, Belsley, et. al. (1980), sugieren la definición siguiente:

DEFINICIÓN: Se define la ij-ésima proporción de descomposición de varianza como la proporción de la varianza del i-ésimo coeficiente de regresión asociado con la j-ésima componente de su descomposición en **(3.4.3.1)**. Sea

$$\varphi_{ij} = v_{ij}^2 / \mu_{ij} \quad y \quad \varphi_i = \sum_{j=1}^{k} \varphi_{ij} \quad i=1,2,...,k \qquad \textbf{(3.4.3.3)}$$

Entonces las proporciones de descomposición de varianza son

$$\pi_{ji} = \varphi_{ij} / \varphi_k \qquad i,j = 1,2,3,4 \qquad \textbf{(3.4.3.4)}$$

Resultará útil pues, revisar la tabla 3.1 en la cual se muestran las proporciones de descomposición de varianza.

Tabla 3.1 Proporciones de descomposción de varianza

Valor singular asociado	Proporciones de descomposición			
	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$		$Var(\hat{\beta}_k)$
μ_1	π_{11}	π_{12}	...	π_{1k}
μ_2	π_{21}	π_{22}	...	π_{2k}
...	...	...	...	...
μ_k	π_{k1}	π_{k2}	...	π_{kk}

Es importante detenernos en este punto, a fin de hacer algunas aclaraciones. En la descomposición de varianza dada en (**3.4.3.2**), valores pequeños de μ_j, manteniendo constantes las demás, conducen a valores grandes de las componentes de $\mathbf{Var}(\hat{\beta}_i)$. Pero no todas las $\mathbf{Var}(\hat{\beta}_i)$ pueden ser afectadas por una μ_j pequeña, ya que el numerador V_{ij}^2 puede ser aún más pequeño. El caso extremo, en el que $V_{ij}=0$ [cuando las columnas i y j de **X** son ortogonales (Belsley y Klema [1974])], no afectaría en lo absoluto. Esto es interesante, ya que nos dice que, en la DVS de $\mathbf{X}=[\mathbf{X}_1\ \mathbf{X}_2]$ con $\mathbf{X}_1'\mathbf{X}_2=0$ es posible siempre encontrar una matriz **V** de la forma

$$\mathbf{V}=\begin{bmatrix} V_{11} & 0 \\ 0 & V_{22} \end{bmatrix}$$

Lo que podría permitir mitigar los efectos de la multicolinealidad, que resultarán en valores de μ relativamente pequeños.

Debe pasar, además, que al menos dos de las variables estén involucradas, esto es, debe requerirse que haya 2 ó más variables para crear una dependencia cercana, por lo que debe pasar que dos ó más varianzas sean afectadas adversamente por altas proporciones de descomposición de varianza asociadas con un valor singular pequeño.

Es interesante analizar el caso en el que se tiene una matriz de datos **X** con columnas mutuamente ortogonales. De los resultados obtenidos anteriormente, se desprende que la matriz **V** de la descomposición en valores singulares de **X**, es diagonal, ya que $V_{ij}=0$, para $i \neq j$. La matriz Π de proporciones de descomposición de varianza tiene la forma

Valor singular asociado	Proporciones de descomposición			
	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	...	$Var(\hat{\beta}_k)$
μ_1	1	0	...	0
μ_2	0	1	...	0
...	...	...	...	...
μ_k	0	0	...	1

Hemos indicado anteriormente que una proporción alta de cualquier varianza asociada con un valor singular es un buen indicador de multicolinealidad. Así pues, necesitaríamos un ejemplo en el cual un valor singular μ_i estuviera asociado con una gran proporción de la varianza de dos o más coeficientes para observar la degradación del estimador de regresión debida a la multicolinealidad.

Belsley, et. al. (1980), sugieren el ejemplo siguiente. Se tiene una matriz **X** con k=5, en la cual las columnas 4 y 5 de **X** están altamente correlacionadas, y todas las columnas restantes son mutuamente ortogonales, por estas razones se podría esperar una matriz Π de descomposición de varianza de la siguiente forma:

Valor singular asociado	Proporciones de varianza				
	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	$Var(\hat{\beta}_3)$	$Var(\hat{\beta}_4)$	$Var(\hat{\beta}_5)$
μ_1	1.0	0	0	0	0
μ_2	0	1.0	0	0	0
μ_3	0	0	1.0	0	0
μ_4	0	0	0	1.0	0.9
μ_5	0	0	0	0	1.0

En este caso, μ_4 está involucrada en las proporciones de $Var(\hat{\beta}_4)$ y $Var(\hat{\beta}_5)$.

Dado que al presentársenos una matriz de datos **X** no tendremos medios racionales -según estos autores - para seleccionar las variables involucradas en un problema de multicolinealidad, podemos evitar el problema haciendo uso de la matriz Π, esta matriz exhibe todas las posibles dependencias. Los auores arriba citados sugieren un procedimiento de diagnóstico, que resumimos en el cumplimiento de las dos condiciones siguientes:

i) Existencia de un valor singular con un alto índice de condición asociado.

ii) Altas proporciones de descomposición de varianza para dos ó más varianzas de coeficientes de regresión estimados

Retomando el procedimiento de diagnóstico propuesto por Belsley, et. al. (1980) , examinemos qué sucede con las dependencias cercanas. Aquí suponemos que las dependencias ya han sido identificadas por medio de las proporciones de descomposición de varianza. La dependencia puede tratarse de dos formas, o bien corriendo una regresión auxiliar ó regresando a la expresión matricial:

$$XV_1 = U_1D_{11}$$

$$XV_2 = 0$$

Si particionamos las matrices **X** y V_2 de la forma siguiente:

$$[\,X_1\ X_2\,]\begin{bmatrix} V_{21} \\ V_{22} \end{bmatrix} = X_1V_{21} + X_2V_{22} = 0$$

V_{21} es cuadrada y no singular. Las dependencias entre las columnas se exhiben como

$$X_1 = -X_2V_{22}V_{21}^{-1}$$

hacemos $G = -V_{22}V_{21}^{-1}$

entonces $X_1 = X_2G$

Así, los elementos de **G**, que se calculan directamente de **V**, proveen estimadores alternativos de la relación lineal entre aquellas variables de $\mathbf{X}_2$ y aquellas en $\mathbf{X}_3$. Estos autores advierten el hecho de utilizar mínimos cuadrados para un subconjunto de las columnas de **X**, señalando que esto no sustituye el procedimiento de diagnóstico sugerido. Los mínimos cuadrados se aconsejan una vez que han sido detectadas las dependencias, a la vez que el número de ellas. Dado que el procedimiento de diagnóstico descubre el número de dependencias cercanas sin requerir mayor información sobre las columnas de **X** es más aconsejable realizar primeramente el diagnóstico, detectar las dependencias y utilizar posteriormente el método de mínimos cuadrados.

Belsley, et. al. (1980), abordan el problema de suministrar evidencias de que la multicolinealidad ha dañado la estimación. Según ellos, uno debe mostrar que un intervalo de predicción que es demasiado grande para un propósito dado podría hacerse más angosto de una forma conveniente si lo hacemos estadísticamente condicional sobre datos mejor condicionados.

Estos autores nos dicen también, que si el investigador tuviera información de que: 1) hay dependencias cercanas a la lineal entre los datos, de tal forma que la colinealidad es potencialmente un problema y 2) que las varianzas de los parámetros (o intervalos de confianza) que son de interés para él tienen grandes proporciones de sus magnitudes

asociadas con la presencia de la relación colineal, entonces diría que los coeficientes de regresión afectados han sido degradados (pero no necesariamente dañados) por la presencia de la multicolinealidad, degradado en el sentido de que la magnitud de la varianza estimada está siendo determinada originalmente por la presencia de una relación cercana a la lineal.

Más interesante resulta el análisis hecho por estos autores sobre los efectos de transformaciones lineales en los datos en los diagnósticos de multicolinealidad. Belsley, et. al. (1980), comienzan con el modelo

$$y = X\beta + \varepsilon$$

la versión equivalente, ya parametrizada es

$$y = (XG^{-1})G\beta + \varepsilon = Z\delta + \varepsilon$$

donde **G** es pxp, no singular. Sin embargo, los valores singulares de **Z** no son necesariamente los mismos que los de **X**, así podemos preguntarnos si los diagnósticos de multicolinealidad perderán validez en el nuevo contexto.

No existe respuesta única a la pregunta formulada, pero puede suponerse que la parametrización no reduce la validez ni la utilidad de los diagnósticos de multicolinealidad.

Belsley , et. al. concluyen este capítulo considerando un caso más general, consideran una matriz de datos **X**. El número de condición de **X'X** es $\mu^2_{X,max} / \mu^2_{X,min}$ y además

$$\mu^2_{X,min} = \min |\mathbf{C'X'XC}|, \quad \mathbf{C'C} = 1$$

$$\mu^2_{X,max} = \max |\mathbf{C'X'XC}|, \quad \mathbf{C'C} = 1$$

Sea $\mathbf{C}^*$ solución al problema de minimización, $\mathbf{C}^0$ solución al problema de maximización, $\mathbf{Z=XG^{-1}}$, $\mathbf{d^*=GC^*}$ y $\mathbf{d^0=GC^0}$. Normalizando $\mathbf{d}^*$, obtenemos:

$$\bar{\mathbf{d}}^* = \mathbf{d}^*/||\mathbf{d}^*|| \text{ y } \bar{\mathbf{d}}^0 = \mathbf{d}^0/||\mathbf{d}^0||.$$

Por lo que tenemos

$$|\bar{\mathbf{d}}^{*\prime}\mathbf{Z'Z}\bar{\mathbf{d}}^*| = \mu^2_{X,min}||\mathbf{d}^*||^{-2} \geq \mu^2_{Z,min}$$

y $$|\bar{\mathbf{d}}^{0\prime}\mathbf{Z'Z}\bar{\mathbf{d}}^0| = \mu^2_{X,max}||\mathbf{d}^0||^{-2} \geq \mu^2_{Z,max}$$

donde $\mu^2_{Z,l}$ son los valores propios de **Z'Z**

Por lo que

$$K(\mathbf{Z}) \equiv \frac{\mu_{Z,max}}{\mu_{Z,min}} \geq \frac{\mu_{X,max}||\mathbf{d}^*||}{\mu_{X,min}||\mathbf{d}^0||} = K(\mathbf{X})\frac{||\mathbf{d}^*||}{||\mathbf{d}^0||}$$

Así pues, si **G** está mal condicionada, puede afectar severamente el problema.

Cuando **G** es ortonormal, entonces $||\mathbf{d}^1|| = ||\mathbf{d}^k||$ y se cumple la igualdad en la desigualdad de arriba. Ahora, si K(**X**) es muy grande, K(**Z**) puede ser pequeño si $||\mathbf{d}^1||/||\mathbf{d}^0||$ es muy pequeño. Así pues, llegamos a la conclusión de que la reparametrización puede mejorar el condicionamiento pero en una forma que depende de los valores singulares de **X**.

Existe otro sentido en el cual el resultado de arriba muestra que la reparametrización tiene poco valor práctico para mejorar el condicionamiento. Por ejemplo, supongamos que $\mathbf{y}=\mathbf{X}\beta + \varepsilon$ es la parametrización deseada, y que **X** está mal condicionada. Suponga además que se determina que la reparametrización $\delta=\mathbf{G}\beta$ provee una matriz de datos $\mathbf{Z}=\mathbf{X}\mathbf{G}^{-1}$ que es bien condicionada . Hemos visto que el mal condicionamiento de la matriz **X** implica el mal condicionamiento de **G** y por ende, el de $\mathbf{G}^{-1}$.

Además no debemos olvidar que las reparametrizaciones ortogonales no cambian el condicionamiento de la matriz de datos, así, los valores propios de **X'X** permanecen invariantes a tales transformaciones. Belsley et. al. (1980), ironizan al concluir esta parte, diciéndonos que las transformaciones lineales de los datos no causan problemas pero tampoco los resuelven.

3.4 Evidencia experimental

En esta sección, haremos explícitas las sugerencias hechas por Belsley, et. al. (1980), para diagnosticar multicolinealidad con un ejemplo específico. Es relevante apuntar, la sugerencia de facto que hacen ellos de adoptar la siguiente regla de decisión: los valores estimados están degradados (esto es, la magnitud de la varianza estimada está siendo determinada en su mayoría por la presencia de una relación cercana a la lineal), cuando dos o más varianzas tienen al menos la mitad de su magnitud asociada a un solo valor singular grande.

En su exposición, estos autores ejemplifican lo anterior haciendo uso de un experimento que reporta el comportamiento de los valores singulares y de las proporciones de descomposición de varianza de una serie de matrices de datos que se convierten sistemáticamente en matrices mal condicionadas. Los experimentos se desarrollan de la forma siguiente:

Se tiene una matriz **X** de datos (nxk_1). El número n es igual a 25. Estos datos son utilizados para construir series adicionales de datos colineales, los cuales exhiben dependencias cada vez más fuertes.

Sea **c** un vector (k_1x1) de constantes, se toma

$$\mathbf{w}_i = \mathbf{Xc} + e_i, \text{ con } e_i \sim N(0,\sigma^2(10^{-i}S^2_{Xc})),\ i=0,1,2,3,4$$

$$\mathbf{S^2_{Xc}} \equiv \mathbf{Var(\ Xc\ )} \qquad (3.5.1)$$

Por construcción cada w_i es una combinación lineal **Xc**, de la serie de datos más un término de error e_i con media cero, cuya varianza es cada vez más pequeña.

Notemos que, por ejemplo, cuando i=0, se tiene

$$e_i \sim N(0, \sigma^2 S_{Xc})$$

esto es la varianza coincide con aquella debida a la parte **Xc**.

En cambio, cuado i=4, solamente 1/104 de la varianza de w_i se debe a e_i, por lo que la dependencia entre w_i y **Xc** es fuerte. El conjunto de matrices de datos para el experimento, se construye aumentando la matriz básica **X** con cada w_i, formando

$$X\{i\} = [\, X w_i \,], \qquad i=0,1,2,3,4 \qquad (3.5.2)$$

Las series de datos elegidas para las matrices **X{i}**, i=1,2,3,4 están constituidas por datos arrojados por series de tiempo económicas, ó variables construidas de tal manera que tengan medias y varianzas similares a las de las series de tiempo económicas. Dado que los índices de condición no suministran información estable al usuario de la regresión lineal sobre el grado de multicolinealidad entre las variables **X**, es necesario, estandarizar a las matrices de datos correspondientes y proveerlas de estructuras equivalentes, de tal forma que los componentes de los índices tengan sentido.

Los datos elegidos para las matrices **X{i}**, y=1,2,3,4 fueron tomados de series de tiempo de variables económicas ó construidos aleatoriamente de tal forma que tuvieran medias y varianzas similares a los que presentaban los datos de las series de tiempo.

Es importante señalar, que las proporciones de descomposición de varianza pueden mostrarnos qué variables están involucradas en las relaciones, y podemos entonces, ajustar egresiones entre estas variables para exhibir las dependencias. Las estadísticas t que resultan de estas regresiones pueden utilizarse en la forma habitual para suministrar evidencia adicional de la "significancia" de cada variable en la dependencia lineal específica.

El conjunto básico de datos empleado (Belsley, et. al., (1980)) es:

$$X = [\, ME, IM, MV \,]$$

donde ME es el total de embarques de manufacturas

IM es el total de inventarios de manufacturas

MV es el total de órdenes no cubiertos de manufacturas

los datos están en millones de unidades monetarias anuales (1947-1970) (n=24). Se generan a partir de **X** dos conjuntos de dependencias adicionales, como sigue:

$$w_i = MV + V_i, \quad i=0,1,...,4 \qquad (3.5.3)$$

con $\mathbf{V_i}$ generada normal con media cero y $Var(\mathbf{V_i})=\sigma^2_i\mathbf{I}=10^{-1}S^2_{MV}\mathbf{I}$ ó denotemos $V_i \sim N(0, 10^{-1}S^2_{MV}\mathbf{I})$, con S^2_{MV} la varianza muestral de las series MV, además

$$\mathbf{Z_j} = 0.8\mathbf{ME} + 0.2\mathbf{IM} + \mathbf{V_j} \qquad (3.5.3a)$$

$$V_j \sim N(0, 10^{-1}S_{ZZ}\mathbf{I})$$

S^2_Z es la varianza muestral de $0.8\mathbf{ME} + 0.2\mathbf{IM}$

Se utilizan las series $\mathbf{w_i}$ y $\mathbf{Z_j}$ para aumentar el conjunto básico de datos y producir tres secuencias de matrices:

$$\mathbf{X1}\{\mathbf{I}\} \equiv [\mathbf{Xw_i}], \quad i=0,1,...,4$$

$$\mathbf{X2}\{\mathbf{J}\} \equiv [\mathbf{XZ_j}], \quad j=0,1,...,4 \qquad (3.5.3b)$$

$$\mathbf{X3}\{\mathbf{I,J}\} \equiv [\mathbf{Xw_iZ_j}] \quad i,j=0,1,...,4$$

Exhibimos los resultados obtenidos para las secuencias, comenzando con la secuencia **X1**. Sea $\mathbf{X1}\{\mathbf{I}\}$, $i=0,1,...,4$. Nuestros datos se exhiben en las columnas de la tabla 3.2, los datos de la serie en la columna 4, denotada por **C4**, se relaciona con aquel de la columna 3, **C3**.

$$\mathbf{C4} = \mathbf{C3} + e_i, \qquad i=0,1,...,4$$

esta es la única dependencia entre las cuatro columnas de **X**.

