Universidad Nacional Autónoma de México

Facultad de Ciencias

ANÁLISIS DE REGRESIÓN SESGADA

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A :

ROSA EDITH MOLINA PADRÓN

DIRECTOR DE TESIS:
MAT. MARGARITA ELVIRA CHÁVEZ CANO

2011

1. Nombre del alumno
   Molina
   Padrón
   Rosa Edith
   56179811
   Universidad Nacional Autónoma de México
   Facultad de Ciencias
   Actuaría
   302227891

2. Datos del tutor
   Mat.
   Margarita Elvira
   Chávez
   Cano

3. Datos del sinodal 1
   Act.
   Jaime
   Vázquez
   Alamilla

4. Datos del sinodal 2
   M. en A.P.
   María del Pilar
   Alonso
   Reyes

5. Datos del sinodal 3
   Dra.
   Ruth Selene
   Fuentes
   García

6. Datos del sinodal 4
   Act.
   Francisco
   Sánchez
   Villarreal

7. Datos del trabajo escrito
   Regresión sesgada
   120 p.
   2011

Agradecimientos

En primer lugar quiero agradecer a Dios por bendecirme para llegar hasta donde he llegado, por darme la fortaleza necesaria para seguir superándome y la perseverancia para alcanzar mis más grandes sueños y anhelos.

A mis padres, por ser ellos quienes a lo largo de toda mi vida han contribuido a la realización de mis logros; con orgullo les comparto este trabajo, ustedes más que nadie saben el esfuerzo, tiempo y dedicación que hay detrás de cada línea. Soy afortunada por contar siempre con su amor, comprensión y ejemplo. Los quiero mucho.

A mi hermano, porque siempre he contado con él en todo momento, gracias por la confianza que siempre nos hemos tenido, por tu apoyo y amistad, porque la vida nos a dado la oportunidad de estar juntos y disfrutar todos y cada uno de nuestros logros.
¡Te quiero!

A Luis Daniel, por compartir sus experiencias y consejos, por ser un gran apoyo y compañía, pero sobre todo por motivarme a hacer las cosas de la mejor manera. Gracias por hacerme feliz y por quererme tanto. Te amo.

A una persona que más que mi primo ha sabido ser mi amigo, mi compañero de juegos y travesuras desde pequeño, por lo que lo hago participe de este gran logro en mi vida. Gracias Félix por tu apoyo.

A mis amigos, porque de alguna manera todos ellos han contribuido en la realización de este trabajo, porque siempre me han brindado su apoyo, tanto personal como profesional y siempre he sentido su cercanía y afecto.

A mis profesores de la carrera por sus enseñanzas, principalmente a la profesora Margarita a quien le agradezco el haber aceptado la dirección de esta tesis y el apoyo mostrado durante la realización de la misma. Gracias por haberme brindado su amistad, paciencia, orientación y estímulo.

Al Act. Jaime Vázquez Alamilla, a la M. en A. P. María del Pilar Alonso Reyes, a la Dra. Ruth Selene Fuentes García y al Act. Franciso Sánchez Villarreal. Gracias por su tiempo y dedicación en la revisión de mi trabajo, así como por sus imprescindibles comentarios para mejorarlo.

A la Universidad Nacional Autónoma de México, porque en sus aulas obtuve los conocimientos que me forjaron como profesionista y porque fuera de ellas me enseñó los valores de la amistad, el respeto y la tolerancia.

Índice general

Índice de figuras

Índice de tablas

Introducción

El análisis de regresión es una de las técnicas estadísticas más usadas en el análisis de datos y en el desarrollo de modelos empíricos, su objetivo es investigar la relación estadística que existe entre una variable dependiente (Y) y una o más variables independientes $(X_1, X_2, X_3, \dots)$, sin embargo tradicionalmente se emplea la técnica de mínimos cuadrados ordinarios para postular una relación funcional entre las variables, la cual enfrenta problemas cuando las variables independientes presentan colinealidad, por lo que el objetivo de este trabajo es introducir las diversas clases de estimadores sesgados que ayuden a manejar dicho problema.

En particular, el análisis de regresión ridge es un procedimiento que se encuentra dentro del grupo de las regresiones sesgadas consideradas como no lineales. Su nombre se debe a los trabajos desarrollados por Hoerl y Kennard en el año de 1970, quienes muestran la existencia de constantes positivas k_i, las cuales siempre reducen el error cuadrático medio del estimador de mínimos cuadrados ordinarios y además proporcionan una regla específica para elegir cada k_i, garantizando la reducción del error cuadrático medio. El análisis de regresión ridge constituye toda una alternativa a la estimación por mínimos cuadrados y además proporciona una evidencia gráfica de los efectos de la colinealidad en la estimación de los coeficientes de regresión.

Son numerosas las aplicaciones de la regresión y se encuentran en casi cualquier campo, por ejemplo, en la investigación social, donde se utiliza para predecir un amplio rango de fenómenos, desde medidas económicas hasta diferentes aspectos del comportamiento humano; en el contexto de la investigación de mercados, puede utilizarse para determinar en cual de los diferentes medios de comunicación puede resultar más eficaz invertir, o para predecir el número de ventas de un determinado producto; en física se utiliza para caracterizar la relación entre variables o para calibrar medidas; etc. Por lo anterior, otro de los objetivos de este trabajo es completar la teoría presentada con el análisis de diversos ejemplos, para que de esta manera el lector cuente con información suficiente para realizar en determinado momento un análisis adecuado de los datos a estudiar.

También es propósito que esta investigación sirva como material de consulta para profesores y/o estudiantes de las carreras de actuaría, matemáticas o alguna carrera afín, para que puedan realizar trabajos y/o prácticas con fluidez y sobre todo entender los métodos empleados para el análisis de datos, así como la utilización del software que se maneja.

Este trabajo está organizado en 5 capítulos, que han sido pensados para llevar una secuencia que permita poder comprender mejor todos los elementos que lo constituyen. Es necesario tener el conocimiento de algunos conceptos de álgebra matricial para poder expresar y manejar información de muchas variables, por lo que en la sección de apéndices se presentan los elementos de álgebra necesarios para la comprensión y aplicación tanto del análisis multivariado como del análisis de regresión ridge. En el primer capítulo se describe el problema de colinealidad en la regresión lineal, así como los procedimientos comúnmente empleados para manejarla o eliminarla. En el segundo capítulo se estudian los estimadores ridge y los estimadores reducidos, los cuales pertenecen a la clase de estimadores sesgados, así como también se proporciona un ejemplo para mostrar su comportamiento. En el tercer capítulo se explica con detalle la validación y contrucción del modelo de regresión ridge y de su traza, la cual muestra en dos dimensiones los efectos de la no ortogonalidad. El capítulo cuarto está dedicado a entender mejor los conceptos de la regresión ridge, así como la relación que presenta con la regresión de la inversa generalizada, mediante el análisis de varios ejemplos. En el quinto capítulo se expone la aplicación del método propuesto, presentando cada uno de los pasos necesarios que fundamentan el análisis. Por último se presentan las conclusiones que han sido obtenidas tanto de la teoría propuesta, como de sus respectivas aplicaciones con conjuntos de datos reales, donde se comenta la utilidad del método propuesto.

Por otra parte, la aplicación práctica de los métodos de regresión requieren la manipulación de muchos datos, a veces en gran cantidad, así como el cálculo de algunas fórmulas matriciales. Para ello en este trabajo se ha optado por incluir algunos ejemplos utilizando el lenguaje **R**. Las principales razones por las que se utilizó, es porque **R** es una implementación libre e independiente del lenguaje de programación **S**, es un conjunto de programas integrados para manejo de datos, simulaciones, cálculos y realización de gráficas, además es un lenguaje de programación orientado a objetos, tiene algunos módulos específicos para los modelos lineales y es programable, también **R** utiliza un lenguaje que al principio puede resultar un tanto difícil de aprender, sin embargo superada la primera etapa de adaptación, su utilización abre todo un mundo de posibilidades, no sólo en los modelos lineales, sino en todo el cálculo estadístico. La sintaxis utilizada en este lenguaje será presentada en el anexo de ápendices para su consulta.

Capítulo 1

Colinealidad

Si no existe una relación lineal entre las variables regresoras, se dice que éstas son ortogonales. Cuando las regresoras son ortogonales, se pueden obtener con relativa facilidad inferencias como: identificación de los efectos de las variables regresoras, predicción y/o estimación y selección de un conjunto adecuado de variables para el modelo, desafortunadamente en la mayoría de las aplicaciones de regresión, las regresoras no son ortogonales.

El analista debe estar consciente que un grave problema que puede impactar seriamente la utilidad e interpretación de los resultados del modelo de regresión es la colinealidad, que se presenta cuando existe casi singularidad entre las columnas de la matriz $X'X$, es decir, determinadas combinaciones lineales de las variables regresoras son casi cero, estas variables son las columnas de la matriz X, así que es claro que una dependencia lineal exacta resultaría en una matriz $X'X$ singular. La presencia de dependencias casi lineales puede influir fuertemente en la precisión con la que se estiman los coeficientes de regresión, e implica que exista redundancia entre las variables independientes, esencialmente que la misma información se esté presentando en varias formas.

Ahora se analizará el efecto que tiene el diferente grado de correlación entre las variables independientes del modelo sobre la estimación por mínimos cuadrados ordinarios. Se distinguen tres casos distintos:

1. **Ausencia de Colinealidad:** Éste sería el caso ideal en cuanto a la interpretación de los coeficientes, en el sentido de que la estimación de cada uno de los parámetros se encuentra libre de las interacciones entre las variables regresoras. Desde el punto de vista práctico no resulta ser un caso muy interesante, ya que no es habitual la ausencia de correlación entre las variables explicativas del modelo.

2. **Colinealidad Perfecta:** Si al menos dos regresoras están perfectamente relacionadas, entonces los coeficientes de regresión por mínimos cuadrados no están definidos y no es posible realizar la estimación de los parámetros.

3. **Colinealidad Aproximada:** Si las variables independientes están correlacionadas, entonces, los valores estimados de los coeficientes de regresión estarán sesgados y serán inestables, de modo que se pueden encontrar anomalías tan graves como un signo contrario al que realmente debería tener.

Correlación entre X_1 y X_2 despreciable

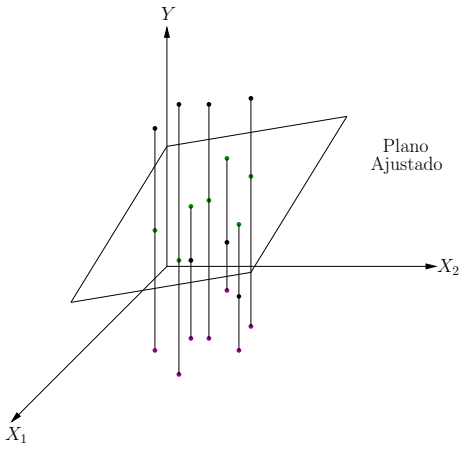


Figura 1.1: Ausencia de colinealidad

$$X_1 = a + bX_2$$

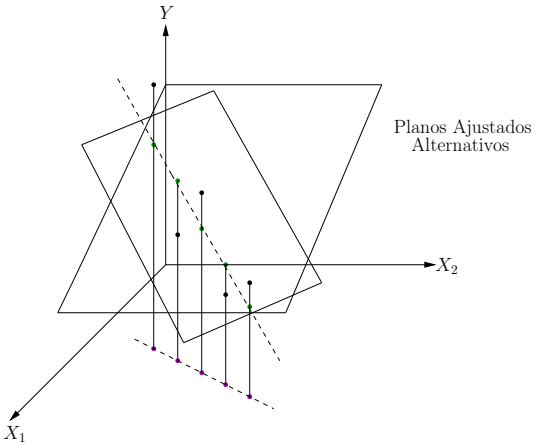


Figura 1.2: Colinealidad perfecta

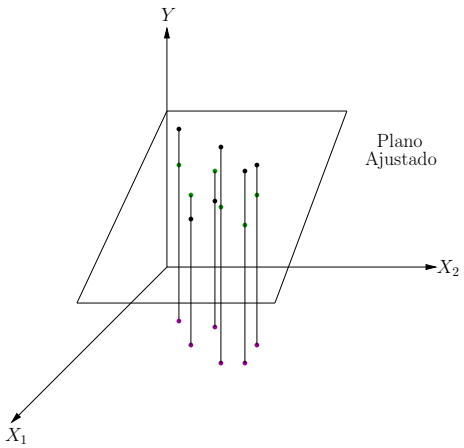


Figura 1.3: Colinealidad aproximada

Por lo anterior, se concluye que la colinealidad está presente cuando se observa inestabilidad, signos incorrectos en los parámetros estimados y frecuentemente elevados errores estándar, lo que conduce a generar modelos con muy poco poder explicativo o de difícil interpretación.

A continuación se explicarán los principales efectos que tiene el fenómeno de la colinealidad en la estimación de parámetros de regresión lineal y cómo puede ésta ser detectada o diagnosticada en las variables independientes, así como los principales procedimientos adoptados para manejarla o eliminarla. La fuente de la multicolinealidad impacta en el análisis, las correcciones y la interpretación del modelo lineal, por lo que es importante tomarla en cuenta.

1.1. Fuentes de detección

Existen cinco fuentes principales de colinealidad, que son:

1. **Recolección de los Datos.** En este caso los datos se han recolectado en un subespacio reducido y limitado de las variables independientes. La colinealidad fue creada por la metodología del muestreo, por lo que obtener más datos en un rango extendido corregiría el problema.

2. **Restricciones Físicas del Modelo Lineal o Poblacional.** Esta fuente de colinealidad existirá de acuerdo a cual técnica de muestreo se utilice. Por ejemplo, muchos procesos de manufactura o servicio tienen restricciones en las variables independientes (así como en su rango), ya sean físicas, políticas o legales, que crean colinealidad.

3. **Modelo sobre-definido.** Hay más variables que observaciones.

4. **Selección o Especificación del Modelo.** Esta fuente de colinelidad proviene de utilizar variables independientes que tienen potencias más altas o interacciones de un conjunto original de variables. Se debe tener en cuenta que si el subespacio de muestreo X_j es reducido, entonces cualquier combinación de variables con x_j incrementará el problema de multicolinealidad aún más.

5. **Outliers.** Los valores extremos, o que se salen del comportamiento en el espacio X pueden ocasionar colinealidad, así como esconderla.

1.2. Principales efectos de la colinealidad

Cuando se sospecha de presencia de colinealidad en las variables independientes, este fenómeno debe ser investigado antes de ajustar un modelo de regresión, ya que puede ocasionar errores en los pronósticos y dificultar la iterpretación del estimador.
Las principales consecuencias de las altas colinealidades entre las variables independientes son:

a) Los coeficientes estimados pueden ser insignificantes o de signo contrario al esperado y consecuentemente ser muy sensibles a los cambios en los datos muestrales. Esto es debido a la colinealidad de las variables independientes, por lo que los errores estándar serán grandes y consecuentemente la estadística de prueba t será pequeña. Los coeficientes estimados con error estándar muy grande presentarán inestabilidad.

b) Cuando existe colinealidad en las variables independientes es díficil estimar adecuadamente la importancia de éstas en el modelo generado, especialmente cuando existe un signo contrario al esperado en uno de los coeficientes estimados.
Por ejemplo, se espera que a mayor calidad de un producto terminado, la demanda se incremente si el precio se mantiene constante; sin embargo, puede encontrarse mediante un modelo de regresión lineal, empleado como pronóstico, que a mayor calidad del producto la demanda disminuya, lo cual es ilógico.

c) La colinealidad de las variables independientes puede sugerir que se excluyan importantes variables en los modelos. Sin embargo este proceso puede generar modelos menos efectivos o que no representan la realidad, dado que estadísticamente no son suficientes.

1.3. Principales técnicas de detección

Existen muchas técnicas para determinar en qué medida la colinealidad afecta gravemente a la estimación y contraste de un modelo, las cuales comprenden desde reglas de eliminación de variables, hasta el cálculo de índices complejos.
Algunas de las técnicas para la detección de colinealidad son:

1. **Cálculo de los coeficientes de correlación.** Estimar los coeficientes de correlación para determinar el grado de colinealidad. En algunos casos la construcción de una matriz de correlación y la representación gráfica es de gran utilidad. Mason y Perreault, recomiendan que sea eliminada una de las variables que tenga un coeficiente de correlación mayor a 0.8.

2. **Inspección de las R^2 y la estadística.** Cuando los valores de R^2 y la estadística F son grandes, esto indica una fuerte relación entre las variables independientes analizadas. Además si algunos de los coeficientes son insignificantes (valores pequeños o muy grandes) y los valores de R^2 y F son grandes, es un indicativo de que algunas variables independientes poseen alta correlación y se puede sospechar de la presencia de colinealidad.

3. **Factor de Inflación de Varianza.** El factor de inflación de varianza **(FIV)** y la tolerancia **(T)** están definidos como:

$$FIV_i = \frac{1}{1 - R_i^2} \qquad\qquad T_i = \frac{1}{FIV_i} = 1 - R_i^2$$

   donde R_i^2 es el coeficiente de determinación obtenido al efectuar la regresión de X_i sobre el resto de las regresoras del modelo.
   El FIV muestra el agrandamiento de la varianza del estimador como consecuencia de la no ortogonalidad de las variables regresoras.
   La mayoría de los autores consideran que existe un problema grave de colinealidad cuando el FIV de algún coeficiente es mayor de 5 o 10, lo que es un indicador de que los coeficientes de regresión asociados tienen una pobre estimación. Análogamente, se puede decir que existe un problema de colinealidad cuando la tolerancia es menor que 0.10 (Tolerancia<0.10).
   Es importante mencionar que puede existir colinealidad con FIV bajos, además de que puede haber colinealidades que no involucren a todas las variables independientes y que, por tanto, no son bien detectadas por el FIV.
   El problema que presenta el FIV o el R_j^2, es que no proporcionan ninguna información que pueda ser utilizada para corregir el problema.

4. **Análisis del sistema de valores propios.** Considera los valores propios de la matriz de correlación de las variables independientes que sean pequeños. Un valor propio de cero o cercano a él, indica que existe una dependencia lineal exacta.

5. **Número condición.** El número condición, $\kappa(X)$, es igual a la raíz cuadrada del valor propio más grande (λ_{max}) entre el valor propio más pequeño (λ_{min}) de la matriz $X'X$, es decir:

$$\kappa(X)=\sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$$

el número condición mide la sensibilidad de las estimaciones por mínimos cuadrados ante pequeños cambios en los datos. Tanto con datos observados, como con datos simulados, el problema de colinealidad es grave cuando el número condición toma un valor entre 20 y 30.

6. **Índice condición.** Como la matriz $X'X$ es de orden pxp se obtienen p raíces características, pudiéndose calcular para cada una de ellas un índice condición definido de la siguiente forma:

$$ic\lambda_i=\sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

Para Belsley, Kuh y Welsch (1980), los índices condición entre 5 y 10 están asociados con una colinealidad débil, mientras que índices condición entre 30 y 100 señalan una colinealidad de moderada a fuerte.
Los índices condición altos (mayores que 30) indican el número de dependencias casi lineales que contribuyen al problema de colinealidad y la magnitud de los mismos mide su importancia relativa.

7. **Proporción de varianza.** Una vez determinada la presencia de colinealidad, es conveniente averiguar qué variables están implicadas en ellas. Usando ciertas propiedades de la matrices se puede calcular la proporción de la varianza de las variables sobre cada componente.
Si dos o más variables tienen una proporción de varianza alta en un componente, ésto indica que esas variables están implicadas en la colinealidad y, por tanto, la estimación de sus coeficientes está degradada por la misma.
Belsley propone usar conjuntamente los índices condición y la proporción de descomposición de varianza, para realizar el diagnóstico de colinealidad, usando como principio de proporción alta 0.5, de modo que, si un componente tiene un índice condición mayor que 30 y dos o más variables tienen una proporción de varianza alta en el mismo, esas variables son colineales.

1.4. Técnicas de corrección o manejo de la colinealidad

Se han propuesto varias reglas para mejorar los problemas de colinealidad.
Entre los métodos más usados se encuentran:

a) Transformación de las variables por diferenciación. En algunos casos, la diferenciación consecutiva de cada variable del conjunto de datos puede reducir el impacto de la multicolinealidad. Por ejemplo, la variable dependiente puede ser expresada como $y_t = Ln(y_t) - Ln(y_{t-1})$ y también para cada una de las variables independientes en la matriz X, $x_t = Ln(x_t) - Ln(x_{t-1})$.

b) Incorporación de información apriori en el modelo. Se pretende incorporar información o valores que han sido estimados en modelos anteriores en el nuevo modelo, la cual puede ser para cualquiera de las regresoras.

c) Agregar datos adicionales o nuevos en la muestra. Algunas veces el problema de colinealidad puede ser eliminado mediante la obtención de una nueva muestra u obteniendo más información para la ya existente.

d) Eliminar variables del análisis. Este procedimiento consiste en eliminar una o más variables correlacionadas del modelo. Para la determinación de las variables que se integran al nuevo modelo, generalmente se emplean técnicas de ánalisis multivariado, como el análisis de factores, donde con base en los valores propios de la matriz X se estima el poder de explicación de cada una de las variables independientes. Este enfoque es aceptado por ser reduccionista y por simplificar el modelo, sin embargo reduce el rango de X y esto lo puede convertir en una técnica que genere un modelo con menor poder explicativo.
Se ha propuesto sacrificar ciertas características de los estimadores obtenidos mediante la técnica de mínimos cuadrados, como es el caso del sesgo. En la Figura 1.4(a) se observa el caso en el que se tiene un parámetro β insesgado pero con un error estándar muy grande, mientras que en la Figura 1.4(b) se observa el caso de un estimador de un parámetro β que es sesgado pero que tiene un menor error estándar.

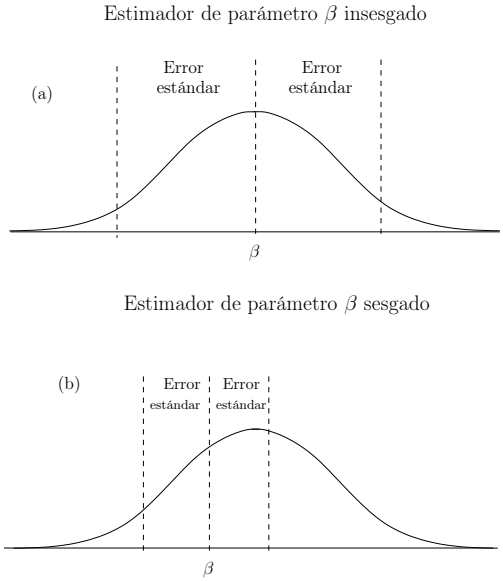


Figura 1.4: Distribuciones muestrales del estimador sesgado e insesgado de β

La principal técnica empleada para obtener estimadores sesgados fue propuesta por Kennard y Hoerl (1962) y es denominada **Regresión Ridge**, donde se agrega un sesgo a los parámetros estimados con la finalidad de reducir el error estándar de éstos, agregando una constante k, con valores que se encuentran entre 0 y 1. En el caso en que los valores de k son 0, entonces la Regresión Ridge coincide con la técnica de mínimos cuadrados. Según el teorema de Gauss-Markov, cuando los parámetros son obtenidos por mínimos cuadrados, éstos son insesgados; entonces a medida que los valores de k crecen, el sesgo de los parámetros estimados por la Regresión Ridge también crece, así que se pretende minimizar los valores de k para minimizar el sesgo y a su vez el error estándar. Los parámetros del modelo pueden ser estimados mediante la siguiente expresión:

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

Otra técnica ha sido propuesta por Liu (2003), donde se mejoran los parámetros estimados por la Regresión Ridge; este método considera que aún después de admitir un sesgo mediante k en la Regresión Ridge, se requiere de un segundo parámetro para disminuir aún más los factores de la colinealidad, al que denomina d. Las expresiones para obtener el estimador de Liu son:

$$\hat{\beta}_{k,d} = (X'X + kI)^{-1}(X'Y - dI)\hat{\beta}; \quad k > 0 \quad \text{y} \quad -\infty < d < \infty$$

$$k = \frac{\lambda_1 - 100\lambda_p}{99}$$

$$d = \frac{\sum_{i=1}^{p}((\lambda_i(\sigma_R^2 - k\alpha_{Ri}^2))/(\lambda_i + k)^3}{\sum_{i=1}^{p}((\lambda_i(\lambda_i\alpha_{Ri}^2 + \sigma_R^2))/(\lambda_i + k)^4}$$

donde $\hat{\beta}$ puede ser cualquier estimador.
Observe que cuando $d = 0$ resulta el estimador ridge.
También se tiene la representación gráfica de datos multivariados, denominada **Biplot** (Gabriel, 1971). De la misma manera que un diagrama de dispersión, éste muestra la distribución conjunta de dos variables, además se observan las similitudes relativas a los puntos de datos individuales y los valores relativos de las observaciones para cada variable independiente.
El Biplot aproxima la distribución de una muestra multivariada en un espacio de dimensión reducida, normalmente de dimensión dos y superpone sobre la misma, representaciones de las variables sobre las que se mide la muestra. Las representaciones de las variables son normalmente vectores y coinciden con las direcciones en las que mejor se muestra el cambio individual de cada variable.

Capítulo 2

Estimación Sesgada en Modelos Lineales

El procedimiento de estimación que usualmente se utiliza para β desconocidas es el de Gauss-Markov, el cual asegura que el estimador de mínimos cuadrados tiene varianza mínima en la clase de los estimadores lineales insesgados. Este procedimiento de estimación es adecuado si $X'X$ (en forma de una matriz de correlación) se aproxima a la matriz identidad, sin embargo si la matriz del problema está mal condicionada, el hecho de que cuente con un estimador de mínima varianza, no garantiza que su varianza sea pequeña, por lo que la varianza total del estimador de mínimos cuadrados puede ser demasiado grande para los propósitos finales, por lo tanto los estimadores de mínimos cuadrados serán sensibles a una serie de errores.

Hoerl y Kennard introducen una clase de estimadores sesgados para los parámetros en un modelo lineal general, resultando los estimadores ridge, los cuales se interpretan en términos de la función de verosimilitud, bajo el supuesto de la teoria normal y son vistos como una clase de transformaciones lineales de los estimadores de mínimos cuadrados. Este procedimiento de estimación está basado en agregar pequeñas cantidades positivas a la diagonal de la matriz $X'X$, lo que da lugar a la traza ridge, un método que muestra en dos dimensiones los efectos de la no ortogonalidad.

Otra clase de estimadores alternativos que serán revisados en este capítulo son los llamados estimadores reducidos, los cuales satisfacen la condición de admisibilidad propuesta por Hoerl y Kennard. Además, tanto los estimadores ridge como los estimadores reducidos, son obtenidos como estimadores de norma mínima en la clase de las transformaciones lineales de los estimadores de mínimos cuadrados, el primero de ellos minimiza la norma euclidiana y el segundo minimiza la norma dependiente del diseño. Por lo que se obtienen estimadores que son transformaciones lineales de varianza mínima de los estimadores de mínimos cuadrados y se muestra que los miembros de esta clase son estimadores estocásticamente reducidos.
En este capítulo también es considerado el problema de elegir un factor de reducción, así como el examinar las diferentes clases de estimadores sesgados.

2.1. Conceptos previos

Considere el modelo estándar de regresión lineal múltiple:

$$Y = X\beta + \varepsilon$$

donde Y y ε son vectores aleatorios de orden n, observables y no observables respectivamente, X es una matriz no-estocástica de nxp valores de entrada conocidos y de rango p ($p \leqslant n$) y β es un vector de parámetros de orden p. Se supone que $E[\varepsilon] = 0$ y $E[\varepsilon\varepsilon'] = \sigma^2 I_n$. Por conveniencia, supone que las variables x están estandarizadas, de manera que $X'X$ tiene la forma de una matriz de correlación.
Si una cierta observación respecto a los factores se denota por $x_v = \{x_{1v}, x_{2v}, x_{3v}, \ldots, x_{pv}\}$, la forma general de $X\beta$ es $\left\{\sum_{i=1}^{p} \beta_i \theta_i(x_v)\right\}$ donde θ_i son funciones libres de parámetros desconocidos.

Se define la suma de cuadrados de residuales de un estimador B de dimensión arbitraria p como:

$$\phi(B) = (Y - XB)'(Y - XB)$$

Usando un estimador lineal insesgado con varianza mínima o un estimador de máxima verosimilitud cuando el vector aleatorio, ε, es normal se tiene:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2.1}$$

como un estimador de β y puesto que la matriz $X'X$ es definida positiva se sigue inmediatamente que:

$$\phi(\hat{\beta}) = \min_B \phi(B)$$

Demostración.

$$\text{P.D} \qquad \hat{\beta} = (X'X)^{-1}X'Y$$

$$\text{Como: } \underline{e} = (Y - \hat{Y}) \qquad \text{y} \qquad \hat{Y} = X\hat{\beta}$$

la suma de cuadrados de residuales está dada por

$$\begin{aligned}
\text{Min} \sum_{i=1}^{n} e_i^2 &= e'e \\
&= (Y - \hat{Y})'(Y - \hat{Y}) \\
&= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\
&= (Y' - \hat{\beta}'X')(Y - X\hat{\beta}) \\
&= Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \\
&= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta}
\end{aligned}$$

ya que, $\hat{\beta}'X'Y$ es una matriz de 1x1 y su transpuesta $(\hat{\beta}'X'Y)' = Y'X\hat{\beta}$ es también un escalar, ahora calculando la derivada

$$\frac{\partial \text{Min} \sum_{i=1}^{n} e_i^2}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta}$$

igualando a cero la expresión anterior se tiene:

$$\begin{aligned} 2(X'X\hat{\beta} - X'Y) &= 0 \\ (X'X)^{-1}X'X\hat{\beta} &= (X'X)^{-1}X'Y \\ \hat{\beta} &= (X'X)^{-1}X'Y \end{aligned}$$

■

Es fácil demostrar que la mínima suma de cuadrados de residuales para un estimador B arbitrario puede ser descompuesta como:

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

Demostración.

$$\begin{aligned} \phi(B) &= (Y - XB)'(Y - XB) \\ &= (Y - XB + X\hat{\beta} - X\hat{\beta})'(Y - XB + X\hat{\beta} - X\hat{\beta}) \\ &= [(Y - X\hat{\beta})' - (XB - X\hat{\beta})'][(Y - X\hat{\beta}) - (XB - X\hat{\beta})] \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) - \underline{(Y - X\hat{\beta})'(XB - X\hat{\beta})} \\ &\quad - \underline{(XB - X\hat{\beta})'(Y - X\hat{\beta})} + (XB - X\hat{\beta})'(XB - X\hat{\beta}) \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (XB - X\hat{\beta})'(XB - X\hat{\beta}) \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (B - \hat{\beta})'X'X(B - \hat{\beta}) \end{aligned}$$

Ahora se demostrará que $\underline{(Y - X\hat{\beta})'(XB - X\hat{\beta})} + \underline{(XB - X\hat{\beta})'(Y - X\hat{\beta})}$ es igual a cero

$$\begin{aligned} &(Y - X\hat{\beta})'(XB - X\hat{\beta}) + (XB - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= (Y' - \hat{\beta}'X')(XB - X\hat{\beta}) + (B'X' - \hat{\beta}'X')(Y - X\hat{\beta}) \\ &= Y'XB - Y'X\hat{\beta} - \hat{\beta}'X'XB + \hat{\beta}'X'X\hat{\beta} + B'X'Y - B'X'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \\ &= Y'XB + B'X'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y - \hat{\beta}'X'XB - B'X'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= B'X'Y + B'X'Y - \hat{\beta}'X'Y - \hat{\beta}'X'Y - B'X'X\hat{\beta} - B'X'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= 2B'X'Y - 2\hat{\beta}'X'Y - 2B'X'X\hat{\beta} + 2\hat{\beta}'X'X\hat{\beta} \\ &= 2B'X'Y - 2B'X'X\hat{\beta} - 2\hat{\beta}'X'Y + 2\hat{\beta}'X'X\hat{\beta} \\ &= 2B'X'Y - 2B'X'X[(X'X)^{-1}X'Y] - 2\hat{\beta}'X'Y + 2\hat{\beta}'X'X[(X'X)^{-1}X'Y] \\ &= 2B'X'Y - 2B'[(X'X)(X'X)^{-1}X'Y] - 2\hat{\beta}'X'Y + 2\hat{\beta}'[(X'X)(X'X)^{-1}X'Y] \\ &= 2B'X'Y - 2B'X'Y - 2\hat{\beta}'X'Y + 2\hat{\beta}'X'Y \\ &= 0 \end{aligned}$$

■

Para cualquier estimador B se considerarán las siguientes propiedades:

$$\begin{aligned} Var(B) &= E(B - E(B))(B - E(B))' \\ V(B) &= trazaVar(B) \\ G(B) &= E(B - \beta)'(B - \beta) \\ D(B) &= (E(B) - \beta)'(E(B) - \beta) \end{aligned}$$

las cuales denotan la matriz de varianza-covarianza, varianza total, error cuadrático medio total y sesgo de B, respectivamente. Note que si $D(B) = 0$ entonces $V(B) = G(B)$ y por lo tanto la varianza y el error cuadrático medio total son idénticos para estimadores insesgados.

El principal problema en regresión múltiple consiste en los casos en que la matriz $X'X$ no se aproxima a la matriz identidad (a menos que se diga lo contrario, el modelo será propuesto para tener una forma de correlación en $X'X$). Para demostrar el efecto de esta condición en el estimador de β, considere las siguientes propiedades de $\hat{\beta}$.

i) Es insesgado, es decir, $E(\hat{\beta}) = \beta$.

