Universidad del Bío-Bío. Sistema de Bibliotecas - Chile





UNIVERSIDAD DEL BÍO-BÍO
FACULTAD DE CIENCIAS

DEPARTAMENTO DE ESTADÍSTICA

TESIS DE MAGÍSTER EN MATEMÁTICA
MENCIÓN ESTADÍSTICA

ESTIMADORES DE CONTRACCIÓN PARA MODELOS DE REGRESIÓN BETA

por Romer Alberto Machaca Aguilar

Profesor Guía: Manuel González Navarrete
Profesor Co-guía: Luis Alberto Firinguetti Limone

Abril del 2022

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Esta tesis está dedicada a mi madre y a mi padre por todo su amor, apoyo, comprensión y sacrificio.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Agradecimientos

A todas las personas que me apoyaron e hicieron posible que este trabajo se realice con éxito.

A mis tutores por el tiempo dedicado y los conocimientos brindados.

A mi familia por confiar, apoyarme incondicionalmente y creer en mí.

A mis amigos por apoyarme y darme fuerzas.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Resumen

En el presente trabajo se realiza un estudio de los estimadores de contracción de un modelo de regresión beta, presentando en primera instancia a la distribución beta y sus propiedades. Posteriormente se realiza la estimación de los parámetros del modelo sin penalizaciones mediante el método de máxima verosimilitud, donde se obtiene un sistema de ecuaciones no lineal, por lo cual, para llegar a un resultado numérico se acude a métodos iterativos (método de Newton Raphson).

En el caso de que las variables explicativas presenten multicolinealidad, se recurre a métodos de contracción que disminuyen la varianza de los estimadores. Sin embargo, también incrementan su sesgo. Los estimadores de contracción que se estudian en esta investigación son el estimador de regresión ridge y el estimador de regresión lasso (least absolute shrinkage and selection operator).

El estimador ridge se calcula mediante dos métodos de estimación: el método de mínimos cuadrados iterativamente ponderados y el método de máxima verosimilitud. Luego se demuestra que existe un parámetro de contracción, de tal manera que el error cuadrático medio del estimador ridge sea menor al del estimador de máxima verosimilitud. En cuanto al estimador lasso, este es calculado minimizando el negativo de la log-verosimilitud penalizada mediante el método descenso coordinado. Su error cuadrático medio sólo es estudiado mediante simulaciones debido a que su estimador no presenta una forma cerrada.

Posteriormente, se realizan simulaciones para comparar el error cuadrático medio de los tres estimadores estudiados: estimador de máxima verosimilitud, estimador ridge y estimador lasso, los cuales son hallados mediante funciones creadas en el software estadístico R. Por último, se realiza una aplicación mostrando algunas medidas de diagnóstico de cada estimador.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Índice general

Agradecimientos **II**
Resumen III

Lista de figuras **VI**

Lista de tablas **VII**

1. Introducción **1**
1.1. Objetivos 3
1.1.1. Objetivo general 3
1.1.2. Objetivos específicos 3

2. Modelo de regresión Beta **4**
2.1. Distribución Beta 4
2.2. Modelo de regresión beta 6
2.3. Función de score y matriz de información de Fisher 7
2.4. Estimador de máxima verosimilitud 10
2.5. Estimación para ϕ conocido 12

3. Estimadores de contracción **13**
3.1. Estimador ridge 14
3.1.1. Estimador mínimos cuadrados ponderados 14
3.1.2. Estimador de máxima verosimilitud penalizada 16
3.1.3. Propiedades del estimador ridge 18
3.2. Estimador lasso 21
3.3. Parámetro de contracción k 25
3.3.1. Métodos de estimación de k para el estimador ridge 26
3.3.2. Validación cruzada 26

4. Simulación y aplicación **28**
4.1. Simulación 28
4.1.1. Programa para relacionar la correlación y el ECM 28
4.1.2. Comparación de los estimadores ridge con distintos parámetros de contracción 31
4.1.3. Comparación de los estimadores de máxima verosimilitud, ridge y lasso 33
4.2. Aplicación 37