Los datos de la tabla 3.2 confirman nuestras sospechas, a medida que i aumenta se tiene un índice de condición más alto que cuenta para

una alta proporción de varianza para dos ó más coeficientes, siendo éstos Var($\hat{\beta}_3$) y Var($\hat{\beta}_4$). Todos los índices de condición excepto uno permanecen virtualmente sin cambios mientras que el índice de condición correspondiente a la dependencia introducida se incrementa rápidamente con cada salto i.

Tabla 3.2 Proporciones de descomposición de varianza e índices de condición. Series X1.
Se construyó una dependencia cercana: C4 = C3 + el

Valor singular asociado	Proporciones de varianza				Índice de condición, η
	Var($\hat{\beta}_1$)	Var($\hat{\beta}_2$)	Var($\hat{\beta}_3$)	Var($\hat{\beta}_4$)	
		X1{ 0 }			
μ_1	0.005	0.012	0.002	0.003	1
μ_2	0.044	0.799	0.004	0.032	5
μ_3	0.906	0.002	0.041	0.238	8
μ_4	0.045	0.187	0.954	0.727	14
		X1{ 1 }			
μ_1	0.005	0.011	0.001	0.001	1
μ_2	0.094	0.834	0.003	0.002	5
μ_3	0.899	0.117	0.048	0.035	9
μ_4	0.002	0.038	0.948	0.962	27
		X1{ 2 }			
μ_1	0.005	0.012	0.000	0.000	1
μ_2	0.086	0.889	0.000	0.000	5
μ_3	0.901	0.083	0.003	0.003	9
μ_4	0.007	0.016	0.997	0.997	95
		X1{ 3 }			
μ_1	0.005	0.012	0.000	0.000	1
μ_2	0.078	0.903	0.000	0.000	5
μ_3	0.855	0.079	0.000	0.000	9
μ_4	0.061	0.006	0.999	0.999	481
		X1{ 4 }			
μ_1	0.005	0.010	0.000	0.000	1
μ_2	0.084	0.792	0.000	0.000	5
μ_3	0.906	0.070	0.000	0.000	9
μ_4	0.004	0.127	1.000	1.000	976

Para una interpretación más clara de esta situación, en la tabla 3.3 se exhiben los coeficientes de correlación entre **C3** y **C4** para cada una de las matrices **X1(i)** y también las regresiones múltiples de **C4** en **C1**, **C2** y **C3**.

Tabla 3.3 Regresión de C4 en C1, C2, C3*

Matriz de datos	r(C3, C4)	C1	C2	C3	R^2
X1(0)	0.766	0.3905	-0.1354	0.9380	0.8229
		[1.11]	[0.91]	[4.76]	
X1(1)	0.931	0.1461	0.0925	0.8852	0.8765
		[0.97]	[1.36]	[10.10]	
X1(2)	0.995	-.0076	0.0142	0.9982	0.9893
		[-0.17]	[0.72]	[38.80]	
X1(3)	0.999	0.0011	0.0015	0.9901	0.9996
		[1.22]	[0.37]	[188.96]	
X1(4)	1.000	-0.0012	0.0033	0.9978	0.9999
		[-0.28]	[1.78]	[400.56]	

Es importante notar, tomando como base los datos de la tabla 3.3 que con cada incremento en i (correspondiente a una reducción de la varianza debida a e_i), las correlaciones simples y los R^2 se incrementan un paso, digamos, añadiendo otro 9, formando la sucesión 0.9, 0.99, 0.999 y así y los índices de condición siguiendo la progresión 10, 30, 100, 300, 1000. Además, con cada incremento en i, la

* Los números entre paréntesis son los valores de las t's

proporción de descomposición de varianza de los coeficientes afectados asociada con el mayor de los η se incrementa notablemente.

Así pues, en virtud del cumplimiento conjunto de 1) altas proporciones de descomposición de varianza para dos o más coeficientes y 2) un alto índice de condición, existe evidencia que señala la presencia de multicolinealidad.

Exhibimos a continuación los resultados abtenidos para la serie **X2{ I }**, en esta serie se construye la dependencia que involucra tres variables, las columnas 1, 2 y 4 con la forma:

$$\mathbf{C4} = 0.8\mathbf{C1} + 0.2\mathbf{C2} + e_i \qquad i=0,1,2,3,4$$

Esperaríamos, entonces, altas proporciones de descomposición de varianza para estas tres variables, asociadas con un elevado índice de condición. En la tabla 3.4, mostramos la matriz Π de proporciones de descomposición de varianza para la serie **X2{ I }** y en la tabla 3.5, las correlaciones simples y las regresiones correspondientes. La regresión es de **C4** en **C1**, **C2** y **C3**.

Como en el caso de las **X1{ I }** haremos notar algunos puntos: Primero, los resultados concuerdan con las prediciones hechas al principio; segundo, en el caso de **X2{ 0 }** la dependencia es débil,

siendo la correlación igual a 0.477, además los índices de condición μ_3=9 y μ_4=11 son similares, esta situación dificulta el diagnóstico y provoca que las proporciones de descomposición de varianza estén arbitrariamente distribuídas entre los índices de condicion similar.

Tabla 3.4 Proporciones de descomposición de varianza e índices de condición. Series X2
Dependencia construida: C4 = 0.8C1 + 0.2C2 +ei

Valor singular asociado	Proporciones de varianza				Índice de condición,
	$Var(\hat{\beta}_1)$	$Var(\hat{\beta}_2)$	$Var(\hat{\beta}_3)$	$Var(\hat{\beta}_4)$	η
		X2(0)			
μ_1	0.003	0.012	0.004	0.005	1
μ_2	0.027	0.735	0.001	0.068	4
μ_3	0.009	0.223	0.836	0.526	9
μ_4	0.960	0.030	0.359	0.401	11
		X2(1)			
μ_1	0.001	0.004	0.003	0.000	1
μ_2	0.021	0.297	0.011	0.001	5
μ_3	0.091	0.026	0.767	0.006	10
μ_4	0.887	0.673	0.219	0.993	31
		X2(2)			
μ_1	0.000	0.001	0.004	0.000	1
μ_2	0.000	0.039	0.012	0.000	5
μ_3	0.004	0.002	0.983	0.002	9
μ_4	0.995	0.958	0.001	0.998	102
		X2(3)			
μ_1	0.000	0.000	0.004	0.000	1
μ_2	0.000	0.002	0.013	0.000	5
μ_3	0.000	0.000	0.938	0.000	9
μ_4	1.000	0.997	0.006	1.000	381
		X2(4)			
μ_1	0.000	0.000	0.004	0.000	1
μ_2	0.000	0.000	0.013	0.000	5
μ_3	0.000	0.000	0.938	0.000	9
μ_4	1.000	1.000	0.046	1.000	1003

Este problema desaparece cuando tomamos **X2{ 1 }**, la correlación simple es ahora igual a 0.934 y el mayor índice de condición es ahora μ_4=31. Recordando la tendencia que seguían los índices de condiciones en las series **X1{ i }**, vemos que aquí se conserva, siendo los saltos en la progresión 10, 30,100, 300, 1000. La multicolinealidad puede intuirse a partir de **X2{ 1 }**, aquí se obtienen índices de condición cercanos a 30, por lo que puede considerarse que los estimados de los coeficientes de la regresión han sido degradados.

Tabla 3.5 Regresión de C4 en C1, C2 y C3

Matriz de datos	$\hat{r}(C4,\hat{C4})$ C4=0.8C1+0.2C2	C1	C2	C3	R^2
X2(0)	0.477	0.8268	-0.0068	0.1089	0.2864
		[3.84]	[-0.07]	[0.88]	
X2(1)	0.934	0.8336	0.1776	0.1032	0.8976
		[10.14]	[6.49]	[2.87]	
X2(2)	0.995	0.8186	0.1879	0.0007	0.9911
		[40.90]	[21.44]	[0.06]	
X2(3)	0.999	0.7944	0.2023	0.0012	0.9993
		[149.03]	[86.69]	[0.40]	
X2(4)	1.000	0.7990	0.1992	0.0012	0.9999
		[393.94]	[224.32]	[1.02]	

CAPÍTULO IV

SOLUCIONES AL PROBLEMA DE LA MULTICOLINEALIDAD

4.1 Revisión de los métodos propuestos para el tratamiento de la multicolinealidad.

Durante el presente trabajo hemos presentado las consecuencias de la multicolinealidad, así como los diferentes procedimientos para detectarla. En este capítulo, retomaremos, la discusión acerca de las fuentes de la multicolinealidad y veremos, también que puede asociarse una solución inmediata a cada una de ellas. Sin embargo, como también apuntan Judge, et. al. (1985), debemos tener cuidado al utilizar los procedimientos sugeridos, dado que la cura podría acarrear consecuencias indeseables. Podemos dividir los procedimientos utilizados para hacerle frente a la multicolinealidad en dos: el primero, aquél que contiene las soluciones tradicionales, entre las cuales se incluyen: la incorporación adicional de datos, la reespecificación del modelo, la introducción de restricciones lineales exactas; y el segundo, soluciones ad-hoc, esto es, aquellos procedimientos cuya información introducido al modelo es específico de la muestra a la mano y cuyos estimadores obtenidos - en su forma convencional - tienen riesgo no acotado y son sesgados. Entre los métodos correspondientes a este tipo de soluciones podemos mencionar, la regresión en componentes

principales, la regresión ridge, regresión ridge generalizada, entre otros. Comenzaremos nuestro estudio de los métodos arriba mencionados comentando las características de las soluciones tradicionales.

4.2 Soluciones tradicionales al problema de la multicolinealidad

Básicamente, tratan de incorporar información adicional a la muestra para eliminar la multicolinealidad entre las columnas de la matriz **X**. Revisaremos por ello, las sugerencias de Montgomery y Peck (1982), Belsley, et. al. (1980) y Silvey (1967) al respecto, poniendo énfasis en el trabajo hecho por Silvey (1967) que trata el problema de la estimación imprecisa.

Esta solución está asociada al método empleado en la recolección de los datos. Montgomery y Peck (1982) sugieren recolectar los datos en tal forma que se elimine la multicolinealidad en los datos, esto es obtener observaciones que se encuentren fuera de la región que limita los datos obtenidos en la muestra original; no obstante, este procedimiento tiene desventajas evidentes: no siempre es posible repetir el experimento, el costo económico involucrado y los datos deben encontrarse siempre dentro de la región de interés delimitada por el investigador.

Como Gunst (1983) advierte, la eliminación de una de las variables predictoras en este caso, puede conducir a un excesivo sesgo de los

coeficientes estimados si la multicolinealidad se debe a deficiencias muestrales. Este autor recomienda que se incorpore información adicional, el uso de estimación sesgada - la cual será tratada más adelante -, es también recomendable cuando la fuente de la multicolinealidad es un muestreo deficiente o pobre.

Es en este punto donde incorporaremos las sugerencias hechas por Silvey (1967) para el tratamiento de la multicolinealidad debida a una base muestral pobre.

Silvey (1967) considera el modelo lineal general

$$\mathbf{y = X\beta + \varepsilon} \qquad (4.2.1.1)$$

con las especificaciones ya conocidas. Este autor, reconoce que la multicolinealidad es extrema si existen dependencias lineales exactas entre las columnas de **X**, lo cual hace imposible la estimación de β.

Si el rango de **X** es menor que k, existe una relación de la forma

$$c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \ldots + c_k\mathbf{x}_k = 0 \qquad (4.2.1.2)$$

donde las c_i son constantes, no todas iguales a cero y $\mathbf{X}_i$ es la i-ésima columna de **X**.

Así, la matriz $\begin{bmatrix} y1 & \\ y2 & \\ . & X \\ . & \\ yn & \end{bmatrix}$

obedece otra relación, además de **(4.2.1.1)**

Ahora, según él, la presencia de relaciones como **(4.2.1.2)**, implica la imposibilidad de funciones paramétricas lineales reales como

$$\mathbf{W'\beta} = w_1\beta_1 + \ldots + w_k\beta_k$$

de poseer estimadores lineales insesgados.

Por ello, afirma que $\mathbf{W'\beta}$ posee un estimador lineal insesgado si y sólo si $\mathbf{W'}$ puede ser expresado como una combinación lineal de los renglones de $\mathbf{X}$. Para aclarar lo anterior nos apoyamos en el siguiente teorema:

TEOREMA 1. Dado el modelo lineal $\mathbf{y = X\beta + \varepsilon}$ y las hipótesis asociadas al mismo, donde $\mathbf{X}$ no tiene necesariamente rango completo por columnas, puede obtenerse un estimador insesgado de cualquier combinación lineal de los parámetros $\mathbf{W'\beta}$ de los elementos del vector de coeficientes β si el vector $\mathbf{W'}$ es una combinación lineal de los renglones de $\mathbf{X}$. Entonces, el estimador lineal insesgado de varianza mínima de $\mathbf{W'\beta}$ es $\mathbf{W'\beta^*}$, donde β^* es cualquier solución de las ecuaciones normales $\mathbf{X'X\beta^* = X'y}$, y el estimador $\mathbf{W'\beta^*}$ existe y es único.

Además propone una reparametrización adecuada para lograr la ortogonalidad de las columnas de $\mathbf{X}$, transformación en el siguiente teorema:

TEOREMA 2. La función paramétrica lineal $\mathbf{W'\beta}$ es estimable si y sólo si $\mathbf{W}$ es una combinación lineal de los vectores propios de $\mathbf{X'X}$ correspondientes a las raíces características distintas.

Sea pues, el modelo

$$\mathbf{y = X\beta + \varepsilon = X(I)\beta + \varepsilon = X(TT^{-1})\beta + \varepsilon}$$

donde $\mathbf{T}$ es una matriz cuadrada no singular elegida de tal forma que las columnas de $\mathbf{XT}$ sean ortogonales. Una elección de $\mathbf{T}$ puede ser $\mathbf{U}$, una matriz ortogonal cuyas columnas son los vectores característicos ortonormales de $\mathbf{X'X}$.

$$\mathbf{UU' = I = U'U}$$

$$\mathbf{y = X\beta + \varepsilon = X(UU')\beta + \varepsilon = (XU)(U'\beta) + \varepsilon}$$

$$\mathbf{= Z\theta + \varepsilon}$$

donde

$$\mathbf{Z = XU}$$

$$\mathbf{\theta = U'\beta}$$

además

$$\mathbf{Z'Z = (XU)'XU = U'X'XU = \Lambda}$$

$\mathbf{\Lambda}$ es una matriz diagonal cuyos elementos son las raíces características de $\mathbf{X'X}$ $\lambda_1,\ldots,\lambda_k$, si $\mathbf{X}$ tiene rango k-j$\Rightarrow$$\mathbf{X'X}$ tiene j raíces características que son cero y j columnas de $\mathbf{XU}$ son cero, digamos, las últimas j.

Se sigue que las últimas j componentes de t son eliminadas del modelo. Por lo tanto θ_{k-j+1}, θ_{k-j+2},..., θ_k no pueden ser estimados a partir de las observaciones **X**.

El análisis no sólo nos permite determinar qué funciones $\mathbf{W'\beta}$ son estimables, sino también cuales de éstas pueden ser estimadas de una forma relativamente precisa y cuáles de manera imprecisa.

Por otra parte $\theta_1,...,\theta_{k-j}$ ó cualquier combinación de estos puede ser estimada. Por lo tanto podemos estimar $\mathbf{W\beta} \Leftrightarrow \mathbf{W'\beta}$ se transforma en una combinación lineal de $\theta_1,..., \theta_{k-j}$ ya que

$$\mathbf{W'\beta = W'UU'\beta = (U'W)'\theta}$$

podemos estimar $\mathbf{W'\beta}$ si y sólo si las últimas j componentes de **U'W** son cero.