Demostración.

$$\begin{aligned} E(\hat{\beta}) &= E[(X'X)^{-1}X'Y] \\ &= (X'X)^{-1}X'E(Y) \\ &= (X'X)^{-1}X'X\beta \\ &= \beta \end{aligned}$$

■

ii) $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

Demostración.

$$\begin{aligned} Var(\hat{\beta}) &= Var[(X'X)^{-1}X'Y] \\ &= (X'X)^{-1}X'Var(Y)[(X'X)^{-1}X']' \\ &= \sigma^2[(X'X)^{-1}X'][(X'X)^{-1}X']' \\ &= \sigma^2[(X'X)^{-1}X'][X((X'X)^{-1})'] \\ &= \sigma^2[(X'X)^{-1}X'][X((X'X)')^{-1}] \\ &= \sigma^2[(X'X)^{-1}X'][X(X'X)^{-1}] \\ &= \sigma^2[(X'X)^{-1}(X'X)(X'X)^{-1}] \\ &= \sigma^2(X'X)^{-1}I_p \end{aligned}$$

■

iii) $L_1 \equiv$ Distancia de $\hat{\beta}$ a β.

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

$$E[L_1^2] = \sigma^2 traza(X'X)^{-1} \tag{2.2}$$

o de forma equivalente:

$$E[\hat{\beta}'\hat{\beta}] = \beta'\beta + \sigma^2 traza(X'X)^{-1} \tag{2.3}$$

Demostración.

Para poder demostrar las igualdades (2.2) y (2.3) se utilizará el siguiente teorema.

Teorema. Sea Q un vector aleatorio de tamaño p, sea M una matriz simétrica de tamaño pxp. Si $E(Q) = \theta$ y $Var(Q) = \sigma^2 I_p$, entonces:

$$\begin{aligned} E[Q'MQ] &= traza(M\sigma^2) + \theta'M\theta \\ &= \sigma^2 traza(M) + \theta'M\theta \end{aligned}$$

Se demostrará la igualdad (2.2) con base en el teorema anterior.

Se sabe que

$$\begin{aligned} L_1 &= (\hat{\beta} - \beta) \\ L_1^2 &= (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \end{aligned}$$

sea

$$\begin{aligned} S &= (\hat{\beta} - \beta) \\ S'S &= (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \\ &= L_1^2 \end{aligned}$$

por lo tanto

$$\begin{aligned} E(S) &= E(\hat{\beta} - \beta) \\ &= E(\hat{\beta}) - \beta \\ &= \beta - \beta \\ &= 0 \end{aligned}$$

y

$$\begin{aligned} Var(S) &= Var(\hat{\beta} - \beta) \\ &= Var(\hat{\beta}) \\ &= \sigma^2 (X'X)^{-1} \end{aligned}$$

entonces

$$\begin{aligned} E[L_1^2] &= E[S' I_p S] \\ &= traza(I_p \sigma^2 (X'X)^{-1}) \\ &= \sigma^2 traza(X'X)^{-1} \end{aligned}$$

A continuación se demostrará la igualdad (2.3) tomando los resultados del teorema anterior.

Ya que

$$E(\hat{\beta}) = \beta$$

y

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

entonces

$$\begin{aligned} E[\hat{\beta}' I_p \hat{\beta}] &= traza(I_p \sigma^2 (X'X)^{-1}) + \beta' I_p \beta \\ &= \beta'\beta + \sigma^2 traza[(X'X)^{-1}] \end{aligned}$$

■

Cuando el error ε se distribuye normal, tenemos:

$$Var(L_1^2) = 2\sigma^4 traza(X'X)^{-2} \tag{2.4}$$

Demostración.

Se utilizará el siguiente resultado.
Sea H un vector aleatorio de tamaño p, sea M una matriz simétrica de tamaño pxp. Si $E(H) = 0$ y $Var(H) = \sigma^2 I_p$, entonces:

$$\begin{aligned} Var(H'MH) &= 2traza(M\sigma^2)^2 \\ &= 2\sigma^4 traza(M^2) \end{aligned}$$

se sabe que:

$$\begin{aligned} S &= (\hat{\beta} - \beta) \\ E(S) &= 0 \\ Var(S) &= \sigma^2 (X'X)^{-1} \end{aligned}$$

entonces

$$\begin{aligned} Var(L_1^2) &= Var(S' I_p S) \\ &= 2traza[\sigma^2 (X'X)^{-1} I_p]^2 \\ &= 2\sigma^4 traza(X'X)^{-2} \end{aligned}$$

■

Estas propiedades muestran la incertidumbre en $\hat{\beta}$ cuando $X'X$ se aleja de ser una matriz identidad para convertirse en una matriz mal condicionada.
Si los valores propios de $X'X$ se denotan por

$$\lambda_{max} = \lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \cdots \geqslant \lambda_p = \lambda_{min} > 0$$

entonces el valor esperado de la distancia cuadrada de $\hat{\beta}$ a β está dado por:

$$E[L_1^2] = G(\hat{\beta}) = V(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i} > \sigma^2/\lambda_p$$

Demostración.

Existe una matriz ortogonal P, tal que $X'X = P\Lambda P'$ donde $\Lambda = (\partial_{i,j}\lambda_i)$ es la matriz de valores propios de $X'X$.
Entonces:

$$\begin{aligned}
E[L_1^2] &= \sigma^2 traza(X'X)^{-1} \\
&= \sigma^2 traza(P\Lambda P')^{-1} \\
&= \sigma^2 traza[(P')^{-1}(P\Lambda)^{-1}] \\
&= \sigma^2 traza[(P')'(P\Lambda)^{-1}] \\
&= \sigma^2 traza[P\Lambda^{-1}P^{-1}] \\
&= \sigma^2 traza[P^{-1}P\Lambda^{-1}] \\
&= \sigma^2 traza[P'P\Lambda^{-1}] \\
&= \sigma^2 traza[I_p\Lambda^{-1}] \\
&= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}
\end{aligned}$$

■

y la varianza cuando el error ε es normal está dada por:

$$Var(L_1^2) = 2\sigma^4 \sum_{i=1}^{p} \left(\frac{1}{\lambda_i}\right)^2$$

Demostración.

$$\begin{aligned} Var(L_1^2) &= 2\sigma^4 traza(X'X)^{-2} \\ &= 2\sigma^4 traza[P\Lambda P']^{-2} \\ &= 2\sigma^4 traza[(P\Lambda P')^{-1}]^2 \\ &= 2\sigma^4 traza[(P')^{-1}\Lambda^{-1}P^{-1}]^2 \\ &= 2\sigma^4 traza[P\Lambda^{-1}P^{-1}]^2 \\ &= 2\sigma^4 traza[P^{-1}P\Lambda^{-1}]^2 \\ &= 2\sigma^4 traza[P'P\Lambda^{-1}]^2 \\ &= 2\sigma^4 traza[I_p\Lambda^{-1}]^2 \\ &= 2\sigma^4 \sum_{i=1}^{p} \left(\frac{1}{\lambda_i}\right)^2 \end{aligned}$$

■

El límite inferior de la esperanza es:

$$\frac{\sigma^2}{\lambda_{min}}$$

El límite inferior de la varianza es:

$$\frac{2\sigma^4}{\lambda_{min}^2}$$

Por lo tanto, cuando $X'X$ se aleja de la matriz identidad, que es cuando ésta tiene uno o más valores propios pequeños, la $E(L_1^2)$ y $Var(L_1^2)$ muestran que la distancia de $\hat{\beta}$ a β tiende a ser grande. Por otro lado los coeficientes de estimación, $\hat{\beta}_i$, que son grandes en valor absoluto son observados en problemas no ortogonales.

2.2. Estimadores sesgados

A. E. Hoerl fue el primero en proponer en 1962 que para controlar la inflación y la inestabilidad general asociada con el estimador de mínimos cuadrados, se puede utilizar:

$$\begin{aligned} \hat{\beta}^* = b_k &= (b_{k1} \dots b_{kp})' \\ &= [X'X + kI_p]^{-1}X'Y; \qquad k \geqslant 0 \\ &= WX'Y \end{aligned}$$

donde k es una constante no negativa de la matriz diagonal y además la familia de los estimadores dados por $k \geqslant 0$, tienen muchas similitudes matemáticas con la representación de las funciones de respuesta cuadrática. Como $\hat{\beta} = b_0$ y $\hat{\beta}$ es insesgado se sigue inmediatamente que para $k \neq 0$, $\hat{\beta}^*$ es sesgado.

Hoerl y Kennard justificaron el uso del estimador ridge en problemas no ortogonales de las siguientes dos formas:

i) muestran que para una k fija, $\hat{\beta}^*$ corresponde a un punto en la elipse que contiene a $\hat{\beta}$, la cual cuenta con una longitud mínima euclidiana y,

ii) muestran que dado cualquier problema, la clase de estimadores ridge satisface la siguiente condición:
Condición de Admisibilidad: Una clase E de estimadores (cuadrado medio) será llamado admisible si para cada problema existe una e en E tal que $G(e) < G(\hat{\beta}) = V(\hat{\beta})$.

Considere una clase de estimadores alternativos, cuyo miembro típico es:

$$c_\lambda = (c_{\lambda 1} \dots c_{\lambda p})' = \lambda (X'X)^{-1} X'Y = \lambda \hat{\beta} \qquad \lambda \epsilon [0, \infty) \tag{2.5}$$

El estimador c_λ es un estimador reducido y λ es el factor de reducción, si λ es un escalar fijo, entonces c_λ es denominado un estimador determinísticamente reducido. De manera alternativa, si $\lambda = f(\hat{\beta}'\hat{\beta})$ es una función escalar de $\hat{\beta}'\hat{\beta}$, entonces c_λ es llamado un estimador estocásticamente reducido y se escribe como $c(f)$.
Se muestran a continuación las propiedades de c_λ (λ fija):

$$Var(c_\lambda) = \lambda^2 \sigma^2 (X'X)^{-1} \tag{2.6}$$
$$V(c_\lambda) = \lambda^2 \sigma^2 traza (X'X)^{-1} \tag{2.7}$$
$$D(c_\lambda) = (1-\lambda)^2 \beta'\beta \tag{2.8}$$

Demostración de la expresión (2.6).

$$\begin{aligned} Var(c_\lambda) &= Var(\lambda \hat{\beta}) \\ &= \lambda^2 Var(\hat{\beta}) \\ &= \lambda^2 \sigma^2 (X'X)^{-1} \end{aligned}$$

■

Demostración de la expresión (2.7).

$$\begin{aligned} V(c_\lambda) &= V(\lambda \hat{\beta}) \\ &= traza Var(\lambda \hat{\beta}) \\ &= traza[\lambda^2 Var(\hat{\beta})] \\ &= \lambda^2 \sigma^2 traza (X'X)^{-1} \end{aligned}$$

■

Demostración de la expresión (2.8).

$$\begin{aligned}
D(c_\lambda) &= D(\lambda\hat{\beta}) \\
&= [E(\lambda\hat{\beta}) - \beta]'[E(\lambda\hat{\beta}) - \beta] \\
&= [\lambda E(\hat{\beta}) - \beta]'[\lambda E(\hat{\beta}) - \beta] \\
&= [\lambda\beta - \beta]'[\lambda\beta - \beta] \\
&= [\lambda\beta' - \beta'][\lambda\beta - \beta] \\
&= \lambda\beta'\lambda\beta - \lambda\beta'\beta - \beta'\lambda\beta + \beta'\beta \\
&= \lambda^2\beta'\beta - \lambda\beta'\beta - \lambda\beta'\beta + \beta'\beta \\
&= \lambda^2\beta'\beta - 2\lambda\beta'\beta + \beta'\beta \\
&= (1 - 2\lambda + \lambda^2)\beta'\beta \\
&= (1 - \lambda)^2\beta'\beta
\end{aligned}$$

■

Los momentos del estimador estocáticamente reducido $c(f)$ dependen de la forma de f y no serán dados de forma general.
Aunque el estimador reducido c_λ puede parecer una alteración bastante simplista de $\hat{\beta}$, a continuación se verifica que éste satisface la condición de admisibilidad.

Proposición 2.1. *Para cada β existe un λ fijo en $[0, 1]$, tal que, $G(c_\lambda) < V(\hat{\beta})$ y por lo tanto la subclase de los estimadores determinísticamente reducidos es admisible.*

Demostración.

De las expresiones (2.7) y (2.8) se sigue que:

$$G(c_\lambda) = \lambda^2 V(\hat{\beta}) + (1 - \lambda)^2\beta'\beta$$

Por lo tanto, $G(c_\lambda) < V(\hat{\beta})$ si y sólo si

$$\lambda > \frac{\beta'\beta - V(\hat{\beta})}{\beta'\beta + V(\hat{\beta})}$$

Esto se puede demostrar de la siguiente manera: dado que $(\beta'\beta - V(\hat{\beta})) < (\beta'\beta + V(\hat{\beta}))$ existe una $\lambda < 1$ tal que $G(\lambda\hat{\beta}) < V(\hat{\beta})$.

■

Por lo tanto, un factor para justificar el uso de los estimadores reducidos es porque satisfacen la condición de admisibilidad. Además, los estimadores determinísticamente reducidos se derivan en la siguiente sección por un método similar al utilizado para la obtención de los estimadores ridge.

2.2.1. Trasformaciones lineales de $\hat{\beta}$

Definición 2.1. *Sea C la clase de trasformaciones lineales de $\hat{\beta}$. Si $b \epsilon C$ entonces $b = A\hat{\beta}$ para alguna matriz A de $p \times p$.*

Si $b(A) = A\hat{\beta}$ para una A fija, entonces

$$\begin{aligned} E[b(A)] &= A\beta \\ Var[b(A)] &= \sigma^2 A(X'X)^{-1}A' \\ G[b(A)] &= \sigma^2 traza(A'(X'X)^{-1}A) + \beta'(A - I_p)'(A - I_p)\beta \end{aligned}$$

Demostración.

$$\begin{aligned} E[b(A)] &= E[A\hat{\beta}] \\ &= AE[\hat{\beta}] \\ &= A\beta \end{aligned}$$

$$\begin{aligned} Var[b(A)] &= Var[A\hat{\beta}] \\ &= AVar(\hat{\beta})A' \\ &= A\sigma^2(X'X)^{-1}A' \\ &= \sigma^2 A(X'X)^{-1}A' \end{aligned}$$

Finalmente, primero se encontrará el valor de $D[b(A)]$, ya que se utilizará para obtener el valor de $G[b(A)]$.

$$\begin{aligned} D[b(A)] &= [E(b(A)) - \beta]'[E(b(A)) - \beta] \\ &= [E(A\hat{\beta}) - \beta]'[E(A\hat{\beta}) - \beta] \\ &= [A\beta - \beta]'[A\beta - \beta] \\ &= [\beta'A' - \beta'][A\beta - \beta] \\ &= \beta'(A - I_p)'(A - I_p)\beta \end{aligned}$$

entonces

$$\begin{aligned} G[b(A)] &= traza(Var(b(A))) + \beta'(A - I_p)(A - I_p)\beta \\ &= \sigma^2 traza(A(X'X)^{-1}A') + \beta'(A - I_p)'(A - I_p)\beta \end{aligned}$$

■

La suma de cuadrados de residuales asociada con $b(A)$ es:

$$\begin{aligned}
\phi(A) &= (Y - Xb(A))'(Y - Xb(A)) \\
&= (Y - X(A\hat{\beta}))'(Y - X(A\hat{\beta})) \\
&= (Y' - \hat{\beta}'A'X')(Y - XA\hat{\beta}) \\
&= Y'Y - Y'XA\hat{\beta} - \hat{\beta}'A'X'Y + \hat{\beta}'A'X'XA\hat{\beta} \\
&= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (A\hat{\beta} - \hat{\beta})'(X'X)(A\hat{\beta} - \hat{\beta}) \\
&= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (A\hat{\beta} - \hat{\beta})'(X'X)(A - I_p)\hat{\beta} \\
&= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + \hat{\beta}'(A' - I_p)(X'X)(A - I_p)\hat{\beta} \\
&= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + \hat{\beta}'(A - I_p)'(X'X)(A - I_p)\hat{\beta} \\
&= \phi(\hat{\beta}) + \phi^*(A)
\end{aligned}$$

$\phi(A)$ se reduce al mínimo si $A = I$, ya que $\phi^*(I) = 0$ y se obtiene el estimador de mínimos cuadrados. Sin embargo, si $\phi^*(A) > 0$ entonces la asignación para el espacio de matrices A de $p \times p$ a la línea real positiva, la cual se define como $\gamma(A) = \phi^*(A)$, asocia una clase completa de matrices para el mismo valor. La pre-imagen de cualquier constante fija τ consiste en todas las matrices de $p \times p$ que satisfacen:

$$\hat{\beta}'(A - I)'(X'X)(A - I)\hat{\beta} = \tau$$

Definición 2.2. *$C(\tau)$ denota la subclase de C tal que $b(A_0)$ está en $C(\tau)$ si y sólo si $\phi^*(A_0) = \tau$.*

$C(\tau)$ es en realidad una clase de equivalencia dentro de la clase C, esta equivalencia se define con respecto a la suma de cuadrados de residuales. En la clase de equivalencia $C(\tau)$, el problema de elegir un estimador es el decidir cual miembro de la clase de equivalencia debe ser utilizado.

Ahora se demostrará que tanto los estimadores ridge como los estimadores determinísticamente reducidos, pueden ser caracterizados como estimadores de norma mínima en la clase C, entonces supóngase que el criterio para seleccionar un estimador para una clase de equivalencia es elegir el estimador que tiene la longitud (norma) mínima euclidiana. Sea

$$m(A) = b'(A)b(A) = \hat{\beta}'A'A\hat{\beta}$$

que denota el cuadrado de la longitud (norma) euclidiana de $b(A)$.

Proposición 2.2. *Si $A_0 = (k(X'X)^{-1} + I_p)^{-1}$ para alguna k y $b(A_0)\ \epsilon\ C(\tau)$ entonces*

$$m(A_0) = min_{C(\tau)}\ m(A)$$

Demostración.

La demostración es directa, ya que Hoerl y Kennard muestran que de todos los estimadores sobre un elipsoide fijo con centro en $\hat{\beta}$, los estimadores ridge tienen longitud mínima y $b(A_0) = b_k$ es un estimador ridge. Sin embargo, otra forma de demostrarlo es mediante la diferenciación de la expresión de Lagrange con respecto a A e igualando a cero:
Minimizar F, donde:

$$\begin{aligned}
F &= \hat{\beta}'A'A\hat{\beta} + k^{-1}[\hat{\beta}'(A - I_p)'(X'X)(A - I_p)\hat{\beta} - \gamma] \\
&= \hat{\beta}'A'A\hat{\beta} + k^{-1}[(\hat{\beta}'A' - \hat{\beta}')((X'X)A\hat{\beta} - (X'X)\hat{\beta}) - \gamma] \\
&= \hat{\beta}'A'A\hat{\beta} + k^{-1}[(\hat{\beta}'A' - \hat{\beta}')((X'X)A\hat{\beta}) - (\hat{\beta}'A' - \hat{\beta}')((X'X)\hat{\beta}) - \gamma] \\
&= \hat{\beta}'A'A\hat{\beta} + k^{-1}[\hat{\beta}'A'(X'X)A\hat{\beta} - \hat{\beta}'(X'X)A\hat{\beta} - \hat{\beta}'A'(X'X)\hat{\beta} + \hat{\beta}'(X'X)\hat{\beta} - \gamma] \\
&= \hat{\beta}'A'A\hat{\beta} + k^{-1}\hat{\beta}'A'(X'X)A\hat{\beta} - k^{-1}\hat{\beta}'(X'X)A\hat{\beta} - k^{-1}\hat{\beta}'A'(X'X)\hat{\beta} + k^{-1}\hat{\beta}'(X'X)\hat{\beta} - k^{-1}\gamma]
\end{aligned}$$

Entonces

$$\begin{aligned}
\frac{\partial F}{\partial A} &= 2A\hat{\beta}\hat{\beta}' + 2k^{-1}(X'X)A\hat{\beta}\hat{\beta}' - 2k^{-1}(X'X)\hat{\beta}\hat{\beta}' \\
&\Rightarrow \quad 2\left[A\hat{\beta}\hat{\beta}' + k^{-1}(X'X)A\hat{\beta}\hat{\beta}' - k^{-1}(X'X)\hat{\beta}\hat{\beta}'\right] = 0
\end{aligned}$$

Por lo que

$$\begin{aligned}
A + k^{-1}(X'X)A - k^{-1}(X'X) &= 0 \\
(I_p + k^{-1}(X'X))A - k^{-1}(X'X) &= 0 \\
(I_p + k^{-1}(X'X))A &= k^{-1}(X'X) \\
(I_p + k^{-1}(X'X))^{-1}(I_p + k^{-1}(X'X))A &= (I_p + k^{-1}(X'X))^{-1}k^{-1}(X'X) \\
A &= (I_p + k^{-1}(X'X))^{-1}k^{-1}(X'X) \\
A &= (I_p + k^{-1}(X'X))^{-1}[(k^{-1}(X'X))^{-1}]^{-1} \\
A &= [(k^{-1}(X'X))^{-1}(I_p + k^{-1}(X'X))]^{-1} \\
A &= [(k^{-1}(X'X))^{-1} + (k^{-1}(X'X))^{-1}(k^{-1}(X'X))]^{-1} \\
A &= [k(X'X)^{-1} + I_p]^{-1}
\end{aligned}$$

■

La proposición 2.2 establece que dentro de su clase de equivalencia el estimador ridge es el estimador de menor longitud, siempre que $m(A)$ sea la norma usada para medir la longitud.

Ahora considere la norma dependiente del diseño

$$m_d(A) = b'(A)(X'X)b(A) = \hat{\beta}'A'(X'X)A\hat{\beta}$$

y suponga que el estimador óptimo en una clase de equivalencia se define como el estimador con una longitud mínima, medida por $m_d(A)$.

Proposición 2.3. *Si $A_1 = \lambda I$ para alguna $\lambda \epsilon [0,1]$ y c_λ pertenece a $C(\tau)$ entonces*

$$m_d(\lambda I) = min\ m_d(A)$$

Demostración.

La demostración se sigue a partir de la diferenciación de la expresión de Lagrange

$$\begin{aligned}
F^* &= \hat{\beta}'A'(X'X)A\hat{\beta} + \lambda[\hat{\beta}'(A - I_p)'(X'X)(A - I_p)\hat{\beta} - \gamma] \\
&= \hat{\beta}'A'(X'X)A\hat{\beta} + \lambda[(\hat{\beta}'A' - \hat{\beta}')((X'X)A\hat{\beta} - (X'X)\hat{\beta}) - \gamma] \\
&= \hat{\beta}'A'(X'X)A\hat{\beta} + \lambda[(\hat{\beta}'A' - \hat{\beta}')((X'X)A\hat{\beta}) - (\hat{\beta}'A' - \hat{\beta}')((X'X)\hat{\beta}) - \gamma] \\
&= \hat{\beta}'A'(X'X)A\hat{\beta} + \lambda[\hat{\beta}'A'(X'X)A\hat{\beta} - \hat{\beta}'(X'X)A\hat{\beta} - \hat{\beta}'A'(X'X)\hat{\beta} + \hat{\beta}'(X'X)\hat{\beta} - \gamma] \\
&= \hat{\beta}'A'(X'X)A\hat{\beta} + \lambda\hat{\beta}'A'(X'X)A\hat{\beta} - \lambda\hat{\beta}'(X'X)A\hat{\beta} - \lambda\hat{\beta}'A'(X'X)\hat{\beta} + \lambda\hat{\beta}'(X'X)\hat{\beta} - \lambda\gamma
\end{aligned}$$

Entonces:

$$\begin{aligned}
\frac{\partial F}{\partial A} &= 2(X'X)A\hat{\beta}\hat{\beta}' + 2\lambda(X'X)A\hat{\beta}\hat{\beta}' - 2\lambda(X'X)\hat{\beta}\hat{\beta}' \\
&\Rightarrow \quad 2\left[(X'X)A\hat{\beta}\hat{\beta}' + \lambda(X'X)A\hat{\beta}\hat{\beta}' - \lambda(X'X)\hat{\beta}\hat{\beta}'\right] = 0
\end{aligned}$$

Por lo que

$$\begin{aligned}
(X'X)A\hat{\beta}\hat{\beta}' + \lambda(X'X)A\hat{\beta}\hat{\beta}' - \lambda(X'X)\hat{\beta}\hat{\beta}' &= 0 \\
(X'X)[A + \lambda A - \lambda I_p] &= 0 \\
(A + \lambda A - \lambda I_p) &= 0 \\
(A + \lambda A) &= \lambda I_p \\
(1 + \lambda)A &= \lambda I_p \\
A &= \left(\frac{\lambda}{1 + \lambda}\right) I_p
\end{aligned}$$

■

Como $b(A_1) = \lambda\hat{\beta} = c_\lambda$ se ha demostrado que tanto los estimadores ridge como los estimadores reducidos son estimadores de norma mínima, ambos contienen un parámetro que debe ser dado antes de que el valor del estimador sea determinado. Hoerl y Kennard trabajaron para mostrar la elección de una k apropiada mediante la graficación de los coeficientes individuales $(b_{k1} \dots b_{kp})$ contra k, observando que los coeficientes se estabilizan lo suficiente para pequeños valores de k. Sin embargo, si se pretende trabajar con

un esquema similar para los estimadores determinísticamente reducidos, no se obtendrán resultados favorables, ya que el valor absoluto de cada elemento en c_λ es linealmente creciente en λ, por lo que en la siguiente sección se proponen algunos métodos para elegir el factor de reducción más apropiado.

Observe que aunque Hoerl y Kennard mostraron un método heurístico para obtener un valor adecuado de k para sus estimadores ridge, si su método es utilizado surgen dos complicaciones:

i) a pesar de que la clase de estimadores ridge es admisible, los autores no pueden garantizar que el estimador elegido por su método, tenga un error cuadrático medio total más pequeño que la varianza del estimador de mínimos cuadrados,

ii) dado que el valor de k elegido por el método de Hoerl y Kennard es una función de $\hat{\beta}$ y por lo tanto una variable aleatoria, los momentos de b_k para una k fija no son los momentos del estimador que se utilizan en la práctica.

Asimismo, al utilizar el estimador reducido uno se enfrenta a los mismos problemas, ya que si λ es una constante fija, entonces los momentos de c_λ son conocidos, pero desafortunadamente no es evidente como fijar λ sin observar el estimador de mínimos cuadrados.

2.2.2. Transformación de mínima varianza de $\hat{\beta}$

El estimador óptimo dentro de una clase de equivalencia no tiene que ser elegido minimizando la norma del estimador. En realidad, puesto que diversas normas conducen a diferentes estimadores, no hay razón evidente para preferir una norma sobre otra, por lo que se buscará un criterio de selección alternativa.
Un estudio realizado muestra que la selección de una norma dependiente del diseño no es una buena alternativa, ya que el diseño es conocido por ser deficiente y si un problema está lo suficientemente mal condicionado, requiere de un estimador sesgado. También se muestra, que dado que la norma impone ciertas restricciones en el sesgo del estimador, la norma óptima se basa en criterios externos a la observación de datos en particular.
Por lo tanto, se considerará el estimador que tiene varianza total mínima entre todos los estimadores en una clase de equivalencia dada, señalando tales estimadores en la siguiente proposición.

Proposición 2.4. *Sea* $A_2 = \delta\hat{\beta}\hat{\beta}'(I + \delta\hat{\beta}\hat{\beta}')^{-1}$ *para alguna* δ, *si* $b(A_2)\ \epsilon\ C(\tau)$ *entonces*

$$V(b(A_2)) = min_{C(\tau)}\ V(b(A))$$

Demostración.

La demostración procede minimizando $\phi(A)$, sujeto a que $traza(A'(X'X)^{-1}A) = \gamma$ utilizando multiplicadores de Lagrange.
La derivada de la expresión de Lagrange es

$$\frac{\partial F}{\partial A} = 2A(X'X) + \delta[2A(X'X)\hat{\beta}\hat{\beta}' - 2(X'X)\hat{\beta}\hat{\beta}'] \tag{2.9}$$

Entonces

$$\begin{aligned}
A(X'X) + \delta A(X'X)\hat{\beta}\hat{\beta}' - \delta(X'X)\hat{\beta}\hat{\beta}' &= 0 \\
A(X'X)(I_p + \delta\hat{\beta}\hat{\beta}') &= \delta(X'X)\hat{\beta}\hat{\beta}' \\
A(X'X)(I_p + \delta\hat{\beta}\hat{\beta}')(I_p + \delta\hat{\beta}\hat{\beta}')^{-1} &= \delta(X'X)\hat{\beta}\hat{\beta}'(I_p + \delta\hat{\beta}\hat{\beta}')^{-1} \\
A(X'X) &= \delta(X'X)\hat{\beta}\hat{\beta}'(I_p + \delta\hat{\beta}\hat{\beta}')^{-1} \\
(X'X)^{-1}((X'X)A')' &= (X'X)^{-1}[(X'X)(\hat{\beta}\hat{\beta}'(I_p + \delta\hat{\beta}\hat{\beta}')^{-1})']'\delta \\
[A'(X'X)(X'X)^{-1}]' &= [(\hat{\beta}\hat{\beta}'(I_p + \delta\hat{\beta}\hat{\beta}')^{-1})'(X'X)(X'X)^{-1}]'\delta \\
[A'I_p]' &= [(\hat{\beta}\hat{\beta}'(I_p + \delta\hat{\beta}\hat{\beta}')^{-1})'I_p]'\delta \\
A &= \delta\hat{\beta}\hat{\beta}'(I_p + \delta\hat{\beta}\hat{\beta}')^{-1}
\end{aligned}$$

■

La proposición 2.4 establece que la clase de estimadores

$$d_\delta = \delta\hat{\beta}\hat{\beta}'(I + \delta\hat{\beta}\hat{\beta}')^{-1}\hat{\beta} \qquad \text{para } \delta \in [0, \infty) \tag{2.10}$$

son de mínima varianza dentro de cada clase de equivalencia.

Observe que, el estimador ridge tiene una menor distancia cuadrática entre todos los estimadores con una determinada suma de cuadrados de residuales, sin embargo, el estimador d_δ cuenta con varianza mínima entre estos estimadores, los cuales son transformaciones lineales del estimador de mínimos cuadrados. Para elegir el estimador con varianza total mínima se debe observar la clase C.
El siguiente resultado mostrará que el estimador d_δ pertenece a la clase de estimadores reducidos.

Proposición 2.5. *$d_\delta = c(f)$ para $f = \delta[\hat{\beta}'\hat{\beta} + (1 + \delta\hat{\beta}'\hat{\beta})^{-1}\delta(\hat{\beta}'\hat{\beta})^2]$ y por lo tanto d_δ es un estimador estocásticamente reducido.*

Lema 2.1. $(I_p + \delta\hat{\beta}\hat{\beta}')^{-1} = [I_p + (1 + \delta\hat{\beta}'\hat{\beta})^{-1}\delta\hat{\beta}\hat{\beta}']$

Demostración de la Proposición 2.5.

$$d_\delta = A\hat{\beta} = \delta\hat{\beta}\hat{\beta}'(I_p + \delta\hat{\beta}\hat{\beta}')^{-1}\hat{\beta}$$

Por el lema 2.1

$$\begin{aligned} d_\delta &= \delta\hat{\beta}\hat{\beta}'[I_p + (1+\delta\hat{\beta}'\hat{\beta})^{-1}\delta\hat{\beta}\hat{\beta}']\hat{\beta} \\ &= \delta[\hat{\beta}'\hat{\beta} + (1+\delta\hat{\beta}'\hat{\beta})^{-1}\delta(\hat{\beta}'\hat{\beta})^2]\hat{\beta} \end{aligned}$$

lo cual completa la demostración.

■

Aunque d_δ es un estimador estocásticamente reducido, en la práctica la diferencia entre d_δ y c_λ (estimador determinísticamente reducido) es mínima, por las siguientes dos razones:

i) Supongamos que la suma de cuadrados de residuales es fija y la clase de equivalencia está determinada, entonces si d_δ y c_λ pertenecen a la clase de equivalencia, se concluye que son idénticas.

ii) Si en la práctica el factor de reducción es elegido después de observar $\hat{\beta}$, entonces el estimador reducido que se utiliza es estocástico, siempre y cuando sea de la forma c_λ o d_δ.

Las proposiciones 2.4 y 2.5, muestran que los estimadores reducidos corresponden a una clase de equivalencia que tiene varianza mínima y por lo tanto son de máxima confianza para ser utilizados en problemas mal condicionados, así como también dichos estimadores pueden ser obtenidos sin elegir una norma a minimizar.

En la Figura 2.1 se muestra que el estimador reducido c_λ o d_δ en una clase de equivalencia dada, corresponde al punto en la elipse que se encuentra en la línea trazada del origen hasta el estimador de mínimos cuadrados $\hat{\beta}$, mientras que el estimador ridge $\hat{\beta}^*$, corresponde al punto en la elipse más cercano al origen, desde el punto de vista euclidiano. Observe, que tanto el estimador ridge como el estimador reducido en una clase de equivalencia dada, son más pequeños que el estimador de mínimos cuadrados, este factor es importante, ya que el estimador de mínimos cuadrados en un problema mal condicionado tiende a superar en longitud al vector real del parámetro.