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

5. Conclusiones y trabajos futuros **42**
5.1. Conclusiones 42
5.2. Trabajos futuros 43

A. Principios fundamentales **44**
A.1. Conceptos básicos para la estimación lasso 44
A.2. Descenso coordinado. 47
A.3. Método lasso para un modelo de regresión clásico 48
A.4. Método de Newton-Raphson 48
A.5. Método de Fisher-Scoring 49

B. Códigos en R **50**
B.1. Funciones utilizadas 50
B.1.1. Simulación de las variables aleatorias 50
B.1.2. Código para el modelo de regresión beta 51
B.1.3. Código para el modelo de regresión beta ridge 54
B.1.4. Código para el modelo de regresión lasso 55
B.1.5. Código para calcular el ECM 57
B.1.6. Código de funciones para estimar k 58
B.2. Creación de tablas de la sección simulación 61
B.3. Código utilizado para la aplicación 63

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Índice de figuras

2.1. Función de densidad de beta para distintas combinaciones de (μ, ϕ) . . . 5

3.1. Relación entre el parámetro de contracción k y el $ECM(\hat{\beta}_{ridge})$ 20
3.2. Relación entre el parámetro de contracción k y el $ECM(\hat{\beta}_{lasso})$ 25

A.1. Operador soft-thresholding 45
A.2. Función valor absoluto y su derivada 46

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Índice de tablas

4.1. Error cuadrático medio de los estimadores de máxima verosimilitud y ridge para distintos estimadores de k, simulaciones de $S = 10000$ repeticiones, para un tamaño de muestra de $n = 30, 50, 100, 200$, $\phi = 2$, $p = 4$ y distintos niveles de colinealidad 32
4.2. Error cuadrático medio de los estimadores de máxima verosimilitud y ridge para distintos estimadores de k, simulaciones de $S = 10000$ repeticiones, para un tamaño de muestra de $n = 30, 50, 100, 200$, $\phi = 4$, $p = 4$ y distintos niveles de colinealidad 32
4.3. Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 30$, $\phi = 2$ y $p = 4$ . . . 34
4.4. Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 50$, $\phi = 2$, $p = 4$ 34
4.5. Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 100$, $\phi = 2$, $p = 4$ . . . 34
4.6. Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 30$, $\phi = 10$, $p = 4$ . . . 35
4.7. Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 50$, $\phi = 10$, $p = 4$ . . . 35
4.8. Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 100$, $\phi = 10$, $p = 4$ . . . 36
4.9. Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para un tamaño de muestra de $n = 30$, $\phi = 2$, $p = 6$ con 2 parámetros iguales a 0 ($\beta_5 = \beta_6 = 0$) 36
4.10. Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para un tamaño de muestra de $n = 50$, $\phi = 2$, $p = 6$ con 2 parámetros iguales a 0 ($\beta_5 = \beta_6 = 0$) 37
4.11. Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para un tamaño de muestra de $n = 100$, $\phi = 2$, $p = 6$ con 2 parámetros iguales a 0 ($\beta_5 = \beta_6 = 0$) 37
4.12. Estimadores de máxima verosimilitud, ridge y lasso para el conjunto de datos 40

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Capítulo 1

Introducción

Los modelos lineales de regresión permiten realizar un análisis entre dos o más variables de las cuales una es dependiente del resto. La variable dependiente se le denomina variable respuesta (llamada también endógena) y las variables de las que depende son las variables explicativas (o exógenas). Usualmente para construir un modelo cualquiera se realizan algunas suposiciones. Una de estas es que las variables explicativas son linealmente independientes, es decir, que no existe una relación entre ellas. Sin embargo, en la mayoría de los casos esta suposición no se cumple (problema de multicolinealidad), resultando en varianzas grandes para los estimadores, lo que a su vez afecta la precisión de estos.

Una solución a este problema son los estimadores de contracción, los cuales sesgan los estimadores de un modelo con el objetivo de disminuir su varianza, de tal manera que su error cuadrático medio sea menor al calculado con estimadores insesgados. Entre este tipo de estimadores están:

- El estimador garotte propuesto por Breiman [2].
- El estimador lasso desarrollado por R. Tibshirani [22] cuyas bases vienen del estimador garotte.
- El estimador ridge desarrollado por Hoerl & Kennard [11].
- El estimador elastic net propuesto por Zou & Hastie [25] el cual generaliza los dos anteriores.