Sea **P** un vector propio normalizado de **X'X** correspondiente a una raíz característica distinta de cero λ, entonces $\mathbf{P'\beta}$ es estimable y por el teorema 1 su estimador insesgado de varianza mínima es $\mathbf{P'\beta^*}$, donde $\mathbf{\beta^*}$ es una solución a las ecuaciones normales. La varianza de $\mathbf{P'\beta^*}$ está dada por σ^2/λ, ya que

$$\mathbf{P'X'y = P'X'X\beta^* = \lambda P'\beta^*}$$

$\mathbf{P'X'X = \lambda P'}$, por lo que

$$\mathbf{\lambda^2 Var(P'\beta^*) = P'X'Var(y)XP = \sigma^2 P'X'XP = \lambda\sigma^2}$$

Similarmente, puede mostrarse que si $\mathbf{P}_1$ y $\mathbf{P}_2$ son vectores característicos de $\mathbf{X'X}$ correspondientes a raíces distintas de cero, entonces

$$Cov(\mathbf{P}_1'\beta^*, \mathbf{P}_2'\beta^*) = 0$$

Si la función $\mathbf{W'}\beta$ es estimable, implica que

$$\mathbf{W} = K_1\mathbf{P}_1 + \ldots + K_{k-j}\mathbf{P}_{k-j}$$

donde k_i son constantes y los $\mathbf{P}_i$ son los vectores propios correspondientes a las raíces distintas de cero $\lambda_1,\ldots,\lambda_{k-j}$ y la varianza del estimador $\mathbf{W'}\beta^*$ es

$$Var(\mathbf{W'}\beta^*) = \sigma^2(K_1^2/\lambda_1 + k_2^2/\lambda_2 + \ldots + K_{k-j}^2/\lambda_{k-j})$$

La expresión anterior resume lo que sabemos acerca de la precisión con la que podemos estimar una función. Silvey (1967) hace notar que la precisión depende de la varianza del error σ^2, de las magnitudes de las constantes k_i y de las magnitudes de las raíces características distintas de cero.

Una vez que el problema ha sido detectado, surge la pregunta natural de cómo solucionarlo. Silvey (1967) analiza detalladamente lo que sucede cuando se añade información adicional. Dado que un diseño muestral pobre es un problema básico, la solución más obvia es obtener datos muestrales adicionales. Este autor considera el

problema de cuáles valores de las variables independientes son óptimos en algún sentido, si vamos a tomar nuevas observaciones.

Supongamos que en el modelo $\mathbf{y= X\beta + \varepsilon}$, se tiene una dependencia exacta. Entonces existe un vector **P** tal que $\mathbf{X'XP=0}$, lo cual significa que $\mathbf{P'\beta}$ no es estimable, ni lo es cualquier función lineal $\mathbf{W'\beta}$ para la cual **W** tenga una componente distinta de cero en la dirección de **P**. Ya que $\mathbf{X'X= 0 \Leftrightarrow Xp = 0}$, el problema de la multicolinealidad existe si cada renglón de **X** es ortogonal a **P**; esto es, para destruir la multicolinealidad debemos elegir una nueva observación que no sea ortogonal a **P**. Una elección obvia es tomar una observación en la dirección de **P**, esto es $\mathbf{X_{T+1} = LP}$, con L un escalar.

Un análisis similar se cumple si **X'X** tiene raíces distintas de cero, pero algunas de ellas son pequeñas. Suponga que **X'X** tiene una raíz pequeña λ correspondiente al vector propio **P**, que define la dirección en la cual la estimación es imprecisa. Tomemos entonces una observación adicional y_{T+1} con valores $\mathbf{X_{T+1}=LP}$. El modelo es:

$$\begin{bmatrix} y \\ y_{T+1} \end{bmatrix} = \begin{bmatrix} X \\ X'_{T+1} \end{bmatrix} \beta + \begin{bmatrix} \varepsilon \\ \varepsilon_{T+1} \end{bmatrix}$$

o $\mathbf{y_* = X_*\beta + \varepsilon_*}$

Entonces

$$\mathbf{X'_*X_*} = \mathbf{X'X} + \mathbf{X_{T+1}X'_{T+1}} = \mathbf{X'X} + L^2\mathbf{PP'}$$

y

$$\mathbf{X'_*XP} = \mathbf{X'XP} + L2\mathbf{PP'P} = \lambda\mathbf{P} + L^2\mathbf{P} = (\lambda + L^2)\mathbf{P}$$

de tal forma que **P** es un vector propio de $\mathbf{X'_*X}$ correspondiente a la raíz $L^2 + \lambda$.

Eligiendo una nueva observación en la dirección de **P**, la precisión de la estimación puede ser mejorada en la que era imprecisa con respecto a la que se obtenía originalmente.

Silvey (1967) procede a investigar dos cuestiones más interesantes, primero ¿ cómo es afectada la precisión de la estimación tomando otra observación $\mathbf{X_{T+1}}$, no necesariamente en la dirección de los vectores característicos de **X'X** ? Para responder esta pregunta, sea $\mathbf{b_T}$ y $\mathbf{b_{T+1}}$ los estimadores de mínimos cuadrados de β basados en las **T** y **T+1** observaciones, respectivamente. Entonces

$$\mathrm{Var}(\mathbf{W'b_T}) - \mathrm{Var}(\mathbf{W'b_{T+1}}) = \frac{\sigma^2\alpha'\Lambda^{-1}\mathbf{ZZ'}\Lambda^{-1}\alpha}{1 + \mathbf{Z'}\Lambda^{-1}\mathbf{Z}}$$

donde

$$\alpha = \mathbf{U'W},\ \mathbf{z} = \mathbf{U'Z_{T+1}}\ \text{y}\ \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_k).$$

Esta expresión provee la mejora de precisión del estimado de cualquier combinación lineal $\mathbf{W'}\beta$ adicionando una nueva observación

X_{T+1}. Segunda, y quizás la más interesante, ¿ cómo debe ser elegida X_{T+1} para mejorar tanto como sea posible la estimación de una función lineal específica de $W'b$? Esta puede surgir cuando se está interesado en una cierta combinación lineal de parámetros o cuando se está interesado en hacer una predicción para un conjunto particular de los valores explicativos **W**. Este autor prueba el siguiente teorema:

TEOREMA 3. Dado el modelo lineal y la condición $X'_{T+1}X_{T+1}=d^2$, con d^2 un escalar positivo, satisfecha por un nuevo conjunto de observaciones X_{T+1} de las variables explicativas, la dirección óptima de X_{T+1} para mejorar la precisión de estimación de $W'\beta$ es la del vector **V**, donde $V=(I+d^{-2}X'X)^{-1}W$

En otras palabras, el teorema significa que la nueva observación debe ser proporcional a **V**. Se supone que el error asociado a la nueva observación no está correlacionado con los errores en el modelo original y tiene la misma varianza que cada uno de ellos. No se impone ninguna condición en el rango de **X**.

4.2.1 Reespecificación del modelo

En el capítulo II, vimos que una de las fuentes de la multicolinealidad era la elección del modelo. Montgomery y Peck (1982) sugieren que la reespecificación de la ecuación de regresión empleada puede disminuir el impacto de la multicolinealidad. Un posible camino es, redefinir las variables regresoras, dado que si un subconjunto de las

variables se encuentra en una dependencia cercana a la lineal, podrá ser posible encontrar una función de estas variables que conservará la información original y que redujera el impacto de la multicolinealidad.

Otro posible camino lo constituye la eliminación de variables. La eliminación de variables es generalmente una técnica muy eficaz, pero - siguiendo la advertencia hecha por Gunst (1983) - se debe tener cuidado de no eliminar una variable que dañe el poder predictivo del modelo, además, no existe ninguna garantía de que el modelo final esté exento del problema.

El tratamiento de la multicolinealidad inherente a la población es similar a aquél para la multicolinealidad surgida de la especificación del modelo. Así pues, si en nuestro diseño se han incluído variables que se esperan siempre con problemas de multicolinealidad, no existe razón alguna para retener alguna de ellas, lo que sugiere una seleción de variables.

4.2.2 Restricciones lineales exactas

En ocasiones, el investigador posee información exacta de un parámetro particular o de una combinación lineal de parámetros. Por ejemplo, al estimar una función log-lineal de producción, podemos tener información disponible de que la empresa está operando bajo la condición de rendimientos constantes de escala. Si existe información

disponible de este tipo, ésta puede establecerse en la forma del siguiente conjunto de ecuaciones lineales ó de desigualdades lineales

$$\mathbf{R}\beta = \mathbf{r}$$

donde **r** es un vector (jx1) de elementos conocidos y **R** es una matriz de diseño conocida de (jxk) que expresa la estructura de la información en los parámetros individuales β_i ó de alguna combinación lineal de los elementos del vector β.

De este modo, las restricciones impuestas describen alguna restricción física en las variables involucradas. Alternativamente, podemos observar relaciones como

$$\mathbf{Xp} \cong \mathbf{0}$$

y usar resultados de muestreo para imponer restricciones en el espacio de parámetros.

Una consecuencia de utilizar restricciones exactas de los parámetros es que reduce la variabilidad muestral de los estimadores, dando fin a uno de los daños ocasionados por la multicolinealidad. La imposición de restricciones sujetas, si bien es incorrecta, puede reducir el errror cuadrático medio del estimador, pero las restricciones incorrectas producen estimadores sesgados de los parámetros. Las restricciones lineales exactas pueden usarse en el caso de que exista multicolinealidad extrema ó severa, esto reducirá la dimensión del

espacio de parámetros en el primer caso, y volverá estimables a las funciones paramétricas $\mathbf{W}'\boldsymbol{\beta}$ en el segundo.

4.3 Soluciones ad-hoc al problema de la multicolinealidad

Hemos notado que los procedimientos convencionales que combinan información muestral y no muestral pueden utilizarse para reducir los efectos de la multicolinealidad. Esto presupone que se dispone de información no muestral, ya sea de la teoría fundamental, de trabajo estadístico previo, o de consideraciones personales. En esta sección consideramos algunos estimadores que pueden mitigar los efectos de la multicolinealidad, pero que son ad-hoc en el sentido de que la información que introducen es específica de la muestra a la mano.

4.3.1 Regresión en componentes principales

Es sabido que al enfrentar datos mal condicionados, los investigadores elijan la reducción de la información considerando subconjuntos del espacio paramétrico k-dimensional. Estos subespacios pueden ser sugeridos por resultados estadísticos previos o procedimientos subjetivos. Esta reducción, si bien tiene sus ventajas, acarrea consecuencias no deseables, una de las cuales es que a menos que el vector de parámetros reales se encuentre en

nuestro subespacio, los parámetros reales serán sesgados; esto es, nos enfrentamos a una decisión entre la reducción de la varianza muestral y el sesgo.

Consideremos la forma canónica del modelo

$$\mathbf{y} = \mathbf{Z}\alpha + \varepsilon \qquad (4.3.1.1)$$

donde

$$\mathbf{Z} = \mathbf{XT},\ \alpha = \mathbf{T}'\beta,\ \mathbf{T}'\mathbf{X}'\mathbf{XT} = \mathbf{Z}'\mathbf{Z} = \Lambda$$

y $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_k)$ es una matriz diagonal (kxk) de los valores propios de $\mathbf{X}'\mathbf{X}$ y $\mathbf{T}$ es una matriz ortogonal (kxk) cuyas columnas son los vectores propios asociados con $\lambda_1, \lambda_2, \ldots, \lambda_k$. Las columnas de $\mathbf{Z}$, que definen un nuevo conjunto de regresoras ortogonales

$$\mathbf{Z} = [\mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_k],$$

serán llamadas las componentes principales.

El estimador de mínimos cuadrados de α es

$$\hat{\alpha} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{y} = \Lambda^{-1}\mathbf{Z}'\mathbf{y} \qquad (4.3.1.2)$$

y la matriz de covarianza de $\hat{\alpha}$ es,

$$\text{Var}(\hat{\alpha}) = \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1} = \sigma^2\Lambda^{-1} \qquad (4.3.1.3)$$

Si $\mathbf{X}'\mathbf{X}$ tiene un valor propio pequeño, entonces se obtendrá una varianza grande del coeficiente de regresión ortogonal asociado

$$\mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{p}\sum_{j=1}^{p}\mathbf{Z}_i\mathbf{Z}_j' = \Lambda$$

esto es, el valor del valor propio λ_j será la varianza de la j-ésima componente principal.

Como ya habíamos visto en el capítulo III, un valor de λ_j cercano al cero implica la presencia de multicolinealidad entre las regresoras originales. Además

$$Var(\hat{\beta}) = Var(T\hat{\alpha}) = T\Lambda^{-1}T'\sigma^2$$

lo cual implica que

$$Var(\hat{\beta}_j) = \sigma^2 \left(\sum_{i=1}^{k} t_{ji}^2 / \lambda_i \right)$$

El enfoque de la regresión en componentes principales combate a la multicolinealidad al utilizar un conjunto menor que el conjunto original de componentes principales en el modelo. Arreglemos los valores propios en orden decreciente

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k > 0$$

Supongamos que los últimos 5 de estos valores propios son iguales a cero. En la regresión en componentes principales los componentes principales correspondientes a los valores propios cercanos a cero son removidas del análisis y se aplican mínimos

cuadrados a las componentes restantes. El estimador de componentes principales, será

$$\hat{\alpha}_{PC} = B\hat{\alpha} \qquad (4.3.1.4)$$

donde

$$b_1 = b_2 = \ldots = b_{k-s} = 1$$

y $$b_{K-s}+1 = b_{k-s+2} = \ldots = b_k = 0$$

La forma del estimador de componentes principales es

$$\hat{\alpha}_{PC} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ . \\ . \\ \hat{\alpha}_{k-s} \\ \text{-----} \\ 0 \\ 0 \\ . \\ . \\ 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{matrix} \begin{matrix} \\ \\ \\ \\ \text{k-s componentes} \\ \\ \\ \\ \\ \text{s componentes} \\ \\ \end{matrix}$$

en términos de las regresoras estandarizados, tenemos

$$\hat{\beta}_{PC} = T\hat{\alpha}_{PC}$$

$$= \sum_{j=1}^{k-s} \lambda_j^{-1} t_j' X' y t_j \qquad (4.3.1.5)$$

Ahora podemos detallar de una manera equivalente el análisis anterior. El estimador de componentes principales de β se obtiene borrando una ó más de las variables z_i , aplicando mínimos cuadrados y haciendo una transformación que nos regresa al espacio paramétrico inicial. Supongamos que $\mathbf{Z}$ ha sido particionada en dos matrices, $\mathbf{Z}_1$, compuesta por las z_i a ser retenidas y $\mathbf{Z}_2$ compuesta por las z_i a ser removidas. Nuestro modelo **(4.3.1.1)** toma la forma

$$\mathbf{y = XT_1\alpha_1 + XT_2\theta_2 + \varepsilon = Z_1\alpha_1 + Z_2\alpha_2 + \varepsilon} \qquad \textbf{(4.3.1.6)}$$

Las propiedades de $\hat{\alpha}_1 = (\mathbf{Z_1'Z_1})^{-1}\mathbf{Z_1y}$, el estimador de mínimos cuadrados de α_1 con $\mathbf{Z}_2$ omitido de **(4.3.1.6)** se obtienen fácilmente. Específicamente, $\hat{\alpha}_1$ es insesgado, esto debido a la ortogonalidad de $\mathbf{Z}_1$ y $\mathbf{Z}_2$ y tiene

$$\text{Var} (\hat{\alpha}_1) = \hat{\sigma}^2(\mathbf{Z_1'Z_1})^{-1}$$

El estimador de componentes principales se obtiene mediante una transformación lineal inversa. Como $\mathbf{b=T\alpha = T_1\alpha_1 + T_2\alpha_2}$, al omitir las componentes en $\mathbf{Z}_2$ hacemos que α_2 se establezca automáticamente igual a cero.

Por ello,

$$\mathbf{T_2\alpha_2 = 0}$$

y el estimador de componentes principales de b es

$$\hat{\beta}_{PC} = T_1\hat{\alpha}_2 = T\hat{\alpha}_{PC}$$

donde $\hat{\alpha}_{PC}=(\hat{\alpha}_1' ,0')'$ con 0 un vector nulo. Las propiedades de $\hat{\beta}_{PC}$ se siguen al ver su equivalencia con el estimador de mínimos cuadrados en el modelo $y = X\beta + \varepsilon$ sujeto a $T_2'\beta = 0$

El estimador de componentes principales, tiene una varianza muestral menor que el estimador de mínimos cuadrados $\hat{\beta}$, pero es sesgado a menos que $T_2'\beta=0$ sea cierta. De forma natural surge la pregunta de como seleccionar las componentes a remover y cuáles serán las consecuencias de ello. La respuesta no es única, pero podemos distinguir dos enfoques principales.