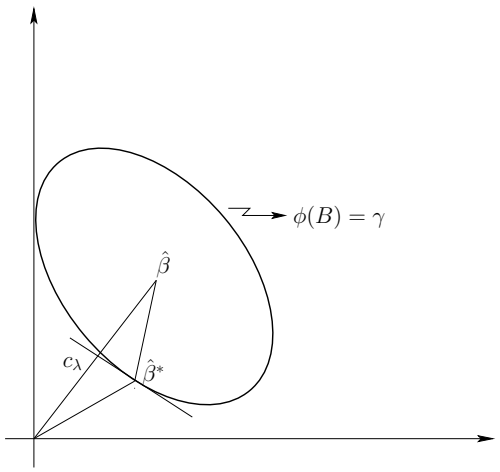


Figura 2.1: Representación geométrica de c_λ y b_k

La siguiente proposición muestra algunos resultados que son inmediatos de la expresión (2.10).

Proposición 2.6.

i) $d_\delta = 0$ *si* $\delta = 0$

ii) $d_\delta = \hat{\beta}$ *si* $\delta = 2^{-1/2}(\hat{\beta}'\hat{\beta})^{-1}$

iii) $|d_{\delta i}|$ *es monotonamente creciente en* δ

Una vez que se decide utilizar un estimador reducido surge el problema de cómo elegir el factor de reducción apropiado, se puede elegir una δ y usar d_δ o elegir una λ y usar c_λ o como alternativa, elegir la clase de equivalencia determinando un valor de la suma de cuadrados de residuales, en cuyo caso se determina el estimador reducido. Después de que $\hat{\beta}$ es observada, el problema de elegir un estimador reducido puede ser abordado de varias maneras. Primero, la δ en d_δ puede ser elegida graficando los elementos de d_δ contra δ y utilizando los criterios de estabilidad empleados por Hoerl y Kennard, o de forma alternativa se puede utilizar el "factor de inflación de varianza máxima", criterio proporcionado por Marquardt (1970) para decidir cuánto sesgo es permitido. Finalmente, se podrá utilizar un estimador reducido con

$$[1 + \xi s^2(\hat{\beta}'\hat{\beta})^{-1}]$$

como el factor de reducción, donde $s^2 = Y'Y - \hat{\beta}'(X'X)\hat{\beta}$, este factor es de especial interés, ya que la clase de estimadores

$$e_\xi = (e_{\xi 1} \dots e_{\xi p})' = [1 + \xi s^2(\hat{\beta}'\hat{\beta})^{-1}]\hat{\beta} \tag{2.11}$$

satisface la condición de admisibilidad más fuertemente que el estimador presentado anteriormente.

En particular, si

$$U(B) = E(B-\beta)'(X'X)(B-\beta)$$

denota el error cuadrático medio total ponderado del estimador B, se sigue la proposición dada por Sclove (1968), con base en resultados de James y Stein (1961).

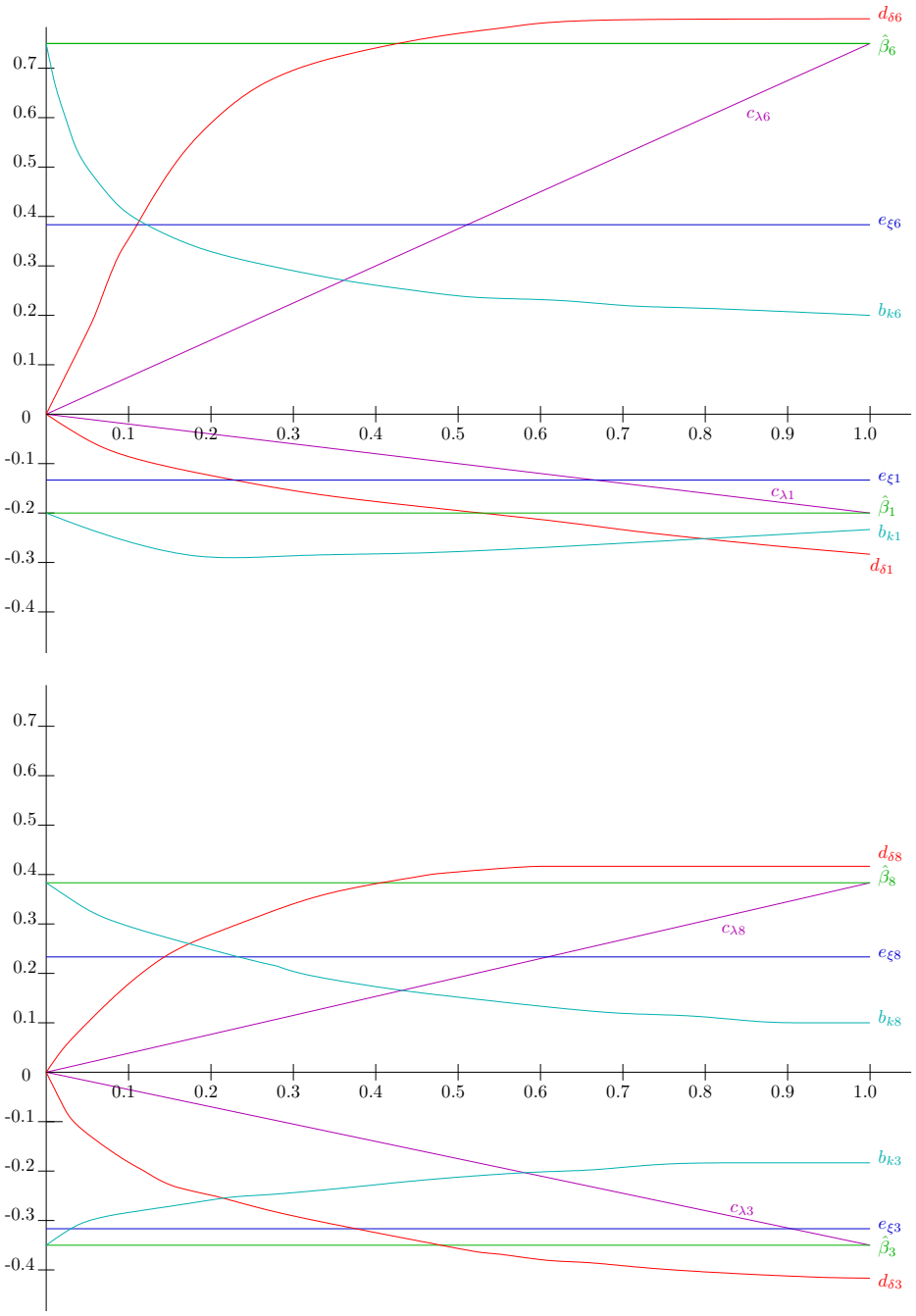
Proposición 2.7. *Si* $p \geqslant 3$ *y* $0 < \xi < 2(p-2)(n-p+2)^{-1}$ *entonces* $U(e_\xi) < U(\hat{\beta})$ *y si* $\xi_0 = (p-2)(n-p+2)^{-1}$ *entonces*

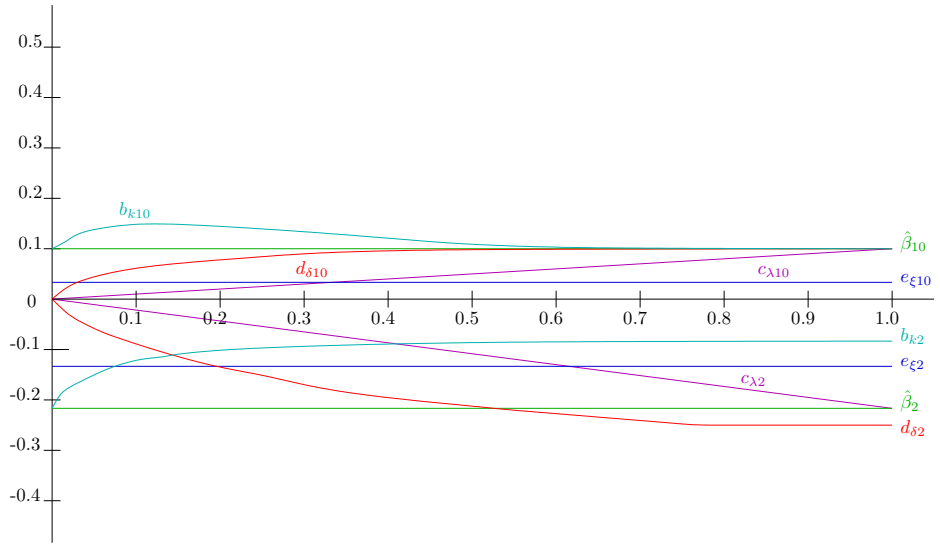
$$U(e_{\xi_0}) = min_\xi \; U(e_\xi)$$

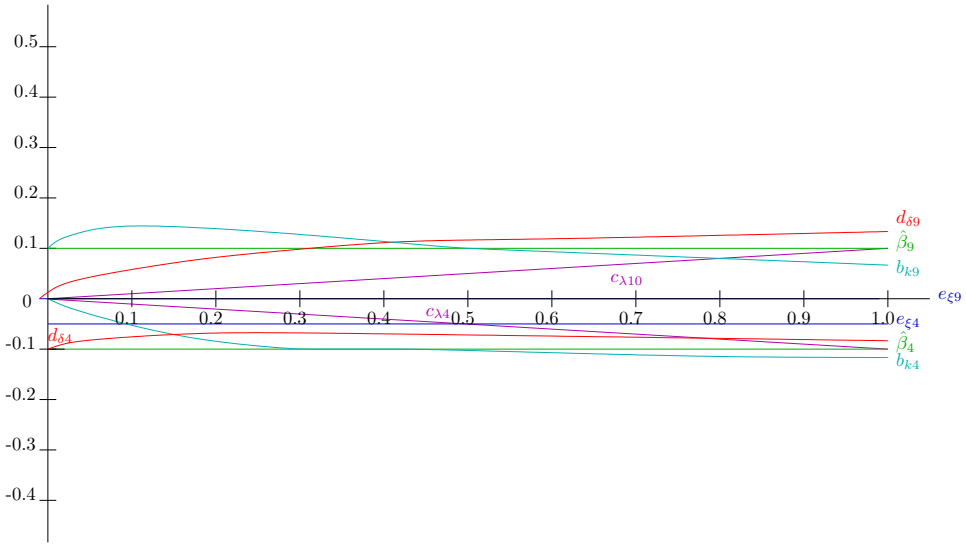
El estimador e_{ξ_0} en cierto sentido es superior a los estimadores ridge o a los estimadores determinísticamente reducidos, ya que un valor determinado de ξ puede garantizar un estimador con menor error cuadrático medio total ponderado que $\hat{\beta}$. Por lo tanto, la clase e_ξ es fuertemente admisible con respecto al error cuadrático medio ponderado, en el sentido de que se sabe con exatitud qué elementos son mejores (en términos del error cuadrático medio total ponderado), que los del estimador de mínimos cuadrados. Sclove también muestra que los estimadores son bastante complejos y por ende difíciles de utilizar en la práctica, ya que ni la distribución, ni los momentos son proporcionados, pero están garantizados para tener un error cuadrático medio total más pequeño que la varianza total de $\hat{\beta}$.

2.3. Ejemplo

Para mostrar el funcionamiento de los estimadores ridge b_k, de los estimadores reducidos c_λ, d_δ y e_ξ se han calculado los estimadores para 36 observaciones, tomando 10 variables de entrada de los datos de Gorman y Toman (1960), las cuales fueron utilizadas por Hoerl y Kennard para mostrar el uso de los estimadores ridge. Explicaron la matriz de correlación de las 10 variables de entrada, así como el rango de los valores propios (eigenvalores) de 3.692 a 0.068, concluyendo que el diseño está mal condicionado.




0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
b_{k10}
$\hat{\beta}_{10}$
$d_{\delta 10}$
$c_{\lambda 10}$
$e_{\xi 10}$
b_{k2}
$e_{\xi 2}$
$c_{\lambda 2}$
$\hat{\beta}_2$
$d_{\delta 2}$


0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
$d_{\delta 9}$
$\hat{\beta}_9$
b_{k9}
$c_{\lambda 10}$
$e_{\xi 9}$
$c_{\lambda 4}$
$e_{\xi 4}$
$d_{\delta 4}$
$\hat{\beta}_4$
b_{k4}

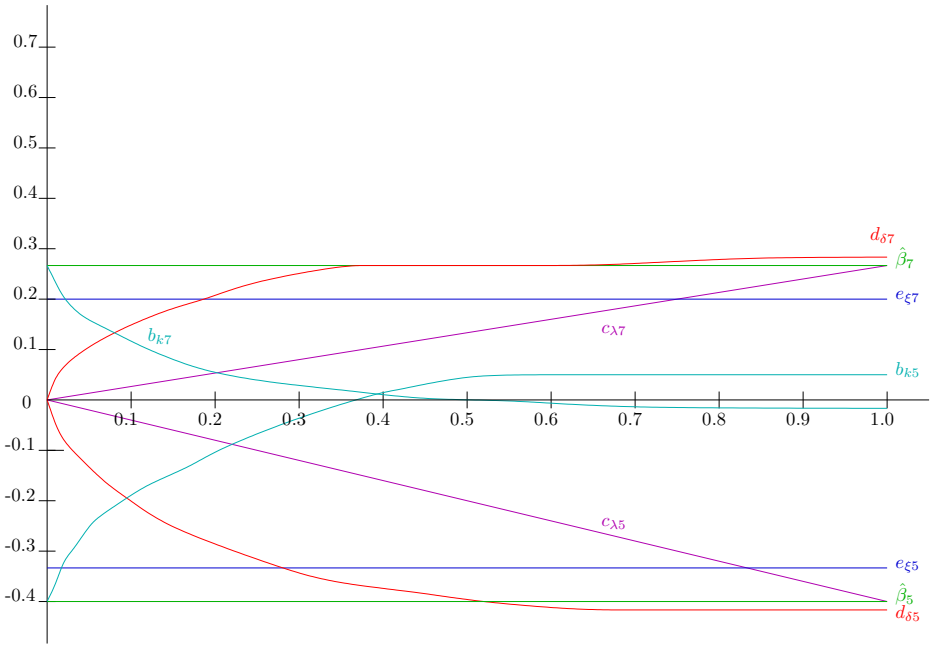


Figura 2.2: Comparación de estimadores sesgados en datos de Gorman y Toman

En la Figura 2.2 se presentan varias estimaciones de los coeficientes de regresión, que muestran diversas características del comportamiento de los estimadores.

a) El estimador insesgado de varianza mínima $\hat{\beta}$ fue calculado usando el programa de regresión BMD y cada $\hat{\beta}_i$ se muestra como una línea horizontal, ya que es una constante sobre la gráfica,

b) el estimador ridge b_k es calculado por la expresión

$$b_k = (b_{k1} \dots b_{kp})' = (X'X + kI_p)^{-1}X'Y$$

y cada b_{ki} se muestra en función de k,

c) el estimador determinísticamente reducido c_λ es calculado por la expresión (2.5) y cada $c_{\lambda i}$ se muestra en función de λ. Observe que $c_{\lambda i}$ es una función lineal simple de λ,

d) el estimador estocásticamente reducido d_δ es calculado por la expresión (2.10) y cada $d_{\delta i}$ se muestra en función de δ y

e) el estimador estocásticamente reducido e_ξ (con $\xi = (p-2)(n-p+2)^{-1}$) se muestra como una función constante.

Tanto el estimador ridge como el estimador estocásticamente reducido se estabilizan a medida que k o δ aumentan, por lo tanto, el análisis dado por Hoerl y Kennard puede

ser usado para elegir un buen valor de k o de δ. Si se va a utilizar un estimador reducido se recomienda usar d_δ y elegir la más pequeña δ para la cual d_δ se estabilice, o el uso del factor de reducción propuesto por Sclove. Por otro lado, el estimador determinísticamente reducido c_λ es lineal en λ y por lo tanto no es estable.
Observe, que el estimador reducido e_ξ con frecuencia corresponde a un estimador ridge, cuyo valor de k es muy diferente al valor en el cual el estimador ridge se estabiliza.

Los estimadores ridge y los estimadores reducido también pueden ser comparados equiparando sus varianzas y despejando λ en términos de k, obteniendo

$$\lambda^2(k) = \frac{traza((X'X)_k^{-1}(X'X)(X'X)_k^{-1})}{traza(X'X)^{-1}}$$

y entonces se compara $c_{\lambda(k)}$ con b_k. Este análisis fue realizado, pero no se presenta debido a que no muestra ningún conocimiento sobre el comportamiento de alguno de los estimadores.

Capítulo 3

Regresión Ridge: Estimación Sesgada para Problemas no Ortogonales

Un estudio de las propiedades del estimador ridge ($\hat{\beta}^*$), muestra que se puede utilizar para mejorar la estimación del error cuadrático medio y se amplía la magnitud de esa mejora, con un incremento en la dispersión de los valores propios. Como ya se sabe, una estimación basada en $\hat{\beta}^*$ es sesgada y el uso de un estimador sesgado implica una cota particular en el vector de regresión β. Sin embargo, los datos en cualquier problema en particular, tienen información que puede mostrar la clase de β generadores que sean razonables. Por lo que en este capítulo se aborda más a detalle el uso de la regresión ridge y de su traza, esta última teniendo como propósito presentar dicha información de manera explícita y de ahí guiar al usuario a un mejor estimador, $\hat{\beta}^*$.

3.1. Regresión ridge

Los estimadores ridge propuestos por Hoerl y Kennard son más confiables que los estimadores de mínimos cuadrados en presencia de una matriz mal condicionada. Para cualquier $k \, \epsilon \, [0, \infty)$ el estimador ridge correspondiente se define como:

$$\begin{aligned} \hat{\beta}^* &= [X'X + kI_p]^{-1} X'Y; \qquad k \geqslant 0 \\ &= WX'Y \end{aligned}$$

En general, existe un valor óptimo de k para cualquier problema, pero es deseable examinar la solución ridge en un rango de valores admisibles de k. Recuerde que, admisibles significa tener errores cuadráticos medios más pequeños en los parámetros, que la solución de mínimos cuadrados y cabe mencionar que el error de cuadrático medio para predicciones futuras también se reduce proporcionalmente.

La relación de un estimador Ridge con un estimador ordinario está dado por:

$$\begin{aligned} \hat{\beta}^* &= [I_p + k(X'X)^{-1}]^{-1} \hat{\beta} \\ &= Z\hat{\beta} \end{aligned}$$

ya que:

$$\begin{aligned}\hat{\beta}^* &= [X'X + kI_p]^{-1}X'Y \\ &= [(X'X)(I_p + k(X'X)^{-1})]^{-1}X'Y \\ &= [I_p + k(X'X)^{-1}]^{-1}(X'X)^{-1}X'Y \\ &= [I_p + k(X'X)^{-1}]^{-1}\hat{\beta}\end{aligned}$$

En el capítulo anterior una de las transformaciones lineales de $\hat{\beta}$ se denotó como $A_0 = (k(X'X)^{-1} + I_p)^{-1}$, la cual es utilizada para establecer la relación anterior.

Hoerl le dio a su procedimiento el nombre de regresión ridge debido a la similitud de los métodos matemáticos con en el análisis ridge, para representar gráficamente las características de las ecuaciones de superficie de respuesta de segundo orden, que resultan de utilizar muchas variables predictoras.

Las propiedades fundamentales aplicables a la regresión ridge son:

a) Si $\hat{\beta}^*$ es la solución de $(X'X + kI_p)\hat{\beta}^* = g$ (donde $g = X'Y$), entonces $\hat{\beta}^*$ minimiza la suma de cuadrados de residuales en la esfera con centro en el origen y radio la longitud de $\hat{\beta}^*$. La suma de cuadrados de residuales es una función creciente de k.

b) La longitud de $\hat{\beta}^*$ es una función decreciente de k.

c) El ángulo γ entre la solución ridge $\hat{\beta}^*$ y el vector gradiente g es una función decreciente de k.

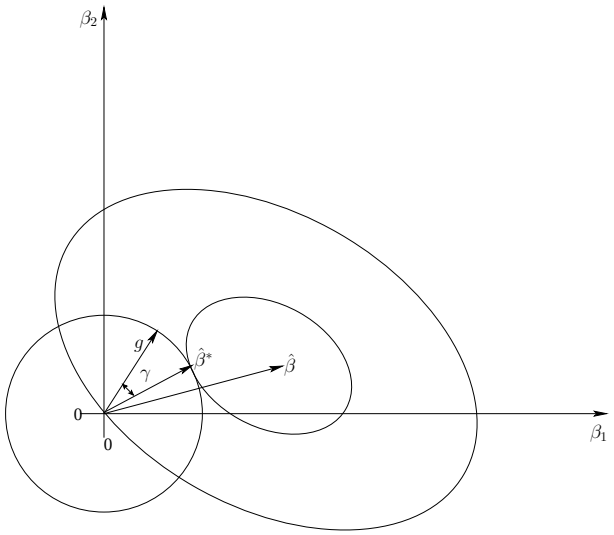


Figura 3.1: Geometría de la regresión ridge

En la Figura 3.1 se ilustra la geometría de la regresión ridge para un problema teórico, involucrando solamente dos parámetros, β_1 y β_2. El punto $\hat{\beta}$ en el centro de la elipse es la solución de mínimos cuadrados y en este punto la suma de cuadrados de residuales, ϕ, alcanza su mínimo. En la elipse pequeña, ϕ es constante con un cierto valor más grande en comparación con el mínimo y el estimador ridge, $\hat{\beta}^*$, es el vector más pequeño que proporcionará una suma de cuadrados de residuales tan pequeña como el valor de ϕ en cualquier lugar de la elipse. El círculo con centro en el origen es tangente a la elipse en $\hat{\beta}^*$, encontrándose siempre el estimador ridge entre $\hat{\beta}$ y g, siendo este gradiente perpendicular con el contorno ϕ a través del origen. Con respecto al ángulo γ, se puede decir que frecuentemente es tan pequeño como k.

Es importante tener en cuenta que el factor de inflación de varianza (FIV), es una medida de cómo los valores propios más pequeños se acercan a cero. Hoerl y Kennard mencionan que cuando en problemas no ortogonales la estimación por mínimos cuadrados es sensible a un número de errores, la longitud cuadrada esperada del vector de coeficientes es:

$$\begin{aligned} E(\hat{\beta}'\hat{\beta}) &= \beta'\beta + \sigma^2 traza(X'X)^{-1} \\ &> \beta'\beta + \sigma^2/\lambda_{min} \end{aligned}$$

Por lo tanto el vector de coeficientes de mínimos cuadrados, $\hat{\beta}$, es mucho más grande en promedio, para datos mal condicionados, ya que $\lambda_{min} < 1$. La solución de mínimos cuadrados produce coeficientes cuyos valores absolutos son demasiado grandes y cuyos signos pueden invertirse con cambios insignificantes en los datos.

Algunas propiedades de $\hat{\beta}^*$, W y Z que se utilizarán son:

i) Sea $\xi_i(W)$ y $\xi_i(Z)$ valores propios de W y Z respectivamente.
Entonces:

$$\begin{aligned} \xi_i(W) &= \frac{1}{(\lambda_i + k)} \\ \xi_i(Z) &= \frac{\lambda_i}{(\lambda_i + k)} \end{aligned}$$

donde λ_i son los valores propios de $X'X$.

Demostración.

Las propiedades de vectores y valores propios que se utilizan para la demostración se pueden consultar en la sección de apéndices.

$$P.D. \qquad \xi_i(W) = \frac{1}{(\lambda_i + k)}$$

$$\text{donde } W = [X'X + kI]^{-1}$$

Utilizando la *Propiedad 1*, se tiene que el i-ésimo valor propio de $[X'X + kI_p]$ es $\lambda_i + k$.
Ahora utilizando la *Propiedad 2*, se tiene que el i-ésimo valor propio de W es $\dfrac{1}{\lambda_i + k}$.

$$P.D. \qquad \xi_i(Z) = \frac{\lambda_i}{(\lambda_i + k)}$$

$$\text{donde } Z = [I_p + k(X'X)^{-1}]^{-1}$$

Utilizando la *Propiedad 1*, se tiene que el i-ésimo valor propio de $[I_p + k(X'X)^{-1}]$ es $\dfrac{\lambda_i + k}{\lambda_i}$.
Ahora utilizando la *Propiedad 2*, se tiene que el i-ésimo valor propio de Z es $\left(\dfrac{\lambda_i + k}{\lambda_i}\right)^{-1} = \dfrac{\lambda_i}{\lambda_i + k}$

■

ii) $Z = I_p - k(X'X + kI_p)^{-1} = I_p - kW$

$$Z = [I_p + k(X'X)^{-1}]^{-1}$$

Demostración.

$$\begin{aligned} Z &= [(X'X)^{-1}(X'X) + k(X'X)^{-1}]^{-1} \\ &= [(X'X)^{-1}(X'X + kI_p)]^{-1} \\ &= [(X'X) + kI_p]^{-1}[(X'X)^{-1}]^{-1} \\ &= [X'X + kI_p]^{-1}(X'X) \\ &= W(X'X) \end{aligned}$$

$$\text{P.D} \qquad Z = I_p - kW = WX'X$$

Se tiene que:

$$\begin{aligned} W &= [X'X + kI_p]^{-1} \\ W^{-1} &= \left[[X'X + kI_p]^{-1}\right]^{-1} \\ &= [X'X + kI_p] \end{aligned}$$

entonces como:

$$\begin{aligned} X'X &= X'X \\ X'X + kI_p - kI_p &= X'X \\ (X'X + kI_p) - kI_p &= X'X \\ W^{-1} - kI_p &= X'X \\ W^{-1} - k(W^{-1}W) &= X'X \\ W^{-1}(I_p - kW) &= X'X \\ WW^{-1}(I_p - kW) &= WX'X \\ I_p - kW &= WX'X \end{aligned}$$

$$\therefore \quad Z = I_p - kW = WX'X$$

■

iii) $\hat{\beta}^*$ para $k \neq 0$ es más pequeño que $\hat{\beta}$, es decir:

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

Por definición $\hat{\beta}^* = Z\hat{\beta}$. De esta definición y de los supuestos de $X'X$, Z es claramente una matriz simétrica y definida positiva. Entonces se tiene la siguiente relación:

$$(\hat{\beta}^*)'(\hat{\beta}^*) \leq \xi^2_{max}(Z)\hat{\beta}'\hat{\beta}$$

Pero $\xi_{max}(Z) = \dfrac{\lambda_1}{(\lambda_1 + k)}$ donde λ_1 es el valor propio más grande de $X'X$. De $\xi_i(Z)$ y de Z se observa que $Z = I$ cuando $k = 0$ y que Z se aproxima a cero cuando $k \to \infty$.
Para un estimador $\hat{\beta}^*$ la suma de cuadrados de residuales es:

$$\phi^*(k) = (Y - X\hat{\beta}^*)'(Y - X\hat{\beta}^*)$$

la cual puede ser escrita en la forma:

$$\phi^*(k) = Y'Y - (\hat{\beta}^*)'X'Y - k(\hat{\beta}^*)'(\hat{\beta}^*)$$

ya que:

$$\begin{aligned} \phi^*(k) &= (Y - X\hat{\beta}^*)'(Y - X\hat{\beta}^*) \\ &= (Y' - \hat{\beta}^{*'}X')(Y - X\hat{\beta}^*) \\ &= Y'Y - Y'X\hat{\beta}^* - \hat{\beta}^{*'}X'Y + \hat{\beta}^{*'}X'X\hat{\beta}^* \\ &= Y'Y - 2\hat{\beta}^{*'}X'Y + \hat{\beta}^{*'}X'Y \end{aligned}$$

y finalmente

$$\phi^*(k) = Y'Y - \hat{\beta}^{*'}X'Y - k(\hat{\beta}^*)'(\hat{\beta}^*)$$

donde $k(\hat{\beta}^*)'(\hat{\beta}^*)$ es una modificación que depende de la longitud cuadrada de $\hat{\beta}^*$. Obsérvese que si $k = 0$ se tiene el modelo visto en regresión lineal múltiple.

R. W. Kennard observó que el estimador ridge tiene una importante interpretación Bayesiana, por lo que hace énfasis en que los mínimos cuadrados suponen la hipótesis de una distribución uniforme no acotada sobre el vector de coeficientes, lo cual puede ser usado en lugar del requisito de insesgamiento en la obtención del estimador de mínimos cuadrados. Es importante tener en cuenta que cuando se selecciona la cantidad de sesgo se utilizan tanto las variables de predicción, como la variable de respuesta, ambas estandarizadas en forma de correlación. El estimador ridge es equivalente a poner requisitos mínimos de acotación en el vector de coeficientes, el cual es finito.
Theobald generaliza las condiciones bajo las cuales la regresión ridge es conocida por hacer una distancia cuadrada esperada más pequeña que en mínimos cuadrados. También se sabe que la mejora esperada de la regresión ridge sobre mínimos cuadrados depende de la orientación del verdadero vector de regresión, relativo a los ejes principales, definidos por los vectores propios de la matriz $X'X$, la mejora esperada es mayor cuando la orientación de β coincide con el vector propio asociado con el mayor valor propio de $X'X$.

3.2. Traza ridge

Cuando la matriz de correlación de las variables de predicción contiene varios coeficientes de correlación grandes, es difícil esclarecer la relación que existe entre las variables de predicción mediante el estudio de los coeficientes de correlación simple. Algunos procedimientos automáticos, como la búsqueda paso a paso, la selección por mejores subconjuntos y la regresión PRESS, tratan de esclarecer la relación entre las variables, mediante la selección de algunos de los mejores subconjuntos de predicción. Sin embargo, estos métodos no proporcionan realmente una idea de la estructura, ni de la sensibilidad de los resultados para el conjunto particular de datos obtenidos. En el problema de regresión con diez variables publicado por Groman y Toman, Hoerl y Kennard mostraron que el procedimiento de selección por mejores subconjuntos, no reduce necesariamente las correlaciones de las variables de pedicción, además las correlaciones pueden ser más grandes que entre las variables originales.

Una de las grandes ventajas de la regresión ridge es que la visualización gráfica, llamada traza ridge, puede ayudar al analista a considerar qué coeficientes son sensibles a los datos. Por lo tanto, el análisis de sensibilidad es uno de los objetivos de la regresión ridge. La traza ridge es una gráfica del valor de cada coeficiente en función de k y tendrá una curva asignada por coeficiente, por lo que, para un mejor análisis se recomienda que estén trazadas menos de diez curvas en un gráfico dado. La varianza de un coeficiente

es una función decreciente y el sesgo es una función creciente con respecto a k, por lo que, a medida que aumenta k, el coeficiente del error cuadrático medio (varianza más sesgo cuadrado) decrece a un mínimo y luego aumenta. El objetivo es encontrar un valor de k que proporcione un conjunto de coeficientes con menor error cuadrático medio que la solución de mínimos cuadrados. Note que a medida que aumente k, la suma de cuadrados de residuales también aumentará, lo cual no debe ser de gran preocupación, porque el objetivo no es obtener el ajuste más cercano a los datos de estimación, sino el desarrollar un conjunto "estable" de coeficientes, que harán un buen trabajo de predicción en observaciones futuras. Por estable se entiende que los coeficientes no son sensibles a pequeños cambios en los datos de estimación. Si las variables de predicción están altamente correlacionadas, los coeficientes cambiarán rápidamente para valores pequeños de k y gradualmente se estabilizarán en los valores más grandes de k, así que, el valor de k en el que los coeficientes se han estabilizado proporciona el deseado conjunto de coeficientes, pero si las variables de predicción son ortogonales, entonces los coeficientes cambiarían muy poco (es decir, los coeficientes ya son estables), indicando que la solución de mínimos cuadrados es un buen conjunto de coeficientes.

3.2.1. Características de la traza ridge

En el capítulo anterior, se definió la suma de cuadrados de residuales de un estimador B como:

$$\begin{aligned}\phi &= (Y - XB)'(Y - XB) \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (B - \hat{\beta})'X'X(B - \hat{\beta}) \\ &= \phi_{min} + \phi(B)\end{aligned} \tag{3.1}$$

Desde el punto de vista gráfico, los contornos de la constante ϕ son las superficies de las elipses con centro en $\hat{\beta}$, el estimador de mínimos cuadrados ordinarios de β. Observe que el valor de ϕ es el valor mínimo, ϕ_{min}, más el valor de la forma cuadrática $(B - \hat{\beta})$ y que existe una continuidad de los valores de B_0 que satisface la relación $\phi = \phi_{min} + \phi_0$, donde $\phi_0 > 0$ es un incremento fijo. Sin embargo, la distancia esperada de $\hat{\beta}$ a β tenderá a ser grande si $X'X$ tiene un valor propio pequeño.

La traza ridge se mostrará a través de la suma de cuadrados de la superficie, así que para un ϕ fijo se elige un valor único de B, el cual es el de menor tramaño. Ésto se puede definir de la siguiente manera:
Minimizar F, donde

$$F = B'B + (1/k)[(B - \hat{\beta})'X'X(B - \hat{\beta}) - \phi_0]$$

donde $(1/k)$ es el multiplicador.