Los estimadores que serán estudiados en este trabajo son lasso (least absolute shrinkage and selection operator por sus siglas en ingles) y ridge o traducido al español cresta.

El estimador ridge se halla minimizando la suma de cuadrados del error sujeto a una restricción cuadrática que depende de los parámetros del modelo y de un parámetro de contracción. Hoerl & Kennard [11] hallan el estimador para el modelo de regresión clásico donde se supone que la variable dependiente tiene distribución normal. Los autores minimizan la longitud del estimador sujeto a una expresión cuadrática que depende del estimador mínimos cuadrados ordinarios (mco). De esta manera, el estimador ridge depende del estimador mco.

Por otro lado, Tibshirani [22] halló el estimador lasso para una variable dependiente que se distribuye normal. Este es hallado minimizando la suma de cuadrados del error

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

sujeto a una restricción que depende del valor absoluto de los parámetros, y al igual que el método ridge, de un parámetro de contracción. Al hallar el estimador lasso, surge el inconveniente de que el valor absoluto no es derivable. Por tal motivo, se hace uso de un algoritmo denominado descenso coordinado (coordinate descent) y del operador umbral suave (soft-thresholding). Este método puede hallar elementos del estimador de β iguales a cero, es decir, lasso también realiza una selección de variables.

Los estimadores antes descritos son hallados suponiendo que la distribución de la variable dependiente es normal. Ante tal limitante, algunos investigadores han realizado un análisis de ridge y lasso con otro tipo de distribuciones. Es así, que se estudia el modelo lineal generalizado (MLG), que es una extensión del modelo lineal clásico, ya que la función de probabilidad de la variable dependiente pertenece a la familia exponencial. Este tipo de modelos se dividen en tres componentes: el componente aleatorio, el componente sistemático y la función de enlace la cual relaciona las dos anteriores.

La estimación de los parámetros del MLG se realiza mediante el método de máxima verosimilitud (emv), dando como resultado un sistema de ecuaciones no lineales, que es resuelto mediante el método iterativo de Newton Raphson. Este método de estimación puede ser consultado en Dobson & Barnett [4].

En el caso de que las variables no sean linealmente independientes (multicolinealidad perfecta), Schaefer & et. al [23] y Le Cessie & Van Houwelingen [14], utilizaron el método ridge para un modelo logístico (un caso particular del MLG). Dicho método consiste en realizar una penalización en los elementos de las variables explicativas de tal forma que la matriz construida con ellas deje de presentar colinealidad.

Friedman et. al. [7] aplicaron el método lasso para variables dependientes con funciones de probabilidad pertenecientes a la familia exponencial. Al igual que en el caso de la distribución normal, se realiza la estimación con ayuda del operador soft-thresholding y el algoritmo descenso coordinado. En este caso no se minimiza el error al cuadrado, sino el negativo de la log-verosimilitud penalizada de la función de probabilidad la cual será aproximada mediante las ecuaciones de Taylor. Estos autores también desarrollaron funciones en el software estadístico R, aplicando los métodos ridge, lasso y elastic net para cualquier distribución que pertenezca a la familia exponencial. El paquete en el que se desarrollan estas funciones es **glmnet**.

Sin embargo, los modelos anteriores (modelo lineal clásico y MLG) no son adecuados cuando la variable dependiente es una proporción o una tasa, dado que estos últimos son acotados. Es así que este trabajo se centrará en un modelo de regresión que modela este tipo de variables.

El modelo de regresión beta tiene el supuesto de que la distribución de la variable dependiente es beta con parámetros μ y ϕ, donde μ es la media y ϕ es el parámetro de precisión. Esta distribución modela el comportamiento de variables que están acotadas. Ferrari & Cribari-Neto [5] fueron los precursores de este modelo, quienes realizaron la estimación de sus parámetros, utilizando la misma metodología que en el MLG, iniciando con la maximización de la log-verosimilitud del modelo y con ayuda del método iterativo de Newton Raphson se obtienen los estimadores del modelo de regresión beta.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Los autores Qasim, Mansson & Kibria [18] además de Abonazel & Taha [1] aplicaron el método ridge al modelo de regresión beta, obteniendo los estimadores de regresión beta ridge minimizando la longitud del estimador de β sujeto a una restricción cuadrática.