Entonces:

$$T_2\alpha_2 = 0$$

y el estimador de componentes principales de b es

$$\hat{\beta}_{PC} = T_1\hat{\alpha}_2 = T\hat{\alpha}_{PC}$$

donde $\hat{\alpha}_{PC}=(\hat{\alpha}_1' ,0')'$ con 0 un vector nulo. Las propiedades de $\hat{\beta}_{PC}$ se siguen al ver su equivalencia con el estimador de mínimos cuadrados en el modelo $y = X\beta + \varepsilon$ sujeto a $T_2'\beta = 0$

El estimador de componentes principales, tiene una varianza muestral menor que el estimador de mínimos cuadrados $\hat{\beta}$, pero es sesgado a menos que $T_2'\beta=0$ sea cierta. De forma natural surge la pregunta de cómo seleccionar las componentes a remover y cuáles serán las consecuencias de ello. La respuesta no es única, pero podemos distinguir dos enfoques principales.

El primero de ellos, incluye en Z_2 a aquellas componentes asociadas con las raíces características pequeñas, esto supondría que las dependencias cercanas se aceptan como exactas. Esta hipótesis no es obviamente válida, no obstante, se justifica sobre la base de mantener tanto como sea posible la varianza en la muestra mientras se reduce la dimensionalidad del problema.

El segundo de ellos, se basa en el uso de pruebas de hipótesis sobre las restricciones de la muestra $T_2'\beta=0$. Este enfoque es equivalente a probar si estas dependencias lineales se cumplen para las variables en la población o no. De hecho, dado que la relación no se cumple exactamente, habrá problemas de interpretación, esto debido a que en algunas ocasiones los estimadores obtenidos serán inadmisibles.

En vista de lo anterior, la regresión en componentes principales no provee estimadores con mejores propiedades muestrales que los estimadores de mínimos cuadrados sobre el espacio de parámetros.

La utilidad del análisis de componentes principales radica en su uso como herramienta exploratoria, sirviendo además para identificar la multicolinealidad en el problema dado.

4.3.2 Regresión Ridge

Puesto que el método de mínimos cuadrados produce estimadores deficientes de los coeficientes de regresión al ser aplicado a datos que obedecen aparentemente a relaciones lineales - esto es, estimadores que son demasiado grandes en valor absoluto en promedio y con varianzas considerablemente infladas -, podemos preguntarnos si los requerimientos del método de mínimos cuadrados pueden relajarse. En particular, nos preguntamos lo que sucedería si el requerimiento de insesgamiento en β se desecha. El Teorema de Gauss-Markov (ver Apéndice) nos asegura que el estimador de mínimos cuadrados tiene varianza mínima en la clase de estimadores lineales insesgados pero no asegura que la varianza será pequeña en comparación con la de otros estimadores.

El artículo clásico al respecto de la regresión ridge es el de Hoerl y Kennard (1970) y a partir de él mucho se ha hablado sobre este procedimiento. La regresión ridge fue sugerida como un procedimiento para investigar la sensibilidad de los estimadores de mínimos cuadrados, basados en datos que exhibían multicolinealidades

cercanas, así que pequeñas perturbaciones en los datos producían grandes cambios en las magnitudes de los coeficientes estimados.

Así, los estimadores obtenidos por regresión ridge se esperaba que mejoraran a los estimadores de mínimos cuadrados en el sentido que tenían riesgos más pequeños que los estimadores convencionales de mínimos cuadrados. En esta sección, investigaremos las propiedades de los estimadores obtenidos por regresión ridge.

Consideremos el modelo

$$y = X\beta + \varepsilon$$

el error cuadrático medio del estimador $\hat{\beta}^*$, un estimador que suponemos sesgado y con varianza menor que el estimador $\hat{\beta}$, está definido como:

$$ECM(\hat{\beta}^*) = E(\hat{\beta}^* - \beta)^2 = V(\hat{\beta}^*) + [E(\hat{\beta}^*) - \beta]^2$$

o

$$ECM(\hat{\beta}^*) = Var(\hat{\beta}^*) + (sesgo\ \hat{\beta}^*)^2$$

Si permitimos un pequeño sesgo de $\hat{\beta}^*$, la varianza de $\hat{\beta}^*$ puede hacerse pequeña de tal forma que $ECM(\hat{\beta}^*)$ sea menor que la varianza del estimador insesgado $\hat{\beta}$. Montgomery y Peck (1982) ilustran lo anterior en la figura 4.1.

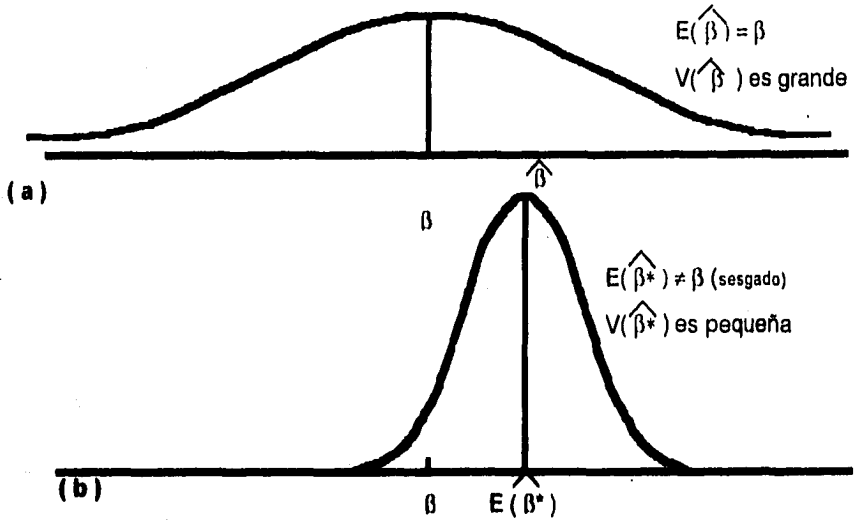

Figura 4.1 Distribución muestral de un estimador a) insesgado, b) sesgado

Es interesante notar que

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta^{(i)} + \sigma^2 \text{traza}(\mathbf{X'X})^{-1} > \beta'\beta + \sigma^2/\lambda_k$$

donde λ_k es la mínima raíz característica de **X'X**. Así, con datos mal condicionados, se tendría una λ_k muy pequeña, lo que implica que la longitud al cuadrado esperada del estimador de mínimos cuadrados es mucho mayor que la longitud al cuadrado del vector de coeficientes reales.

Hoerl y Kennard (1970) sugieren que si deseamos controlar la inflación de los coeficientes y la inestabilidad asociada con el estimador de mínimos cuadrados, hagamos uso de

$$\hat{\beta}_R = (X'X + KI)^{-1}X'y$$

donde $K \geq 0$ define a la familia de estimadores ridge, de la cual el estimador de mínimos cuadrados se obtiene haciendo k=0.

Como Montgomery y Peck (1982) señalan, el estimador ridge es una transformación lineal del estimador de mínimos cuadrados

$$\begin{aligned}\hat{\beta}_R &= (X'X + KI)^{-1}X'y \\ &= (X'X + KI)^{-1}(X'X)\hat{\beta} \\ &= Z_K\hat{\beta}\end{aligned}$$

nos referiremos a K como el parámetro de sesgo. La matriz de varianzas y covarianzas de $\hat{\beta}_R$ es

$$V(\hat{\beta}_R) = \sigma^2(X'X + KI)^{-1}X'X(X'X + KI)^{-1}$$

y el error cuadrático medio del estimador ridge es

$$\begin{aligned}ECM(\hat{\beta}_R) &= V(\hat{\beta}_R) + (sesgo\ \hat{\beta}_R)^2 \\ &= \sigma^2 Tr[(X'X + KI)^{-1}X'X(X'X + KI)^{-1}] \\ &\quad + K^2\beta'(X'X+KI)^{-2}\beta \\ &= \sigma^2 \sum_{j=1}^{k} \lambda_j/(\lambda_j + k_j)^2 + K^2\beta'(X'X + KI)^{-2}\beta\end{aligned}$$

donde $\lambda_1, \lambda_2, \ldots, \lambda_k$ son los valores propios de X'X.

Las propiedades del estimador ridge son las siguientes:

1) $\hat{\beta}_R$ minimiza la suma de cuadrados de los residuales en la esfera centrada en el origen y cuyo radio es la longitud de $\hat{\beta}_R$. Para una suma dada de cuadrados de los residuales, es el vector de coeficientes con longitud mínima. La interpretación geométrica es interesante, como lo muestra la figura 4.2., el punto $\hat{\beta}$ en el centro de la elipse corresponde a la solución de mínimos cuadrados (para el caso de un problema con dos regresoras), donde la suma de cuadrados de los residuales toma su mínimo valor. La elipse menor representa el lugar geométrico de puntos en el plano β_1, β_2 donde la suma de cuadrados de los residuales es constante con un valor mayor que el mínimo.

Hocking et. al. (1976), han observado que los estimadores ridge encogen al estimador de mínimos cuadrados, con respecto a los contornos de $X'X$, $\hat{\beta}_R$ es pues, la solución de

$$\min \ (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})$$

$$\text{s.a.} \ \beta'\beta \leq d^2$$

donde **d depende de K**

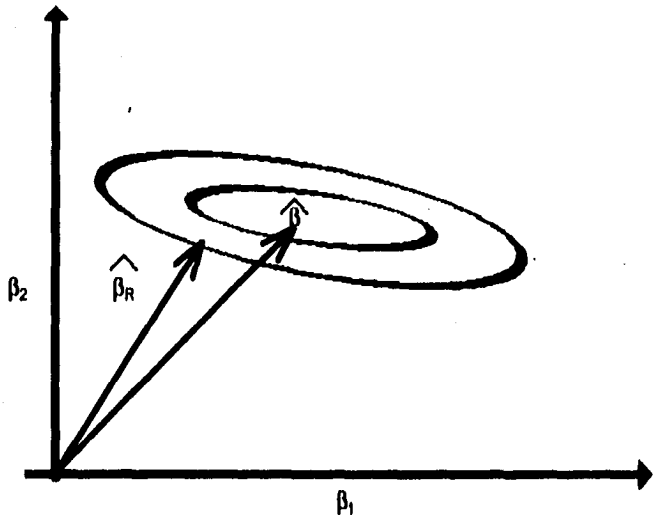

FIGURA 4.2 Interpretación geométrica del estimador ridge

2) La suma de cuadrados de los residuales es una función creciente de K. En efecto, dado que la expresión de la suma de cuadrados de los residuales es

$$SC_E = (y - X\hat{\beta}_R)'(y - X\hat{\beta}_R)$$
$$= (y - X\hat{\beta})'(y - X\hat{\beta}) + (\hat{\beta}_R + \hat{\beta})'X'X(\hat{\beta}_R - \hat{\beta})$$

el primer término de esta última expresión es la suma de cuadrados de los residuales para los estimados $\hat{\beta}$, así que si K aumenta, la suma de cuadrados de los residuales aumenta.

3) $\hat{\beta}_R'\hat{\beta}_R < \hat{\beta}'\hat{\beta}$, y $\hat{\beta}_R'\hat{\beta}_R \to 0$ si $K \to \infty$

$\hat{\beta}_R = Z_K\hat{\beta}$, donde $Z_K = (X'X + KI)^{-1}(X'X)$

$\hat{\beta}_R = (Z_K\hat{\beta})' = \hat{\beta}Z_K'$

$\hat{\beta}_R\hat{\beta}_R = \hat{\beta}'Z_k'Z\hat{\beta}$

4) La razón de la raíz característica más grande de la matriz Z_K a la raíz más pequeña, llamada el número de condición, es $(\lambda_1 + K)/(\lambda_K + K)$, donde λ_1, λ_2, ..., λ_K son las raíces características en orden descendente, y es una función decreciente de K

5) El estimador ridge

$$\hat{\beta}_R = [X'X + KI]^{-1}X'y$$

es una transformación lineal del estimador de mínimos cuadrados.

6) El error cuadrático medio de $\hat{\beta}_R$ es

$$ECM = [\sigma^2] \sum_{j=1}^{k} \lambda_j/(\lambda_j + K)^2 + K^2\beta'(X'X + KI)^{-2}\beta$$

8) Hoerl y Kennard (1970), mostraron que siempre existe una k>0 tal que $\hat{\beta}_R$ tiene un error cuadrático medio menor que $\hat{\beta}$. Esta propiedad es la propiedad de la familia de estimadores ridge que proporciona una esperanza de mejorar la estimación. Debe aclararse que $\hat{\beta}_R$ mejora $\hat{\beta}$

solo en un rango limitado del espacio de parámetros y que esta región depende de los parámetros desconocidos β y σ^2.

Como Montgomery y Peck (1982) señalan, la controversia en el uso de este estimador se centra alrededor de la elección del parámetro de sesgo K. Los métodos de elección son algunos de naturaleza subjetiva como la traza ridge y otros de naturaleza más analítica.

4.3.2.a Traza ridge

Es una gráfica bidimensional de los valores de los coeficientes ridge estimados $\hat{\beta}_R$ y la suma de cuadrados de los residuales para un número de valores de K. Se traza una curva de traza para cada coeficiente. Hoerl y Kennard (1975) sugieren seleccionar el valor de K sobre la base de criterios como la estabilidad de los coeficientes estimados cuando K se incrementa, signos razonables, magnitudes de los coeficientes y factores de inflación de varianza máxima. Otra regla de decisión es la propuesta por O'Hagan et. al. (1974), que sugieren seleccionar una valor de K tal que ahí alcance el último estimado ridge su magnitud absoluta máxima después de alcanzar su "último" signo, donde "último" signo se define se define como el signo en, por decir, K=0.9.

Este método exhibe la sensitividad de los coeficientes a pequeñas perturbaciones de los datos.

4.3.2.b El estimador de Hoerl - Kennard y Baldwin

Hoerl, Kennard y Baldwin (1975) sugirieron adoptar K como

$$K_{HKB} = K\sigma^2 / (\hat{\beta}'\hat{\beta}) \qquad (4.3.2.2)$$

proponiendo como procedimiento iterativo de estimación de **(4.3.2.2)** la siguiente secuencia de estimados de β y K_{HKB}.

$$\hat{\beta} \qquad K^0_{HKB} = K\sigma^2 / (\hat{\beta}'\hat{\beta})$$

$$\hat{\beta}_R(K_0) \qquad K^1_{HKB} = K\sigma^2 / [\hat{\beta}_R'(K_0)\hat{\beta}_R(K_0)]$$

$$\hat{\beta}_R(K_1) \qquad K^2_{HKB} = K\sigma^2 / [\hat{\beta}_R'(K_1)\hat{\beta}_R(K_1)]$$

$$\vdots$$

El cambio relativo en K^j_{HKB} se utiliza para terminar las iteraciones. Si

$$\frac{K_{j+1} - K_j}{K_j} > 20(T^{-1.3}) \text{ entonces continuamos}$$

donde $T = Tr(X'X)^{-1} / K$. Esta elección está justificada ya que **T** se incrementa con la desviación en los valores propios de **X'X**, lo cual provoca una mayor reducción de algunos coeficientes al incrementarse el mal condicionamiento en los datos.

4.3.2.c El estimador Mc.Donald-Galarneau

Dado que

$$E (\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 Tr(X'X)^{-1}$$

la longitud al cuadrado esperada del estimador de mínimos cuadrados es mayor que la del vector de coeficientes reales. McDonald y Galarneau (1975) notaron que

$$E(\hat{\beta}'\hat{\beta} - \sigma^2 Tr(X'X)^{-1}) = \beta'\beta$$

lo que sugiere que k se elija tal quede tal forma que:

$$\hat{\beta}_R(K)\hat{\beta}_R(K) = \hat{\beta}'\hat{\beta} - \hat{\sigma}^2 Tr(X'X)^{-1} \qquad (4.3.2.3)$$

esto significa que el estimador ridge adaptado tiene la misma longitud al cuadrado esperada que el vector de parámetros reales.

4.3.2.d Regresión ridge generalizada.