Se tiene:

$$\begin{aligned}
F &= B'B + (1/k)[(B-\hat{\beta})'X'X(B-\hat{\beta}) - \phi_0] \\
&= B'B + (1/k)[(B'-\hat{\beta}')X'X(B-\hat{\beta}) - \phi_0] \\
&= B'B + (1/k)[(B'X'-\hat{\beta}'X')(XB-X\hat{\beta}) - \phi_0] \\
&= B'B + (1/k)[B'X'XB - B'X'X\hat{\beta} - \hat{\beta}'X'XB + \hat{\beta}'X'X\hat{\beta} - \phi_0] \\
&= B'B + (1/k)B'X'XB - (1/k)B'X'X\hat{\beta} - (1/k)\hat{\beta}'X'XB + (1/k)\hat{\beta}'X'X\hat{\beta} - (1/k)\phi_0
\end{aligned}$$

Entonces:

$$\begin{aligned}
\frac{\partial F}{\partial B} &= 2B + 2(1/k)X'XB - (1/k)(X'X)\hat{\beta} - (1/k)(X'X)\hat{\beta} \\
\Rightarrow \quad & 2B + \frac{1}{k}\left[2(X'X)B - 2(X'X)\hat{\beta}\right] = 0
\end{aligned}$$

Por lo que:

$$\begin{aligned}
2\left[B + (1/k)\left[(X'X)B - (X'X)\hat{\beta}\right]\right] &= 0 \\
B + (1/k)\left[(X'X)B - (X'X)\hat{\beta}\right] &= 0 \\
B + (1/k)(X'X)B - (1/k)(X'X)\hat{\beta} &= 0 \\
[I + (1/k)(X'X)]B - (1/k)(X'X)\hat{\beta} &= 0
\end{aligned}$$

$$\begin{aligned}
[I + (1/k)(X'X)]B &= (1/k)(X'X)\hat{\beta} \\
[I + (1/k)(X'X)]^{-1}[I + (1/k)(X'X)]B &= [I + (1/k)(X'X)]^{-1}[(1/k)(X'X)\hat{\beta}] \\
B &= [I + (1/k)(X'X)]^{-1}[(1/k)(X'X)\hat{\beta}] \\
&= [I + (1/k)(X'X)]^{-1}(1/k)(X'X)[(X'X)^{-1}X'Y] \\
&= [I + (1/k)(X'X)]^{-1}(1/k)X'Y \\
&= k[kI + (X'X)]^{-1}(1/k)X'Y \\
&= [kI + (X'X)]^{-1}X'Y
\end{aligned}$$

$$\therefore \quad B = \hat{\beta}^* = [(X'X) + kI]^{-1}X'Y$$

A esta expresión se le da el nombre de **estimador ridge**. Hay que tener en cuenta que en la práctica es más fácil elegir una $k \geqslant 0$ y después calcular ϕ_0.
En términos de $\hat{\beta}^*$ la suma de cuadrados de residuales es:

$$\begin{aligned}
\phi^*(k) &= (Y - X\hat{\beta}^*)'(Y - X\hat{\beta}^*) \\
&= \phi_{min} + k^2\hat{\beta}^{*'}(X'X)^{-1}\hat{\beta}^*
\end{aligned}$$

Por lo tanto, si la distancia al cuadrado del vector de regresión B se fija en R^2, entonces se dice que el valor de B es $\hat{\beta}^*$ y así se obtiene la suma mínima de cuadrados. Es decir, $\hat{\beta}^*$ es el valor de B que minimiza la función:

$$F_1 = (Y - XB)'(Y - XB) + (1/k)(B'B - R^2)$$

3.2.2. Características de probabilidad de la traza ridge

Tomando en cuenta la hipótesis de que el vector de errores presenta una distribución normal $(0, \sigma^2 I_n)$ con

$$\Theta = \{(\beta_0, \beta_1, \beta_2, \ldots, \beta_p, \sigma^2) \mid \beta_i \in \mathbb{R}; i = 1, 2, \ldots, p; 0 < \sigma^2 < \infty\}$$

entonces la función de probabilidad conjunta es:

$$\begin{aligned} L(\Theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\{(1/2\sigma^2)(Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip})^2\}} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{\{-(1/2\sigma^2)(Y - X\beta)'(Y - X\beta)\}} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{\{-(1/2\sigma^2)(Y'Y - 2\beta' X'Y + \beta' X'X\beta)\}} \end{aligned}$$

El kernel de esta función es la forma cuadrática de la exponencial, que puede ser escrito de la siguiente manera:

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

La expresión (3.1) muestra que un incremento en la suma de cuadrados de residuales es equivalente a un decremento en el valor de la función de probabilidad. Así que los contornos de igual probabilidad también se encuentran en la superficie de las elipses con centro en $\hat{\beta}$.

La traza ridge puede ser interpretada como una trayectoria a través del espacio de probabilidad y surge la pregunta de por qué esta trayectoria en particular puede ser de especial interés. El razonamiento es el mismo que para la suma de cuadrados. A pesar de que los vectores grandes proporcionen los mismos valores de probabilidad que los vectores más cortos, no tendrán siempre el mismo significado físico. Lo que implica una restricción en los valores posibles de $\hat{\beta}$, que no es evidente en la formulación del modelo lineal general dado al inicio, por lo que esta implicación será analizada en la siguiente sección.

3.3. Propiedades del error cuadrático medio de la regresión ridge

3.3.1. Sesgo y varianza de un estimador ridge

Para estudiar a $\hat{\beta}^*$ desde el punto de vista del error cuadrático medio es necesario obtener una expresión para $E[L_1^2(k)]$, por lo que se plantea la siguiente igualdad:

$$\begin{aligned} E[L_1^2(k)] &= E[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)] \\ &= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^2} + k^2 \beta'(X'X + kI)^{-2}\beta \\ &= \gamma_1(k) + \gamma_2(k) \end{aligned} \tag{3.2}$$

Demostración.

$$E[L_1^2(k)] = E[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)]$$

Como $\hat{\beta}^* = Z\hat{\beta}$, entonces se sustituye en la igualdad anterior

$$\begin{aligned} E[L_1^2(k)] &= E[(Z\hat{\beta} - \beta)'(Z\hat{\beta} - \beta)] \\ &= E[(Z\hat{\beta} - Z\beta + Z\beta - \beta)'(Z\hat{\beta} - Z\beta + Z\beta - \beta)] \\ &= E\left[[(Z\hat{\beta} - Z\beta)' + (Z\beta - \beta)'][(Z\hat{\beta} - Z\beta) + (Z\beta - \beta)]\right] \\ &= E[(Z\hat{\beta} - Z\beta)'(Z\hat{\beta} - Z\beta) + (Z\hat{\beta} - Z\beta)'(Z\beta - \beta) \\ &\quad + (Z\beta - \beta)'(Z\hat{\beta} - Z\beta) + (Z\beta - \beta)'(Z\beta - \beta)] \\ &= \underbrace{E[(Z\hat{\beta} - Z\beta)'(Z\hat{\beta} - Z\beta)]}_{A} + \underbrace{E[(Z\hat{\beta} - Z\beta)'(Z\beta - \beta)]}_{B} \\ &\quad + \underbrace{E[(Z\beta - \beta)'(Z\hat{\beta} - Z\beta)]}_{C} + \underbrace{E[(Z\beta - \beta)'(Z\beta - \beta)]}_{D} \end{aligned}$$

desarrollando A se obtiene:

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

desarrollando B se obtiene:

$$\begin{aligned} E[(Z\hat{\beta} - Z\beta)'(Z\beta - \beta)] &= E\left[[(Z\hat{\beta})' - (Z\beta)'](Z\beta - \beta)\right] \\ &= E[(\hat{\beta}'Z' - \beta'Z')(Z\beta - \beta)] \\ &= E[(\hat{\beta}'Z' - \beta'Z')(Z\beta - \beta)] \\ &= [E(\hat{\beta}')Z' - \beta'Z'][Z\beta - \beta] \\ &= (\beta'Z' - \beta'Z')(Z\beta - \beta) \\ &= 0 \end{aligned}$$

desarrollando C se obtiene:

$$\begin{aligned} E[(Z\beta-\beta)'(Z\hat{\beta}-Z\beta)] &= E[(\beta'Z'-\beta')(Z\hat{\beta}-Z\beta)] \\ &= (\beta'Z'-\beta')E[Z\hat{\beta}-Z\beta] \\ &= (\beta'Z'-\beta')[E(Z\hat{\beta})-Z\beta] \\ &= (\beta'Z'-\beta')[Z\beta-Z\beta] \\ &= 0 \end{aligned}$$

desarrollando D se obtiene:

$$E[(Z\beta-\beta)'(Z\beta-\beta)] = (Z\beta-\beta)'(Z\beta-\beta)$$

por lo anterior se tiene:

$$E[L_1^2(k)] = E[(\hat{\beta}-\beta)'Z'Z(\hat{\beta}-\beta)] + (Z\beta-\beta)'(Z\beta-\beta)$$

A continuación se demostrará que $E[(\hat{\beta}-\beta)'Z'Z(\hat{\beta}-\beta)] = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+k)^2}$.

Como:

$$\begin{aligned} E(Q) &= E(\hat{\beta}-\beta) = 0 \\ Var(Q) &= \sigma^2(X'X)^{-1} \end{aligned}$$

Entonces, si $A = Z'Z$ una matriz de tamaño pxp.

$$\begin{aligned} E[(\hat{\beta}-\beta)'Z'Z(\hat{\beta}-\beta)] &= traza[A\sigma^2(X'X)^{-1}] \\ &= \sigma^2 traza[A(X'X)^{-1}] \\ &= \sigma^2 traza[Z'Z(X'X)^{-1}] \\ &= \sigma^2 traza[Z(X'X)^{-1}Z'] \end{aligned}$$

Del inciso ii) de la sección 3.1 se tiene que:

$$\begin{aligned} Z' &= (X'X)[(X'X)+kI]^{-1} \\ Z &= I - k(X'X+kI)^{-1} \end{aligned}$$

Entonces:

$$\begin{aligned}
E[(\hat{\beta}-\beta)'Z'Z(\hat{\beta}-\beta)] &= \sigma^2 traza[Z(X'X)^{-1}(X'X)[X'X+kI]^{-1}] \\
&= \sigma^2 traza[Z(X'X+kI)^{-1}] \\
&= \sigma^2 traza[[I-k(X'X+kI)^{-1}][X'X+kI]^{-1}] \\
&= \sigma^2 traza[[X'X+kI]^{-1}-k[X'X+kI]^{-1}[X'X+kI]^{-1}] \\
&= \sigma^2 traza[[X'X+kI]^{-1}-k[X'X+kI]^{-2} \\
&= \sigma^2\left[traza[X'X+kI]^{-1}-ktraza[X'X+kI]^{-2}\right] \\
&= \sigma^2\left[\sum_{i=1}^{p}\frac{1}{\lambda_i+k}-k\sum_{i=1}^{p}\left(\frac{1}{\lambda_i+k}\right)^2\right] \\
&= \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2}
\end{aligned}$$

Ahora se demostrará que $(Z\beta-\beta)'(Z\beta-\beta)=k^2\beta'(X'X+kI)^{-2}\beta$, utilizando que

$$Z=I-k(X'X+kI)^{-1}$$

Entonces:

$$\begin{aligned}
(Z\beta-\beta)'(Z\beta-\beta) &= \beta'(Z-I)'(Z-I)\beta \\
&= \beta'\left[I-k(X'X+kI)^{-1}-I\right]'\left[I-k(X'X+kI)^{-1}-I\right]\beta \\
&= \beta'\left[-k(X'X+kI)^{-1}\right]'\left[-k(X'X+kI)^{-1}\right]\beta \\
&= \beta'\left[-k\left[(X'X+kI)^{-1}\right]'\right]\left[-k(X'X+kI)^{-1}\right]\beta \\
&= \beta'\left[-k\left[(X'X+kI)'\right]^{-1}\right]\left[-k(X'X+kI)^{-1}\right]\beta \\
&= \beta'\left[-k(X'X+kI)^{-1}\right]\left[-k(X'X+kI)^{-1}\right]\beta \\
&= \beta'\left[k^2(X'X+kI)^{-2}\right]\beta \\
&= k^2\beta'(X'X+kI)^{-2}\beta
\end{aligned}$$

$$\begin{aligned}
\therefore \quad E[L_1^2(k)] &= \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2}+k^2\beta'(X'X+kI)^{-2}\beta \\
&= \gamma_1(k)+\gamma_2(k)
\end{aligned}$$

■

El segundo elemento de la expresión anterior, $\gamma_2(k)$, es la distancia al cuadrado de $Z\beta$ a β y tendrá un valor nulo cuando k sea igual a cero, ya que Z será equivalente a la matriz identidad. Entonces, $\gamma_2(k)$ puede ser considerado como el sesgo al cuadrado, cuando $\hat{\beta}^*$ sea utilizado en lugar de $\hat{\beta}$. El primer término, $\gamma_1(k)$, puede ser mostrado como la suma de varianzas (varianza total) de los parámetros estimados. Por otra parte, en términos de la variable aleatoria Y,

$$\hat{\beta}^* = Z\hat{\beta} = Z(X'X)^{-1}X'Y$$

entonces

$$\begin{aligned} Var(\hat{\beta}^*) &= Z(X'X)^{-1}X'Var(Y)X(X'X)^{-1}Z' \\ &= \sigma^2 Z(X'X)^{-1}Z' \end{aligned} \tag{3.3}$$

Note que la suma de varianzas de todas las $\hat{\beta}_i^*$, son la suma de los elementos de la diagonal de la igualdad (3.3).

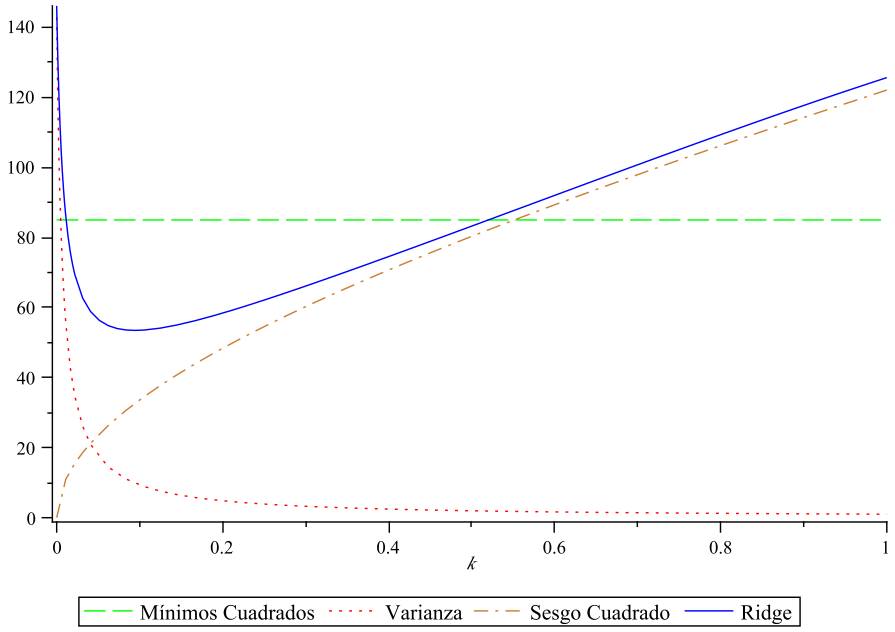


Figura 3.2: Funciones del error cuadrático medio

En la Figura 3.2 se muestra de forma cualitativa la relación entre las varianzas, el sesgo al cuadrado y el parámetro k. Observe que a medida que aumenta el valor de k, la varianza total decrece, mientras que con el sesgo cuadrado ocurre lo contrario, ya que éste crece a medida que k aumenta. En la gráfica se muestra una línea continua, que es la suma de $\gamma_1(k)$ y $\gamma_2(k)$ y por lo tanto es $E[L_1^2(k)]$, existe la posibilidad de que haya valores de k para los cuales el error cuadrático medio sea menor para $\hat{\beta}^*$ que para la solución usual de $\hat{\beta}$. Esta posibilidad es apoyada por las propiedades de $\gamma_1(k)$ y $\gamma_2(k)$.

La función $\gamma_1(k)$ es una función monótonamente decreciente en k, mientras que $\gamma_2(k)$ es monótonamente creciente. Sin embargo, la característica más significativa es el valor de la derivada de cada función alrededor del origen.

Los valores de estas derivadas son:

$$\lim_{k\to 0^+}\left(\frac{\partial\gamma_1}{\partial k}\right) = -2\sigma^2\sum_{i=1}^{p}(1/\lambda_i^2)$$

$$\lim_{k\to 0^+}\left(\frac{\partial\gamma_2}{\partial k}\right) = 0$$

Demostración.

$$\text{Se tiene que } \gamma_1(k) = \sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^2}$$

entonces

$$\begin{aligned}\frac{\partial\gamma_1}{\partial k} &= \sum_{i=1}^{p}\frac{(\lambda_i+k)^2(0)-\sigma^2\lambda_i[2(\lambda_i+k)]}{((\lambda_i+k)^2)^2}\\ &= \sum_{i=1}^{p}\frac{-2\sigma^2\lambda_i(\lambda_i+k)}{(\lambda_i+k)^4}\\ &= -2\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3}\end{aligned}$$

por lo tanto

$$\begin{aligned}\lim_{k\to 0^+}\frac{\partial\gamma_1}{\partial k} &= \lim_{k\to 0^+} -2\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3}\\ &= -2\sigma^2\sum_{i=1}^{p}\lim_{k\to 0^+}\frac{\lambda_i}{(\lambda_i+k)^3}\\ &= -2\sigma^2\sum_{i=1}^{p}\frac{\lambda_i}{\lambda_i^3}\\ &= -2\sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i^2}\end{aligned}$$

La demostración de que $\lim_{k\to 0^+}\left(\frac{\partial\gamma_2}{\partial k}\right) = 0$ se verá mas adelante.

■

Observe que $\gamma_1(k)$ tiene una derivada negativa, la cual se aproxima a $-2p\sigma^2$ cuando $k \to 0^+$ para una matriz $X'X$ ortogonal y se aproxima a $-\infty$ cuando $X'X$ esta mal condicionada y $\lambda_p \to 0$. Por otro lado cuando $k \to 0^+$, el $\lim_{k\to 0^+}\frac{\partial\gamma_2}{\partial k}$ muestra que $\gamma_2(k)$ es

monótona y vale cero en el origen. Con estas propiedades se puede concluir que es posible modificar a $k > 0$, tomando un sesgo pequeño y principalmente reduciendo la varianza, mejorando así la estimación y predicción del error cuadrático medio. Para verificar lo anterior, se revisará más adelante el teorema de existencia.

3.3.2. Teoremas sobre la función error cuadrático medio

Teorema 3.1. *La varianza total $\gamma_1(k)$ es una función continua y monótonamente decreciente en función de k.*

Corolario 3.1. *La primera derivada con respecto a k de la varianza total, $\gamma_1'(k)$, se aproxima a $-\infty$ cuando $k \to 0^+$ y $\lambda_p \to 0$.*

Corolario 3.2. *La primera derivada de la varianza total, $\gamma_1'(k)$, se aproxima a $-\infty$ cuando $k \to 0^+$ y la matriz $(X'X)$ se convierte en singular.*

Teorema 3.2. *El sesgo cuadrado $\gamma_2(k)$ es una función continua y monótonamente creciente en función de k.*

Demostración.

De la expresión (3.2) se tiene que $\gamma_2(k) = k^2\beta'(X'X + kI)^{-2}\beta$. Si Λ es la matriz de valores propios de $(X'X)$ y P es una transformación ortogonal, tal que $X'X = P'\Lambda P$, entonces

$$\gamma_2(k) = k^2 \sum_{i=1}^{p} \frac{\alpha_i^2}{(\lambda_i + k)^2} \qquad \text{donde } \sum_{i=1}^{p} \alpha_i^2 = \alpha'\alpha \quad \text{y} \quad \alpha = P\beta \tag{3.4}$$

ya que $\lambda_i > 0\ \forall i$ y $k \geq 0$, cada elemento de $(\lambda_i + k)$ es positivo y la suma es no singular. Es claro que $\gamma_2(0) = 0$, por lo que $\gamma_2(k)$ es una función continua para $k \geq 0$.
Para $k > 0$ la ecuación (3.4) se puede escribir como

$$\gamma_2(k) = \sum_{i=1}^{p} \frac{\alpha_i^2}{[1 + (\lambda_i/k)]^2} \qquad \lambda_i > 0\ \forall i$$

ya que

$$\begin{aligned}\gamma_2(k) &= \sum_{i=1}^{p} \frac{k^2\alpha_i^2}{(\lambda_i+k)^2} \\ &= \sum_{i=1}^{p} \left(\frac{k\alpha_i}{\lambda_i+k}\right)^2 \\ &= \sum_{i=1}^{p} \alpha_i^2 \left(\frac{1}{(\lambda_i+k)/k}\right)^2 \\ &= \sum_{i=1}^{p} \alpha_i^2 \left(\frac{1}{\lambda_i/k+1}\right)^2 \\ &= \sum_{i=1}^{p} \frac{\alpha_i^2}{(1+\lambda_i/k)^2}\end{aligned}$$

como $\lambda_i > 0\ \forall i$, las funciones de la forma λ_i/k son claramente monótonamente decrecientes para cada incremento en k y cada término de $\gamma_2(k)$ es monótonamente creciente. Por lo tanto, $\gamma_2(k)$ es monótonamente creciente.

■

Corolario 3.3. *El sesgo cuadrado de $\gamma_2(k)$ se aproxima a $\beta'\beta$ con el límite superior.*

Demostración.

$$\text{Se tiene que } \gamma_2(k) = \sum_{i=1}^{p} \frac{\alpha_i^2}{(1+\lambda_i/k)^2}$$

entonces

$$\lim_{k\to\infty} \sum_{i=1}^{p} \frac{\alpha_i^2}{(1+\lambda_i/k)^2} = \sum_{i=1}^{p} \alpha_i^2 \lim_{k\to\infty} \frac{1}{(1+\lambda_i/k)^2} = \sum_{i=1}^{p} \alpha_i^2$$

y como $\sum_{i=1}^{p} \alpha_i^2 = \alpha'\alpha$ y $\alpha = P\beta$

$$\begin{aligned}\sum_{i=1}^{p} \alpha_i^2 &= (P\beta)'(P\beta) \\ &= \beta' P' P \beta \\ &= \beta' I \beta \\ &= \beta'\beta\end{aligned}$$

■

Corolario 3.4. *La derivada $\gamma_2'(k)$ se aproxima a cero cuando $k \to 0^+$.*

Demostración.

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

$$\begin{aligned}\frac{\partial \gamma_2(k)}{\partial k} &= \sum_{i=1}^{p} \frac{(\lambda_i + k)^2 2\alpha_i^2 k - 2\alpha_i^2 k^2 (\lambda_i + k)}{(\lambda_i + k)^4} \\ &= \sum_{i=1}^{p} \frac{(\lambda_i + k) 2\alpha_i^2 k[(\lambda_i + k) - k]}{(\lambda_i + k)^4} \\ &= \sum_{i=1}^{p} \frac{2\alpha_i^2 k \lambda_i}{(\lambda_i + k)^3} \\ &= 2k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3}\end{aligned} \tag{3.5}$$

Cada término en la suma $\dfrac{2k\lambda_i\alpha_i^2}{(\lambda_i + k)^3}$ es una función continua y el límite de cada término cuando $k \to 0^+$ es cero.

■

Teorema 3.3 (Teorema de existencia)**.** *Siempre existe una $k > 0$ tal que:*

$$E[L_1^2(k)] < E[L_1^2(0)] = \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

Demostración.

De las igualdades (3.2), (3.4) y (3.5) se tiene

$$\begin{aligned}\frac{\partial E[L_1^2(k)]}{\partial k} &= \frac{\partial \gamma_1(k)}{\partial k} + \frac{\partial \gamma_2(k)}{\partial k} \\ &= -2\sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + k)^3} + 2k \sum_{i=1}^{p} \frac{\lambda_i \alpha_i^2}{(\lambda_i + k)^3}\end{aligned} \tag{3.6}$$

Observe que

$$\gamma_1(0) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i + 0)^2}$$
$$= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{\lambda_i^2}$$
$$= \sigma^2 \sum_{i=1}^{p} \frac{1}{\lambda_i}$$

$$\gamma_2(0) = 0$$

En los Teoremas 3.1 y 3.2 se menciona que $\gamma_1(k)$ y $\gamma_2(k)$ son monótonamente decrecientes y crecientes, respectivamente. La primera derivada de $\gamma_1(k)$ es siempre no positiva, mientras que la de $\gamma_2(k)$ es siempre no negativa. Por lo tanto, esto es lo que se necesita para demostrar que siempre existe una $k > 0$ tal que

$$\frac{\partial E[L_1^2(k)]}{\partial k} < 0$$

La condición para esto es mostrada en la expresión (3.6) por ser

$$k < \frac{\sigma^2}{\alpha_{max}^2}$$

■

3.3.3. Algunos comentarios sobre la función del error cuadrático medio

Las propiedades de $E[L_1^2(k)] = \gamma_1(k) + \gamma_2(k)$ muestran que pasará por un mínimo. Y como el límite de $\gamma_2(k)$ se aproxima a $\beta'\beta$ cuando $k \to \infty$, este mínimo se moverá hacia $k = 0$ a medida que la magnitud de $\beta'\beta$ se incremente. Ya que $\beta'\beta$ es la longitud cuadrada del vector de regresión desconocido, parece imposible elegir un valor de $k \neq 0$ y lograr un error cuadrático medio más pequeño sin asignar una cota superior a $\beta'\beta$. En la práctica, $\beta'\beta$ no llega a ser infinito y uno debe ser capaz de encontrar el valor o valores de k que pongan a $\hat{\beta}^*$ más cerca de β en comparación con $\hat{\beta}$.

3.4. Una forma general de la regresión ridge

Existe una transformación ortogonal P, tal que $X'X = P'\Lambda P$ donde $\Lambda = (\partial_{i,j}\lambda_i)$ es la matriz de valores propios de $X'X$.
Sean

$$X = X^*P$$
$$Y = X^*\alpha + \varepsilon$$

donde

$$\alpha = P\beta, \quad (X^*)'(X^*) = \Lambda \quad \text{y} \quad \alpha'\alpha = \beta'\beta$$

Entonces, el procedimiento general de la estimación ridge se define de la siguiente forma:

$$\alpha^* = [(X^*)'(X^*) + K]^{-1}(X^*)'Y$$

donde

$$K = (\partial_{i,j}k_i), \qquad k_i \geqslant 0$$

Todos los resultados básicos vistos en la sección anterior se pueden demostar para esta forma más general.
Tenga en cuenta que se busca un k_i para cada variable canónica definida por X^*. Definiendo $(L_1^*)^2 = (\hat{\alpha}^* - \alpha)'(\hat{\alpha}^* - \alpha)$ se puede mostrar que los valores óptimos para las k_i serán $k_i = \sigma^2/\alpha_i^2$. No existe una equivalencia gráfica de la traza ridge, pero en su lugar se puede utilizar un procedimiento iterativo, iniciando en $\hat{k}_i = \hat{\sigma}^2/\hat{\alpha}_i^2$.

3.5. Selección de una mejor estimación de β

Se ha mostrado que el estimador por mínimos cuadrados ordinarios del vector de regresión β sufre una serie de deficiencias, cuando $X'X$ no tiene valores propios uniformes. Al aumentar pequeñas cantidades positivas a la diagonal de la matriz $X'X$, se obtienen los estimadores sesgados $\hat{\beta}^*$, los cuales se han introducido tanto para describir la sensibilidad de la solución de $X'X$, como para obtener un estimador de β con un error cuadrático medio más pequeño. Examinando las propiedades de $\hat{\beta}^*$, se puede mostrar que su uso es equivalente a hacer acotaciones certeras, ya sea en relación con las coordenadas individuales de β o con su longitud al cuadrado, $\beta'\beta$. Aunque el investigador Barnard señala que una alternativa para el insesgamiento del estimador por mínimos cuadrados $\hat{\beta}$, es limitar al error cuadrático medio sin considerar la hipótesis de acotación en β. Si es posible hacer supuestos matemáticos concretos sobre β, entonces es posible limitar el procedimiento de estimación para reflejar estas hipótesis.

Las hipótesis de acotación relacionadas al uso de $\hat{\beta}^*$ hacen evidente que no será posible una contrucción bien definida, sin embargo, esto no es un inconveniente para su uso, ya que con cualquier conjunto de datos no es díficil seleccionar una $\hat{\beta}^*$ que sea mejor que $\hat{\beta}$.

De hecho, cualquier conjunto de datos que es candidato para el análisis mediante regresión lineal, tiene implícito en él restricciones sobre los posibles valores de las estimaciones, que pueden ser compatibles con las propiedades conocidas de los datos generados, sin embargo, es difícil ser explícito sobre estas restricciones, en especial de forma matemática. Clutton-Brock [3] muestra que para el problema de calcular la media μ de una distribución, un conjunto de datos tiene restricciones implícitas en los valores de σ, que pueden ser candidatos lógicos como generadores. Por supuesto, en regresión lineal el problema es mucho más difícil, ya que el número de posibilidades tiende a aumentar. Primero, está el número de parámetros involucrados. Es común tener 10 o 20 coeficientes de regresión. Y deben ser considerados sus signos. Entonces, se tiene la matriz $X'X$, la combinación de $\binom{p}{2}$ diferentes factores de correlación y las formas en las cuales pueden estar relacionados. Sin embargo, lo anterior se puede integrar para realizar una evaluación de si los valores estimados son consistentes con los datos y con las propiedades de los datos generados.

Conforme a la experiencia, el mejor método para alcanzar un estimador $\hat{\beta}^*$ deseable es utilizando $k_i = k$ para toda i y usar la traza ridge para seleccionar un único valor de k y una única $\hat{\beta}^*$. Los siguientes puntos se pueden tomar en cuenta para una mejor elección:

a) En un cierto valor de k el sistema se estabilizará y tendrá las características generales de un sistema ortogonal.

b) Los coeficientes no tendrán valores absolutos grandes, con respecto a los factores para los que ellos representan la velocidad de cambio.

c) Los coeficientes con signos incorrectos en $k = 0$ deberán ser corregidos.

d) La suma de cuadrados de residuales no será inflada a un valor excesivo. No será relativamente grande para la suma mínima de cuadrados de residuales o grande en relación a lo que sería una varianza razonable para el proceso de generación de datos.

Otra perspectiva es usar las estimaciones de los valores óptimos de k_i desarrollados en la sección 3.4. Un enfoque usual sobre ésto es el siguiente:

a) Reducir el sistema canónico mediante las transformaciones $X = X^*P$ y $\alpha = P\beta$.

b) Determinar las estimaciones de las k_i's óptimas, utilizando $\hat{k}_{i0} = \hat{\sigma}^2/\hat{\alpha}_i^2$. Usando a $\hat{k}_{i0}$ para obtener $\hat{\beta}^*$.

c) Las $\hat{k}_{i0}$ tenderán a ser demasiado pequeñas debido a la tendencia de sobreestimar $\alpha'\alpha$. Ya que el uso de las $\hat{k}_{i0}$ disminuirán la longitud del vector de regresión estimado, k_{i0} puede volverse a estimar usando $\hat{\alpha}_i^*$. Esta nueva estimación puede continuar hasta alcanzar una estabilidad en $(\hat{\alpha}^*)'(\alpha^*)$ y en $\hat{k}_{i0} = \hat{\sigma}^2/(\hat{\alpha}_i^*)^2$.

Capítulo 4

Regresión Ridge en la Práctica

En este capítulo se discute el uso de la estimación sesgada en el análisis de datos, los procedimientos tanto para la selección de variables, como para el cálculo de la regresión ridge y de la inversa generalizada, así como la relación entre estas dos regresiones, ya que se ha observado que existe una gran similitud geométrica entre la solución de la inversa generalizada, expresada como una función de rango q asignada a la matrix $X'X$ y la solución ridge, la cual está expresada como una función del parámetro de sesgo k añadido a los elementos de la diagonal de $X'X$. También se analizan los resultados de un experimento de simulación y tres ejemplos del uso de la regresión ridge en la práctica. Los ejemplos que se presentan muestran que cuando las variables explicativas del modelo están altamente correlacionadas, la regresión ridge produce coeficientes que predicen y extrapolan mejor que el método de mínimos cuadrados, así como también es un procedimiento seguro para la selección de variables.

4.1. Ejemplos teóricos e ilustrativos

En esta sección se proporcionan diversos ejemplos que ayudarán a entender mejor los conceptos abordados anteriormente y se enfatiza tanto en la regresión ridge, como en la relación que ésta presenta con la regresión de la inversa generalizada.

Los conjuntos de datos con los que se trabajan pueden presentar variables de predicción que estén correlacionadas, ya que los datos históricos se obtuvieron sin la ayuda de un diseño experimental. Además, las restricciones físicas y matemáticas pueden necesitar variables de predicción correlacionadas, incluso cuando un diseño experimental es utilizado. La presencia de errores graves, valores omitidos, errores de correlación, varianza no constante y otros problemas, pueden crear resultados sin sentido, aún cuando se empleen sofisticadas técnicas de regresión para tratar el problema de correlación. Por lo que se supone que ninguno de estos problemas está presente.

Cabe destacar que la estimación sesgada es solamente una de las herramientas que se utilizan en el análisis de un conjunto de datos. Para realizar dicho análisis, primero

se empieza por comprender los antecedentes técnicos del problema y la definición de las variables candidatas, después se considera la forma y la necesidad de transformaciones del modelo. Entonces, los datos son examinados para valores anormales, se construyen diagramas de dispersión para buscar relaciones y posteriormente son examinados los residuales. Si los factores de inflación de varianza (FIV) de los estimadores por mínimos cuadrados son grandes, entonces se considera un procedimiento de estimación sesgada, tal como la regresión ridge, con la finalidad de reducir los efectos de correlación de la variable predictora y desarrollar un conjunto de coeficientes estables.