En el presente trabajo, se realiza la estimación de los parámetros del modelo de regresión beta ridge mediante la maximización de la log-verosimilitud penalizada, la cual se compara con el estimador beta ridge hallado con la minimización de la longitud del estimador de β. Se aplica además el método lasso en el modelo de regresión beta y se compara este estimador con el estimador ridge y el estimador sin penalizar.

En el capítulo 2 de este trabajo se estudia la distribución beta además de sus propiedades. Se desarrolla también el estimador del modelo de regresión beta sin penalizar, aplicando el método de máxima verosimilitud. En el capítulo 3 se hallan los estimadores mediante los métodos ridge y lasso, posteriormente se halla el estimador de contracción mediante distintas metodologías. En el capítulo 4, se realizan simulaciones para la comparación del error cuadrático medio de los estimadores antes expuestos y se presenta un ejemplo con datos reales. Por último, se escriben algunas conclusiones y recomendaciones del presente estudio.

1.1. Objetivos

1.1.1. Objetivo general

Estimar los parámetros del modelo de regresión beta bajo el supuesto de multicolinealidad mediante los métodos de regresión ridge y lasso.

1.1.2. Objetivos específicos

- Hallar los estimadores de regresión beta ridge, estimadores de regresión lasso y los estimadores de máxima verosimilitud sin penalizar.
- Evaluar las propiedades de los estimadores de regresión beta ridge y lasso mediante simulaciones.
- Comparar los estimadores ridge, lasso y el estimador de máxima verosimilitud sin penalizar, mediante el error cuadrático medio.
- Estimar el parámetro de contracción k del método ridge y lasso mediante diferentes metodologías.
- Programar en el software estadístico R la estimación e inferencia de los parámetros del modelo de regresión.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Capítulo 2

Modelo de regresión Beta

2.1. Distribución Beta

Las funciones de probabilidad muestran el comportamiento que tienen los diferentes tipos de variables. Por ejemplo, la temperatura ambiental (medida en grados centígrados) puede ser analizada con una distribución normal. La cantidad de caras en 10 lanzamientos puede ser visto mediante la distribución binomial. La proporción de estudiantes que pasaron cierta materia puede ser analizada mediante la distribución beta. El presente trabajo realiza un modelo para variables que estén acotadas, por tal motivo, en esta sección se realiza un estudio de la última distribución mencionada.

La distribución beta nos permite representar variables aleatorias cuyos valores están restringidos a un intervalo de longitud finita. Una variable aleatoria Y se dice que sigue una distribución beta de parámetros $a > 0$ y $b > 0$, $(Y \sim beta(a, b))$, si la función de densidad tiene la siguiente forma:

$$f(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1} = \frac{1}{B(a,b)} y^{a-1}(1-y)^{b-1}, \tag{2.1}$$

donde $\Gamma(.)$ es la función gamma y $B(a, b)$ es la función beta:

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Observe que si $a = b = 1$ la distribución uniforme es un caso particular de la distribución beta.

La esperanza y varianza de la distribución beta están dados por

$$E(Y) = \frac{a}{a+b} \qquad \text{y} \qquad V(Y) = \frac{ab}{(a+b)^2(a+b+1)}.$$

Para fines prácticos, la función de densidad beta definida en la ecuación (2.1) se reformula realizando una reparametrización de la siguiente manera, $\mu = \dfrac{a}{a+b}$ y $\phi = a+b$, por lo cual $a = \mu\phi$ y $b = (1-\mu)\phi$. Entonces, la esperanza y la varianza de la variable respuesta se pueden escribir en función de μ y ϕ

$$E[Y] = \mu \qquad\qquad V[Y] = \frac{\mu(1-\mu)}{1+\phi}.$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

La función de densidad escrita con la nueva reparametrización es:

$$f(y,\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, 0 < y < 1, \tag{2.2}$$

donde $0 < \mu < 1$ es la media de la distribución y $\phi > 0$ es el parámetro de precisión.