Hoerl y Kennard (1970) propusieron una extensión del procedimiento de regresión ridge que permite separar los parámetros de sesgo para cada regresor. Es más cómodo en este caso, trabajar con el modelo canónico u ortogonal. Sea **P** una matriz cuyas columnas son vectores característicos ortonormales de **X'X**. Entonces

$$y = XPP'\beta + \varepsilon = Z\theta + \varepsilon$$

donde $Z = XP$ y $\theta = P'\beta$. Además $Z'Z = P'X'XP = \Lambda$ es una matriz diagonal cuyos elementos λ_i son las raíces características de **X'X**. El estimador ridge generalizado se define como

$$\hat{\theta}_{RG} = [Z'Z + D]^{-1}Z'y = [\Lambda + D]^{-1}\Lambda\hat{\theta} \qquad (4.3.2.4)$$

donde

$$\hat{\theta} = (Z'Z)^{-1}Z'y$$

En el espacio β,

$$\beta_{RG} = P\theta_{RC}$$

$$= (X'X + PDP')^{-1}X'y \qquad (4.3.2.5)$$

donde **D** es una matriz diagonal con elementos $d_i \geq 0$, i=1,2,...,k. Por esto, se elige d_i pequeño si λ_i es grande y d_i grande si λ_i es pequeño, de tal manera que la mayoría de los elementos estimados de θ que han sido estimados de manera precisa serían reducidos en menor grado que los elementos estimados imprecisamente. Para el estimador ridge generalizado el valor óptimo de d_i es σ^2/θ_i^2 bajo la medida de pérdida con error cuadrático. Los procedimientos considerados para seleccionar los valores d_i cuando θ y σ^2 no son conocidos son de naturaleza iterativa, el propuesto por Hoerl y Kennard (1970) es el siguiente

$$d_i = \hat{\sigma}^2/\hat{\theta}_i^2$$

y se obtiene un estimador para

$$\hat{\theta}_{RG} = [\Lambda + D]^{-1}\Lambda\hat{\theta}$$

el cual se denota $\hat{\theta}_{RG}^{1}$, las nuevas d_i's se construyen a partir de $\hat{\theta}_{RG}^{1}$.

Hemmerle (1975) y Teeckens (1977) ha demostrado que la iteración no es necesaria ya que el valor límite puede ser determinado analíticamente. El método iterativo provee la solución de las ecuaciones

$$\theta_{i,RG} = \frac{c_i}{\lambda i + d_i} \qquad (4.3.2.6)$$

$$d_i = \frac{\hat{\sigma}^2}{[\hat{\theta}^*_{i,RG}]^2} \qquad (4.3.2.7)$$

y donde c_i es el i-ésimo elemento de **Z'y**

Sustituyendo **(4.3.2.7)** en **(4.3.2.6)** se tiene una ecuación cuadrática en $\hat{\theta}^*_{i,RG}$ con soluciones

$$\hat{\theta}^*_{i,RG} = \frac{1}{2\lambda_i}[c_i + (c_i^2 - 4\lambda_i\hat{\sigma}^2)^{1/2}] \qquad (4.3.2.8)$$

siempre que $c_i^2 \geq 4\lambda_i\hat{\sigma}^2$. Los valores correspondientes de d_i resultan de la sustitución de **(4.3.2.5)** en **(4.3.2.7)**.

Es importante considerar la efectividad de los métodos de estimación que fueron expuestos en esta sección dado que la naturaleza de ellos es distinta. Como Montgomery y Peck (1982) señalan, existe una tendencia a considerar como superiores a los estimadores obtenidos al abandonar el requerimiento de insesgamiento. Algunos investigadores considerarán los procedimientos gráficos de gran utilidad, mientras que otros, dada la subjetividad de los mismos, requerirán otros procedimientos como la selección de variables, por citar alguno. Dado que la estimación sesgada es apropiada cuando se adiciona información externa al problema, el problema es visto por algunos investigadores - desde una

perspectiva Bayesiana -como uno que refleja el conocimiento a priori de los coeficientes por parte del analista.

Esto es, al enfrentarnos con muestras que no son suficientemente ricas en información surge la pregunta natural sobre cómo mejorar los estimadores "deficientes" que son obtenidos por el método tradicional de mínimos cuadrados. Así, para aquellos que desearan estimadores puntuales únicamente, existe un gran avance al tratar de obtener medidas que resuman los efectos de la multicolinealidad ó su severidad. La pregunta relevante es: pueden ser estimadas de forma relativamente precisa las funciones paramétricas de interés usando los estimadores convencionales únicamente con la información disponible de la muestra.

CAPÍTULO V

EVIDENCIA EXPERIMENTAL

Con el fin de mostrar el efecto de la multicolinealidad e ilustrar los diferentes métodos de solución al problema de la multicolinealidad que fueron exhibidos en ésta tesis, trabajaremos con el siguiente modelo lineal, propuesto por Judge, et. al. (1985). Estos autores proponen el modelo:

$$y = 10.0X_1 + 0.4X_2 + 0.6X_3 + 0.6X_4 + 2X_5 + \varepsilon$$

donde **X** es la siguiente matriz de diseño

Tabla 5.1 Matriz de diseño de Judge, et. al. (1985)

$$\mathbf{X} = \begin{bmatrix} X_1 & X_2 & X_3 & X_4 & X_5 \\ 1.0000 & 0.6930 & 0.6930 & 0.6684 & 2.0546 \\ 1.0000 & 1.7330 & 0.6930 & 2.1371 & 3.1229 \\ 1.0000 & 0.6930 & 1.3860 & 1.1698 & 2.5122 \\ 1.0000 & 1.7330 & 1.3860 & 1.9562 & 4.0698 \\ 1.0000 & 0.6930 & 1.7920 & 0.7942 & 3.4415 \\ 1.0000 & 2.3400 & 0.6930 & 2.5684 & 3.9941 \\ 1.0000 & 1.7330 & 1.7920 & 2.0581 & 3.9859 \\ 1.0000 & 2.3400 & 1.3860 & 2.5217 & 3.7654 \\ 1.0000 & 2.3400 & 1.7920 & 2.6976 & 4.5443 \\ 1.0000 & 0.6930 & 0.6930 & 0.7015 & 1.6239 \\ 1.0000 & 0.6930 & 1.3860 & 1.1059 & 2.1703 \\ 1.0000 & 1.7330 & 0.6930 & 1.7434 & 3.4031 \\ 1.0000 & 1.7330 & 1.3860 & 1.9974 & 3.8763 \\ 1.0000 & 0.6930 & 1.7920 & 0.8544 & 3.3523 \\ 1.0000 & 2.3400 & 0.6930 & 2.4991 & 3.7269 \\ 1.0000 & 1.7330 & 1.7920 & 2.1302 & 3.5307 \\ 1.0000 & 2.3400 & 1.3860 & 2.6212 & 4.4914 \\ 1.0000 & 2.3400 & 1.7920 & 2.4942 & 4.9053 \\ 1.0000 & 1.7330 & 1.3860 & 1.6995 & 3.5624 \\ 1.0000 & 0.6930 & 0.6930 & 0.8516 & 1.5174 \end{bmatrix}$$

y las variables X_4 y X_5 poseen una característica especial. Ellas fueron construidas como:

$$X_4 = X_2 + X_3 + \omega_1$$

$$X_5 = X_2 + \omega_2$$

donde ω_1 es un vector aleatorio de una distribución uniforme sobre el intervalo [0,1] y ω_2 es un vector aleatorio de una distribución uniforme sobre el intervalo [0,½], e es un vector de variables aleatorias normales, independientes e idénticamente distribuidas con media cero y varianza constante 1.0.

Usando este modelo fueron generadas cinco muestras de datos las cuales se presentan en la tabla 5.2

Tabla 5.2 Matriz de diseño de Judge, et. al. (1985)

$$y = \begin{array}{c} \begin{matrix} y_1 & y_2 & y_3 & y_4 & y_5 \end{matrix} \\ \begin{bmatrix} 16.64 & 15.81 & 15.48 & 16.97 & 15.97 \\ 17.77 & 16.76 & 19.36 & 16.35 & 16.94 \\ 16.98 & 17.31 & 16.70 & 16.92 & 15.38 \\ 22.40 & 20.02 & 20.73 & 22.41 & 22.44 \\ 19.19 & 16.51 & 17.79 & 16.58 & 19.33 \\ 22.03 & 21.66 & 19.72 & 20.70 & 21.75 \\ 22.22 & 22.67 & 20.96 & 23.66 & 23.03 \\ 20.65 & 22.01 & 20.77 & 21.14 & 22.29 \\ 22.88 & 22.49 & 23.05 & 21.50 & 23.33 \\ 16.24 & 14.29 & 14.52 & 14.87 & 14.71 \\ 17.85 & 17.17 & 16.63 & 17.15 & 17.77 \\ 18.95 & 21.98 & 19.97 & 20.37 & 19.99 \\ 21.12 & 20.76 & 22.29 & 21.72 & 19.98 \\ 19.85 & 19.45 & 19.24 & 18.29 & 20.27 \\ 21.34 & 23.66 & 22.17 & 20.85 & 21.08 \\ 22.74 & 20.99 & 21.37 & 19.15 & 19.58 \\ 24.36 & 24.18 & 21.34 & 24.10 & 24.14 \\ 24.70 & 25.19 & 23.71 & 24.81 & 23.27 \\ 18.92 & 20.59 & 20.81 & 21.00 & 20.17 \\ 15.31 & 13.12 & 16.08 & 15.58 & 13.16 \end{bmatrix} \end{array}$$

5.1 Definición

Con las muestras anteriormente obtenidas, procedemos a utilizar los principales métodos de detección que fueron expuestos en el capítulo III.

5.1.1 Examen de la matriz de correlación

En el capítulo III, dijimos que el examen de la matriz de correlación constituía una medida sencilla de multicolinealidad, de tal forma que, al analizar los elementos fuera de la diagonal principal de **X'X** - suponiendo que esta ha sido centrada y cambiada de escala previamente -, se tiene que éstos coinciden con los r_{ij}, donde r_{ij} es la correlación entre X_i y X_j. Dillon y Goldsmith (1984) detallan el procedimiento a seguir para obtener tal matriz de correlación, procedimiento que detallamos en el apéndice en A-25 a A-30. La matriz de datos de nuestro problema tendrá la siguiente forma, después de haber sido centrada y de haber eliminado la columna de ceros.

$$\mathbf{X} = \begin{bmatrix} -0.6581 & -0.5722 & -1.1151 & -1.3260 \\ 0.1619 & -0.5722 & 0.3536 & -0.2597 \\ -0.6581 & 0.1207 & -0.6137 & -0.8704 \\ 0.1619 & 0.1207 & 0.1727 & 0.6672 \\ -0.6581 & 0.5267 & -0.9893 & 0.0589 \\ 0.7689 & -0.5722 & 0.7849 & 0.6114 \\ 0.1619 & 0.5267 & 0.2746 & 0.6033 \\ 0.7689 & 0.1207 & 0.7382 & 0.3826 \\ 0.7689 & 0.5267 & 0.9141 & 1.1617 \\ -0.6581 & -0.5722 & -1.0820 & -1.7567 \\ -0.6581 & 0.1207 & -0.8776 & -1.2124 \\ 0.1619 & -0.5722 & -0.0401 & 0.0204 \\ 0.1619 & 0.1207 & 0.2139 & 0.4937 \\ -0.6581 & 0.5267 & -0.9291 & -0.0304 \\ 0.7689 & -0.5722 & 0.7156 & 0.3463 \\ 0.1619 & 0.5267 & 0.3467 & 0.1481 \\ 0.7689 & 0.1207 & 1.0377 & 1.1067 \\ 0.7689 & 0.5267 & 0.7107 & 1.5226 \\ 0.1619 & 0.1207 & 0.1160 & 0.1798 \\ -0.6581 & -0.5722 & -0.9319 & -1.8652 \end{bmatrix}$$

Cuando los datos han sido centrados (por columnas), el vector constante $\hat{\beta}_1$ es igual a **0**, sin importar que valores tomen $\hat{\beta}_2$, $\hat{\beta}_3,\ldots,\hat{\beta}_k$, así nuestro modelo puede escribirse como:

$$y = X\beta + \varepsilon$$

Una vez obtenida **X**, obtenemos entonces **S**, donde

$$S = \frac{1}{(n-1)} X'X$$

y en nuestro caso

$$1/19(X'X) = \begin{bmatrix} 0.4800 & 0.0279 & 0.5035 & 0.5475 \\ 0.0279 & 0.2136 & 0.0550 & 0.2284 \\ 0.5035 & 0.0550 & 0.5506 & 0.5854 \\ 0.5475 & 0.2284 & 0.5854 & 0.9109 \end{bmatrix}$$

donde s_{ij}^2 es la covarianza entre X_i y X_j para la matriz **X** que ha sido centrada y cambiado de escala y que tiene una columna menos que la matriz de diseño original.

Sea **D** la siguiente matriz diagonal cuyos elementos en la diagonal son los mismos que los que componen la diagonal de **S**.

$$D = \begin{bmatrix} 0.4800 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.2136 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.5506 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.9109 \end{bmatrix}$$

y

$$D^{-1/2} = \begin{bmatrix} 1/(.4800)^{.1/2} & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 1/(.2136)^{.1/2} & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1/(.5506)^{.1/2} & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 1/(.9109)^{.1/2} \end{bmatrix}$$

donde $\mathbf{D}^{-1/2}$ es una matriz diagonal, cuyos elementos de la diagonal principal son los inversos de las raíces cuadradas de los elementos de la diagonal principal de **D**.

$$\mathbf{D}^{-1/2} = \begin{bmatrix} 1.4434 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 2.1640 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1.3477 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 1.0477 \end{bmatrix}$$

por lo que **R**, la matriz de correlación está dada por

$$\mathbf{R} = \mathbf{D}^{-1/2}\,\mathbf{S}\,\mathbf{D}^{-1/2}$$

y obtenemos finalmente la expresión para **R**

$$\mathbf{R} = \begin{bmatrix} 1.0000 & 0.0871 & 0.9795 & 0.8280 \\ 0.0871 & 1.0005 & 0.1603 & 1.0001 \\ 0.9795 & 0.1603 & 1.0001 & 0.8265 \\ 0.8280 & 0.5179 & 0.8265 & 0.9999 \end{bmatrix}$$

Una vez obtenida **R**, podemos emplear los resultados del capítulo **XIX**. Como vimos en dicho capítulo, la multicolinealidad entre dos variables estaría dada por valores de r_{ij} cercanos al uno, este es nuestro caso, ya que r_{13}=0.9795, recordemos que se está utilizando una forma especial de la matriz **X** que no incluye los elementos iguales a uno de la columna $\mathbf{X}_1$ de la matriz de diseño original **X**; por esta causa, r_{13}=0.9795 se interpreta como una elevada correlación entre $\mathbf{X}_{1+1}=\mathbf{X}_2$ y $\mathbf{X}_{3+1}=\mathbf{X}_4$, lo cual está de acuerdo con nuestra construcción original, en la cual hacíamos $\mathbf{X}_4$ =

$X_2 + X_3 + \omega_1$. Es así que esperaríamos una elevada correlación entre X_4 y X_3, esto es, un valor de r_{23} cercano al uno, sin embargo, este último es sólo de 0.1603, lo cual, si no conociéramos la relación existente entre X_4, X_2, X_3 no constituiría evidencia sólida de que las variables X_4 y X_3 exhiben una dependencia cercana a la lineal.

Parece ser más sugerente que r_{14}=0.8280, esto es un valor cercano al uno por lo que puede sospecharse que $X_{1+1}=X_2$ y $X_{4+1}=X_5$ se encuentran involucradas en alguna relación lineal, lo cual es cierto, dada la construcción del problema. Una relación no tan evidente es la que proporciona el valor r_{34}, este valor es de 0.8265 , también cercano al uno, por lo que puede existir una relación entre $X_{4+1}=X_5$ y $X_{3+1}=X_4$.

Notemos que

$$X_4 = X_2 + X_3 + \omega_1$$

y

$$X_5 = X_2 + \omega_2$$

por lo que podremos sospechar que existe alguna relación entre X_3 y X_5. El valor r_{24}= 0.5179 no es muy cercano al uno, pero si lo suficientemente grande para no abandonar tal sospecha.

5.2 Factores de inflación de varianza

Como expusimos en el capítulo III, los elementos de la diagonal principal de la matriz $(X'X)^{-1}=S$, esto es, los factores de inflación de varianza (**VIF**), son de gran utilidad en la detección de la multicolinealidad . La matriz $S=(X'X)^{-1}$ es la siguiente:

$$S = (X'X)^{-1} = \begin{bmatrix} 44.1715 & 8.0682 & -34.3244 & -12.3833 \\ 8.0682 & 3.4517 & -4.5984 & -4.6675 \\ -34.3245 & -4.5984 & 31.2402 & 4.9821 \\ -12.3832 & -4.6675 & 4.9821 & 9.5529 \end{bmatrix}$$

como habíamos visto en el capítulo III, $S_{jj}=(1-R_j^2)^{-1}$ donde R_j^2 era el coeficiente de determinación obtenido cuando se ajusta una regresión de X_j sobre el conjunto de regresoras restantes. De hecho, en el mismo capítulo vimos que VIF mayores de 5 o 10 indicaban que los coeficientes de regresión obtenidos por mínimos cuadrados estaban siendo estimados pobremente debido a la multicolinealidad.