4.1.1. Comentarios sobre algunas prácticas comunes

Una práctica común que se observa en el análisis de regresión, es el no poder eliminar el mal condicionamiento, a través del uso de variables de predicción estandarizadas, observe que dicha estandarización es adecuada siempre que el término constante esté presente en el modelo. El mal condicionamiento que resulta de la falta de estandarización, no se debe a un defecto real en los datos, sino sólo en los orígenes arbitrarios de las escalas en las que las variables predictoras son expresadas. En la estandarización de las variables de predicción, a cada variable se le resta la media (centrar) y después, la variable centrada se dividide entre su desviación estándar (estandarizar), por lo tanto, al centrar se elimina el mal condicionamiento, reduciendo así la inflación de la varianza en los coeficientes estimados y al estandarizar, se expresa la ecuación en una forma que se preste a una interpretación y a un uso más directo.

Tabla 4.1: Datos de acetileno

x_1 Temperatura del Reactor (°C)	x_2 Relación de H_2 a n-heptano (relación molar)	x_3 Tiempo de Contacto (segundos)	y Conversión de n-heptano en Acetileno (%)
1300	7.5	0.0120	49.0
1300	9.0	0.0120	50.2
1300	11.0	0.0115	50.5
1300	13.5	0.0130	48.5
1300	17.0	0.0135	47.5
1300	23.0	0.0120	44.5
1200	5.3	0.0400	28.0
1200	7.5	0.0380	31.5
1200	11.0	0.0320	34.5
1200	13.5	0.0260	35.0
1200	17.0	0.0340	38.0
1200	23.0	0.0410	38.5
1100	5.3	0.0840	15.0
1100	7.5	0.0980	17.0
1100	11.0	0.0920	20.5
1100	17.0	0.0860	29.5

En un modelo lineal centrado, se elimina la correlación entre el término constante y todos los términos lineales, mientras que en un modelo cuadrático centrado, se reduce y en

ciertas situaciones se elimina completamente la correlación entre los términos lineales y cuadráticos.

Los datos del primer ejemplo a analizar son mostrados en la Tabla 4.1. Este es un conjunto típico de datos de un proceso químico, para los que se suele considerar que una superficie de respuesta totalmente cuadrática en las tres variables regresoras es un modelo tentativo adecuado.

En la Figura 4.1 se grafica el tiempo de contacto en función de la temperatura del reactor, como estas dos variables regresoras están altamente correlacionadas, causan en el modelo de mínimos cuadrados la inflación de la varianza de los coeficientes estimados.

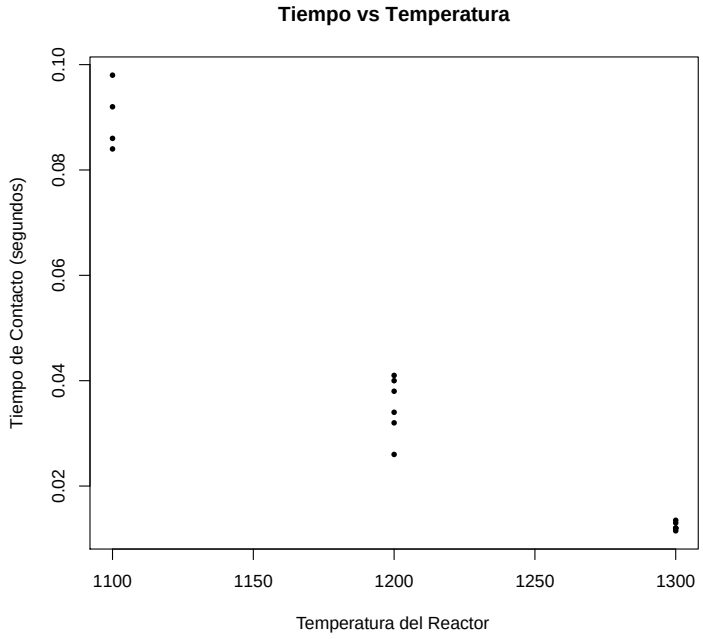


Figura 4.1: Datos de acetileno

El factor de inflación de varianza para cada término en el modelo, mide el impacto colectivo de estas correlaciones simples sobre la variación del coeficiente de ese término. Los factores de inflación de varianza, se definen como los elementos de la diagonal de la matriz inversa de correlaciones simples. Se debe tener en cuenta que cuando la correlación múltiple de cualquier predictor con otros se aproxima a la unidad, el FIV se hace infinito.

Los factores de inflación de varianza para los datos de acetileno son mostrados en la Tabla 4.2. El factor de inflación de varianza máximo es la mejor medida del condicionamiento de los datos. En el caso de que las variables predictoras sean ortogonales, los FIV son todos igual a uno.
Observe que en la tabla se presentan los factores de inflación de varianza, tanto para un

modelo no estadarizado, como para el que lo es, donde en este último el FIV máximo es de 6,563.35, por lo que se concluye que existe un problema de multicolinealidad, además al observar los factores de inflación de varianza de las demás variables cuadráticas y de productos cruzados, se puede decir que en los casos en los que intervienen x_1 y x_3 los FIV son grandes, por lo que los factores de inflación de varianza ayudan a identificar cuáles regresoras intervienen en la multicolinealidad, así como también muestran que el estandarizar no los afecta, pero en cambio el centrar sí lo hace. Ahora, una vez que ya se estandarizó, ¿cómo se puede llevar a cabo un análisis significativo de datos con un factor de inflación de varianza de más de seis mil?. Antes de contestar a la pregunta, se tienen que analizar las limitaciones que existen en algunas metodologías clásicas empleadas en el análisis de datos mal condicionados, una de estas metodologías es la de mínimos cuadrados, ya que en estas circunstancias no proporciona buenos estimadores. Para lograr un ajuste óptimo a los datos estimados, los mínimos cuadrados con frecuencia eliminan la buena predicción de los nuevos datos.
Otra de las metodologías es la selección de variables como una técnica para reducir el mal condicionamiento, la selección de dichas variables debe ser clasificada como significativa o no significativa, por lo que los sesgos de predicción grandes con frecuencia resultan de la eliminación de predictores no significativos.
Los estimadores sesgados pueden aliviar estas dos limitaciones, ya que es mejor usar un poco de todas las variables que todo de algunas y nada de las restantes (ésto es lo que hacen los estimadores sesgados).

Tabla 4.2: Resultados de regresión de los datos de acetileno

			Coeficiente Básico de Correlación			
	Mínimos Cuadrados - FIV			Ridge		
Términos	No Estandarizado	Estandarizado	Mín. Cuadrados	$k = 0.01$	$k = 0.05$	Inv. Generalizada $q = 3.8$
x_1=Temperatura	2, 856, 748.93	375.25	0.336	0.589	0.522	0.507
x_2=H_2/n-Heptano	10, 956.14	1.74	0.233	0.216	0.209	0.108
x_3=Tiempo de Contacto	2, 017, 162.52	680.28	-0.676	-0.327	-0.379	-0.414
x_1x_2	9, 802.90	31.04	-0.480	-0.326	-0.202	-0.095
x_1x_3	1, 428, 091.88	6, 563.35	-2.034	-0.094	-0.061	-0.051
x_2x_3	240.36	35.61	-0.266	-0.083	0.042	0.123
x_1^2	2, 501, 944.59	1, 762.58	-0.835	0.126	0.125	0.165
x_2^2	65.73	3.16	-0.090	-0.054	-0.047	-0.063
x_3^2	12, 667.10	1, 156.77	-1.001	-0.069	-0.024	-0.053
FIV Máximo			6, 563.35	12.38	2.63	0.460
R_A^2			0.994	0.990	0.983	0.973

Un comentario final sobre las prácticas en regresión, es que la mayoría de los estadísticos se limitan a utilizar modelos lineales en los parámetros, pero frecuentemente los antecedentes del problema sugieren una función con parámetros no lineales que pueda proporcionar un modelo más simple y natural.

4.1.2. Análisis de los datos de acetileno

Para el ejemplo de los datos de acetileno el modelo cuadrático completo estandarizado es:

$$E[Y] = \beta_0 + \sum_{j=1}^{3} \beta_j x_j + \sum_{1 \leq j < j'}^{3} \beta_{jj'} x_j x_{j'} + \sum_{j=1}^{3} \beta_{jj} x_j^2$$

donde

$$\begin{aligned} Y &= \text{porcentaje de conversión} \\ x_1 &= (\text{temperatura} - 1212.50)/80.623 \\ x_2 &= [H_2/(\text{n-Heptano}) - 12.44]/5.662 \\ x_3 &= (\text{tiempo de contacto} - 0.0403)/0.03164 \end{aligned}$$

Note que cada variable de predicción está estandarizada, pero los términos cuadráticos y de productos cruzados son creados directamente de los términos lineales estandarizados, además el modelo es no estandarizado con respecto a Y, por lo que, la evaluación numérica en esta forma es exacta y la interpretación de los coeficientes es sencilla.

Sin embargo, para la selección de la cantidad de sesgo, es necesario examinar la ecuación y su ajuste con todas las variables estandarizadas en forma de correlación, incluyendo a Y, a los predictores lineales y a la expansión de las variables explicativas.
Se hace referencia a los coeficientes de regresión obtenidos como los coeficientes de regresión básicos de correlación. En el uso regular de la regresión ridge, se muestran dichos coeficientes en tablas y/o gráficos para 25 valores de k, espaciados de manera logarítmica sobre el intervalo [0,1].

Tabla 4.3: Resultados de regresión de los datos de acetileno (cinco coeficientes que reducen el modelo cuadrático)

		Coeficiente Básico de Correlación			
	FIV		Ridge		
Términos	Mínimos Cuadrados	Mínimos Cuadrados	$k = 0.01$	$k = 0.05$	Inv. Generalizada $q = 4.0$
x_1=Temperatura	43.11	0.602	0.557	0.514	0.518
x_2=H_2/n-Heptano	1.07	0.194	0.192	0.187	0.193
x_3=Tiempo de Contacto	53.52	-0.323	-0.368	-0.391	-0.417
x_1x_2	1.09	-0.273	-0.270	-0.258	-0.272
x_1^2	4.68	0.173	0.180	0.169	0.197
FIV Máximo		53.52	13.63	1.72	1.15
R_A^2		0.991	0.990	0.989	

La Tabla 4.2 muestra los coeficientes de regresión básicos de correlación por mínimos cuadrados y por ridge, éste con dos valores de k, donde las matrices $X'X$ y $X'Y$ están

en forma de correlación, por lo tanto, el coeficiente básico de correlación r, es el cambio esperado en Y (medido en Y desviaciones estándar), dado un incremento en x por cada desviación estándar, por ejemplo, si r=0.5 implica que Y se incrementa 0.5 desviaciones estándar cuando x se incrementa una desviación estándar. En el problema se observa que los coeficientes se han estabilizado en k=0.05 y que el factor de inflación de varianza es razonable. Note que los coeficientes de mínimos cuadrados son grandes para las interacciones x_1x_3 y para x_3^2, pero en el modelo ridge desaparece dicha relación, además, los coeficientes son pequeños para x_2x_3 y x_2^2. Resultados similares son obtenidos con el modelo de la inversa generalizada, mostrada para $q = 3.8$, para el cual la longitud del vector de regresión es la misma que en el modelo ridge. Ahora, supongamos que estos cuatro términos son eliminados. La selección de variables es la estrategia más segura en este problema, ya que el sesgo ha eliminado la mayor parte del mal condicionamiento. Para un modelo casi ortogonal en la estadarización básica de correlación, todos los coeficientes tendrán varianzas casi iguales, por lo tanto, la selección de variables puede hacerse con base en los valores absolutos de los coeficientes. El componente de sesgo moderado de los errores cuadráticos medios de los coeficientes es ignorado para este propósito. Finalmente la Tabla 4.3 muestra que el sesgo futuro de los cinco términos en el modelo no cambia mucho los coeficientes.

La Figura 4.2 muestra las predicciones del modelo completo de mínimos cuadrados (es decir, con los nueve términos), del modelo ridge con los nueve términos y del modelo de mínimos cuadrados con cinco términos. Observe que los puntos de predicción son los puntos extremos de los datos, los cuales definen los límites de la región. Para objetivos prácticos, los tres modelos de predicción son igualmente buenos.

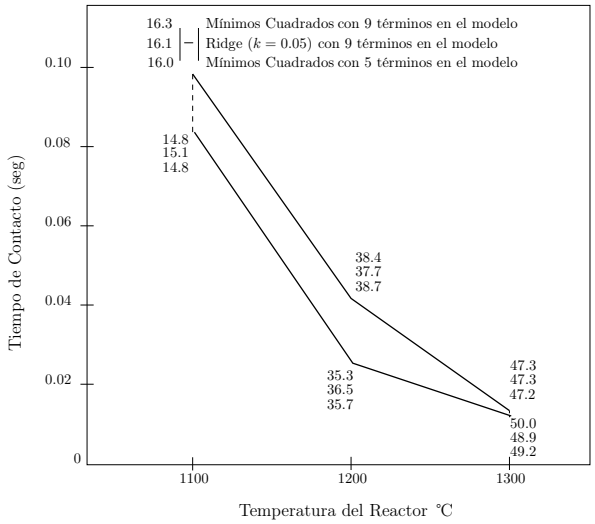


Figura 4.2: Datos de acetileno (predicciones en los datos)

La Figura 4.3 muestra predicciones en las esquinas de la región, definida por los márgenes de los límites de las variables de predicción individuales, esos puntos representan extrapolaciones relativamente suaves, ya que los intervalos originales de las regresoras no se han rebasado. Se puede notar que el modelo completo de mínimos cuadrados predice

un porcentaje de conversión de -86.2%, lo que es prácticamente imposible, mientras que los otros modelos tienen el 32.9% y el 38%, mostrando un porcentaje de predicción mucho más realista. Una situación similar se presenta en la parte inferior izquierda de la gráfica.

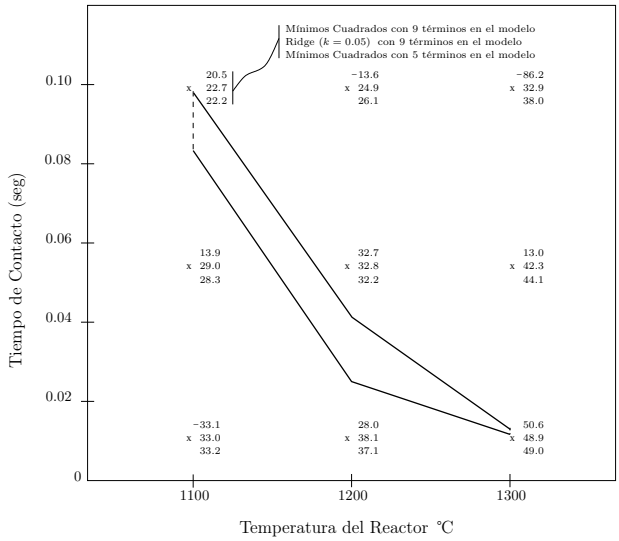


Figura 4.3: Datos de acetileno (predicciones fuera de los datos)

En resumen, el modelo completo de mínimos cuadrados se ajusta bien a los datos, pero es muy malo para extrapolaciones, una causa probable de esto es la presencia de multicolinealidad. Por otro lado, el modelo ridge hace predicciones tan buenas como el modelo de mínimos cuadrados, sin embargo produce predicciones mucho más realistas cuando se extrapola, aún usando los nueve términos. Por lo que de este ejemplo se concluye que tanto el modelo de mínimos cuadrados con eliminación de variables, como el modelo de regresión ridge, producen un mejor modelo en comparación con el de mínimos cuadrados original.

4.1.3. ¿Cuándo la selección de variables es una buena estrategia?

La selección de variables es una buena estrategia cuando:

a) las variables candidatas son ortogonales o casi ortogonales.

b) las variables se hicieron ortogonales o casi ortogonales, mediante la introducción de sesgo en el estimador.

c) la elección de los subconjuntos candidatos esté fuertemente guiada por los antecedentes del problema y las propiedades de las variables.

La selección de variables es una mala estrategia cuando:

a) las variables candidatas están altamente correlacionadas.

b) se calculan las estimaciones por mínimos cuadrados en presencia de inflación extrema de varianza, ya que existe un desequilibrio en todos los criterios, lo que conduce a una selección de subconjuntos muy inestable.

c) las variables candidatas incluyen efectos curvilíneos de otras variables candidatas.

Este último punto se ilustra en la Figura 4.4.

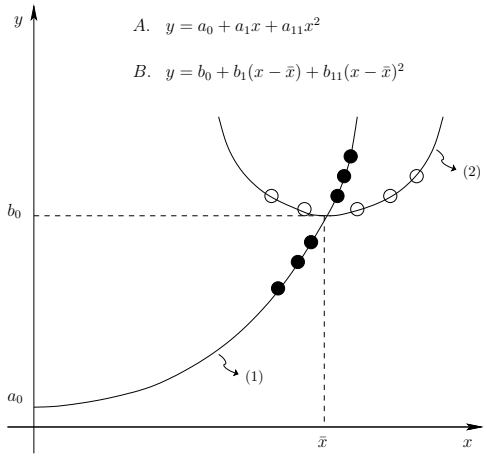


Figura 4.4: Selección de la variable con modelos curvilíneos

Considérense los modelos A y B como candidatos. Ambos pueden representar la fución cuadrática (1) (datos de puntos negros), en donde $y = a_0$ y $x = 0$, así como la función (2) (datos de puntos blancos), cuyo mínimo se encuentra en $x = \bar{x}$ y $y = b_0$. Ahora bien, suponga que los términos lineales son descartados o no seleccionados por el procedimiento del subconjunto, entonces, si los datos son como en la función (2), el modelo A sería totalmente un desastre, mientras que el B haría un buen trabajo; pero si los datos son como en la función (1), se observa un comportamiento opuesto. Así, el comportamiento operacional de todos los procedimientos de la selección de subconjuntos no es invariante con la elección arbitraria de calcular el origen.

En muchas ocasiones donde los polinomios de segundo grado o superior son aplicados, el modelo funciona como extensión de una función en la región de los datos. Normalmente son solamente dos puntos los apropiados en el espacio de predicción sobre el cual dicha extensión puede ser natural, éstos son, el punto medio de los datos y el origen del espacio de predicción. Esto implica que cualquier procedimiento de selección de variables debería ser hecho al menos dos veces con modelos curvilíneos, una vez calculado sobre el punto medio y otra calculado sobre el origen, a fin de saber si tiene un buen comportamiento el modelo.

4.1.4. Experimiento de simulación

A continuación se ilustra cómo ambos estimadores, tanto el de ridge como el de la inversa generalizada, son mejores que cualquiera de los de mínimos cuadrados o cualquier subconjunto de éste. También se ilustra la mecánica de la utilización de los procedimientos de estimación sesgada.

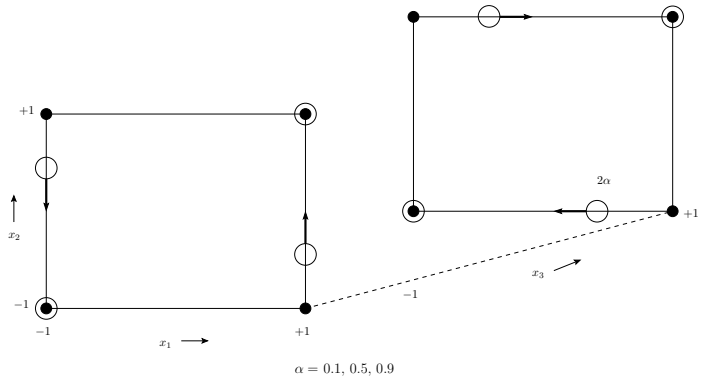


Figura 4.5: Ejemplo con tres predictoras (estructura de datos)

Para este ejemplo con tres predictoras, la estructura de datos es como se muestra en la Figura 4.5. Hay ocho puntos de datos estimados, mostrados por los círculos blancos y ocho puntos de predicción, mostrados por los círculos negros. Si el parámetro α es cero, el diseño es 2^3 factorial [1]. Ahora obsérvese que como α se aproxima a 1, cuatro de los puntos se mueven en las direcciones de las flechas y la correlación de x_1 y x_2 se hace cada vez más grande, teniendo como límite una correlación de valor uno.

Se han generado datos en α igual a 0.1, 0.5 y 0.9, donde se observa que x_3 es ortogonal a x_1 y a x_2 para todos los valores de α. El modelo verdadero es $E(Y) = x_1 + x_2 + x_3$, donde todos los coeficientes tienen valor uno y la constante de regresión es cero. A este modelo se le han introducido adicionalmente errores seleccionados aleatoriamente de una distribución normal con media cero y desviación estándar $\sigma = 0.8$.
Los datos reales son mostrados en la siguiente tabla:

i	x_1	x_2	x_3	$E(Y)$	ε_i
1	-1	-1	-1	-3	-0.305
2	1	1	-1	1	-0.321
3	-1	-1	1	-1	1.900
4	1	1	1	3	-0.778
5	-1	(1 - 2α)	-1	(- 1 - 2α)	0.617
6	1	-(1 - 2α)	-1	(- 1 + 2α)	-1.430
7	-(1 - 2α)	1	1	(1 + 2α)	0.267
8	(1 - 2α)	-1	1	(1 - 2α)	0.978

[1] Con el diseño factorial 2^3 se estudian tres factores en dos niveles cada uno. Consta de $2^3 = 2\times2\times2 = 8$ tratamientos diferentes, los cuales pueden identificarse con los puntos de la Figura 4.5.

La correlación entre x_1 y x_2 es $r_{12} = \alpha/(1-\alpha+\alpha^2)$, por lo tanto $r_{12} = 0.110, 0.667, 0.989$ para $\alpha = 0.1, 0.5, 0.9$ respectivamente.

El modelo de regresión es $\bar{Y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$. En todos los casos se incluyó un término constante, bajo el supuesto de que el analista no conozca que el verdadero valor de la constante de regresión es cero. El criterio por el cual se considera la calidad de los modelos de regresión, es por el error de predicción estándar en las ocho esquinas del cubo.

La Tabla 4.4 muestra los resultados de la regresión ridge en $\sigma = 0.8$, para cinco valores de k y para tres valores de α. Las cantidades mostradas en la tabla son:

$$
\begin{aligned}
S_e &= \text{error estándar de residuales de los datos de estimación} \\
R_A^2 &= \text{valor ajustado de } R^2 = 1 - S_e^2/S_Y^2, \text{ donde} \\
&\qquad S_Y^2 = \text{varianza de } Y_i \\
&\qquad Y_i = E(Y_i) + \varepsilon_i \\
FIV &= \text{máximo factor de inflación de varianza} \\
S_p &= \text{predicción de la desviación estándar en los ocho puntos de predicción} \\
&= \left(\sum_{i=1}^{8} [\hat{Y}_i - E(Y_i)]^2/8 \right)^{1/2}
\end{aligned}
$$

Comenzando por examinar los resultados para $\alpha = 0.9$. Observe como la estimación del error de residuales cuando $k = 0$ es de 0.537 y dicho valor aumenta a medida que el sesgo k se incrementa, como consecuencia de ello se observa que la R^2 ajustada disminuye, sin embargo, tenga en cuenta la reducción de la inflación de varianza cuando k se incrementa. Ahora, en la predicción del error de residuales, S_p, que pasa por un mínimo en $k = 0.2$, muestra que un estimador ridge con este sesgo da predicciones en las esquinas del cubo con un error estándar de sólo 0.536, dicho valor es pequeño en comparación con el error de residuales de predicción de mínimos cuadrados que es de 1.972. Note que resultados similares, aunque menos dramáticos ocurren para pequeños valores de α.

En la parte inferior de la tabla se encuentra una línea adicional, estos valores son los factores de inflación de varianza para las predictoras ortogonales, en función de k. Tenga en cuenta que el error de predicción residual mínimo, aquí se produce para valores de k, en o apenas por encima del valor donde el factor de inflación de varianza máxima es de aproximadamente el mismo tamaño que si los factores fueran ortogonales.

La Tabla 4.5 muestra resultados usando la regresión de la inversa generalizada. Analizando primero $\alpha = 0.9$, note como los residuales de regresión aumentan a medida que el rango asignado de la matriz $X'X$ disminuye, también observe como disminuye la R^2 ajustada, al igual que el factor de inflación de varianza máximo. Al igual que en la tabla anterior, el error de los residuales de predicción pasa por un mínimo y dicho mínimo ocurre cuando el rango asignado es de 1.5. En este ejemplo el residual de predicción mínimo mediante la inversa generalizada es aún más pequeño que en la regresión ridge.

Tabla 4.4: Ejemplo con tres predictoras ($\sigma = 0.80$)

Ridge k

α		0	0.1	0.2	0.4	0.8
0.1	S_e	0.729	0.759	0.828	0.995	1.286
	R^2_A	0.856	0.843	0.813	0.730	0.550
	FIV	1.012	0.833	0.698	0.511	0.309
	S_P	0.609	0.598	0.619	0.702	0.878
0.5	S_e	0.591	0.626	0.701	0.876	1.173
	R^2_A	0.906	0.895	0.868	0.794	0.630
	FIV	1.800	1.155	0.825	0.510	0.309
	S_P	0.713	0.651	0.636	0.674	0.817
0.9	S_e	0.537	0.621	0.699	0.878	1.189
	R^2_A	0.936	0.914	0.891	0.829	0.685
	FIV	45.751	0.826	0.694	0.510	0.309
	S_P	1.972	0.560	0.536	0.583	0.738
$(FIV)_o$		1.000	0.826	0.694	0.510	0.309

Tabla 4.5: Ejemplo con tres predictoras ($\sigma = 0.80$)

Inversa Generalizada q

α		3.0	2.5	2.0	1.5	1.0
0.1	S_e	0.729	0.741	0.776	1.247	2.101
	R^2_A	0.856	0.851	0.836	0.577	0.000
	FIV	1.012	1.000	1.000	0.500	0.450
	S_P	0.609	0.581	0.571	0.527	1.087
0.5	S_e	0.591	0.606	0.649	1.172	2.057
	R^2_A	0.906	0.901	0.887	0.630	0.000
	FIV	1.800	1.050	1.000	0.500	0.300
	S_P	0.713	0.634	0.605	0.563	1.105
0.9	S_e	0.537	0.553	0.599	1.146	2.042
	R^2_A	0.936	0.932	0.920	0.708	0.072
	FIV	45.751	23.001	1.000	0.500	0.251
	S_P	1.972	1.100	0.563	0.518	1.083

No se debe interpretar esto como un resultado general, solamente como un indicador de que cualquier estimador sesgado puede proporcionar resultados esencialmente mejores que el de mínimos cuadrados.

Finalmente en la Tabla 4.6 se muestra el resultado correspondiente para todos los modelos de subconjuntos posibles. Tomando de nuevo $\alpha = 0.9$, se observa que el error de residuales de predicción varía entre 1.15 y 2.18, por lo tanto, los dos mejores modelos de subconjuntos son la participación de x_1 con x_3 y de x_2 con x_3, ambos son mejores que el modelo completo de mínimos cuadrados, donde el error de residuales de predicción fue de 1.972, pero note que todos estos modelos de mínimos cuadrados son mucho más pobres que los modelos ridge o los de la inversa generalizada.

Tabla 4.6: Ejemplo con tres predictoras ($\sigma = 0.80$)

Modelos de Subconjuntos

α		x_1	x_2	x_3	x_1, x_2	x_1, x_3	x_2, x_3
0.1	S_e	1.942	1.807	1.320	1.864	1.214	0.934
	R^2_A	0.000	0.110	0.525	0.054	0.598	0.763
	FIV	1.000	1.000	1.000	1.010	1.000	1.000
	S_P	1.460	1.420	1.470	1.110	1.130	1.080
0.5	S_e	1.870	1.693	1.340	1.825	0.933	0.624
	R^2_A	0.121	0.229	0.517	0.104	0.766	0.895
	FIV	1.000	1.000	1.000	1.800	1.000	1.000
	S_P	1.420	1.430	1.470	1.170	1.070	1.090
0.9	S_e	1.687	1.656	1.643	1.811	0.603	0.492
	R^2_A	0.367	0.389	0.399	0.270	0.919	0.946
	FIV	1.000	1.000	1.000	45.750	1.000	1.000
	S_P	1.470	1.480	1.470	2.180	1.150	1.160

4.1.5. Validación del modelo

Después de que se ha desarrollado una ecuación de predicción, es fundamental que una medida de la exactitud de los coeficientes y de las predicciones del modelo sea obtenida. Una forma de lograr esto, es mediante el estudio de la naturaleza física y de las bases teóricas del sistema que es analizado. Por ejemplo, en los datos de acetileno que anteriormente se estudiaron, el porcentaje de conversión negativo fue predicho en algunas partes del espacio factor por los 10 coeficientes del modelo de mínimos cuadrados y como la conversión negativa es prácticamente imposible, se observó que el modelo asociado no proporcionaba una correcta descripción del sistema que generó los datos.

Otro método para la validación del modelo es recolectar datos adicionales y analizar que tan bien el modelo predice los nuevos datos. Esto con frecuencia no es posible, por lo que existe una forma para simular su recolección, la cual consiste en dividir los datos disponibles en dos subconjuntos. Un subconjunto llamado datos de estimación, se utiliza para estimar los coeficientes en el modelo. El subconjunto restante, llamado datos de predicción, es usado para medir la exactitud de predicción del modelo. Cuando los datos son ordenados con respecto al tiempo pueden ser usados para dividirse en los subconjuntos mencionados. Por ejemplo, los datos de rendimiento de maíz de Laird y Cady, que serán discutidos más adelante, fueron recolectados en un periodo de cuatro años, Laird y Cady utilizaron los datos de los primeros tres años como datos de estimación y los datos del cuarto año como datos de predicción. Por otra parte, el procedimiento CADEX de Kennard y Stone, es otra manera de dividir los datos y será usado en el análisis del modelo GC-ASTM.

Sin embargo, otro método de validación es la comparación con un modelo teórico, el cual también se utilizará con los datos GC-ASTM.

4.2. Uso de la estimación sesgada en el análisis de datos

En la primera sección se presentaron los resultados de una pequeña simulación experimental y se ilustró el uso de la regresión ridge en el desarrollo del modelo para los datos de acetileno. Ahora se describirán dos conjuntos de datos, los cuales ilustrarán el uso de la estimación sesgada en problemas con un gran número de variables.

4.2.1. Interpretación de la traza ridge

En la traza ridge se representan los valores de los estimadores dependiendo del valor de k. En él se pone de manifiesto la inestabilidad de los coeficientes de regresión y el incremento de la suma de cuadrados. La k seleccionada a través de la traza ridge es técnicamente una variable aleatoria, aunque esta circunstancia no es de interés práctico en la selección de los estimadores de regresión, hace complicada la teoría de los límites de confianza y pruebas de hipótesis, debido a la introducción de sesgo. Por este sesgo, los errores cuadráticos medios son todos ligeramente dependientes de los verdaderos coeficientes del vector β, el cual es desconocido.

Analizando la parte práctica observe que los modelos que carecen de término constante ($\beta_0 = 0$), normalmente requieren un valor más pequeño de k (muchas veces ≤ 0.01) que los modelos con término constante ($\beta_0 \neq 0$). Además, los modelos con bajo valor de la estadística R_A^2 por lo general requieren valores más grandes de k, que los modelos con alto valor de la estadística R_A^2. El aumento de k conducirá en última instancia a todos los coeficientes a cero, pero para los valores más pequeños de k no es extraño observar el incremento del coeficiente en valor absoluto (quizás después de un cambio en el signo inicial), a medida que k aumenta. En esta situación se ha encontrado con frecuencia, que los buenos resultados se obtienen utilizando un valor de k donde el coeficiente pase a través del valor absoluto máximo. Este procedimiento se utilizó para seleccionar el valor de k para el modelo GC-ASTM que se discutirá más adelante.

Las gráficas de los coeficientes de la inversa generalizada pueden ser construidas e interpretadas de la misma manera que la traza ridge, utilizando el rango asignado q como el eje de las abscisas en lugar de k.

A continuación se realiza un análisis de la traza ridge para los datos de acetileno previamente estudiados, tanto para el modelo completo como para el modelo con cinco variables regresoras.

En la Figura 4.6 se presenta la traza ridge y en la Tabla 4.7 se muestran los coeficientes ridge para diferentes valores de k, así como el error cuadrático medio (MSE) y el coeficiente de determinación múltiple para cada modelo ridge. Note que a medida que se incrementa k, MSE incrementa y R^2 disminuye. En la traza ridge se observa la inestabilidad de la solución por mínimos cuadrados, debido a que existen grandes cambios en los coeficientes de regresión para pequeños valores de k, sin embargo los coeficientes se

estabilizan rápidamente a medida que k incrementa.
La gráfica logra una estabilidad razonable de coeficientes en la región $0.008 < k < 0.064$, sin presentar un grave incremento del error cuadrático medio o una pérdida en R^2, por lo que se elige $k = 0.032$, sin embargo como los coeficientes de mínimos cuadrados x_1x_3 y x_3^2 son grandes, tienden rápidamente a cero a medida que k incrementa. Además observe que con dicho valor de k los coeficientes son pequeños para las interacciones x_2x_3 y x_2^2. Ahora se eliminarán estos cuatro términos debido a los pequeños valores que tienen sus coeficientes de regresión en el modelo ridge.