Las siguientes transformaciones de la variable Y serán necesarias para operaciones posteriores en la construcción del modelo:

- $$E[1-Y] = 1-\mu. \tag{2.3}$$
- $E[\log(Y)] = \psi(\mu\phi) - \psi(\phi)$, donde $\psi(.)$ es la función digamma definida como la derivada logarítmica de la función gamma, es decir:

$$\psi(z) = \frac{\partial \log\Gamma(z)}{\partial z} = \frac{\Gamma'(z)}{\Gamma(z)}. \tag{2.4}$$

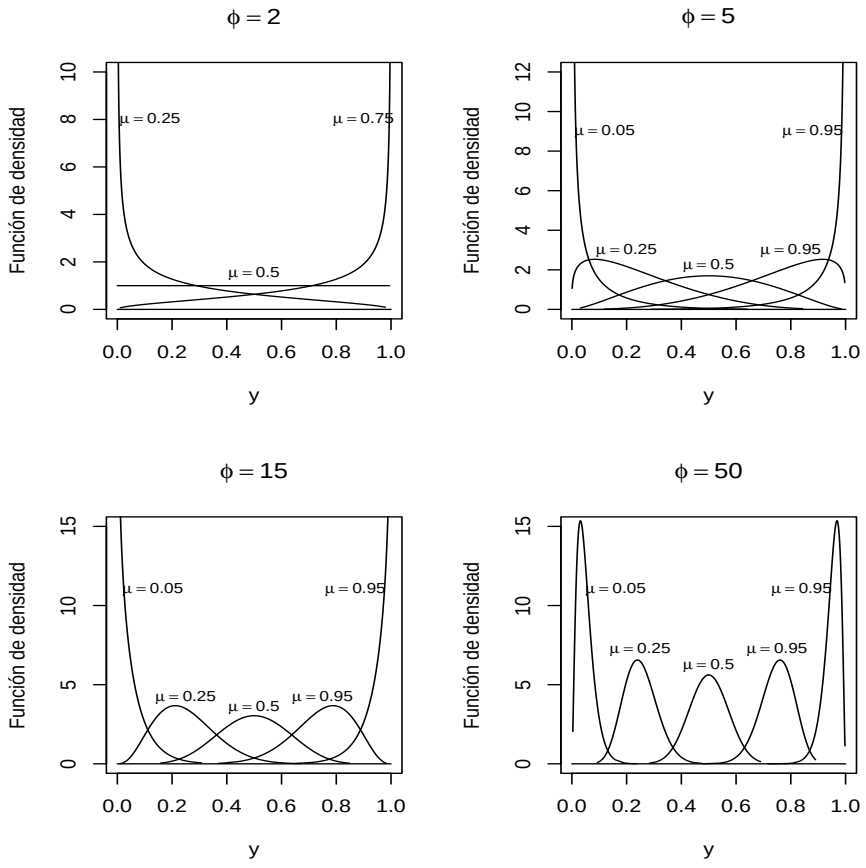


Figura 2.1: Función de densidad de beta para distintas combinaciones de (μ, ϕ)

La figura (2.1) muestra diferentes densidades de la distribución beta para distintas combinaciones de los parámetros (μ, ϕ). En el primer gráfico de densidad, se observa la distribución uniforme con $\mu = 0{,}5$ y $\phi = 2$ que equivale a $a = b = 1$; en este gráfico, como en los demás se aprecia que según μ va aumentando, la densidad de la función va

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

tomando forma de una exponencial, por el contrario, cuando μ disminuye, la densidad toma la forma de una exponencial invertida, por lo cual, μ estaría controlando la simetría de la distribución llegando a una función simétrica cuando $\mu = 0{,}5$.

Por otro lado, el parámetro de precisión ϕ cumple un papel inverso a la varianza: mientras más grande es ϕ, los datos están más concentrados hacia la media μ. Por ejemplo, se observa en la figura 2.1 que con un $\phi = 5$ los datos se distribuyen casi de manera uniforme en el espacio muestral de Y, mientras que para un ϕ más grande ($\phi = 15; 50$) los datos se van concentrando hacia un punto que llega a ser μ.