En este problema $C_{11}=44.1715$, $C_{33}=31.2402>10$ y $C_{44}=9.5529>5$, por lo que los coeficientes de las variables X_2, X_4, X_5 estan siendo afectados por la multicolinealidad , lo cual concuerda con el análisis anterior para los valores de la matriz de correlación.

5.3 Valores del determinante $|X'X|$

Un valor cercano a cero para $|X'X|$ corroboraría la información hasta ahora recabada, para este problema, tenemos que

$$\det X'X = |X'X| = 0.0016$$

el cual, a pesar del redondeo hasta la cuarta cifra decimal, resulta ser un valor muy bajo, es por ello que podemos afirmar que existe dependencia lineal entre las columnas 2, 3, 4, 5 ó algún subconjunto de ellas.

5.4 Análisis de valores y vectores propios

Una virtud del análisis de valores y vectores propios es que no tenemos que realizar todos los cálculos basados en la matriz reducida **X**, sino que podemos trabajar de inmediato con la matriz de diseño original **X**. En particular, la descomposición en valores singulares de la matriz **X'X** arroja los resultados siguientes:

$$X'X = \begin{bmatrix} 20.0000 & 31.0220 & 25.3050 & 35.6698 & 67.6529 \\ 31.0220 & 57.2384 & 39.7606 & 64.8947 & 115.3389 \\ 25.3050 & 39.7606 & 36.0763 & 46.1755 & 89.9381 \\ 35.6698 & 64.8947 & 46.1755 & 74.0490 & 131.7801 \\ 67.6529 & 115.3389 & 89.9381 & 131.7801 & 246.1525 \end{bmatrix}$$

$$U = \begin{bmatrix} 0.2092 & -0.2257 & 0.8530 & -0.4159 & 0.0688 \\ 0.3607 & 0.5151 & -0.0529 & -0.0789 & 0.7717 \\ 0.2767 & -0.6333 & -0.4903 & -0.4879 & 0.2099 \\ 0.4122 & 0.4863 & -0.1616 & -0.4835 & -0.5777 \\ 0.7614 & -0.2151 & 0.0563 & 0.5907 & -0.1480 \end{bmatrix}$$

$$T = \begin{bmatrix} 423.3963 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 7.8327 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1.2377 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.9547 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.1249 \end{bmatrix}$$

$$V = \begin{bmatrix} 0.2092 & -0.2257 & 0.8530 & -0.4159 & 0.0688 \\ 0.3607 & 0.5151 & -0.0529 & -0.0769 & 0.7717 \\ 0.2767 & -0.6333 & -0.4903 & -0.4679 & 0.2099 \\ 0.4122 & 0.4863 & -0.1616 & -0.4635 & -0.5777 \\ 0.7614 & -0.2151 & 0.0563 & 0.5907 & -0.1460 \end{bmatrix}$$

donde **X'X=UTV'** con **U'U=I**, **V'V=I** y **T** es una matriz diagonal cuyos elementos de la diagonal principal son los valores propios de **X'X**, en orden descendente. El número de condición de la matriz es

$$k = \frac{\lambda_{max}}{\lambda_{min}} = \frac{423.3963}{0.1249} = 3389.882306$$

como vimos en el capítulo **III**, números de condición de una magnitud así (>1000) indicaban severa multicolinealidad entre las columnas de **X**, lo cual es evidente en este problema.

Además, por ser **X'X** real y simétrica **U=V**, por lo que **X'X** puede escribirse como

$$X'X = UTU'$$

La descomposición en valores singulares de **X** arrojó los siguientes resultados:

$$T^{1/2} = \begin{bmatrix} 20.5766 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 2.7987 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1.1125 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.9971 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.3534 \end{bmatrix}$$

$$M = \begin{bmatrix} 0.1210 & -0.1517 & 0.4352 & 0.0820 & 0.1665 \\ 0.2082 & 0.2128 & 0.2265 & -0.0796 & -0.4107 \\ 0.1573 & -0.2566 & 0.0802 & -0.2291 & -0.4332 \\ 0.2490 & -0.0482 & 0.0047 & 0.2300 & -0.0999 \\ 0.1897 & -0.4851 & 0.0028 & 0.3049 & 0.0328 \\ 0.2597 & 0.3325 & 0.1790 & 0.1793 & -0.1550 \\ 0.2534 & -0.1159 & -0.2027 & -0.0678 & 0.0098 \\ 0.2597 & 0.1851 & -0.1311 & -0.2727 & 0.4286 \\ 0.2975 & 0.0640 & -0.2962 & -0.0952 & 0.0561 \\ 0.1058 & -0.1128 & 0.4086 & -0.1892 & 0.2927 \\ 0.1434 & -0.2414 & 0.0721 & -0.4007 & -0.1856 \\ 0.2107 & 0.1228 & 0.2978 & 0.2773 & 0.1155 \\ 0.2426 & -0.0262 & -0.0205 & 0.0953 & -0.0862 \\ 0.1878 & -0.4878 & -0.0104 & 0.2228 & -0.0283 \\ 0.2485 & 0.3408 & 0.1757 & 0.0558 & 0.0693 \\ 0.2380 & -0.0684 & -0.2362 & -0.3725 & 0.0825 \\ 0.2925 & 0.1814 & -0.1379 & 0.0122 & -0.3650 \\ 0.3067 & 0.0009 & -0.2484 & 0.2173 & 0.2374 \\ 0.2290 & -0.0191 & -0.0222 & -0.0432 & 0.2052 \\ 0.1048 & -0.0786 & 0.3814 & -0.3251 & 0.0919 \end{bmatrix}$$

$$U = \begin{bmatrix} 0.2092 & -0.2257 & 0.8530 & -0.4159 & 0.0688 \\ 0.3607 & 0.5151 & -0.0529 & -0.0789 & 0.7717 \\ 0.2767 & -0.8333 & -0.4952 & -0.4879 & 0.2099 \\ 0.4122 & 0.4863 & -0.1618 & -0.4835 & -0.5777 \\ 0.7614 & -0.2151 & 0.0563 & 0.5907 & -0.1480 \end{bmatrix}$$

donde

$$X = M\, T^{1/2}\, U'$$

La inspecciónlos elementos de la matriz diagonal $T^{1/2}$ son los valores singulares de **X**, esto es, los μ_j (j=1,2,3,4,5) que están relacionados con los valores y vectores propios de la matriz **X'X** ya que:

$$X'X = (M T^{1/2} U')' (M T^{1/2} U') = UTU'$$

Como Belsley, et. al., señalan, el mal condicionamiento de la matriz **X** se refleja en el tamaño de los valores singulares. Según ellos, habrá un valor singular pequeño para cada dependencia lineal cercana, la "pequeñez", como discutimos en el capítulo III , se basa en la comparación de cada uno de los valores singulares de **X** con el mayor de los valores singulares de la matriz **X**, este último hecho junto con la información de la matriz $D = T^{1/2}$

$$T^{1/2} = D = \begin{bmatrix} 20.5768 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 2.7987 & 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 1.1125 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.7771 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.3534 \end{bmatrix}$$

nos permite calcular con facilidad los índices de condición de la matriz **X**, siendo éstos:

$$\eta_j = \frac{\mu_{max}}{\mu_j}, \qquad j=1,2,3,4,5$$

por ello

$$\eta_1 = 1 \qquad \text{ya que } M'M=I$$

$\eta_2 = 7.3521$

$\eta_3 = 18.4958$

$\eta_4 = 21.0588$

$\eta_5 = 58.2246$

Retomaremos estos valores cuando analicemos las proporciones de descomposición de varianza.

La matriz de varianza de $\hat{\beta}$ es,

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} = \sigma^2 U T^{-1} U'$$

y la varianza del j-ésimo coeficiente

$$V(\hat{\beta}_j) = \sigma^2 \sum_{i=1}^{5} U_{ji}^2 / \mu j$$

haciendo uso de los valores obtenidos anteriormente, obtenemos que:

$$V(\hat{\beta}_1) = \sigma^2 (0.8135)$$

$$V(\hat{\beta}_2) = \sigma^2 (4.8112)$$

$$V(\hat{\beta}_3) = \sigma^2 (0.8477)$$

$$V(\hat{\beta}_4) = \sigma^2 (2.9687)$$

$$V(\hat{\beta}_5) = \sigma^2 (0.55069)$$

Nótese, que ninguna de estas varianzas es sospechosamente elevada, sin embargo, al centrar y cambiar de escala la matriz **X**,

obtuvimos factores de inflación de varianza elevados, algunos de ellos con valores superiores a 10.

Siguiendo la sugerencia hacha por Belsey, et. al., encontraremos ahora la matriz de proporciones de descomposición de varianza Π, con el elemento π_{ij} dado por

$$\pi_{ij} = \frac{\mu_{ij}^2 / \mu_i^2}{VIF_j} \qquad i,j=1,2,...,p$$

en este caso usaremos la matriz **X**, que no ha sido centrada ni cambiada de escala.

Los resultados aparecen exhibidos en la tabla 5.3

Tabla 5.3 Proporciones de descomposición de varianza para los datos de Judge, et al.

Variables regresoras no centradas

Número	Valor Propio	Indice de condición	Proporción de descomposición de varianza				
			X_1	X_2	X_3	X_4	X_5
1	423.6963	1.0000	0.000127	0.000063	0.000213	0.000135	0.002486
2	7.8327	7.5521	0.007994	0.007040	0.060403	0.010170	0.010726
3	1.2377	18.4958	0.722645	0.000469	0.229121	0.007107	0.004685
4	0.9547	21.0588	0.222716	0.001355	0.294139	0.082482	0.663681
5	0.1249	58.2246	0.046586	0.991017	0.416121	0.900069	0.345970

La tabla 5.3 exhibe los índices de condición de $X(\eta_i)$ y las proporciones de descomposicón de varianza (π_{ij}) para los datos de Judge, Hill, et. al.; es importante recordar que estos datos no han sido centrados ni cambiados de escala. En esta tabla tenemos tres índices de condición η_1, η_4, η_5 mayores que 10; además, las

proporciones π_{31}, π_{45}, π_{52}, π_{54}, exceden 0.5, pero hay otras como π_{41}, π_{43}, π_{53}, con valores entre 0.2 y 0.5, por lo que no puede descartarse la existencia de relaciones lineales directas entre las variables 1 y 3. Los valores $\pi 31$, $\pi 33$, (0.722645 y 0.229121), respectivamente, parecen así confirmarlo. Si analizamos el cuarto renglón, nos daremos cuenta que las proporciones más grandes están dadas por π_{45}, π_{43}, π_{41} en ese orden respectivamente, por lo que parecería que estas tres regresoras están involucradas en una relación de multicolinealidad . El análisis del quinto renglón proporciona evidencia más fuerte, las proporciones más elevadas en este caso son π_{52} y π_{54}, esto concuerda con el hecho de que $\mathbf{X_5 = X_2 + \omega_2}$, por el elevado valor de π_{52}=0.991017, podemos decir que la mayor cantidad de la variabilidad en β_5 proviene de la influencia de $\mathbf{X_2}$, siendo ínfima la contribución a tal variabilidad del vector ω_2. Esto último es congruente con el análisis hecho para el cuarto índice de condición, η_4=21.0588, ya que el renglón correspondiente de la matriz Π arrojaba el valor mayor π_{45}, siendo la relación exacta para $\mathbf{X_4}$, $\mathbf{X_4 = X_2 + X_3 + \omega_6}$. Recordemos que $\mathbf{X_5}$ estaba fuertemente correlacionada con $\mathbf{X_2}$, y que el valor de π_{43} es 0.294139, un valor que, si bien no es muy elevado, debe ser tomado en cuenta para explicar la variabilidad en $\widehat{\beta_4}$.

5.5 Soluciones propuestas

En esta sección propondremos dos soluciones al problema de la multicolinealidad que se presenta en las columnas de la matriz **X** propuesta por Judge, et. al. (1985). Estas soluciones fueron explicadas junto con otros métodos expuestos en el capítulo **IV**, y son: el análisis de componentes principales y la regresión ridge. Comenzaremos con el análisis de componentes principales utilizando los resultados del capítulo **IV** de este trabajo que conciernen a la regresión ridge.

5.5.1 Regresión ridge aplicada a los datos de la tabla 5.1.

Como mencionamos en el capítulo anterior, la técnica conocida como regresión ridge intenta estabilizar los estimadores de los parámetros sacrificando la propiedad de un estimador de ser insesgado a cambio de una reducción de la varianza de los estimadores. Comenzamos nuestro análisis con el modelo de regresión lineal

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

por ello, es conveniente que utilicemos el procedimiento iterativo de Hoerl y Kennard (1970) para calcular los valores las k's mencionadas en el capítulo anterior. Para poder realizar cada uno de los pasos del algoritmo es necesario conocer los valores de $\hat{\sigma}^2$ y $\bar{k}$, la varianza estimada y el número de variables en el modelo, respectivamente.

El valor de $\bar{k}=5$, mientras que $\hat{\sigma}^2$ puede calcularse a partir de la matriz de diseño de acuerdo a

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-\bar{k}} = \frac{y'y - \hat{\beta}X'y}{15}$$

donde

$$\hat{\beta} = \begin{bmatrix} 10 \\ 0.4 \\ 0.6 \\ 1 \\ 2 \end{bmatrix}$$

obteniendo un valor igual a

$$\hat{\sigma}^2 = 4.7837$$

En el procedimiento propuesto por Hoerl y Kennard (1976), se debe calcular

$$\hat{\beta}_R = (X'X + KI)^{-1}X'y$$

donde

$$K_R = \frac{K\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$$

de aquí, obtenemos

$$K_R = 0.2267$$

esto es, el estimador ridge es resultado de

$$\hat{\beta}_R = (X'X + 0.2267I)^{-1}X'y$$

de donde resulta el valor

$$\hat{\beta}_R = \begin{bmatrix} 9.4524 \\ 0.2401 \\ 1.2692 \\ 0.6831 \\ 2.1733 \end{bmatrix}$$

esto es, una solución ridge inicial sería

$$y = 9.4524X_1 + 0.2401X_2 + 1.2692X_3 + 0.6831X_4 + 2.1733X_5 \qquad (5.6.1.1)$$

El procedimiento iterativo, como sabemos nos aconsejaría comenzar ahora con esta $\hat{\beta}_R$, la cual llamaremos $\hat{\beta}_{R(0)}$ de tal forma que el número $K_{R(1)}$ a insertar en la ecuación

$$\hat{\beta}_{R(1)} = (X'X + K_{R(1)}I)^{-1}X'y$$

está dado por $K_{R(1)} = \dfrac{5\hat{\sigma}^2}{\hat{\beta}_{R(0)}'\hat{\beta}_{R(0)}}$

y en general:

$$\hat{\beta}_{R(l)} = (X'X + K_{R(l)}I)^{-1} X'y \qquad (5.6.1.2)$$

$$K_{R(l)} = \frac{5\hat{\sigma}^2}{\hat{\beta}_{R(l-1)}'\hat{\beta}_{R(l-1)}} \qquad (5.6.1.3)$$

El procedimiento terminará hasta que

$$\frac{K_{R(l+1)} - K_{R(l)}}{K_{R(l)}} \leq 20T^{-1.3}$$

donde

$$T = \frac{\text{Traza}(X'X)^{-1}}{5} = \frac{9.9955}{5} = 1.9999$$

Esto es, nos detendremos cuando

$$\frac{K_{R(i+1)} - K_{R(i)}}{K_{R(i)}} \leq 8.1230$$

Con esta regla de término, los cálculos son llevados a cabo, obteniéndose los siguientes resultados

$$K_{R(1)} = \frac{5\hat{\sigma}^2}{\hat{\beta}_{R(0)}'\hat{\beta}_{R(0)}} = 0.2486$$

$$\beta_{R(1)} = (X'X + K_{R(1)}I)^{-1}X'y$$

$$= (9.3279, 0.2273, 1.2810, 0.6858, 2.2075)'$$

$$\frac{K_{R(1)} - K_{R(0)}}{K_{R(0)}} = \frac{0.2486 - 0.2267}{0.2267} = 0.0966 < 8.1230$$

Así, resulta muy conveniente tomar como solución ridge $\hat{\beta}_{R(1)}$ esto es

$$\hat{\beta}_{R(1)} = (9.3279, 0.2273, 1.2810, 0.6858, 2.2075)'$$

El modelo toma la forma

$$y = 9.3279X_1 + 0.2273X_2 + 1.2810X_3 + 0.6858X_4 + 2.2075X_5 \qquad (5.8.1.4)$$

En la tabla 5.4 se hace una comparación de los modelo ridge para $K_{(0)}$ y $K_{(1)}$ así como la solución propuesta originalmente. Las conclusiones que se obtienen de la tabla son las siguientes:

i) El coeficiente de X_1 ha sido reducido entre 5.5 y casi 7% respectivamente al comparar las soluciones ridge con la original

ii) Ningún coeficiente ha cambiado de signo

iii) Las variables X_3 y X_5 aumentan su participación, siendo el caso de la X_3 significativo ya que pasa de un coeficiente de 0.6 en el modelo original a un valor de 1.2810 para la solución ridge obtenida en la segunda iteración

iv) Las variables X_1, X_2, X_4 disminuyen significativamente su participación en el valor de y, al comparar los coeficientes del modelo normal con los de K_R=0.2267 de la tabla 5.4

Tabla 5.4 Comparación entre el modelo original y el modelo ridge calculado en dos valores de K_R

Término	(1) K_R = 0.2267	(2) K_R = 0.2486	(3) Modelo Normal
X_1	9.4524	9.3279	10.0
X_2	0.2401	0.2273	0.4
X_3	1.2692	1.2810	0.6
X_4	0.6631	0.6858	1.0
X_5	2.1733	2.2075	2.0
MSE	2.689	2.6311	4.7637
R^2	0.9950	0.9947	0.9912

Basados en los datos que arroja la tabla 5.4 y en las conclusiones alcanzadas se recomienda que se utilice el modelo

$$\hat{y} = 9.4524X_1 + 0.2401X_2 + 1.2692X_3 + 0.6831X_4 + 2.1733X_5$$

que fue producido al hacer uso de la regresión ridge con un valor K=0.2267. Debemos notar que este modelo no deja de ajustar adecuadamente los datos, para ello compárese el valor R^2=0.9912 de la regresión original con el valor R^2=0.9950 producido por el modelo que se obtiene por regresión ridge. La disminución del error cuadrático medio es sensible, ya que pasa de un valor de 4.7837 a uno de 2.689 en el modelo obtenido por el método conocido como regresión ridge.