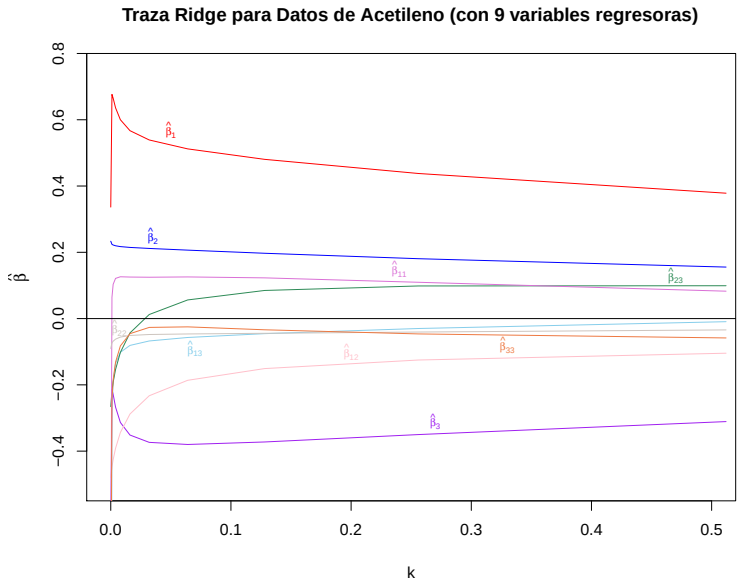


Figura 4.6: Traza ridge

Tabla 4.7: Coeficientes para valores diferentes de k

k	0	0.001	0.002	0.004	0.008	0.016	0.032	0.064	0.128	0.256	0.512
$\hat{\beta}_1^*$	0.3364	0.6765	0.6649	0.6359	0.6000	0.5670	0.5390	0.5121	0.4805	0.4378	0.3783
$\hat{\beta}_2^*$	0.2334	0.2239	0.2220	0.2197	0.2172	0.2147	0.2116	0.2065	0.1971	0.1806	0.1554
$\hat{\beta}_3^*$	-0.6758	-0.2132	-0.2286	-0.2671	-0.3134	-0.3514	-0.3735	-0.3799	-0.3723	-0.3500	-0.3108
$\hat{\beta}_{12}^*$	-0.4799	-0.4479	-0.4258	-0.3913	-0.3437	-0.2879	-0.2329	-0.1862	-0.1508	-0.1250	-0.1045
$\hat{\beta}_{13}^*$	-2.0339	-0.2775	-0.1888	-0.1351	-0.1017	-0.0809	-0.0675	-0.0570	-0.0454	-0.0299	-0.0093
$\hat{\beta}_{23}^*$	-0.2657	-0.2173	-0.1920	-0.1535	-0.1019	-0.0433	0.0122	0.0562	0.0848	0.0984	0.0991
$\hat{\beta}_{11}^*$	-0.8345	0.0641	0.1034	0.1212	0.1261	0.1253	0.1247	0.1256	0.1228	0.1095	0.0826
$\hat{\beta}_{22}^*$	-0.0903	-0.0731	-0.0681	-0.0620	-0.0558	-0.0509	-0.0480	-0.0464	-0.0444	-0.0406	-0.0341
$\hat{\beta}_{33}^*$	-1.0008	-0.2450	-0.1852	-0.1313	-0.0825	-0.0455	-0.0266	-0.0250	-0.0338	-0.04629	-0.0584
MSE	0.00038	0.00047	0.00049	0.00054	0.00062	0.00074	0.00094	0.00127	0.00206	0.00425	0.01002
R^2	0.998	0.997	0.997	0.997	0.996	0.996	0.994	0.992	0.988	0.975	0.940

Si se aplica la regresión ridge a las cinco variables regresoras restantes $(x_1, x_2, x_3, x_1x_2, x_1^2)$, se obtiene la traza ridge de la Figura 4.7. Esta nueva traza es mucho más estable que cuando se consideran las nueve variables regresoras, lo que significa que la introducción de más sesgo al aumentar k no cambia mucho los coeficientes de regresión, además hay pequeños cambios en MSE y en R^2 (Tabla 4.8).

Por lo anterior se concluye que la eliminación de algunas regresoras mejoró el condicionamiento de los datos, por lo que el modelo con cinco variables regresoras es razonablemente bueno.

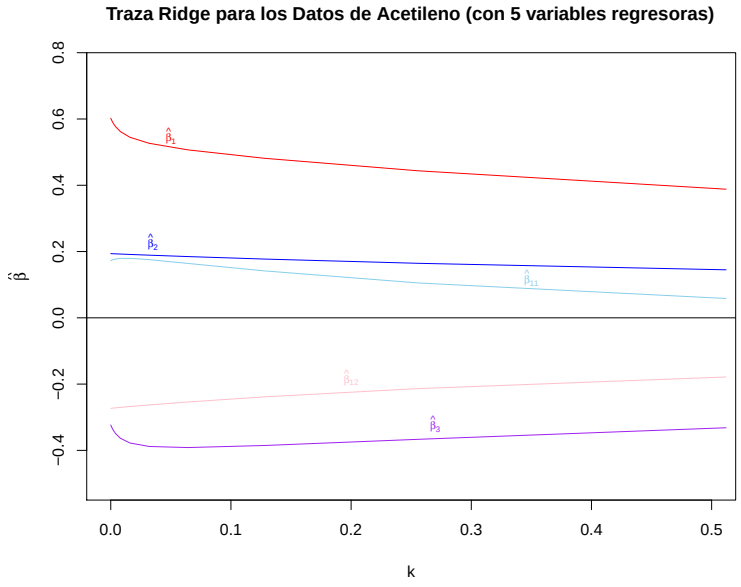


Figura 4.7: Traza ridge

Tabla 4.8: Coeficientes para valores diferentes de k

k	0	0.001	0.002	0.004	0.008	0.016	0.032	0.064	0.128	0.256	0.512
$\hat{\beta}_1^*$	0.6023	0.5943	0.5876	0.5769	0.5620	0.5447	0.5266	0.5069	0.4813	0.4436	0.3881
$\hat{\beta}_2^*$	0.1936	0.1935	0.1933	0.1930	0.1924	0.1912	0.1890	0.1848	0.1772	0.1644	0.1448
$\hat{\beta}_3^*$	-0.3234	-0.3316	-0.3385	-0.3492	-0.3633	-0.3777	-0.3882	-0.3915	-0.3852	-0.3666	-0.3317
$\hat{\beta}_{12}^*$	-0.2733	-0.2728	-0.2724	-0.2717	-0.2703	-0.2677	-0.2628	-0.2540	-0.2385	-0.2135	-0.1785
$\hat{\beta}_{11}^*$	0.1725	0.1742	0.1755	0.1774	0.1792	0.1795	0.1756	0.1642	0.1416	0.1052	0.0583
MSE	0.00063	0.00063	0.00063	0.00063	0.00064	0.00065	0.00069	0.00082	0.00127	0.00266	0.00632
R^2	0.994	0.994	0.994	0.994	0.994	0.994	0.993	0.992	0.987	0.973	0.937

4.2.2. Datos de rendimiento de maíz de Lair y Cady

El primer ejemplo a analizar de un conjunto de datos grande, son los datos de rendimiento de maíz publicados por Laird y Cady. Es importante mencionar que Cady y Allen más tarde usaron estos datos para ilustrar el procedimiento PRESS.
En este ejemplo, la variable de respuesta es el rendimiento de maíz en toneladas por hectárea, medido en cada uno de cuatro niveles de nitrógeno en cada uno de los 72 sitios experimentales, lo cual produjo 288 datos durante un período de cuatro años.

Las 11 variables de predicción correspondientes al modelo, son:

N	Nitrógeno Aplicado
A	Nitrógeno en el Suelo
B	Cultivo Anterior
C	Exceso de Humedad
D	Sequía
E	Profundidad de las Raíces en la Zona
F	Pendiente del Suelo
G	Textura del Suelo
H	Lluvia
J	Plaga
L	Maleza

Cuatro de estas variables se miden (N, A, E, F); una está expresada como un índice (D) y al resto les ha sido asignado un valor en una escala subjetiva. Las variables que no contengan nitrógeno serán llamadas "variables de sitio".

El modelo utilizado por Cady y Allen para describir estos datos fue un subconjunto de la ecuación cuadrática completa, la cual contiene un término constante, 11 términos lineales, 18 términos de productos cruzados o de interacción y 4 términos cuadrados, para un total de 33 términos (Tabla 4.9). Observe que 13 de los 18 términos involucran la interacción de las dos variables de nitrógeno (N y A).

Cady y Allen optaron por dividir los datos en dos conjuntos, donde los 228 datos recolectados en los primeros tres años fueron utilizados para estimar los coeficientes en el modelo y las 60 observaciones obtenidas en el cuarto año son usadas para probar la previsibilidad del modelo, estos conjuntos de datos serán llamados "datos de estimación" ($n = 228$) y "datos de predicción" ($n = 60$), respectivamente.

Tabla 4.9: Datos de rendimiento de maíz de Laird y Cady (33 coeficientes en el modelo)

Rango	Variable	Regresión Ridge ($k = 0.3$)	Inv. Generalizada ($q = 9.5$)	Rango	Variable	Regresión Ridge ($k = 0.3$)	Inv. Generalizada ($q = 9.5$)
1	$^*N$	0.249	0.179	18	A^2	0.035	0.040
2	BN	0.188	0.166	19	JN	0.033	-0.009
3	AN	0.182	0.181	20	BL	0.031	0.000
4	$^*J$	-0.119	-0.113	21	AL	0.027	0.017
5	$^*AC$	-0.112	-0.096	22	A	0.026	0.042
6	AJ	-0.101	-0.108	23	G^2	0.025	0.039
7	AH	-0.099	-0.069	24	BH	-0.017	-0.078
8	$^*DN$	-0.096	0.028	25	HN	0.017	0.008
9	BJ	-0.091	-0.109	26	LN	0.015	0.086
10	$^*N^2$	0.091	0.175	27	$^*B^2$	-0.014	-0.011
11	C	-0.082	-0.095	28	D	-0.013	-0.062
12	$^*H$	-0.074	-0.069	29	CN	-0.007	-0.006
13	$^*F$	-0.072	-0.066	30	L	-0.005	0.006
14	BD	-0.069	-0.057	31	AD	-0.005	-0.053
15	$^*AB$	0.056	0.050	32	E	0.003	0.054
16	BC	-0.050	-0.088	33	B	0.002	0.013
17	G	0.243	0.360	Tamaño del vector		0.239	0.236

* Selección de Variables por el Método PRESS

Se encontró que la desviación estándar de residuales de los modelos completos, de búsqueda paso a paso y de la regresión PRESS son de 1.03, 0.84, 0.72 para los datos de predicción y de 0.59, 0.62 y 0.65 para los datos de estimación. Por lo anterior, se concluye que el modelo PRESS hizo el mejor trabajo de predicción, lo cual era de esperarse ya que las variables en el modelo completo y en el de búsqueda paso a paso están altamente correlacionadas, con un FIV máximo de 180 y de 122, respectivamente, mientras que las variables en el modelo PRESS están menos correlacionadas, con un FIV máximo de 12.

¿Qué hace ridge con este problema?, se encontró que la traza ridge calculada para los datos de los primeros tres años se estableció alrededor de $k = 0.3$.

La desviación estándar de los datos de predicción para el cuarto año están graficados en la Figura 4.8 en función de k (utilizado en la elaboración del modeldo ridge). Observe como la desviación estándar de los datos de predicción decrece a medida que k aumenta, alcanzando un mínimo muy plano de 0.71 en $k = 0.6$. Un sesgo de $k = 0.01$ reduce la desviación estándar de la predicción de 1.03 a 0.82. Ahora, para $k = 0.3$ que se seleccionó de la traza ridge, la desviación estándar de la predicción es de 0.72, la cual es idéntica a la desviación estándar del modelo PRESS. Los coeficientes ridge para $k = 0.3$ se muestran en la Tabla 4.9, los cuales están ordenados en valor absoluto de forma descendente. En esta tabla, note que los tres coeficientes más grandes son N, BN, AN y todos ellos cuentan con nitrógeno aplicado (N). Los nueve términos del modelo PRESS están identificado por un asterisco (*), se pueden encontrar en los primeros 27 coeficientes, 8 de estos términos se encuentran en los primeros 15, mientras que el noveno, B^2, ocupa el lugar 27. Además, observe que fuera de los 15 primeros coeficientes de la regresión ridge, nueve son términos

de productos cruzados y también nueve cuentan con nitrógeno aplicado o con nitrógeno del suelo.

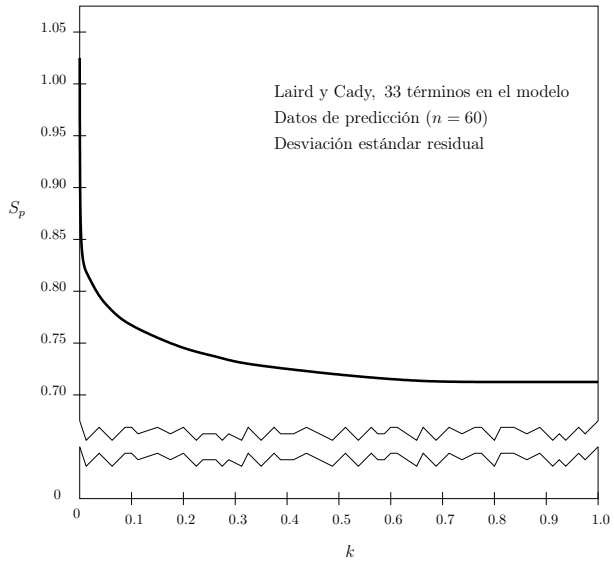


Figura 4.8: Desviación estándar de los datos de predicción

Esto muestra que el nitrógeno aplicado y el nitrógeno en el suelo, ahora denominados solamente "nitrógeno", es la variable dominante. Cuando el nitrógeno está ausente (es decir igual a 0) no puede haber ninguna producción de maíz, lo cual sugiere una intercepción en cero con respecto al nitrógeno, sin embargo las otras variables no tienen una intercepción en ese punto. Por lo que se concluye que un modelo multiplicativo sería más natural para estos datos, lo que es congruente con el predominio que tienen la mayoría de los términos de productos cruzados con el nitrógeno.
El modelo multiplicativo postulado es:

$$\begin{aligned} E(Y) &= (\text{nitrógeno aplicado} + \text{nitrógeno en el suelo}) \times (\text{variables de sitio}) \\ &= Z_1 Z_2 \end{aligned}$$

donde

$$\begin{aligned} Z_1 &= (N + \beta_2 A) + \beta_{12}(N + \beta_2 A)^2 \\ Z_2 &= \beta_1 + \beta_2 B + \beta_4 C + ... + \beta_{11} L \end{aligned}$$

La producción del maíz se modela como un producto de dos factores, primero el nitrógeno aplicado más el nitrógeno en el suelo, el cual se denota por Z_1; el segundo factor contiene las variables de sitio, denotadas por Z_2. El factor Z_1 es una función cuadrática de $(N+\beta_2 A)$ y Z_2 es una función lineal de las variables de sitio, las cuales no contienen nitrógeno. El coeficiente β_2 es necesario porque el nitrógeno aplicado (N) y el nitrógeno en el suelo (A) son medidos en diferentes unidades. El modelo contiene doce coeficientes en comparación con los 10 coeficientes que contiene el modelo PRESS, el cual se ajusta a los datos de

estimación y a los de predicción con una desviación estándar de residuales de 0.65 y 0.72 respectivamente, mientras que el modelo multiplicativo con 12 coeficientes, ajustado por mínimos cuadrados no lineales, tiene desviaciones estándar de residuales de 0.72 y 0.75 para los datos de estimación y predicción, dichos datos se observan en la Tabla 4.10.

Tabla 4.10: Ajuste de los modelos PRESS y multiplicativo (datos de Laird y Cady)

		Desviación estándar de residuales	
Modelo	Número de Coeficientes	Estimación ($n = 228$)	Predicción($n = 60$)
PRESS	10	0.65	0.72
Multiplicativo	12	0.72	0.75
Multiplicativo	9	0.73	0.64

Analizando los límites de confianza de los coeficientes, se llegó a la conclusión que los que corresponden al cultivo anterior (B), a la profundidad de las raíces en la zona (E) y a la maleza (L), no eran significativos. Cuando estas variables son eliminadas, el modelo resultante multiplicativo con nueve coeficientes ajusta los datos de estimación y los de predicción con desviaciones estándar de residuales de 0.73 y 0.64, respectivamente, dando un poco de mejor ajuste a los datos de predicción que el modelo PRESS (Tabla 4.10).

Existen otras diferencias entre el modelo PRESS y el multiplicativo que deben tenerse en cuenta:

i) El modelo PRESS contiene un término constante y el modelo multiplicativo no lo tiene.

ii) El cultivo anterior (B) es incluido en el modelo PRESS, pero no en el modelo multiplicativo.

iii) La textura del suelo (G) se incluye en el modelo multiplicativo y no en el modelo PRESS.

Para resumir este problema, se considera que al mantener todos los términos en el modelo y reduciendo las variables de correlación con ayuda de la regresión ridge, se ha sido capaz de

i) obtener un modelo que predice bien, y

ii) aprender más acerca de los roles de todas las variables en el modelo.

En este caso los resultados de la regresión ridge sugirieron un modelo alternativo no lineal. Si bien esto puede no ser el último modelo, es coherente con los antecedentes físicos del problema, como lo describen en su documento Laird y Cady, dando al científico una forma diferente de pensar sobre el mecanismo de estudio.

4.2.3. Modelo GC-ASTM

El segundo ejemplo a analizar, se refiere a la relación que existe entre la destilación ASTM y la cromatografía de gases, también llamado GC, para la destilación de una muestra de gasolina. Una de las propiedades que determina la calidad de la gasolina es la volatilidad, la cual es medida por el porcentaje de la mezcla evaporada a diferentes temperaturas ($^{\circ}F$). El método estándar para medir la volatilidad es por medio de la destilación ASTM, en el cual la muestra de la gasolina se calienta y los vapores pasan a través de un baño de hielo y se condensan, de manera que el porcentaje acumulativo evaporado en diferentes temperaturas es registrado. En esta destilación algunos de los componentes de mayor punto de ebullición "contienen" a los componentes de ebullición más bajos, pero en cambio, la destilación GC es mucho más exacta y cada componente se "desprende" en su verdadero punto de ebullición.

Aunque las curvas de destilación ASTM y GC no son idénticas, las especificaciones de la volatilidad de la gasolina son escritas en términos de ASTM. Para que una refinería utilice GC para el control en la línea de volatilidad, es necesario un modelo para predecir la destilación ASTM de una mezcla a partir de una destilación GC de la mezcla. Este ejemplo es típico de las situaciones donde la propiedad de un material se mide por una serie de puntos que forman la curva y donde es importante usar un número suficiente de puntos para describirla, pero se debe tener cuidado de no sobredefinirla, debido a que el uso de muchos puntos da lugar a información redundante. En este ejemplo, los puntos en la curva GC fueron seleccionados por un ingeniero, para dar una descripción adecuada de la misma, mientras que los puntos en la curva ASTM, corresponden a las especificaciones de diversas gasolinas.

El rango de temperatura se dividió en quince cortes: $0-15$, $15-40$, . . ., $414-487$ $^{\circ}F$ (Tabla 4.11). Las 15 variables de predicción en el modelo, $x_1, ..., x_{15}$, son la fracción de volumen de la mezcla evaporada en cada uno de los cortes y las variables de respuesta a predecir, $Y_1, ..., Y_{14}$, son el porcentaje acumulado evaporado de la mezcla en cada una de las 14 temperaturas de ASTM. Si bien se desarrolló un modelo para cada una de las 14 temperaturas de ASTM, sólo se analizarán en detalle las siguientes tres especificaciones, que son consideradas las más importantes: $Y_4 = ASTM$ 158, $Y_6 = ASTM$ 212 y $Y_{10} = ASTM$ 302.
El modelo postulado es:

$$E(Y) = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{15} x_{15}$$

donde Y es el porcentaje acumulado evaporado en una temperatura dada de ASTM y x_j es la fracción evaporada en el corte $j-$ésimo de GC. El término constante (β_0) ha sido eliminado, ya que la suma de las x_j es 1 y porque produce una matriz $X'X$ singular si β_0 fuera incluido.

Tabla 4.11: Modelo para los datos de ASTM a una temperatura de 158°

		Datos de Coeficientes de Estimación			
Corte GC	Rango de Temperatura $^\circ F$	Mínimos Cuadrados	Ridge ($k = 0.006$)	Inv. Generalizada ($q = 7$)	Coeficientes Teóricos de Destilación de Mínimos Cuadrados
1	$0 - 15$	224	126	120	125
2	$15 - 40$	87	94	105	132
3	$40 - 87$	110	104	109	102
4	$87 - 126$	116	75	74	84
5	$126 - 145$	-92	45	46	42
6	$145 - 175$	80	26	45	20
7	$175 - 198$	96	38	21	0
8	$198 - 220$	54	14	-8	-9
9	$220 - 237$	-125	-62	-45	-15
10	$237 - 285$	-30	-15	-20	-21
11	$285 - 300$	-65	-19	1	-29
12	$300 - 333$	21	6	-4	-25
13	$333 - 376$	181	36	5	-25
14	$376 - 414$	-217	25	24	-25
15	$414 - 487$	22	-40	-10	-24
Longitud del Vector			0.36	0.37	
Desv. Est. de los Datos de Predicción		1.28	1.01	0.96	

Los principales usos del modelo serían la predicción de las destilaciones ASTM para la mezcla de gasolina (donde la desviación estándar de predicción tiene que ser $\leqslant 1.5\,\%$) y la introducción a los cálculos de programación lineal para determinar la volatilidad óptima en los procedimientos de mezclado. Por lo tanto, las predicciones de ASTM son responsables de los cambios en la curva GC en cualquier rango de temperatura y de que los coeficientes de estimación en el modelo sean realistas a la luz de los conocimientos disponibles para la ingeniería. Es importante mencionar que se hizo el desarrollo de los coeficientes por mínimos cuadrados y por la regresión paso a paso, pero éstos fueron inaceptables desde un punto de vista físico.

Los datos de destilación GC y ASTM estaban disponibles en 59 mezclas. Se consideró que sería ventajoso tener un estimador independiente de la desviación estándar del modelo de predicción. Se utilizó el algoritmo CADEX de Stone y Kennard para dividir los datos en dos conjuntos, 29 mezclas de estimación y 30 mezclas de predicción (se describe el análisis de la forma en la que en realidad fue conducido, pero puede no ser necesariamante la mejor manera de dividir los datos). De la misma forma que en el ejemplo del rendimiento de maíz de Laird y Cady, los coeficientes en el modelo fueron estimados a partir de las 29 mezclas de estimación, mientras que las 30 mezclas de predicción se utilizaron para obtener una medida de la desviación estándar del modelo de predicción.

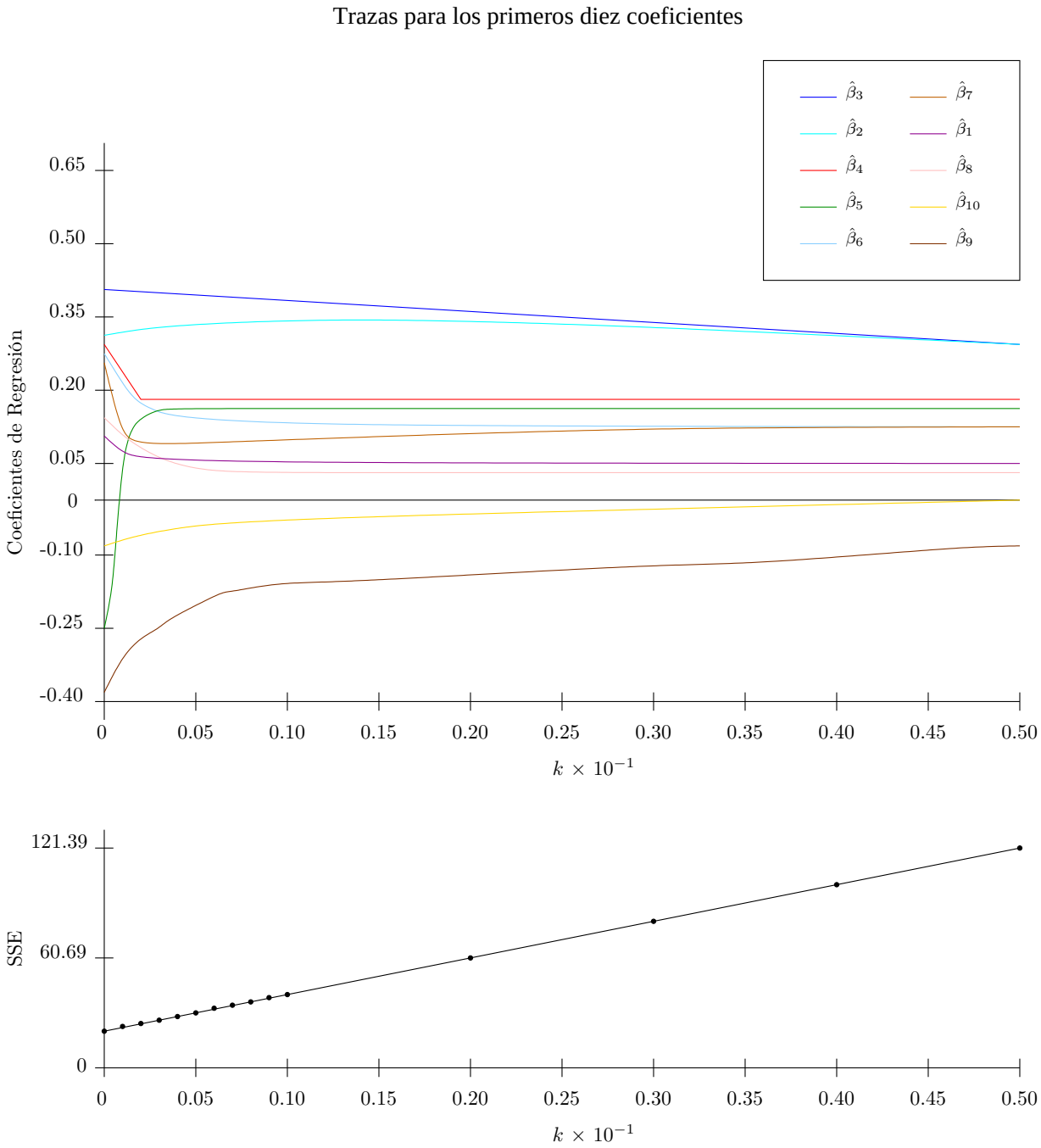


Figura 4.9: Modelo GC - ASTM (Y4: ASTM 158)

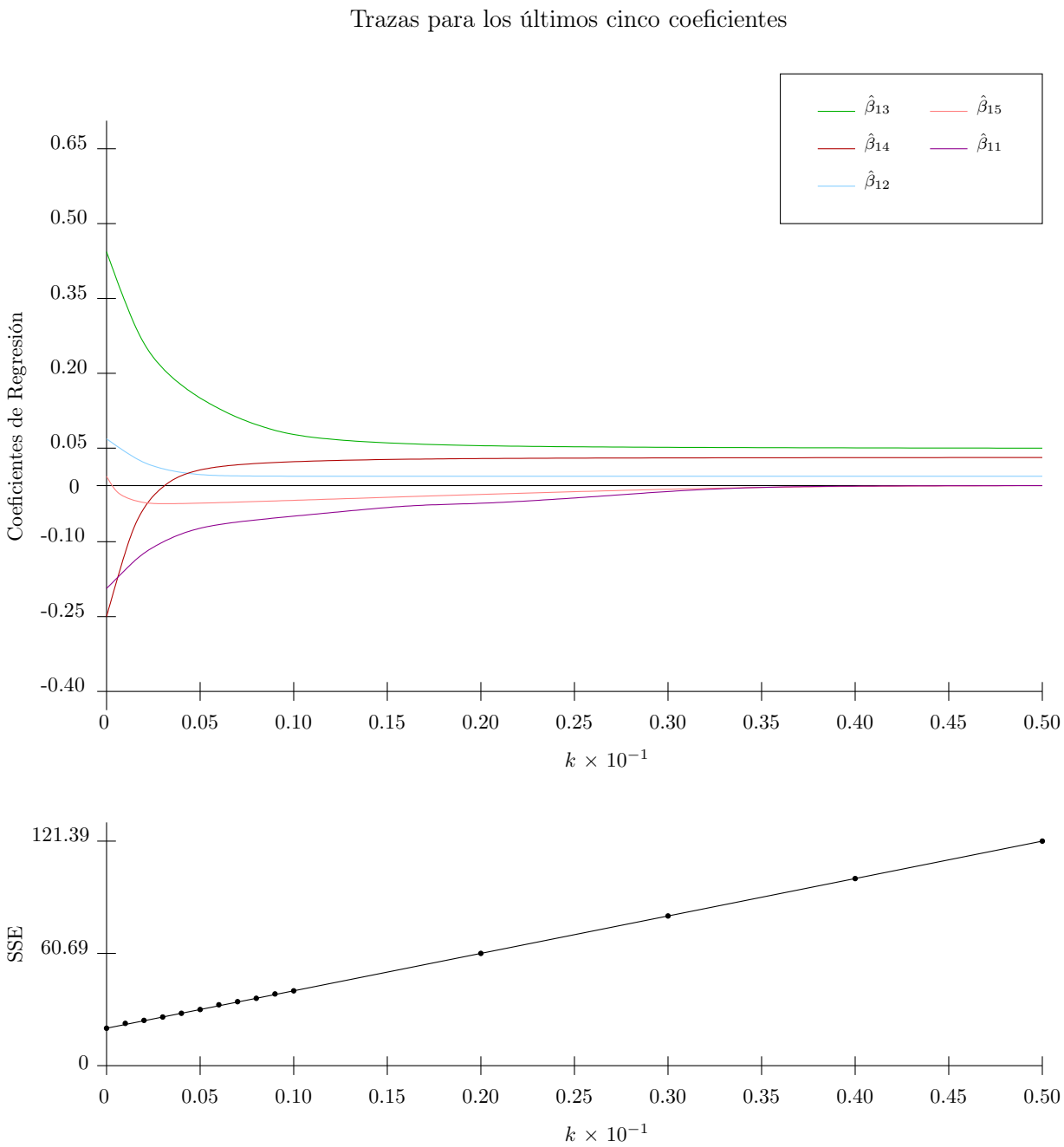


Figura 4.10: Modelo GC - ASTM (Y4: ASTM 158)

La traza ridge para $Y_4 = ASTM\ 158$ es mostrada en las Figuras 4.9 y 4.10, donde se grafican los coeficientes de regresión contra los valores de $k \times 10^{-1}$. Las trazas para los primeros 10 coeficientes son mostradas en la Figura 4.9 y las trazas para los últimos 5 coeficientes se encuentran en la Figura 4.10, observe que el sistema se estabiliza rápidamente alrededor de $k = 0.005$ o 0.01. Por lo que, se ha determinado utilizar los coeficientes cuando $k = 0.006$.
Antes de continuar analizando estos coeficientes se revisará la gráfica de la Figura 4.11, donde la desviación estándar de predicción se encuentra en función de cada valor de k para los 30 datos de predicción. Para $Y_4 = ASTM\ 158$ en la solución de mínimos cuadrados, la desviación estándar tiene un valor de 1.28 y decrece a medida que k aumenta, alcanzando un mínimo cerca de $k = 0.006$, dicho valor se elige de la traza ridge. La curva $ASTM$ 212 sigue un patrón similar, alcanzando un mínimo alrededor de $k = 0.003$, pero en las temperaturas 302 y 375 el error experimental es más pequeño. La desviación estándar de predicción aumenta a medida que se incrementa k, aunque en la curva de ASTM 375 se identifica un intervalo en el que se presenta un comportamiento plano hasta $k = 0.006$.

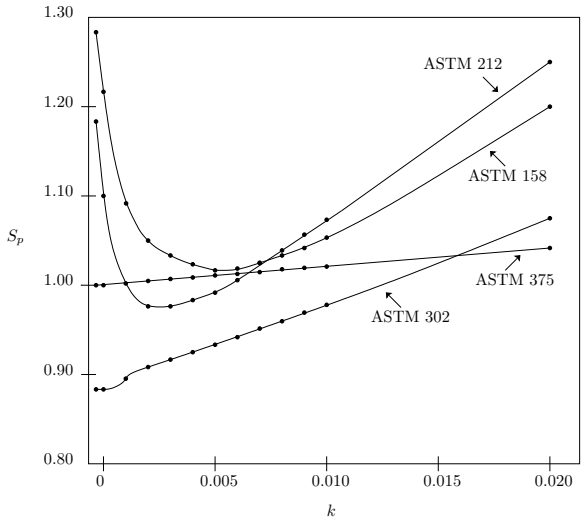


Figura 4.11: Desviación estándar de la predicción vs k ($n = 30$) [modelo GC - ASTM]

Ahora, siguiendo con el análisis de los coeficientes en el modelo ASTM 158 mostrado en la Tabla 4.11. Note que el error de predición estándar se encuentra en la parte inferior de las columnas, donde se observa que el modelo de regresión ridge es un mejor predictor, ya que tiene un menor error estándar que el modelo de mínimos cuadrados, el cual es de 1.01 % y de 1.28 %, respectivamente (estos dos números fueron calculados de las 30 mezclas no utilizadas en el cálculo de los coeficientes). Examinando ambos modelos, se reveló que los coeficientes en el modelo de mínimos cuadrados son en general más grandes que los coeficientes en el modelo de regresión ridge, en el primero los coeficientes más grandes en valor absoluto se encuentran alrededor de 200 (en cortes 1 y 13 y 14), mientras que en el segundo están alrededor de 100 (en cortes 1 y 3). Además, los coeficientes de mínimos cuadrados no tienen un buen comportamiento con respeto al signo. Estos puntos se pueden ver con facilidad cuando los modelos son gráficamente comparados.