2.2. Modelo de regresión beta

Sean $y_1, y_2, \ldots, y_n$ variables aleatorias independientes entre sí, donde y_i, $i = 1, 2, \ldots, n$ sigue la función de densidad beta definida en la ecuación (2.2) con media μ y parámetro de precisión ϕ, la cual será denominada como variable dependiente. Sean además los vectores $x_1, x_2, \ldots, x_n$, con, $x_i^{tp} = \begin{pmatrix} x_{i1} & x_{i2} & \ldots & x_{ip} \end{pmatrix}$, donde p es el número de variables explicativas y tp denota la transpuesta. Luego, la matriz de diseño está definida por:

$$X = \begin{pmatrix} x_1^{tp} \\ x_2^{tp} \\ \vdots \\ x_n^{tp} \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

El conjunto de parámetros que se desea estimar en un modelo de regresión, está dado por $\beta_1, \beta_2, \ldots, \beta_p$. Otra de las suposiciones que se hace en este trabajo es que el número de variables debe ser menor al número de observaciones ($p < n$). Sin embargo, cabe señalar que el método lasso y el método ridge también fueron estudiados para tratar con problemas de alta dimensión, es decir, con problemas del tipo $p \geq n$.

La función que relaciona a y y X es el tercer componente de un MLG, la función de enlace, que se denota con g y es definido mediante:

$$g(\mu_t) = x_t^{tp}\beta,$$

la cual, es monótona y al menos dos veces diferenciable. Existe una variedad de funciones de enlace, cuya utilidad dependerá de la función de densidad que tiene el modelo de regresión; entre estas están: la función de enlace de probabilidad lineal, función de enlace probit, función de enlace log-log complementario, función de enlace logit, entre otros. La función de enlace que se utiliza para analizar el modelo de regresión beta en este trabajo es la función logit:

$$g(\mu_t) = \log\left(\frac{\mu_t}{1-\mu_t}\right). \tag{2.5}$$

Despejando μ_t se tiene que

$$\mu_t = \frac{\exp\left(x_t^{tp}\beta\right)}{1 + \exp\left(x_t^{tp}\beta\right)}.$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Una vez construida la estructura del modelo, se procede a realizar la estimación del vector de parámetros β mediante el método de máxima verosimilitud. Sin embargo, como se verá al finalizar la estimación, este estimador depende del parámetro de precisión ϕ. Por lo tanto, es necesario estimar también este último.

2.3. Función de score y matriz de información de Fisher

Considerando la función de densidad de la variable y, definida en la ecuación (2.2), y tomando una muestra de n observaciones distribuidas según una distribución beta se tiene que la función de verosimilitud está dada por:

$$L(\beta, \phi) = \prod L_t(\mu_t, \phi),$$

al aplicar logaritmo, se tiene que la función de log-verosimilitud es dada por:

$$l(\beta, \phi) = \sum_{t=1}^{n} l_t(\mu_t, \phi), \tag{2.6}$$

donde

$$\begin{aligned} l_t(\mu_t, \phi) &= \log\Gamma(\phi) - \log\Gamma(\mu_t\phi) - \log\Gamma((1-\mu_t)\phi) + (\mu_t\phi - 1)\log y_t + \\ &\quad [(1-\mu_t)\phi - 1]\log(1-y_t). \end{aligned} \tag{2.7}$$

Derivando primero respecto de β_i se tiene que:

$$\frac{\partial l(\beta,\phi)}{\partial \beta_i} = \sum_{t=1}^{n} \frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\frac{\partial \mu_t}{\partial \eta_t}\frac{\partial \eta_t}{\partial \beta_i}, \tag{2.8}$$

donde

$$\begin{aligned} \frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t} &= -\phi\psi(\mu_t\phi) + \phi\psi((1-\mu_t)\phi) + \phi\log y_t - \phi\log(1-y_t) \\ &= \phi\left[\log y_t - \log(1-y_t) - \psi(\mu_t\phi) + \psi((1-\mu_t)\phi)\right] \\ &= \phi\left[\log\left(\frac{y_t}{1-y_t}\right) - \psi(\mu_t\phi) + \psi\left((1-\mu_t)\phi\right)\right]. \end{aligned}$$

Denotando con $y_t^* = \log\left(\frac{y_t}{1-y_t}\right)$ y $\mu_t^* = \psi(\mu_t\phi) - \psi\left((1-\mu_t)\phi\right)$, se tiene que

$$\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t} = \phi(y_t^* - \mu_t^*). \tag{2.9}$$