Ahora nos toca investigar lo que pasaría con la solución originalmente planteada si utilizáramos el análisis de componentes principales. Como sabemos, los estimadores obtenidos por este método comparten con los estimadores ridge la propiedad de ya no ser insesgados. Resultará de mucha utilidad la descomposición en valores singulares de la matriz **X** hecha en 5.4 en el análisis posterior. Consideremos el modelo lineal en la forma canónica

$$y = Z\alpha + \varepsilon$$

donde

$$Z = XV,\ \alpha = V'\beta,\ V'X'XV = Z'Z = T$$

en la cual, **T=diag(λ_1, λ_2, ...,λ_5)** es una matriz diagonal de (5x5) cuyos elementos son los valores propios de la matriz **X'X**. **V** es una matriz ortogonal cuyas columnas son los vectores propios asociados con λ_1, λ_2, λ_3, λ_4 y λ_5 respectivamente

Las columnas de **Z**, que definen un nuevo conjunto de regresores ortogonales

$$Z = [\, Z_1, Z_2, Z_3, Z_4, Z_5 \,]$$

son los componentes principales. Realizando aparte la operación **XV** obtenemos:

$$Z = XV = \begin{bmatrix}
2.4908 & -0.4245 & 0.4842 & 0.0817 & 0.0588 \\
4.2847 & 0.5956 & 0.2520 & -0.0793 & -0.1451 \\
3.2376 & -0.7179 & 0.0691 & -0.2284 & -0.1530 \\
5.1228 & -0.1348 & -0.0052 & 0.2293 & -0.0353 \\
3.9027 & -1.3576 & 0.0031 & 0.3040 & 0.0115 \\
5.3447 & 0.9306 & 0.1992 & 0.1788 & -0.0548 \\
5.2133 & -0.3244 & -0.2254 & -0.0675 & 0.0034 \\
5.3431 & 0.5182 & -0.1458 & -0.2717 & 0.1514 \\
6.1210 & 0.1791 & -0.3294 & -0.0948 & 0.0197 \\
2.1785 & -0.3157 & 0.4546 & -0.1886 & 0.1034 \\
2.9509 & -0.6755 & 0.0802 & -0.3995 & -0.0656 \\
4.3367 & 0.3439 & 0.3314 & 0.2765 & 0.0407 \\
4.9925 & -0.0732 & -0.0227 & 0.0951 & -0.0305 \\
3.8596 & -1.3091 & -0.0116 & 0.2222 & -0.0099 \\
5.1143 & 0.9539 & 0.1955 & 0.0557 & 0.0244 \\
4.8954 & -0.1929 & -0.2625 & -0.3710 & 0.0266 \\
6.0193 & 0.5077 & -0.1533 & 0.0122 & -0.1290 \\
6.3120 & 0.0025 & -0.2782 & 0.2167 & 0.0838 \\
4.7131 & -0.0533 & -0.0246 & -0.0429 & 0.0725 \\
2.1572 & -0.2198 & 0.4243 & -0.3241 & 0.0325
\end{bmatrix}$$

El estimador de mínimos cuadrados de α es

$$\hat{\alpha} = (Z'Z)^{-1}Z'y = T^{-1}Z'y$$

con matriz de varianza

$$V(\hat{\alpha}) = \sigma^2(Z'Z)^{-1} = \sigma^2 T^{-1}$$

Como habíamos obtenido en el capítulo IV, un valor propio de **X'X** muy pequeño nos haría pensar que la varianza del correspondiente coeficiente de regresión sera grande. Además

$$Z'Z = \sum_{i=1}^{5}\sum_{i=1}^{5} Z_i Z_j' = T$$

por lo que nos referimos al valor propio λ_j como la varianza de la j-ésima componente principal. Habíamos obtenido también que

$\lambda_1 = 423.3963$

$\lambda_2 = 7.8327$

$\lambda_3 = 1.2377$

$\lambda_4 = 0.9547$

$\lambda_5 = 0.1249$

El enfoque de regresión en componentes principales ataca la multicolinealidad usando un subconjunto del conjunto de componentes principales en el modelo. Los valores propios han sido ordenados de tal forma que

$$\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5$$

y de estos cinco, los valores propios más cercanos al cero son λ_4 y λ_5, al análisis de componentes principales sugiere remover estos valores y aplicar mínimos cuadrados a las componentes restantes. De esta forma obtenemos

$$\hat{\alpha}_{PC} = b\hat{\alpha}$$

donde $b_1=b_2=b_3=1$ y $b_4=b_5=0$. El estimador de componentes principales es:

$$\hat{\alpha}_{PC}^{(0)} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \hat{\alpha}_3 \\ 0 \\ 0 \end{bmatrix}$$

o en términos de las regresoras

$$\hat{\beta}_{PC}^{(0)} = V\hat{\alpha}_{PC}^{(0)} = \sum_{j=1}^{3} \lambda_j v'_j X' y \, v_j$$

Para obtener dicho resultado, debemos comenzar con la transformación lineal $Z=XV$ que transforma las regresoras originales en un conjunto ortogonal de variables (las componentes principales). La matriz V ya había sido calculada y es:

$$V = \begin{bmatrix} 0.2092 & -0.2257 & 0.8530 & -0.4159 & 0.0688 \\ 0.3607 & 0.5151 & -0.0529 & -0.0789 & 0.7717 \\ 0.2767 & -0.6333 & -0.4903 & -0.4879 & 0.2099 \\ 0.4122 & 0.4863 & -0.1816 & -0.4835 & -0.5777 \\ 0.7614 & -0.2151 & 0.0563 & 0.5907 & -0.1480 \end{bmatrix}$$

mientras que

$$\hat{\alpha} = (Z'Z)^{-1}Z'y$$

$$= (4.35237, -3.04320, 8.82418, -4.44596, 0.05604)'$$

por lo que

$$\hat{\alpha}_{PC}^{(0)} = (4.35237, -3.04320, 8.82418, 0, 0)'$$

obteniendo finalmente

$$\hat{\beta}_{PC}^{(0)} = V\hat{\alpha}_{PC}^{(0)}$$

$$= (9.12439, -0.4644, -1.1949, -1.1118, 4.4652)'$$

Es interesante analizar también el caso en el que únicamente se remueve la quinta componente principal, en este caso nuestros estimadores tendrán ahora las siguientes expresiones

$$\hat{\alpha}_{PC}^{(1)} = (4.35237, -3.04320, 8.82418, -4.44586, 0)'$$

con su respectiva

$$\hat{\beta}_{PC}^{(1)} = V\hat{\alpha}_{PC}$$

$$= (10.9734, -0.1136, 0.9742, 1.0377, 1.8390)'$$

Esto es, en el primer caso hemos considerado el modelo de regresión lineal

$$y = \alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Z_3 + \varepsilon$$

mientras que en el segundo (considerando las primeras cuatro componentes principales)

$$y = \alpha_1 Z_1 + \alpha_2 Z_2 + \alpha_3 Z_3 + \alpha_4 Z_4 + \varepsilon$$

(La tabla 5.5 muestra los coeficientes de regresión resultantes no estandarizados), así como los coeficientes de regresión en términos de las regresoras originales. En la misma tabla mostramos los resultados de añadir la cuarta y la quinta componentes principal al modelo.

Tabla 5.5 Regresión en componentes principales para los datos de la tabla 5.1

Parámetro	A Estimado original Z_1, Z_2, Z_3 $\hat{\beta}_{PC}^{(0)}$	B Estimado original Z_1, Z_2, Z_3, Z_4 $\hat{\beta}_{PC}^{(1)}$	C Estimado original Z_1, Z_2, Z_3, Z_4, Z_5 $\hat{\beta}$
β_1	9.12439	10.0734	10.0
β_2	-0.4644	-0.1136	0.4
β_3	-1.1949	0.9742	0.6
β_4	-1.1118	1.0377	1.0
β_5	4.4652	1.8390	2.0
R^2	0.9954	0.9977	0.9912
MSE	2.4691	1.2176	4.7837

Los modelos ajustados son

$$\hat{y} = 9.12439\, X_1 - 0.4644\, X_2 - 1.1949\, X_3 - 1.1118\, X_4 + 4.4652\, X_5$$

en el caso de las tres primeras componentes, y

$$\hat{y} = 10.9734\, X_1 - 0.1136\, X_2 + 0.9742\, X_3 + 1.0377\, X_4 + 1.8390\, X_5$$

en el caso de la regresión sobre las cuatro primeras componentes principales.

Es interesante notar que el hecho de utilizar un número distinto de componentes principales (3, 4 y 5 en este caso) tiene como efecto producir valores de los estimadores de los coeficientes de la regresión muy distintos. Los estimadores obtenidos por regresión en componentes principales difieren sensiblemente de aquellos de mínimos cuadrados (véase la tercera columna de la tabla 5.5). La diferencia entre dichos estimadores se suaviza cuando añadimos componentes adicionales, en este caso es ilustrativo realizar la comparación entre $\hat{\beta}_{PC}^{(0)}$, $\hat{\beta}_{PC}^{(1)}$ y $\hat{\beta}$. Mientras que $\hat{\beta}$ no tiene ninguna componente negativa, $\hat{\beta}_{PC}^{(0)}$ tiene 3, las correspondientes a X_2, X_3 y X_4; $\hat{\beta}_{PC}^{(1)}$ mientras tanto, sólo tiene 1, que corresponde al coeficiente de X_2.

La regresión en componentes principales reduce en ambos casos la participación de X_2; el impacto de las demás variables - excepto X_1, que mantiene un valor cercano al 10 - es ambigüo. Mientras que cuando se consideran tres componentes principales los coeficientes de X_2, X_3, X_4 son reducidos en valor absoluto, el coeficiente de la variable X_5 se dispara de un valor de 2.0 en la regresión original a uno de 4.4652. Cuando incorporamos una

componente adicional, solamente un coeficiente permanece con signo negativo, el correspondiente a la variable X_2, y en términos de valores absolutos, representa casi una cuarta parte de su valor original; en este mismo caso, el coeficiente de X_5 ha sido ligeramente encogido, aunque la disminución de su contribución para explicar y es compensada por los ligeros incrementos en los coeficientes de X_4 y X_1.

En este problema, se recomienda como una posible solución la regresión en cuatro componentes principales, esto es utilizar, el modelo

$$\hat{y} = 10.9734\,X_1 - 0.1136\,X_2 + 0.9742\,X_3 + 1.0377\,X_4 + 1.3390\,X_5$$

La justificación del uso del modelo anterior surge del hecho de que el error cuadrático medio ha sido reducido dramáticamente de un valor de 4.7637 a uno de 1.2178, por otra parte,la regresión en cuatro componentes principales no parece haber degradado en lo absoluto el ajuste con los datos originales, ya que de hecho, el valor de R^2 ha subido de 0.9912 en el modelo original a 0.9977 en el caso de cuatro componentes principales.

Podemos concluir con base en estos datos, que la regresión en cuatro componentes principales provee un modelo veraz y plausible para los datos propuestos por Judge, et. al. La regresión en componentes principales produjo un modelo que aparentemente ajusta de una mejor manera los datos que el producido por la regresión ridge. De hecho, si denotamos por $\widehat{y}_R$ el modelo obtenido por regresión ridge y por $\widehat{y}_{PC}$ aquel obtenido por medio de componentes principales, resulta que

$$R^2_{\widehat{y}_R} = 0.950 < R^2_{\widehat{y}_{PC}} = 0.9974$$

y respecto al error cuadrático medio, tenemos que

$$MSE_{\widehat{y}_R} = 2.689 > MSE_{\widehat{y}_{PC}} = 1.2178$$

esto es, el error cuadrático medio que arroja el modelo obtenido por componentes pricipales es mucho menor que el que arroja aquel obtenido por regresión ridge.

RESUMEN DEL APÉNDICE

Este apéndice presenta las operaciones matriciales fundamentales que se utilizan a lo largo de ésta tesis. En particular, se enfatiza el uso de:

i) Los conceptos de rango y transpuesta de una matriz.

ii) Los distintos tipos de matrices: cuadrada, simétrica, diagonal, idéntica e idempotente.

iii) Las descomposiciones matriciales en valores singulares y su aplicacion en el análisis de componentes principales.

APÉNDICE

Aquí se presentan los resultados acerca de matrices y operadores lineales que proporcionan la base técnica que está detrás de los métodos propuestos en los capítulos anteriores de ésta tesis.

Es conveniente incluir a su vez, a las operaciones matriciales que desempeñan una labor fundamental en el análisis de regresion, por ello se da una breve exposicion de las operaciones matriciales fundamentales.

A.1 Una matriz es un arreglo rectangular de elementos previamente ordenados dispuestos en renglones y columnas, los elementos son llamados escalares. Las matrices se denotarán con letras mayúsculas en negritas. Los números que forman la matriz serán llamados los elementos de la matriz a_{ij} donde j se refiere a la columna e i se refiere al renglon. Una matriz en se denotará general como:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Donde los subíndices indican el renglon y la columna respectivamente.

A.2 El orden de una matriz es el par ordenado (k,l) donde k es el número de renglones y l el número de columnas.

A.3 El rango de una matriz se define como el número de columnas (o renglones) linealmente independientes en la matriz. Cualquier subconjunto de columnas de una matriz es linealmente independiente si ninguna columna de este subconjunto puede ser expresada como una combinacion lineal de las otras.

A.4 Si no existen dependencias lineales entre las columnas de una matriz, se dice que la matriz es de rango completo, o no singular.

A.5 Un vector es una matriz que tiene un solo renglon o una columna, y es llamado vector renglon o vector columna, respectivamente.

A.6 Una matriz cuadrada tiene igual número de renglones que de columnas.

A.7 Una matriz diagonal es una matriz cuadrada en la cual todos los elementos son cero excepto los elementos de la diagonal. Esto es, A es diagonal si $a_{ij}=0$ para $i\neq j$ y $a_{ij}\neq 0$ para al menos una $i=j$

A.8 Una matriz identidad es una matriz diagonal que tiene todos los elementos de la diagonal iguales a 1; se denota como $\mathbf{I}_N$. Donde N es el número de renglones de la matriz.

Operaciones matriciales

A.9 La transpuesta de una matriz **A**, denotada por **A'**, es la matriz obtenida usando los renglones de **A** como columnas de **A'**.