En la Figura 4.12, los coeficientes son graficados contra los cortes de temperatura GC, mostrando en la primera gráfica a los coeficientes de mínimos cuadrados y en la segunda a los coeficientes de regresión ridge. Al realizar una comparación entre estos dos modelos, note que los coeficientes 5 y 14 y 15 han cambiado de signo y los coeficientes 1, 4, 9 y 13 son considerablemente más pequeños.

Conocimientos previos indicaron que todos los coeficientes en las temperaturas más altas deberían ser negativos, por lo que para conocer mejor este problema, se diseño un experimento de destilación teórico centrado alrededor de las 59 mezclas. La cantidad de cada uno de los 15 cortes de GC fue variada de acuerdo a un simple pseudocomponente diseñado para las mezclas, integrado por 15 componentes puros y 105 mezclas binarias. La destilación para estas 120 mezclas fue calculada utilizando la ley de Raoult, con coeficientes de actividad de uno y presión atmosférica. Los cortes de 15 GC fueron tratados como hidrocarbonos puros, alcanzando su verdadero punto de ebullición en el punto medio del corte. Esta destilación teórica tiene gran importancia, ya que es similar a la destilación ASTM y puede proporcionar corolarios de información sobre la relación entre los coeficientes del modelo y la temperatura GC. En la parte inferior de la Figura 4.12 se observa que los coeficienes de la destilación teórica disminuyen de tamaño con el aumento de la temperatura y finalmente serán negativos en la temperatura más alta, lo que confirma la teoría inicial. Es evidente que los coeficientes ridge llevan una semejanza más cercana a los coeficientes teóricos de destilación que a los coeficientes de mínimos cuadrados, por lo que los coeficientes de regresión ridge y los coeficientes de destilación teóricos siguen un patrón uniforme a medida que aumenta la temperatura GC; sin embargo los coeficientes de mínimos cuadrados no siguen esta buena relación.

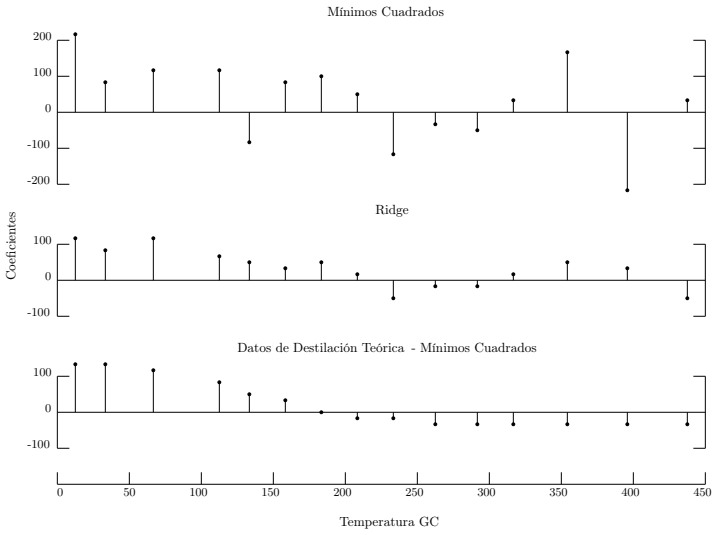


Figura 4.12: Modelo ASTM 158

La regresión ridge dio coeficientes igualmente significativos en todos los modelos de las diferentes temperaturas de ASTM, según lo mostrado por los coeficientes en los modelos

ASTM 212, 302 y 375°F presentados en la Figura 4.13. Como los antecedentes científicos del problema indican, el número de coeficientes positivos grandes en el modelo aumenta a medida que la temperatura ASTM se incrementa.

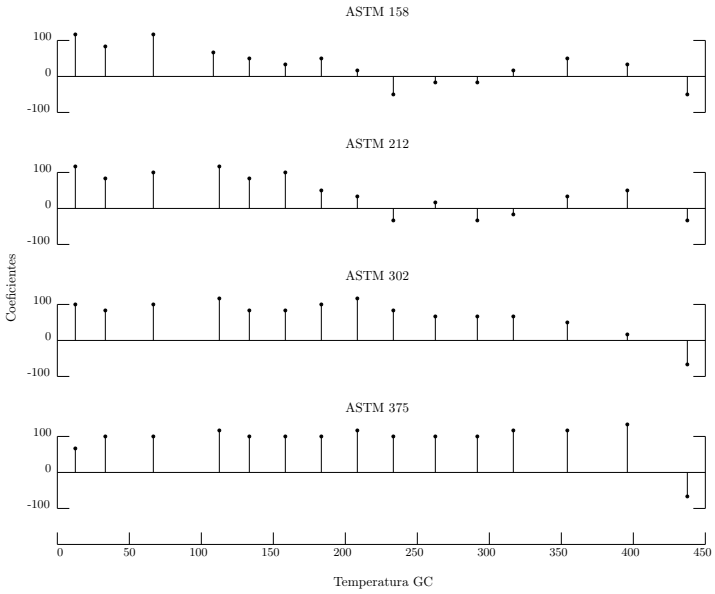


Figura 4.13: Ecuaciones de predicción GC-ASTM

Se termina con este ejemplo concluyendo que tanto el modelo ridge como el modelo de destilación teórica funcionan en la práctica, aunque por otro lado, los modelos de regresión de mínimos cuadrados y búsqueda paso a paso no lo hacen.

4.2.4. Resultados de la inversa generalizada

El análisis de los datos de rendimiento de maíz de Laird y Cady y del modelo de destilación GC-ASTM, se ha enfocado solamente en los estimadores ridge, pero ahora se estudiarán los resultados correspondientes a los estimadores de la inversa generalizada, los cuales se muestran en las Tablas 4.9 y 4.11. Dichos resultados, corresponden a un valor de q para el cual la longitud del vector de regresión es aproximadamente igual a la longitud para la selección ridge de sesgo k. Un análisis detallado de los coeficientes muestra que los relacionados con la regresión ridge y con la inversa generalizada son extraordinariamente similares. Si en la Tabla 4.11 se compara el modelo ridge y el modelo de la inversa generalizada, se observa que este último logra una reducción en la desviación estándar de la predicción.

En la Tabla 4.12 se muestra la correlación que existe entre los coeficientes de la regresión ridge y los de la inversa generalizada (elegido por tener la misma longitud aproximada del vector) para ocho conjuntos de datos, incluyendo los tres ejemplos mencionados en este trabajo. Observe que todos los coeficientes de correlación son muy altos.

Tabla 4.12: Correlación entre los coeficientes de la regresión ridge y de la inversa generalizada

		Ridge		Inversa Generalizada		
Conjunto de Datos	Número de Coeficientes	Sesgo (k)	Longitud del Vector ($\hat{\beta}'\hat{\beta}$)	Sesgo (q)	Longitud del Vector ($\hat{\beta}'\hat{\beta}$)	Coeficiente de Correlación*
Datos de Acetileno (7)	9	0.050	0.524	3.8	0.522	0.98
Gorman & Toman (6)	10	0.260	0.373	6.6	0.384	0.92
Laird & Cady (4)	33	0.300	0.239	9.5	0.236	0.89
GC - ASTM 158°	15	0.006	0.358	7.0	0.374	0.96
Datos de Secuencia (5)	8	0.300	0.329	3.5	0.318	0.96
Motor Espacial (5)	13	0.100	0.817	9.0	0.811	0.90
McDonald & Schwing (2)	15	0.180	0.376	8.5	0.384	0.91
Cirrosis del Higado (2)	4	0.300	0.239	1.0	0.262	0.98

*Coeficiente de Correlación Lineal

Capítulo 5

Aplicaciones

El marco teórico expuesto en los capítulos anteriores muestran que aunque la regresión ridge es sesgada, puede ser usada para el ajuste del modelo en presencia de multicolinealidad, por lo que en este capítulo se estudia en detalle una base de datos obtenida mediante la recopilación de datos en un estudio estadístico, para complementar así la teoría desarrollada y tener todas las herramientas necesarias para comprender mejor el análisis.

Se usará el lenguaje **R** para la obtención de resultados que se requieren en el problema y toda la sintaxis utilizada se podrá consultar en la sección de apéndices.

5.1. Planteamiento del problema

El mezquite es un árbol o arbusto espinoso, que llega a medir hata 10 metros de altura, de acuerdo con la profundidad del suelo, se encuentra principalmente en los Estados Unidos de América, desde la frontera con México al sudoeste de Kansas y del sudeste de California al sudoeste de Utah y en el límite sur del desierto de Sonora. Aunque el mezquite tiene diversos usos como alimento, combustible, sombra, planta medicinal, apicultura y material para la construcción de viviendas, la gente de campo considera a este árbol una molestia, debido a su capacidad de competir y ganarles a los pastos de calidad por humedad, es decir a los pastos del ganado. La erradicación del mezquite es difícil debido a que el árbol puede regenerarse de un pedazo de raíz dejado en el suelo. Algunos herbicidas no son efectivos o lo son parcialmente. La técnica de remoción fuerte del terreno muestra su efectividad contra el rebrote, pero es costosa: más de 280 dólares por hectárea (en Estados Unidos). Por lo que es necesario ajustar un modelo de regresión para poder obtener predicciones sobre la biomasa[1] total de mezquite en el pastizal para muestras futuras y así poder iniciar la aplicación de técnicas silvícolas[2] que permitan su aprovechamiento racional y sostenible.

[1]Suma total de la materia de los seres que viven en un ecosistema determinado, expresada habitualmente en peso estimado por unidad de área o de volumen.

[2]Silvicultura: Conjunto de reglas y técnicas que permiten la explotación racional de los bosques, así como su conservación y regeneración.

En la Tabla 5.1 se presentan los datos para estimar la biomasa, dados los incrementos de cinco variables regresoras[3]:

Tabla 5.1: Datos para diecinueve arbustos

Observación	x_1 Diámetro más ancho	x_2 Diámetro más estrecho	x_3 Altura total	x_4 Altura de la copa	x_5 Densidad del arbusto	Y Medida de biomasa
1	2.50	2.3	1.70	1.40	5	723.0
2	2.00	1.6	1.70	1.40	1	345.0
3	1.60	1.6	1.60	1.30	1	330.9
4	1.40	1.0	1.40	1.10	1	163.5
5	3.20	1.9	1.90	1.50	3	1160.0
6	1.90	1.8	1.10	0.80	1	386.6
7	2.40	2.4	1.60	1.10	3	693.5
8	2.50	1.8	2.00	1.30	7	674.4
9	2.10	1.5	1.25	0.85	1	217.5
10	2.40	2.2	2.00	1.50	2	771.3
11	2.40	1.7	1.30	1.20	2	341.7
12	1.90	1.2	1.45	1.15	2	125.7
13	2.70	2.5	2.20	1.50	3	462.5
14	1.30	1.1	0.70	0.70	1	64.5
15	2.90	2.7	1.90	1.90	1	850.6
16	2.10	1.0	1.80	1.50	2	226.0
17	4.10	3.8	2.00	1.50	2	1745.1
18	2.80	2.5	2.20	1.50	1	908.0
19	1.27	1.0	0.92	0.62	1	213.5

5.2. Análisis del problema

En la Figura 5.1 se muestra la gráfica de dispersión matricial, entre las cinco regresoras y se revelan correlaciones moderadas entre las variables, por lo que es posible que existan problemas de multicolinealidad en los datos.
El modelo de regresión para los datos del problema es:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

donde

$$\begin{aligned} Y &= \text{Medida de biomasa} \\ X_1 &= \frac{x_1 - \bar{x}_1}{s_{x_1}} = \frac{x_1 - 2.28789}{0.69833} \\ X_2 &= \frac{x_2 - \bar{x}_2}{s_{x_2}} = \frac{x_2 - 1.87368}{0.72329} \\ X_3 &= \frac{x_3 - \bar{x}_3}{s_{x_3}} = \frac{x_3 - 1.61684}{0.42374} \\ X_4 &= \frac{x_4 - \bar{x}_4}{s_{x_4}} = \frac{x_4 - 1.25368}{0.33054} \\ X_5 &= \frac{x_5 - \bar{x}_5}{s_{x_5}} = \frac{x_5 - 2.10526}{1.59494} \end{aligned}$$

[3]Los datos del problema fueron obtenidos del libro **Freund, R. J. and Wilson, W. J. and Sa, P** (2006), *"Regression Analysis"*, p. 210 [5]

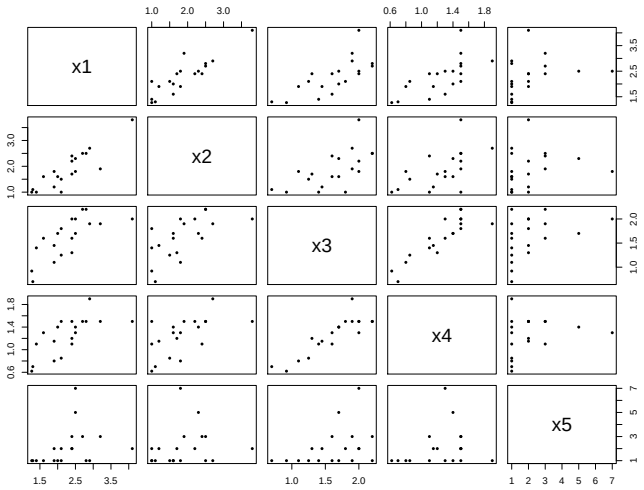


Figura 5.1: Correlación de las variables independientes

Note que cada variable regresora original ha sido escalada (restando la media y dividiendo entre su desviación estándar). El centrar los términos lineales ayuda a eliminar el mal condicionamiento cuando el modelo es ajustado.
El modelo ajustado por mínimos cuadrados es:

$$\hat{Y} = 547.5421 + 294.1492X_1 + 129.4845X_2 + 5.5582X_3 - 36.5011X_4 - 0.3034X_5$$

Tabla 5.2: Resultados de regresión para los datos del mezquite

Análisis de Varianza

Fuente de Variación	Grados de Libertad	Suma de Cuadrados	Cuadrados Medios	F_0
Regresión	5	2774583	554916.6	17.1423
Error	13	420823.1	32371.01	
Total	18	3195406		

$\sqrt{CMError}$	179.9195
R^2	0.8683
R_A^2	0.8177

Parámetros Estimados

Término	Coeficiente de Regresión	Error Estándar	t_0	Coeficiente de Regresión Estandarizado
β_0	547.5421	41.2764	13.265	
β_1	294.1492	102.9725	2.857	0.6981
β_2	129.4845	90.7239	1.427	0.3073
β_3	5.5582	104.0463	0.053	0.0131
β_4	-36.5011	95.1212	-0.384	-0.0866
β_5	-0.3034	49.5593	-0.006	-0.0007

La Tabla 5.2 muestra que el 87 % de la variación en la medida de biomasa está explicada por las cinco regresoras ($R^2 = 0.8683$). A pesar de que este coeficiente puede ser

engañoso (se incrementa en forma artificial con cada témino que se agrega al modelo), el coeficiente de determinación ajustado ($R^2_A = 0.8177$) confirma estos valores. El valor de la estadística F_0 es igual a 17.14 y como $F_0 > F_{.05,(5,13)} = 3.03$ se concluye que al menos una de las variables regresoras contribuye significativamente al modelo. Sin embargo, la mayoría de los valores individuales de t_0 son pequeños. Es decir, observe que las cinco regresoras son importantes, pero a partir de los valores de t_0 al menos una variable regresora puede ser eliminada del modelo y las demás conservarse.

Como se puede notar en la columna de coeficientes estandarizados, x_1, es decir la variable que mide el diámetro más ancho del arbusto, es la que tiene mayor influencia en la medida de la biomasa, mientras que la variable x_5, que mide la densidad del arbusto, presenta un menor dominio en la variable dependiente.

Ahora se analizará si la relación estimada está o no afectada por el problema de multicolinealidad.
A continuación se presenta la matriz de correlación de las variables independientes, que revela fuertes correlaciones entre el diámetro más ancho (x_1) y el diámetro más estrecho (x_2), ya que su correlación es de 0.8767 y entre la altura total (x_3) y la altura de la copa (x_4), cuya correlación es de 0.8794. También hay presencia de correlaciones más bajas entre las variables de diámetro y altura, mientras que la densidad parece no estar correlacionada con ninguna otra variable.

Tabla 5.3: Matriz de correlación

	x_1	x_2	x_3	x_4	x_5
x_1	1.0000	0.8767	0.7254	0.6834	0.3219
x_2	0.8767	1.0000	0.6354	0.5795	0.1999
x_3	0.7254	0.6354	1.0000	0.8794	0.3942
x_4	0.6834	0.5795	0.8794	1.0000	0.2236
x_5	0.3219	0.1999	0.3942	0.2236	1.0000

En la Tabla 5.4 se muestra información de la tolerancia y del FIV de cada variable. Dichas estadísticas indican que la multicolinealidad tiene un cierto grado de importancia en las variables x_1 (diámetro más ancho del arbusto), x_3 (altura total del arbusto) y x_4 (altura de la copa). Es importante mencionar, que tanto los FIV como la tolerancia no presentan valores elevados, por lo que los coeficientes de regresión asociados no indican un problema grave de estimación. Los FIV en el modelo son afectados por los términos lineales centrados y obtenidos de los elementos de la diagonal de la matriz $(X'X)^{-1}$.

Tabla 5.4: Estadísticos de colinealidad

Variable	Tolerancia	FIV
x_1	0.1696	5.8960
x_2	0.2184	4.5767
x_3	0.1661	6.0196
x_4	0.1987	5.0311
x_5	0.7322	1.3657

En la siguiente tabla se presentan los valores y vectores propios de la matriz $X'X$, el índice condición y la participación de cada variable en las varianzas asociadas a las distintas raíces características. Los valores propios cercanos a cero indican dependencia casi lineal en los datos, mientras que los elementos asociados con los vectores propios (τ_j) describen la naturaleza de esa dependencia lineal.

Tabla 5.5: Análisis de componentes principales

			Proporciones de la varianza				
Número	Valores Propios	Índice Condición	x_1	x_2	x_3	x_4	x_5
1	3.3330	1.0000	0.0128	0.0140	0.0125	0.0134	0.0125
2	0.8887	1.9365	0.0029	0.0168	0.0003	0.0042	0.7372
3	0.5722	2.4134	0.0434	0.1213	0.0517	0.1172	0.0255
4	0.1154	5.3720	0.5718	0.4210	0.3916	0.1978	0.0044
5	0.0904	6.0711	0.3689	0.4267	0.5436	0.6672	0.2201

Vectores Propios				
τ_1	τ_2	τ_3	τ_4	τ_5
-0.5019	-0.1242	0.3827	0.6240	-0.4435
-0.4634	-0.2614	0.5636	-0.4717	0.4202
-0.5013	0.0457	-0.4222	-0.5218	-0.5440
-0.4740	-0.1385	-0.5809	0.3390	0.5509
-0.2391	0.9460	0.1413	0.0263	0.1648

Como se comentó anteriormente, las variables afectadas por la multicolinealidad son x_1, x_3 y x_4, que son las que tienen mayor proporción de varianza asociada con los dos valores propios más pequeños.
Al analizar el número condición de $X'X$, que es:

$$\kappa = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} = \sqrt{\frac{3.3330}{0.0904}} = 6.0711$$

se observa que hay presencia de colinealidad débil, por lo que seguramente se corregirá al aplicar la regresión ridge al modelo y no será necesario eliminar alguna variable.

5.3. Solución ridge

Para obtener la solución ridge de los datos del mezquite, se debe resolver la ecuación $(X'X+kI)\hat{\beta}^* = X'Y$ para diversos valores de k $(0 \leqslant k \leqslant 1)$, con las matrices $X'X$ y $X'Y$ en forma de correlación. La traza ridge es mostrada en la Figura 5.2 y los coeficientes ridge para los valores de k se presentan en la Tabla 5.6. En la tabla también se encuentra el cuadrado medio residual y R^2 para cada modelo ridge.

Tabla 5.6: Coeficientes para valores diferentes de k

k	0	0.002	0.004	0.008	0.016	0.032	0.064	0.128	0.256	0.512
$\hat{\beta}_1^*$	0.69813	0.69234	0.68674	0.67608	0.65662	0.62370	0.57393	0.50860	0.43344	0.35447
$\hat{\beta}_2^*$	0.30732	0.31033	0.31319	0.31849	0.32763	0.34131	0.35690	0.36547	0.35443	0.31715
$\hat{\beta}_3^*$	0.01319	0.01327	0.01338	0.01371	0.01468	0.01741	0.02418	0.03773	0.05850	0.08134
$\hat{\beta}_4^*$	-0.08663	-0.08448	-0.08242	-0.07849	-0.07137	-0.05931	-0.04079	-0.01515	0.01662	0.04957
$\hat{\beta}_5^*$	-0.00072	0.00003	0.00075	0.00210	0.00450	0.00839	0.01387	0.02039	0.02715	0.03347
MSE	0.01013	0.01022	0.01013	0.01013	0.01015	0.01021	0.01040	0.01086	0.01195	0.01427
$FIV_{\text{Máximo}}$	6.01963	5.79768	5.58853	5.20477	4.55325	3.58668	2.43687	1.38577	0.68630	0.43496
R^2	0.86830	0.86829	0.86827	0.86820	0.86795	0.86714	0.86478	0.85872	0.84452	0.81441

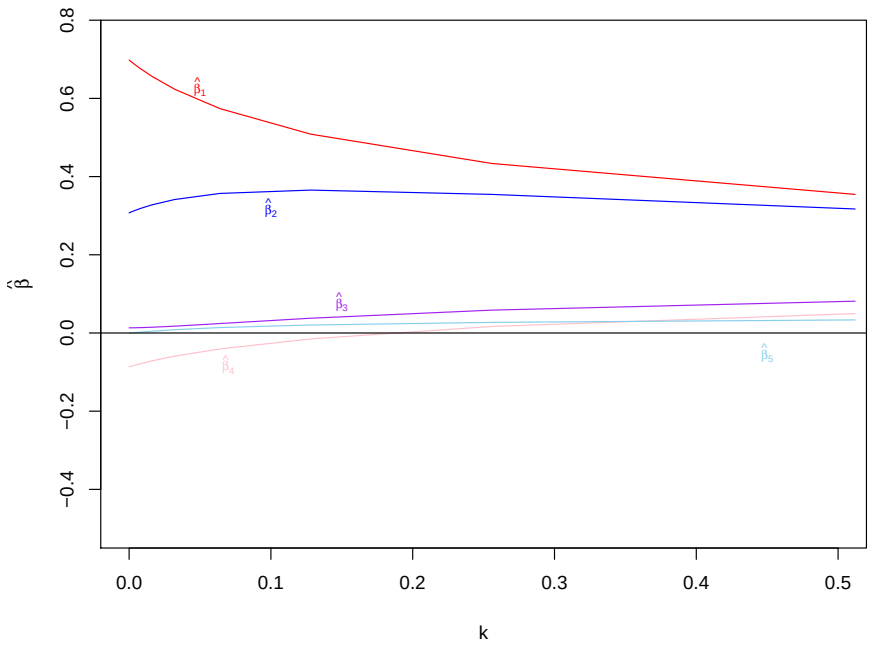


Figura 5.2: Traza ridge

La traza ridge muestra que no cambian drásticamente los coeficientes de regresión con el incremento de sesgo, es decir, del valor de k. Además, el cuadrado medio de residuales (MSE) y el coeficiente de determinación (R^2) presentan cambios muy pequeños.

En la Figura 5.2 observe que se presenta una estabilidad razonable para los coeficientes en $k = 0.128$, sin un aumento grave en el cuadrado medio de residuales y en R^2. Además tenga en cuenta que para este valor de k, el $FIV_{\text{Máximo}}$ disminuyó más de la mitad, en comparación con el obtenido en el modelo de mínimos cuadrados. Por lo que el modelo de regresión ridge es:

$$\hat{y} = 0.50860 + 0.36547x_2 + 0.03773x_3 - 0.01515x_4 + 0.02039x_5$$

El modelo expresado en términos de las regresoras originales es:

$$\hat{Y} = 543.8780 + 285.8301X_1 + 133.7174X_2 + 5.7436X_3 - 33.4322X_4 + 0.7632X_5$$

El modelo propuesto tiene una desviación estándar de residuales de 190.5618, la cual comparada con 179.9195 (regresión por mínimos cuadrados) es un poco mayor, pero esto se justifica, ya que los coeficientes presentan una mejor estabilidad.

Los FIV de los coeficientes para el nuevo modelo presentan un decremento en cada variable regresora.

Tabla 5.7: FIV de los coeficientes para el nuevo modelo

Variable	FIV
x_1	1.3857
x_2	1.2576
x_3	1.3711
x_4	1.2722
x_5	0.8615

Note que las dos variables regresoras más significativas en el modelo son en las que interviene el diámetro.

Conclusiones

Se sabe que uno de los problemas más comunes que se pueden presentar en los modelos de regresión, es la presencia de colinealidad, es decir que dos o más variables del modelo mantienen una relación lineal, lo que conduce a generar modelos sin sentido, por lo que en este trabajo se explicaron las principales técnicas de diagnóstico, así como también los procedimientos que se pueden utilizar para manejarla o eliminarla. Se observó que existen diferentes procedimientos para su detección, pero sin embargo no es una situación que tenga un fácil tratamiento, excepto cuando se haya producido por el uso de datos u observaciones erróneas, en cuyo caso se puede resolver omitiéndolas.
Algunas de las diferentes técnicas para manejar los problemas causados por la colinealidad, son la transformación de datos (por ejemplo, transformaciones logarítmicas), la utilización de la regresión polinomial y la posibilidad de introducir nuevos datos o de seleccionar otro subgrupo de predictoras, lo anterior quizá sea la mejor solución, pero en la mayoría de las ocasiones no es posible dada la situación experimental. Sin embargo, uno de los métodos alternativos es la regresión sesgada que permite utilizar la información original y que hace posible mitigar dicho efecto.

A lo largo de este trabajo se ha mostrado que cuando la matriz $X'X$ tiene uno o más valores propios no uniformes, los estimadores de β en $Y = X\beta + \varepsilon$, basados en el criterio de la suma mínima de cuadrados de residuales, pueden tener una alta probabilidad de estar alejados de β. Sin embargo, al añadir una pequeña cantidad positiva a cada elemento de la diagonal de la matriz $X'X$, el sistema $[X'X + K]\hat{\beta}^* = X'Y$ actúa como un sistema ortogonal. Cuando K es igual a kI y todas las soluciones en el intervalo $0 \leq k \leq 1$ son obtenidas, es posible conseguir una caracterización bidimensional del sistema y mostrar los tipos de dificultades ocasionadas por la intercorrelación entre las predictoras. Así, los estimadores ridge han sido vistos como una subclase de la clase de las transformaciones lineales del estimador de mínimos cuadrados $\hat{\beta}$, lo que muestra que la condición de admisibilidad utilizada por Hoerl y Kennard para justificar en parte el uso de los estimadores ridge, también se satisface para la clase de estimadores (determinísticamente) reducidos.

Se observó que para la regresión ridge, se emplean de 10 a 30 cálculos para obtener la matriz $(X'X + kI)$, uno para cada valor de k, que son generalmente suficientes para determinar la región donde la traza ridge se estabiliza, pero lo interesante, es que los errores de redondeo no se pueden acumular como en algunos algoritmos de búsqueda paso a paso.

Otra ventaja de los estimadores sesgados, es que se puede utilizar la matriz inversa para un k dado o un rango q, para obtener los coeficientes en los modelos de todas las variables de respuesta, pero en cambio en los mejores subconjuntos y algoritmos de búsqueda paso a paso se requiere correr un cálculo separado para cada respuesta.

Se recurrió a la representación de varios ejemplos, para que de esta manera se pueda realizar un análisis más sencillo en presencia colinealidad en problemas mal condicionados, así como también se tenga un conocimiento más claro de la teoría de los métodos de regresión sesgada, en especial de la regresión ridge.

Apéndice A

Álgebra Matricial

Es cierto que en la actualidad hay una gran gamma de paquetes estadísticos que han impulsado y siguen impulsando enormemente la labor de los investigadores en los cálculos de técnicas multivariadas, ya que por ejemplo, les permite una mejor precisión numérica y un aumento considerable en las posibilidades de análisis, sin embargo, es importante primero entender los conceptos del álgebra matricial para poder comprender e interpretar adecuadamente los resultados que se obtengan. Por lo anterior, en este capítulo se dará una breve introducción a la notación y a las operaciones que se pueden realizar con matrices, con las cuales se simplifica la notación y existe una mejor manipulación algebraica.

A.1. Definiciones básicas

Definición 1: Una **matriz** es un arreglo rectangular de elementos ordenados en renglones (m) y columnas (n). Las matrices se denotan con letras mayúsculas y los elementos de las mismas con letras minúsculas y subíndices que indican el lugar que ocupan, es decir, el término a_{ij} representa el elemento de la i-ésima fila y la j-ésima columna de una matriz A de tamaño $m \times n$, es decir, $A = (a_{ij})_{m\times n}$.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & \vdots & & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mn} \end{bmatrix}$$

Definición 2: El **rango** de un matriz se define como el número de columnas (o renglones) linealmente independientes de la matriz. Cualquier subconjunto de columnas de una matriz es linealmente independiente si sus columnas no pueden se expresadas como una combinación lineal de las otras.

$$\begin{aligned} Rango(A) &= \text{número de renglores linealmente independientes de } A \\ &= \text{número de columnas linealmente independientes de } A \end{aligned}$$

Definición 3: Se dice que una matriz A de tamaño $n \times p$ es de **rango completo** si su rango es igual al valor más pequeño de n y p.

Definición 4: Igualdad de matrices. Sean A y B dos matrices, $A = B$ si y sólo si A y B tienen el mismo tamaño y además cada elemento de a_{ij} de A es igual al correspondiente b_{ij} de B, es decir:

$$A = B \Leftrightarrow a_{ij} = b_{ij} \ \forall i, j$$

A.2. Algunos tipos de matrices

Definición 5: Un **vector** es una matriz que tiene solamente un renglón o una columna y es llamado vector renglón o vector columna, respectivamente.

Definición 6: Una **matriz cuadrada** tiene el mismo número de renglones que de columnas.

Definición 7: La **matriz diagonal** es una matriz cuadrada en donde todos los elementos fuera de la diagonal principal son cero, es decir, $a_{ij} = 0$ si $i \neq j$.

Definición 8: La **matriz identidad** es una matriz diagonal donde todos los elementos de la diagonal son uno y se denota por I_n, donde el subíndice indica el orden de la matriz.

$$I_n = \begin{cases} a_{ij} = 0 & \text{si } i \neq j \\ a_{ij} = 1 & \text{si } i = j \end{cases}$$

Definición 9: Una **matriz simétrica** es una matriz cuadrada done $a_{ij} = a_{ji}$, es decir, los elementos son simétricos con respecto a la diagonal principal.

A.3. Operaciones con matrices

A.3.1. Suma y resta de matrices

La suma (o resta) de matrices está bien definida, si y sólo si las matrices son de la misma dimensión. Entonces, la suma (o resta) consiste en sumar (o restar) los correspondientes elementos de las dos matrices.

$$A_{m\times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \qquad B_{m\times n} = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix}$$

Entonces $C_{m\times n} = A + B$

$$C_{m\times n} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1n}+b_{1n} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2n}+b_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1}+b_{m1} & a_{m2}+b_{m2} & \cdots & a_{mn}+b_{mn} \end{bmatrix}$$

Propiedades de suma de matrices:

- ▷ Conmutativa: $A + B = B + A$
- ▷ Asociativa: $A + (B + C) = (A + B) + C$
- ▷ Elemento neutro: $A + 0 = A$ (matriz cero $0_{m\times n}$)
- ▷ Elemento simétrico: $A + (-A) = 0$

A.3.2. Multiplicación de una matriz por un escalar

Para multiplicar una matriz (A) por un escalar(λ), se tiene que multiplicar el escalar por todos los elementos de la matriz, obteniéndose otra matriz de la misma dimensión.

$$A_{m\times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \qquad \lambda A_{m\times n} = \begin{bmatrix} \lambda a_{11} & \lambda a_{12} & \cdots & \lambda a_{1n} \\ \lambda a_{21} & \lambda a_{22} & \cdots & \lambda\lambda a_{2n} \\ \vdots & \vdots & & \vdots \\ \lambda a_{m1} & \lambda a_{m2} & \cdots & \lambda a_{mn} \end{bmatrix}$$

Propiedades del producto de una matriz por un escalar (sean λ y α escalares):

- ▷ Conmutativa: $\lambda A = A\lambda$
- ▷ Asociativa: $\lambda(\alpha A) = (\lambda\alpha)A$
- ▷ Distributiva: $\lambda(A + B) = \lambda A + \lambda B$

  $$(\lambda + \alpha)A = \lambda A + \alpha A$$

- ▷ Elemento neutro escalar: $1(A) = A$

A.3.3. Multiplicación de matrices

La multiplicación de dos matrices está bien definida si y sólo si, el número de columnas en la primera matriz es igual al número de renglones en la segunda matriz. Sea A una matriz de tamaño $m \times n$ y B una matriz de tamaño $n \times p$, entonces el producto AB es una matriz C de dimensión $m \times p$, donde cada elemento (c_{ij}) es igual a $\sum_{k=1}^{n} a_{ik}b_{kj}$.