A.10 Una matriz simétrica es una matriz cuadrada (kxk) en la cual el elemento $a_{ij}=a_{ji}\ \forall\ i, j$ (i,j=1,2,...,k).

A.11 La suma de dos matrices está definida si y solo si las matrices son del mismo orden. Esta matriz se obtiene al sumar los elementos correspondientes de cada una de las matrices.

A.12 La multiplicacion de dos matrices está definida si y solo si el número de columnas en la primera matriz es igual al número de renglones en la segunda matriz. Así, si $\mathbf{A}=[\ a_{ij}\ ]$ y $\mathbf{B}=[\ b_{ij}\ ]$

$$\mathbf{AB}=[\ \sum_{k=1}^{m} a_{ki}\ b_{jk}\ ] = [\ c_{ij}\ ]$$

donde $\mathbf{A}_{(mxn)}$ y $\mathbf{B}_{(nxp)}$.

A.13 La transpuesta de un producto es el producto en el orden inverso de las transpuestas de las dos matrices. Esto es:

$$(\mathbf{AB})' = \mathbf{B'A'}$$

siempre que las dimensiones permitan que se lleve a cabo el producto.

A.14 La multiplicacion escalar es la multiplicacion de una matriz por un número real.

A.15 Dos vectores V_1 y V_2 del mismo orden son ortogonales si el producto punto de vectores

$$V_1'V_2 = 0$$

A.16 Una matriz cuadrada se llama idempotente si permanece sin cambios cuando es multiplicada por si misma. Esto es, la matriz **A** es idempotente si

$$AA = A$$

La suma de elementos en la diagonal de una matriz cuadrada se llama la traza de la matriz y se denota por tr(**A**).

A.17 El determinante de una matriz es un escalar calculado a partir de los elementos de la matriz de acuerdo a reglas bien definidas. Los determinantes están definidos solo para matrices cuadradas y se denotan por $|A|$, donde **A** es una matriz cuadrada. Si el determinante de una matriz es cero, la matriz es singular, esto es equivalente a que la matriz no sea de rango completo.

A.18 La inversa de una matriz **A**, denotada por A^{-1}, se define como la matriz que produce a la matriz identidad cuando es multiplicada por **A**.

Esto es

$$A^{-1}A = AA^{-1} = I$$

La inversa de una matriz puede no existir. Una matriz tiene inversa única si y solo si la matriz es cuadrada y no singular.

Transformaciones ortogonales y proyecciones

La transformacion lineal del vector **x** que da como resultado el vector **y**, ambos de orden n, se escribe como **y=Ax**, donde **A** es una matriz de coeficientes reales de (nxn), dichos coeficientes efectúan la transformacion. La transformacion es una transformacion uno a uno solo si **A** es no singular.

Entonces, la transformacion inversa de **y** a **x** es $\mathbf{x=A^{-1}y}$

A.19 Una transformacion lineal **y=Ax**, $\mathbf{x \in V}$, **V** un espacio vectorial dimensionalmente finito es una transformacion ortogonal si

$$\mathbf{A'A = I}$$

Esta condicion implica que los vectores renglon de **A** son ortogonales y de longitud unitaria. Una transformacion ortogonal mantiene invariantes las distancias y ángulos entre vectores.

A.20 Una proyeccion de un vector en un subespacio es un caso especial de una transformacion. El objetivo de una proyeccion es transformar y en un espacio n-dimensional a un vector $\hat{y}$ en un subespacio, tal que $\hat{y}$ sea tan cercano a y como sea posible. Una transformacion de y a $\hat{y}$, $\hat{y} = \mathbf{Py}$, es una proyeccion si y solo si $\hat{P}$ es

idempotente y simétrica, en cuyo caso **P** se conoce como una matriz de proyeccion.

Valores y vectores propios

Los valores propios y vectores propios de matrices son necesarios para algunos de los métodos discutidos en el presente trabajo, incluyendo el análisis de componentes principales y el diagnostico de multicolinealidad . La determinacion de los valores y vectores propios de una matriz es un problema computacional difícil. Nos vemos limitados a limitar la discusion del análisis de los valores y vectores propios a matrices reales, simétricas y definidas no negativas.

A.21 Puede mostrarse que para una matriz real, simétrica **A** de oreden (nxn) existe un conjunto de escalares no negativos, λ_i^2, y n vectores distintos de cero, $\mathbf{z}_i$, i=1,2,...,n, tales que

$$\mathbf{A}\mathbf{z}_i = \lambda_i^2 \mathbf{z}_i$$

o $$\mathbf{A}\mathbf{z}_i - \lambda_i^2 \mathbf{z}_i = 0$$

o $$(\mathbf{A} - \lambda_i^2 \mathbf{I})\,\mathbf{z}_i = 0, \qquad i=1,2,...,n \qquad (\mathbf{A.21.2})$$

Los valores λ_i^2 son los valores propios (raíces características) de la matriz **A** y las $\mathbf{z}_i$ son los vectores propios correspondientes

(vectores característicos). Las raíces cuadradas positivas de los valores propios son llamadas los valores singulares y serán denotadas por λ_i.

Existen soluciones a la ecuacion (**A.21.2**) solo si la matriz ($\mathbf{A} - \lambda_i^2 \mathbf{I}$) no es de rango completo, esto es, solo si $|(\mathbf{A} - \lambda_i^2 \mathbf{I})| = 0$.

Las λ_i^2 se obtienen resolviendo la ecuacion:

$$|(\mathbf{A} - \lambda_i^2 \mathbf{I})| = 0$$

Esta ecuacion proporciona un polinomio de grado n en λ^2, y al resolverlo, obtenemos los n valores de λ^2 los cuales no son necesariamente distintos.

Por convencion, cada vector propio se define como el vector solucion cambiando la escala para tener longitud unitaria, esto es $z_i'z_j = 0$, cuando $i \neq j$. Así pues, si $\mathbf{Z} = (z_1, z_2, \ldots, z_n)$ es la matriz de vectores propios, entonces $\mathbf{Z'Z} = \mathbf{I}$. Esto implica que $\mathbf{Z'}$ es la inversa de $\mathbf{Z}$.

A.22. Usando **Z** y **L**,con Z definida como en **A.21** y **L** definida como la matriz diagonal de las λ_i^2 también de **A.21** podemos escribir las ecuaciones iniciales $\mathbf{Az_i} = \lambda_i^2 \mathbf{z_i}$ como

$$\mathbf{AZ} = \mathbf{ZL}$$

o $$\mathbf{Z'AZ} = \mathbf{Z'ZL} = \mathbf{IL} = \mathbf{L}$$

o $$\mathbf{A} = \mathbf{ZLZ'} \qquad \text{(A.22.a)}$$

(**A.22.a**) muestra que **A** puede transformarse en una matriz diagonal multiplicando a **L** por **Z'** a la derecha y por **Z** a la izquierda. Como **L** es una matriz diagonal, entonces (**A.22.a**) muestra que **A** puede expresarse como la suma de matrices

$$\mathbf{A} = \mathbf{ZLZ'} = \Sigma \lambda_i^2 (\mathbf{z_i z_i'})$$

Descomposicion en valores singulares de una matriz rectangular

El análisis de valores y vectores propios anterior se aplica a las matrices cuadradas. Utilizaremos ahora los resultados acerca de los valores y vectores propios para desarrollar una descomposicion similar, llamada descomposicion en valores singulares para una matriz rectangular. La descomposicion en valores singulares se usará para producir el análisis de componentes principales.

A.23 Sea **X** una matriz nxp con n>p. Entonces **X'X** es una matriz cuadrada, simétrica de orden pxp. De **A.22**, **X'X** puede expresarse en términos de sus valores propios **L** y vectores propios **Z** como

$$\mathbf{X'X = Z L Z'}$$

Similarmente, **XX'** es una matriz cuadrada simétrica pero de orden nxn. El rango de **XX'** será de a lo más p, asi que habrá a lo más p valores propios distintos de cero, los cuales son los mismos que los de **X'X**. Además **X'X** tendrá (n-p) valores propios que son cero. Denotemos por U la matriz de vectores propios de **XX'** que corresponde a los p valores propios comúnes a **X'X**. Cada vector propio u_i será de orden (nx1). Entonces

$$\mathbf{X'X = U L U'}$$

De forma que

$$\mathbf{X = U L^{1/2} Z'}$$

donde $\mathbf{L^{1/2}}$ es la matriz diagonal cuyos elementos son las raíces cuadradas positivas de los p valores propios de **X'X**. Esto es, $\mathbf{L^{1/2}L^{1/2} = L}$. La ecuacion (**A.23.a**) proporciona la descomposicon en valores singulares de la matriz rectangular **X**. Los elementos de los $\mathbf{L^{1/2}}$, λ_i, son llamados los valores singulares y los vectores columna en **u** y **z** son los vectores singulares derechos e izquierdos respectivamente.

Análisis de componentes principales

A.24 La descomposicion en valores singulares es el primer paso en el análisis de componentes principales. Usando el resultado $\mathbf{X=U\,L^{1/2}\,Z'}$ y la propiedad $\mathbf{Z'Z=I}$, puede definirse la matriz **W**(nxp) como

$$\mathbf{W = XZ = U\,L^{1/2}}$$

La primera columna de **Z** es es primero de los vectores singulares derechos de **X**, o el primer vector propio de **X'X**. Así, los coeficientes en el primer vector propio definen una funcion lineal de las columnas de **X** (las variables originales) que genera a la primera columna de **W**. La segunda columna de **W** es obtenida usando el segundo vector propio de **X'X** y así en adelante. Note que **W'W=L**. Esto es, **W** es una matriz (nxp) con la propiedad de que todas sus columnas son ortogonales. La suma de cuadrados de la i-ésima columna de **W** es λ_i^2, el i-ésimo elemento de la diagonal de **L**.

Esto es, si **X** es una matriz (nxp) de observaciones en las p-variables, cada columna de **W** es una nueva variable definida como una transformacion lineal de las variables originales. Este análisis

se llama análisis de componentes principales de **X**, y las columnas de **W** son las componentes principales.

Preparacion de los datos. Matrices centradas. Matrices que son cambidas de escala

A.25 Sea **X** una matriz de orden nxk, donde n se refiere al número de observaciones y k al número de variables. El vector columna de medias de **X**, denotado por $\bar{\mathbf{x}}$, se conoce como centroide: $\bar{\mathbf{x}}$ se calcula por la siguiente operacion matricial

$$\bar{\mathbf{x}}' = (\,^1/_n\,)\,\mathbf{1}'\,\mathbf{X}$$

donde **1'** denota un vector (**1xn**) cuyos elementos son todos iguales a uno

A.26 Los elementos corregidos por la media pueden ser obtenidos una vez que se ha hallado $\bar{\mathbf{x}}$. Denotando por $\mathbf{X_d}$ la matriz de orden nxk de elementos corregidos por la media, tenemos

$$\mathbf{X_d} = \mathbf{X} - \mathbf{1}\,\bar{\mathbf{x}}'$$

A.27 Si denotamos por $\mathbf{X_d}$ un vector columna de elementos corregidos por la media, podemos calcular la varianza σ^2, que es un estimador de la varianza poblacional σ^2 por medio de

$$\sigma^2 = \frac{1}{n-1} X_d' X_d$$

Una vez que hemos obtenido las varianzas de X_1, X_2, ..., X_k y Y podemos colocarlos en una matriz diagonal

$$D = \begin{bmatrix} s_1^2 & 0 & \dots & 0 & 0 \\ 0 & s_2^2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & s_k^2 & \\ 0 & 0 & \dots & 0 & s_y^2 \end{bmatrix}$$

y los datos estandarizados pueden obtenerse fácilmente por

$$X_s = X_d D^{-1/2}$$

donde $D^{-1/2}$ es una matriz diagonal cuyo i-ésimo elemento en la diagonal es igual a la raíz cuadrada del i-ésimo elemento en la diagonal $(S_j)^{-1}$.

A.28 En ocasiones es conveniente expresar las sumas de cuadrados y productos cruzados en términos de los elementos corregidos por la media. La matriz de sumas de cuadrados y productos cruzados corregidos por la media se denota por **S**, y está dada por

$$S = X'X - (\,^1/_n\,)(\,X'1\,)(\,1'X\,) \quad \text{o}$$

$$S = X_d' X_d$$

A.29 Una vez que la matriz de suma de cuadrados y productos cruzados corregidos por la media **S** ha sido hallada, es fácil obtener la matriz de varianza-covarianza, denotada por **C**. La matriz de varianza-covarianza se obtiene tomando cada elemento de la matriz **S** y dividiendo por el escalar (n-1), el tamaño de la muestra menos uno:

$$\mathbf{C} = \frac{1}{(n-1)}\,\mathbf{S}$$

A.30 La matriz de correlacion, denotado por **R**, puede encontrarse, a partir de la matriz **S** o de la matriz de varianza-covarianza **C**. Denotando por $\mathbf{D}^{-1/2}$ la matriz cuyas entradas en la diagonal principal son los recíprocos de las desviaciones estándar de las variables en **X**, la matriz de correlacion **R**, puede obtenerse multiplicando **S** a la izquierda y a la derecha por la matriz $\mathbf{D}^{-1/2}$:

$$\mathbf{R} = \frac{1}{(n-1)}\left(\mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}\right)$$

BIBLIOGRAFÍA

Belsley, D., Kuh E., and Welsh R. E. (1980). *Regression diagnostics.* John Wiley and sons. New York.

Bring, John. (1994). *How to standardize regression coefficients.* The American Statistician, 48, 209-213.

Dillon, William R., and Goldstein, Matthew. (1984). *Multivariate analysis. Methods and applications.* John Wiley and sons. New York, pp. 4-22.

Dudewicz J. Edward, Mishra N. Satya. (1988). *Modern Mathematical Statistics.* John Wiley and sons. New York.

Farrar, D. E., and Glauber R. R. (1967). *Multicollinearity in regression analysis: The problem revisited.* Review of Economics and Statistics, 49, 92-107.

Gunst F. Richard. (1983). *Regression analysis with multicollinear predictor variables: definition, detection and effects.* Communications in Statistics. 12(19), 2217-2260.

Hemmerle, W. Y. (1975). *An Explicit Solution for Generalized Ridge Regression*, Technometrics, 17, 309-314.

Hocking R. R., Pendleton O. J. (1983). *The regression dilemma.* Communications in Statistics. 12(5), 497-527.

Hocking, R. R., F. H. Speed, and H. J. Lynn (1976). *A Class of Brased Estimators in linear regression,* Technometrics, 18, 425 - 437.

Hoerl, A. E., R. W. Kennard and K. F. Baldwin (1975). *Ridge Regression: Some Simulations*, Communications in Statistics, A, 4, 105-123.

Hoerl, A. E. and R. W. Kennard (1970a). *Ridge Regression: Brased Estimation of Nonortthogonal Problems*, Technometrics; 12, 55-67.

Judge G. George, Griffiths E. W., Hill R. Carter, et. al. (1985). *The theory and practice of Econometrics.* Second edition. John Wiley and sons. New York, pp. 896-936.

Kumar T. Krishna. (1975). *Multicollinearity in regression analysis. Review of Economics and Statistics*, 57, 365-366.

Marquardt, D. W. and R. D. Snee (1975). *Ridge Regression in Practice,*" American Statiscian, 29, 3-20.

Mason, R. L., R. F. Gunst, and J. T. Webster (1975). *Regression Analysis and Problems of Multicollinearity*, Communications in Statistics.

McDonald, G. C. and D. Y. Galarneau (1975). *A Monte Carlo Evaluation of Some Ridge-Type Estimators*, Journal of the American Statiscal Association, 70, 407-416.

Montgomery C. Douglas and Peck Elizabeth A. (1982). *Introduction to Linear Regression Analisys*. John Wiley and sons. New York.

O'Hagan John and McCabe Brendan. (1974). *Tests for the severity of multicollinearity in regression analysis: a comment.* Review of Economics and Statistics, 57, 368-370.

Silvey D. S., (1967). *Multicollinearity and imprecise estimation.* J. R. Statistical Society, 13, 31, 539-552.

Smith, G. (1974). *Multicollinearity and forecasting.* Cowles Foundation Discussion Paper No. 33.

Smith, G. and F. Campbell (1980). *A critique of some Ridge Regression Methods,* Journal of the American Statistical Association, 75, 74-103.

Teekens, R. (1977) *The Exact MSE-Efficiency of the General Ridge Estimator Relative to OLS,* presented at the summer 1977 meeting of the Econometric Society.

Thisted, R. (1978a). *Multicollinearity, Information and Ridge Regression,* Technical Report No. 66, Department of Statistics, University of Chicago.