El producto de dos matrices, generalmente no es conmutativo, es decir, $AB \neq BA$, de hecho, que el producto de AB esté definido no implica que lo esté BA.

Propiedades del producto de matrices:

- ▷ Asociativa: $(AB)C = A(BC)$
- ▷ Distributiva: $A(B + C) = AB + AC$

$$(A + B)C = AC + BC$$

- ▷ $AI = A$

A.3.4. Matriz transpuesta

La transpuesta de la matriz $A = (a_{ij})_{m\times n}$, es la matriz A' de tamaño $n\times m$ representada por

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

Es decir, los renglones de A se convierten en las columnas de su transpuesta, A'.

Propiedades de la matriz transpuesta:

- ▷ Si A es simétrica $\Rightarrow A' = A$
- ▷ $(A')' = A$
- ▷ $(A + B)' = A' + B'$
- ▷ $(AB)' = B'A'$
- ▷ $(\lambda A)' = \lambda A'$, donde λ es un escalar
- ▷ $Rango(A) = Rango(A'A) = Rango(AA')$

A.3.5. Matriz inversa

Sea A una matriz de $n \times n$. Si existe una matriz B de $n \times n$ tal que $AB = BA = I_n$, en donde I_n es la matriz identidad, entonces B es la inversa de A y se representa como $B = A^{-1}$. Es decir

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

Definición 10: Si A es una matriz cuadrada y de rango completo, entonces A tiene inversa y es una **matriz no-singular**.

Definición 11: Si A es una matriz cuadrada y no tiene rango completo, entonces no existe la inversa de A y se dice que es una **matriz singular**.

Propiedades de la matriz inversa:

- ▷ $(\lambda A)^{-1} = \frac{1}{\lambda}A^{-1}$, donde λ es un escalar
- ▷ $(A^{-1})^{-1} = A$
- ▷ $AA^{-1} = A^{-1}A = I$
- ▷ $(A')^{-1} = (A^{-1})'$
- ▷ $(AB)^{-1} = B^{-1}A^{-1}$
- ▷ Si A es ortogonal, entonces $A^{-1} = A'$
- ▷ Si A es no-singular y simétrica, entonces A^{-1} es también simétrica

Definición 12: Dos vectores a y b, de la misma dimensión son **ortogonales**, si el producto $a'b = 0$.

Definición 13: La **matriz ortogonal**, es aquella cuya transpuesta es igual a su inversa, es decir,

$$A \text{ es ortogonal} \Leftrightarrow AA' = I \Leftrightarrow A' = A^{-1}$$

Definición 14: Una matriz cuadrada es llamada **idempotente** si $AA = A$. El rango de una matriz idempotente es igual a la suma de los elementos de la diagonal.

A.3.6. Matriz definida positiva

Sea $A_{n\times n}$ una matriz simétrica, tal que si $x'Ax > 0 \ \forall \ x \neq 0$ en $\mathbb{R}^n$, entonces A es una matriz definida positiva.

Propiedades de las matrices definidas positivas:

- ▷ Todos los elementos de la diagonal de la matriz deben ser positivos.
- ▷ Si $A_{n\times m}$ es una matriz de rango completo con $n > m$, entonces $A'A$ es definida positiva
- ▷ Si A es definida positiva, entonces existe su inversa A^{-1} que también es definida positiva.
- ▷ Si A es definida positiva y B es una matriz no-singular, entonces $B'AB$ es una matriz definida positiva.

A.3.7. Traza de una matriz

Sea A una matriz de $m \times m$, la traza de A se define como la suma de los elementos de la diagonal principal:

$$traza(A) = \sum_{i=1}^{m} a_{ii} = a_{11} + a_{22} + \cdots + a_{mm}$$

Propiedades de la traza de una matriz:
Sean A y B matrices de $m \times m$, λ y α escalares:

- $traza(I_n) = n$
- $traza(\lambda A) = \lambda traza(A)$
- $traza(A + B) = traza(A) + traza(B)$
- $traza(A') = traza(A)$
- $traza(\lambda A + \alpha B) = \lambda traza(A) + \alpha traza(B)$
- $traza(ABC) = traza(CAB)$
- $traza(AB) = traza(BA)$
- Si C es una matriz ortogonal, entonces $traza(C'AC) = traza(A)$

A.4. Valores y vectores propios

A.4.1. Ecuación característica

Muchas veces resulta útil encontrar un escalar λ y un vector x tales que para que una matriz $A_{n\times n}$ se cumpla

$$Ax = \lambda x$$

que es equivalente a

$$(A - \lambda I)x = 0$$

El sistema homogéneo de ecuaciones $(A - \lambda I)x = 0$ tendrá una solución diferente a la trivial siempre y cuando la matriz $(A - \lambda I)$ sea singular, o cuando $|A - \lambda I| = 0$, esta ecuación se conoce como ecuación caractrística de A.

A.4.2. Valores y vectores propios

Las soluciones de la ecuación $(A - \lambda I)x = 0$ son los valores propios λ y los vectores propios x. Éstos también se conocen como raíces y vectores característicos o eigenvalores y eigenvectores, pero en éste trabajo se refiere a ellos como valores y vectores propios.

Definición 15: Los **valores propios** se obtienen al resolver la ecuación $|A - \lambda I| = 0$, cabe señalar que las raíces de este polinomio no forzosamene son reales. Si la matriz A es de dimensión $n \times n$, entonces A tendrá n valores propios $(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

Definición 16: Una vez encontrados los valores propios, retomando la ecuación $(A - \lambda I)x = 0$ es posible encontrar los **vectores propios** (diferentes de cero) correspondientes a λ.

Propiedad 1: Si tenemos la matriz $B = A + kI$, donde k es un escalar, entonces los vectores propios de B son los mismos que los de A y el i-ésimo valor propio de B es: $\lambda_i + k$, donde λ_i es el i-ésimo valor propio de A.

Propiedad 2: Si A^{-1} existe, entonces A^{-p} tiene los mismos vectores propios que A y λ_i^{-p} es el valor propio de A^{-p} correspondiente a el i-ésimo valor propio de A. En particular, $1/\lambda_i$ es el valor propio de A^{-1} correspondiente a λ_i, el i-ésimo valor propio de A.

Para entender mejor estas definiciones, analicemos un ejemplo.
Sea A una matriz de 2×2

$$A = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}$$

$$|A - \lambda I| = \begin{vmatrix} 4-\lambda & 2 \\ 2 & 4-\lambda \end{vmatrix} = (4-\lambda)(4-\lambda) - 4 = \lambda^2 - 8\lambda + 12$$

resolviendo la ecuación obtenemos los dos valores propios de A

$$\lambda_1 = 6$$
$$\lambda_2 = 2$$

retomando la ecuación $(A - \lambda I)x = 0$ y reemplazando los valores propios de λ, tenemos

$$\begin{bmatrix} 4-\lambda & 2 \\ 2 & 4-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

Para $\lambda_1 = 6$

$$\begin{bmatrix} -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$
$$\Rightarrow x_1 = x_2$$

Para $\lambda_2 = 2$

$$\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\Rightarrow x_1 = -x_2$$

Por lo tanto, los vectores propios de A son,

$$x_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad x_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

sin embargo, es conveniente que los vectores propios estén normalizados, es decir, que cumplan con la condición $x'x = 1$. Por lo que normalizando los vectores tenemos:

$$x_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix} \qquad x_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$$

Cabe señalar que los vectores propios normalizados de una matriz son ortogonales entre sí, es decir, $x_i' x_j = 0 \; \forall \; i \neq j$.

Apéndice B

Regresión Lineal Múltiple

La regresión lineal múltiple explica el comportamiento de la variable dependiente Y con más de una variable regresora X, lo que ofrece la ventaja de utilizar más información en la construcción del modelo y en consecuencia realizar estimaciones más precisas.
A continuación se proporciona un breve repaso de algunos de los resultados de la regresión múltiple, los cuales se utilizan a lo largo del trabajo.

B.1. El modelo de regresión lineal múltiple

El modelo de regresión múltiple con p variables regresoras y basado en n observaciones está dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i \qquad \text{para } i = 1, 2, \ldots, n$$

Escribiendo el modelo para cada una de las observaciones, éste puede ser considerado como un sistema de lineales de la forma,

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_{11} X_1 + \beta_2 X_{12} + \cdots + \beta_p X_{1p} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_{21} X_1 + \beta_2 X_{22} + \cdots + \beta_p X_{2p} + \varepsilon_2 \\
Y_3 &= \beta_0 + \beta_{31} X_1 + \beta_2 X_{32} + \cdots + \beta_p X_{3p} + \varepsilon_3 \\
&\vdots \\
Y_n &= \beta_0 + \beta_{n1} X_1 + \beta_2 X_{n2} + \cdots + \beta_p X_{np} + \varepsilon_n
\end{aligned}$$

que puede ser escrito en forma matricial como,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} & \ldots & X_{1p} \\ 1 & X_{21} & X_{22} & X_{23} & \ldots & X_{2p} \\ 1 & X_{31} & X_{32} & X_{33} & \ldots & X_{3p} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} & \ldots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

es decir:

$$Y = X\beta + \varepsilon$$

donde Y es un vector columna de tamaño $n \times 1$ de las observaciones, X es una matriz de orden $n \times (p+1)$ de los niveles de las variables regresoras, β es el vector de los coeficientes de regresión a ser estimados de $(p+1) \times 1$ y ε es un vector columna aleatorio de orden n. Las suposiciones usuales acerca de ε_i se expresan ahora en términos del vector ε, el cual tiene una distribución normal multivariada con $E(\varepsilon) = 0$ y $Var(\varepsilon) = \sigma^2 I_n$, además los errores no están correlacionados.
El modelo estimado también puede expresarse en forma matricial:

$$\hat{Y} = X\hat{\beta}$$
$$Y - \hat{Y} = \varepsilon$$

B.2. Ecuaciones normales y su solución

En notación matricial, las ecuaciones normales son:

$$X'X\hat{\beta} = X'Y$$

cuya única solución es

$$\hat{\beta} = (X'X)^{-1}(X'Y)$$

donde $\hat{\beta}$ es una matriz de tamaño $(p+1) \times 1$. El producto de $X'X$ generá una matriz en cuya diagonal principal se encuentran las sumas de cuadrados de cada una de las variables regresoras y fuera de la diagonal las sumas de los productos cruzados entre ellas.
Los elementos del producto de $X'Y$ son las sumas de productos entre cada variable regresora y la variable dependiente.
Entonces, la forma general de $X'X$ y de $X'Y$es:

$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i2} & \dots & \sum_{i=1}^{n} X_{ip} \\ \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i2}^2 & \sum_{i=1}^{n} X_{i1} \sum_{i=1}^{n} X_{i2} & \dots & \sum_{i=1}^{n} X_{i1} \sum_{i=1}^{n} X_{ip} \\ \sum_{i=1}^{n} X_{i2} & \sum_{i=1}^{n} X_{i1} \sum_{i=1}^{n} X_{i2} & \sum_{i=1}^{n} X_{i2}^2 & \dots & \sum_{i=1}^{n} X_{i2} \sum_{i=1}^{n} X_{ip} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} X_{ip} & \sum_{i=1}^{n} X_{i1} \sum_{i=1}^{n} X_{ip} & \sum_{i=1}^{n} X_{i2} \sum_{i=1}^{n} X_{ip} & \dots & \sum_{i=1}^{n} X_{ip}^2 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1}Y_i \\ \sum_{i=1}^{n} X_{i2}Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{ip}Y_i \end{bmatrix}$$

Las ecuaciones normales tienen solución única solamente si la inversa de la matrix $X'X$ existe, por lo que la matriz X debe ser de rango completo. La implicación práctica de esta condición es que no debe existir redundancia en la información contenida en X.
El estimador $\hat{\beta}$ por mínimos cuadrados cumple:

$$E(\hat{\beta}) = \beta, \text{ es decir, } \hat{\beta} \text{ es insesgado}$$
$$Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

Además:

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

B.3. Coeficiente de determinación múltiple

El coeficiente de determinación múltiple R^2 se define como:

$$R^2 = \frac{SC_R}{SC_T} = 1 - \frac{SC_{RES}}{SC_T}$$

donde

$$SC_R = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2 = \hat{\beta}'X'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \rightarrow \text{ Suma de Cuadrados de la Regresión}$$

$$SC_T = \sum_{i=1}^{n} (Y_i - \bar{Y}_i)^2 = Y'Y - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \rightarrow \text{ Suma de Cuadrados Totales}$$

$$SC_{RES} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = Y'Y - \hat{\beta}'X'Y \rightarrow \text{ Suma de Cuadrados de Residuales}$$

De lo anterior, se obtine la siguiente expresión:

$$SC_T = SC_{RES} + SC_R$$

R^2 es la proporción de la variación explicada por el modelo. Como $0 \leqslant SC_{RES} \leqslant SC_T$, entonces $0 \leqslant R^2 \leqslant 1$. Los valores de R^2 cercanos a uno implican que la mayor parte de la variabilidad de Y está explicada por el modelo de regresión.
El coeficiente de determinación así definido presenta el inconveniente de que al incluir nuevas variables al modelo, aumenta su valor, incluso cuando éstas no resultan significativas. Por lo que para evitar un aumento injustificado de este coeficiente se introduce el coeficiente de determinación ajustado R_A^2, reemplazando SC_{RES} y SC_T en R^2 por sus correspondientes cuadrados medios ($s = p + 1$), es decir

$$\begin{aligned} R_A^2 &= 1 - \frac{SC_{RES}/(n-s)}{SC_T/(n-1)} \\ &= 1 - \frac{CM_{RES}}{CM_T} \\ &= 1 - \left(\frac{n-1}{n-s}\right)\left(\frac{SC_{RES}}{SC_T}\right) \\ &= 1 - \left(\frac{n-1}{n-s}\right)\left(1 - R^2\right) \end{aligned}$$

Cuando el valor R^2 y de R_A^2 difieran drásticamente, entonces es muy probable que existan variables regresoras que no contribuyan significativamente al modelo ajustado.

Apéndice C

Algunos Resultados Básicos de Estadística

C.1. Esperanza

Sean Y y X variables aleatorias.
Sean λ y α constantes.

- $E(\lambda) = \lambda$
- $E(X + \lambda) = E(X) + \lambda$
- $E(X + Y) = E(X) + E(Y)$
- $E(\lambda X) = \lambda E(X)$
- $E(\lambda + \alpha Y) = \lambda + \alpha E(Y)$

Sea A una matriz de $n \times n$ de constantes.
Sea a un vector de constantes de $n \times 1$.
Sea y un vector aleatoria de $n \times 1$ con media μ y matriz V de varianzas y covariazas no sigular.

- $E(a'y) = a'\mu$
- $E(Ay) = A\mu$
- $E(y'Ay) = traza(AV) + \mu' A\mu$

  Nota: Si $V = \sigma^2 I_n$, entonces $E(y'Ay) = \sigma^2 traza(A) + \mu' A\mu$

C.2. Varianza

La varianza de una variable aleatoria Y se define como:

$$\begin{aligned} Var(Y) &= E\left(Y - E(Y)\right)^2 \\ &= E(Y^2) - \left(E(Y)\right)^2 \end{aligned}$$

Sean λ y α constantes.

- $Var(\alpha) = 0$
- $Var(\lambda + \alpha Y) = \alpha^2 Var(Y)$
- $Var(\lambda + Y) = Var(Y)$
- $Var(\alpha Y) = \alpha^2 Var(Y)$

Sea A una matriz de $n \times n$ de constantes.
Sea a un vector de constantes de $n \times 1$.
Sea y un vector aleatoria de $n \times 1$ con media μ y matriz V de varianzas y covariazas no sigular.

- $Var(a'y) = a'Va$
- $Var(Ay) = AVA'$

  Nota: Si $V = \sigma^2 I_n$, entonces $Var(Ay) = \sigma^2 AA'$

Apéndice D

Sintaxis en R

En este apéndice se presenta la sintaxis utilizada para la obtención de resultados de la sección de aplicaciones, utilizando el software **R**.

######## **Ejemplo del capítulo 6** ########

```
MESQUITE=read.table("MESQUITE.txt", header=TRUE) # Lee el archivo MESQUITE.txt y lo guarda con el nombre MESQUITE.
attach(MESQUITE) # Toma los nombres de las variables que se utilizan en el archivo de texto.

plot(MESQUITE) # Gráfica de correlación de las variables regresoras.

X1=(x1-mean(x1))/(sd(x1)) # Se le resta la media a cada variable y se divide entre su desviación estándar.
X2=(x2-mean(x2))/(sd(x2))
X3=(x3-mean(x3))/(sd(x3))
X4=(x4-mean(x4))/(sd(x4))
X5=(x5-mean(x5))/(sd(x5))
A=matrix(c(matrix(c(1),19),X1,X2,X3,X4,X5),19)
B=solve(t(A) %* %A) %* %t(A) %* %y # Se obtienen los coeficientes por mínimos cuadrados.
SCR=t(B) %* %t(A) %* %y-(sum(y)^2)/19 # Suma de cuadrados de la regresión.
SCE=t(y) %* %y-t(B) %* %t(A) %* %y # Suma de cuadrados del error.
SCT=t(y) %* %y-(sum(y)^2)/19 # Suma de cuadrados totales.
MSR=SSR/5 # Cuadrado medio de la regresión.
MSE=SSE/(19-6) #Cuadrado medio del error.
F=MSR/MSE # Distribución F.
R2=SSR/SYY # Coeficiente de determinación.
R2ajust=1-((19-1)/(19-6))*(1-R2) # Coeficiente de determinación ajustado.
summary(lm(y X1+X2+X3+X4+X5)) #Imprime un resumen estadístico de los resultados del análisis de regresión.

######## Cálculos para obtener los coeficientes de regresión estandarizados ########

X_1=x1-mean(x1)
X_2=x2-mean(x2)
X_3=x3-mean(x3)
X_4=x4-mean(x4)
X_5=x5-mean(x5)
X=matrix(c(X_1,X_2,X_3,X_4,X_5),19)
Y=y
S11=sum((X_1-mean(X_1))^2)
S22=sum((X_2-mean(X_2))^2)
S33=sum((X_3-mean(X_3))^2)
S44=sum((X_4-mean(X_4))^2)
S55=sum((X_5-mean(X_5))^2)
SYY=sum((Y-mean(Y))^2)
S1Y=sum((X_1-mean(X_1))*(Y-mean(Y)))
S2Y=sum((X_2-mean(X_2))*(Y-mean(Y)))
S3Y=sum((X_3-mean(X_3))*(Y-mean(Y)))
S4Y=sum((X_4-mean(X_4))*(Y-mean(Y)))
S5Y=sum((X_5-mean(X_5))*(Y-mean(Y)))
r1Y=S1Y/(sqrt(S11*SYY))
r2Y=S2Y/(sqrt(S22*SYY))
r3Y=S3Y/(sqrt(S33*SYY))
r4Y=S4Y/(sqrt(S44*SYY))
r5Y=S5Y/(sqrt(S55*SYY))
Ycero=matrix(c(r1Y,r2Y,r3Y,r4Y,r5Y),5)
b=solve(cor(X)) %* %Ycero # Coeficientes de regresión estandarizados.

cor(X) #Matriz de correlación.

FIV =solve(cor(X))
TOL=1/FIV
```

```
val_prop=eigen(cor(X), only.values = TRUE)$values # Valores propios.
vec_prop=eigen(cor(X)) # Vectores y valores propios
sqrt(val_prop[1]/val_prop) # Índice condición.

######## Calcular la matriz de proporciones de varianza########

matriz_vectores=matrix(c(-0.5019211, -0.12420127, 0.3827655, 0.62404516, -0.4435172,-0.4634085, -0.26147864, 0.5636324, -0.47179715 , 0.4202469,-
0.5013931, 0.04573042, -0.4222154, -0.52183054, -0.5440044,-0.4740638, -0.13850994, -0.5809933 , 0.33907404 , 0.5509574,-0.2391574, 0.94600565,
0.1413866 , 0.02639623 , 0.1648942),5,5)
vector=t(matriz_vectores) # Matriz de vectores propios.
vector2=vector^2
l1=matrix(c(3.33309263,3.33309263,3.33309263,3.33309263,3.33309263),5)
l2=matrix(c(0.88876733,0.88876733,0.88876733,0.88876733,0.88876733),5)
l3=matrix(c(0.57221330,0.57221330,0.57221330,0.57221330,0.57221330),5)
l4=matrix(c(0.11549788,0.11549788,0.11549788,0.11549788,0.11549788),5)
l5=matrix(c(0.09042886,0.09042886,0.09042886,0.09042886,0.09042886),5)
L=matrix(c(l1,l2,l3,l4,l5),5,5)
cociente=vector2/L
FIV1=sum(cociente[1,])
FIV2=sum(cociente[2,])
FIV3=sum(cociente[3,])
FIV4=sum(cociente[4,])
FIV5=sum(cociente[5,])
P1=cociente[1,]/FIV1
P2=cociente[2,]/FIV2
P3=cociente[3,]/FIV3
P4=cociente[4,]/FIV4
P5=cociente[5,]/FIV5
P=matrix(c(P1,P2,P3,P4,P5),5,5)# Matriz de proporción de varianza.

######## Solución Ridge ########

k0=solve(cor(X)+(0*diag(5))) %*% Ycero # Coeficientes de correlación ridge, para k = 0.
k1=solve(cor(X)+(0.002*diag(5))) %*% Ycero # Coeficientes ridge, para k = 0.002.
k2=solve(cor(X)+(0.004*diag(5))) %*% Ycero # Coeficientes ridge, para k = 0.004.
k3=solve(cor(X)+(0.008*diag(5))) %*% Ycero # Coeficientes ridge, para k = 0.008.
k4=solve(cor(X)+(0.016*diag(5))) %*% Ycero # Coeficientes ridge, para k = 0.016.
k5=solve(cor(X)+(0.032*diag(5))) %*% Ycero # Coeficientes ridge, para k = 0.032.
k6=solve(cor(X)+(0.064*diag(5))) %*% Ycero # Coeficientes ridge, para k = 0.064.
k7=solve(cor(X)+(0.128*diag(5))) %*% Ycero # Coeficientes ridge, para k = 0.128.
k8=solve(cor(X)+(0.256*diag(5))) %*% Ycero # Coeficientes ridge, para k = 0.256.
k9=solve(cor(X)+(0.512*diag(5))) %*% Ycero # Coeficientes ridge, para k = 0.512.
k=matrix(c(0,0.002,0.004,0.008,0.016,0.032,0.064,0.128,0.256,0.512),10)
R=matrix(c(k0,k1,k2,k3,k4,k5,k6,k7,k8,k9),5) # Matriz de coeficientes ridge.

######## Calcular MSE para cada valor de k ########

Ynueva=(y-mean(y))/(sqrt(SYY)) SCE0=(t(Ynueva) %*% Ynueva)-(t(k0) %*% Ycero)-(0*(t(k0) %*% k0))
MSE0=SCE0/13
SCE1=(t(Ynueva) %*% Ynueva)-(t(k2) %*% Ycero)-(0.002*(t(k1) %*% k1))
MSE1=SCE1/13
SCE2=(t(Ynueva) %*% Ynueva)-(t(k2) %*% Ycero)-(0.004*(t(k2) %*% k2))
MSE2=SCE2/13
SCE3=(t(Ynueva) %*% Ynueva)-(t(k3) %*% Ycero)-(0.008*(t(k3) %*% k3))
MSE3=SCE3/13
SCE4=(t(Ynueva) %*% Ynueva)-(t(k4) %*% Ycero)-(0.016*(t(k4) %*% k4))
MSE4=SCE4/13
SCE5=(t(Ynueva) %*% Ynueva)-(t(k5) %*% Ycero)-(0.032*(t(k5) %*% k5))
MSE5=SCE5/13
SCE6=(t(Ynueva) %*% Ynueva)-(t(k6) %*% Ycero)-(0.064*(t(k6) %*% k6))
MSE6=SCE6/13
SCE7=(t(Ynueva) %*% Ynueva)-(t(k7) %*% Ycero)-(0.128*(t(k7) %*% k7))
MSE7=SCE7/13
SCE8=(t(Ynueva) %*% Ynueva)-(t(k8) %*% Ycero)-(0.256*(t(k8) %*% k8))
MSE8=SCE8/13
SCE9=(t(Ynueva) %*% Ynueva)-(t(k9) %*% Ycero)-(0.512*(t(k9) %*% k9))
MSE9=SCE9/13
MSE_RIDGE=matrix(c(MSE0,MSE1,MSE2,MSE3,MSE4,MSE5,MSE6,MSE7,MSE8,MSE9),10) # Matriz de valores del cuadrado medio de
residuales (MSE).

######## Calcular los FIV's para los diferentes valores de k ########

FIV_k0=(solve(cor(X)+diag(5)*0)) %*% (cor(X)) %*% (solve(cor(X)+diag(5)*0))
FIV_k1=(solve(cor(X)+diag(5)*0.002)) %*% (cor(X)) %*% (solve(cor(X)+diag(5)*0.002))
FIV_k2=(solve(cor(X)+diag(5)*0.004)) %*% (cor(X)) %*% (solve(cor(X)+diag(5)*0.004))
FIV_k3=(solve(cor(X)+diag(5)*0.008)) %*% (cor(X)) %*% (solve(cor(X)+diag(5)*0.008))
FIV_k4=(solve(cor(X)+diag(5)*0.016)) %*% (cor(X)) %*% (solve(cor(X)+diag(5)*0.016))
FIV_k5=(solve(cor(X)+diag(5)*0.032)) %*% (cor(X)) %*% (solve(cor(X)+diag(5)*0.032))
FIV_k6=(solve(cor(X)+diag(5)*0.064)) %*% (cor(X)) %*% (solve(cor(X)+diag(5)*0.064))
FIV_k7=(solve(cor(X)+diag(5)*0.128)) %*% (cor(X)) %*% (solve(cor(X)+diag(5)*0.128))
FIV_k8=(solve(cor(X)+diag(5)*0.256)) %*% (cor(X)) %*% (solve(cor(X)+diag(5)*0.256))
FIV_k9=(solve(cor(X)+diag(5)*0.512)) %*% (cor(X)) %*% (solve(cor(X)+diag(5)*0.512))

######## Calcular R^2 para cada valor de k ########

SCT0=(t(Ynueva) %*% Ynueva)-(((sum(Ynueva))^2)/19)
SCE0=(t(Ynueva) %*% Ynueva)-(t(k0) %*% Ycero)-(0*(t(k0) %*% k0))
SCR0=SCT0-SCE0
r0=SCR0/SCT0
```

```
SCT1=(t(Ynueva) %* %Ynueva)-(((sum(Ynueva))^2)/19)
SCE1=(t(Ynueva) %* %Ynueva)-(t(k1) %* %Ycero)-(0.002*(t(k1) %* %k1))
SCR1=SCT1-SCE1
r1=SCR1/SCT1
SCT2=(t(Ynueva) %* %Ynueva)-(((sum(Ynueva))^2)/19)
SCE2=(t(Ynueva) %* %Ynueva)-(t(k2) %* %Ycero)-(0.004*(t(k2) %* %k2))
SCR2=SCT2-SCE2
r2=SCR2/SCT2
SCT3=(t(Ynueva) %* %Ynueva)-(((sum(Ynueva))^2)/19)
SCE3=(t(Ynueva) %* %Ynueva)-(t(k3) %* %Ycero)-(0.008*(t(k3) %* %k3))
SCR3=SCT3-SCE3
r3=SCR3/SCT3
SCT4=(t(Ynueva) %* %Ynueva)-(((sum(Ynueva))^2)/19)
SCE4=(t(Ynueva) %* %Ynueva)-(t(k4) %* %Ycero)-(0.016*(t(k4) %* %k4))
SCR4=SCT4-SCE4
r4=SCR4/SCT4
SCT5=(t(Ynueva) %* %Ynueva)-(((sum(Ynueva))^2)/19)
SCE5=(t(Ynueva) %* %Ynueva)-(t(k5) %* %Ycero)-(0.032*(t(k5) %* %k5))
SCR5=SCT5-SCE5
r5=SCR5/SCT5
SCT6=(t(Ynueva) %* %Ynueva)-(((sum(Ynueva))^2)/19)
SCE6=(t(Ynueva) %* %Ynueva)-(t(k6) %* %Ycero)-(0.064*(t(k6) %* %k6))
SCR6=SCT6-SCE6
r6=SCR6/SCT6
SCT7=(t(Ynueva) %* %Ynueva)-(((sum(Ynueva))^2)/19)
SCE7=(t(Ynueva) %* %Ynueva)-(t(k7) %* %Ycero)-(0.128*(t(k7) %* %k7))
SCR7=SCT7-SCE7
r7=SCR7/SCT7
SCT8=(t(Ynueva) %* %Ynueva)-(((sum(Ynueva))^2)/19)
SCE8=(t(Ynueva) %* %Ynueva)-(t(k8) %* %Ycero)-(0.256*(t(k8) %* %k8))
SCR8=SCT8-SCE8
r8=SCR8/SCT8
SCT9=(t(Ynueva) %* %Ynueva)-(((sum(Ynueva))^2)/19)
SCE9=(t(Ynueva) %* %Ynueva)-(t(k9) %* %Ycero)-(0.512*(t(k9) %* %k9))
SCR9=SCT9-SCE9
r9=SCR9/SCT9
r_RIDGE=matrix(c(r0,r1,r2,r3,r4,r5,r6,r7,r8,r9),10) #Matriz de valores del coeficiente de determinación (R^2).

######## Gráfica de la Traza Ridge ########

plot(k,R[1,],type="l",xlab=expression(k),ylab=expression(hat(beta)),col=red",xlim=c(0,1/2),ylim=c(-1/2,3/4))
par(new=TRUE)
plot(k,R[2,],type="l",xlab=,ylab=,col="blue",xlim=c(0,1/2),ylim=c(-1/2,3/4),axes=FALSE)
par(new=TRUE)
plot(k,R[3,],type="l",xlab=,ylab=,col="purple",xlim=c(0,1/2),ylim=c(-1/2,3/4),axes=FALSE)
par(new=TRUE)
plot(k,R[4,],type="l",xlab=,ylab=,col="pink",xlim=c(0,1/2),ylim=c(-1/2,3/4),axes=FALSE)
par(new=TRUE)
plot(k,R[5,],type="l",xlab=,ylab=,col="skyblue",xlim=c(0,1/2),ylim=c(-1/2,3/4),axes=FALSE)
abline(a=0,b=0)
title(main="Traza Ridge")

text(.05,.63,expression(hat(beta)[1]),col=red",cex=.7)
text(.1,.32,expression(hat(beta)[2]),col="blue",cex=.7)
text(.15,.08,expression(hat(beta)[3]),col="purple",cex=.7)
text(.07,-.08,expression(hat(beta)[4]),col="pink",cex=.7)
text(.45,-.05,expression(hat(beta)[5]),col="skyblue",cex=.7)

######## Calcular el modelo de regresión ridge en términos de las regresoras originales ########

B_RIDGE=solve((t(A) %* %A)+(diag(6)*.128)) %* %t(A) %* %y
SCE_RIDGE=t(y) %* %y-t(B_RIDGE) %* %t(A) %* %y
CME_RIDGE=SCE_RIDGE/13
sqrt(CME_RIDGE) # Desviación estándar de residuales.
```

Bibliografía

[1] Armitage, P. & Berry, G. L. (1997), *"Estadística para la Investigación Biomédica"*, Tercera Edición, Harcourt Brace.

[2] Carter, E. M. & Srivastava, M.S. (1983), *"An Introduction to Applied Multivariate Statistics"*, New York: North Holland.

[3] Clutton-Brock, M. (1965), *"Using the observations to estimate prior distribution"*, Journal of the Royal Statistical Society, Series B 27, 17-27.

[4] Dickey, D. A. & Pantula, S.G. & Rawlings, J. O. (1998), *"Applied Regression Analysis: A Research Tool"*, Second Edition, Springer.

[5] Freud, R. J. & Wilson, W. J. & Sa, P. (2006), *"Regression Analysis: Statistical Modeling of a Response Variable"*, Second Edition, Elsevier Inc.

[6] Hoerl, A. E. & Kennard, R. W. (1970), *"Ridge Regression: Biased Estimation for Nonorthogonal Problems"*, Technometrics Vol. 12 No. 1, 55-67.

[7] Kleinbaum, D. G. & Kupper, L. L. & Muller, K.E. (1998), *"Applied Regression Analysis and Other Multivariable Methods"*, Second Edition, Duxbury Press.

[8] Marquardt, D. W. & Snee, R. D. (1975), *"Ridge Regression in Practice"*, The American Statistician Vol. 29 No. 1, 3-20.

[9] Mayer, L. S. & Willke, T. A. (1973), *"On Biased Estimation in Linear Models"*, Technometrics Vol. 15 No. 3, 497-508.

[10] Montgomery, D. C. & Peck, E. A. (1992), *"Introduction to Linear Regression Analysis"*, Second Edition, John Wiley & Sons.

[11] Rencher, A. C. (2002), *"Methods of Multivariate Analysis"*, Second Edition, John Wiley & Sons.