Por otro lado,

$$\begin{aligned} g(\mu_t) &= x_t'\beta = \eta_t, \\ g'(\mu_t) &= \frac{\partial \eta_t}{\partial \mu_t}, \\ \frac{\partial \mu_t}{\partial \eta_t} &= \frac{1}{g'(\mu_t)}. \end{aligned} \tag{2.10}$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Y por último, se tiene que $\dfrac{\partial \eta_t}{\partial \beta_i} = x_{ti}$. Entonces, uniendo las ecuaciones (2.9), (2.10) y $\dfrac{\partial \eta_t}{\partial \beta_i}$ se obtiene:

$$\frac{\partial l(\beta,\phi)}{\partial \beta_i} = \sum_{t=1}^{n} \phi(y_t^* - \mu_t^*)\frac{1}{g'(\mu_t)}x_{ti}.$$

Al expresar en términos matriciales se obtiene la función score para el vector de parámetros β es dada por

$$U_\beta(\beta,\phi) = \phi X^{tp} T(y^* - \mu^*), \tag{2.11}$$

donde X es una matriz de $n \times p$, y^* y μ^* son vectores dados por $y^* = (y_1^* \dots y_n^*)^{tp}$ y $\mu^* = (\mu_1^* \dots \mu_n^*)^{tp}$ y T es una matriz diagonal definida de la siguiente manera

$$T = \text{Diag}\left(\frac{1}{g'(\mu_1)} \dots \frac{1}{g'(\mu_n)}\right),$$

en que Diag(.) denota la construcción de una matriz diagonal con el vector que está dentro del paréntesis.

Ahora derivando (2.6) respecto a ϕ:

$$\begin{aligned}
\frac{\partial l_t(\mu_t,\phi)}{\partial \phi} &= \sum_{t=1}^{n} \psi(\phi) - \mu_t\psi(\mu_t\phi) - (1-\mu_t)\psi((1-\mu_t)\phi) + \mu_t \log y_t + (1-\mu_t)\log(1-y_t) \\
&= \sum_{t=1}^{n} \psi(\phi) - \psi((1-\mu_t)\phi) + \mu_t\left[-\psi(\mu_t\phi) + \psi((1-\mu_t)\phi)\right. \\
&\quad \left. + \log y_t - \log(1-y_t)\right] + \log(1-y_t) \\
&= \sum_{t=1}^{n} \psi(\phi) - \psi((1-\mu_t)\phi) + \mu_t(y_t^* - \mu_t^*) + \log(1-y_t).
\end{aligned}$$

Entonces, la función de score respecto de ϕ está dada por

$$U_\phi(\beta,\phi) = \sum_{t=1}^{n} \psi(\phi) - \psi((1-\mu_t)\phi) + \mu_t(y_t^* - \mu_t^*) + \log(1-y_t), \tag{2.12}$$

de esta manera, se obtiene la función de score $U = (U_\beta(\beta,\phi), U_\phi(\beta,\phi))^{tp}$.

Luego se halla la matriz de información de Fisher calculando la segunda derivada de $l(\beta,\phi)$ respecto a los parámetros β y ϕ. Para esto, primero se deriva la ecuación (2.11) respecto de β_j

$$\begin{aligned}
\frac{\partial^2 l(\beta,\phi)}{\partial \beta_j \beta_i} &= \sum_{t=1}^{n} \frac{\partial}{\partial \beta_j}\left(\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\frac{\partial \mu_t}{\partial \eta_t}\frac{\partial \eta_t}{\partial \beta_i}\right) \\
&= \sum_{t=1}^{n} \frac{\partial}{\partial \mu_t}\left(\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\frac{\partial \mu_t}{\partial \eta_t}x_{ti}\right)\frac{\partial \mu_t}{\partial \eta_t}\frac{\partial \eta_t}{\partial \beta_j} \\
&= \sum_{t=1}^{n} \left(\frac{\partial^2 l_t(\mu_t,\phi)}{\partial^2 \mu_t}\frac{\partial \mu_t}{\partial \eta_t} + \frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\frac{\partial^2 \mu_t}{\partial \mu_t \eta_t}\right)\frac{\partial \mu_t}{\partial \eta_t}x_{ti}x_{tj}, \\
E\left(\frac{\partial^2 l(\beta,\phi)}{\partial \beta_j \beta_i}\right) &= \sum_{t=1}^{n} E\left(\frac{\partial^2 l_t(\mu_t,\phi)}{\partial^2 \mu_t}\left(\frac{\partial \mu_t}{\partial \eta_t}\right)^2 x_{ti}x_{tj}\right) + E\left(\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\frac{\partial^2 \mu_t}{\partial \mu_t \eta_t}\frac{\partial \mu_t}{\partial \eta_t}x_{ti}x_{tj}\right) \\
E\left(\frac{\partial^2 l(\beta,\phi)}{\partial \beta_j \beta_i}\right) &= \sum_{t=1}^{n} E\left(\frac{\partial^2 l_t(\mu_t,\phi)}{\partial^2 \mu_t}\right)\left(\frac{\partial \mu_t}{\partial \eta_t}\right)^2 x_{ti}x_{tj} + E\left(\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\right)\frac{\partial^2 \mu_t}{\partial \mu_t \eta_t}\frac{\partial \mu_t}{\partial \eta_t}x_{ti}x_{tj}.
\end{aligned}$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Analizando el segundo término del lado derecho de la ecuación y por propiedades de la distribución beta (ecuaciones 2.3 y 2.4) se tiene que:

$$\begin{aligned} E\left(\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\right) &= E\left(\phi(y_t^* - \mu_t^*)\right) \\ &= \phi\left[E\left(\log\left(\frac{y_t}{1-y_t}\right)\right) - \psi(\mu_t\phi) + \psi((1-\mu_t)\phi)\right] \\ &= \phi\left[E\left(\log\left(y_t\right)\right) - E\left(\log\left(1-y_t\right)\right) - \psi(\mu_t\phi) + \psi((1-\mu_t)\phi)\right] \\ &= \phi\left[\psi(\mu\phi) - \psi(\phi) - \psi((1-\mu)\phi) + \psi(\phi) - \psi(\mu_t\phi) + \psi((1-\mu_t)\phi)\right] \\ E\left(\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\right) &= 0. \end{aligned} \tag{2.13}$$

De esta manera,

$$E\left(\frac{\partial^2 l(\beta,\phi)}{\partial\beta_j\beta_i}\right) = \sum_{t=1}^{n} E\left(\frac{\partial^2 l_t(\mu_t,\phi)}{\partial^2\mu_t}\right)\left(\frac{\partial\mu_t}{\partial\eta_t}\right)^2 x_{ti}x_{tj},$$

derivando la ecuación (2.9) respecto de μ_t

$$\begin{aligned} \frac{\partial^2 l_t(\mu_t,\phi)}{\partial^2\mu_t} &= -\phi\{\psi'(\mu_t\phi)\phi - \psi'((1-\mu_t)\phi)(-\phi)\} \\ &= -\phi^2\{\psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)\}. \end{aligned} \tag{2.14}$$

Denotando $w_t = \phi\left\{\psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)\right\}\dfrac{1}{(g'(\mu_t))^2}$, se tiene que

$$E\left(\frac{\partial^2 l(\beta,\phi)}{\partial\beta_j\partial\beta_i}\right) = -\phi\sum_{t=1}^{n} w_t x_{ti}x_{tj},$$

y en forma matricial se tiene que:

$$E\left(\frac{\partial^2 l(\beta,\phi)}{\partial\beta\partial\beta^{tp}}\right) = -\phi X^{tp}WX, \tag{2.15}$$

donde $W = \text{Diag}(w_1,\ldots,w_n)$.

Por otra parte, la derivada de (2.11) respecto de ϕ es dada por

$$\frac{\partial^2 l(\beta,\phi)}{\partial\phi\partial\beta_i} = \sum_{t=1}^{n}\left[(y_t^* - \mu_t^*) - \phi\frac{\partial\mu_t^*}{\partial\phi}\right]\frac{1}{g'(\mu_t)}x_{ti},$$

con ayuda de la ecuación (2.3) se puede llegar a la igualdad $E(y_t^*) = \mu_t^*$. Además, $\dfrac{\partial\mu_t^*}{\partial\phi} = \psi'(\mu_t\phi)\mu_t - \psi'((1-\mu_t)\phi)(1-\mu_t)$.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Entonces,

$$\begin{aligned} E\left(\frac{\partial^2 l(\beta,\phi)}{\partial\phi\partial\beta_i}\right) &= E\left[\sum_{t=1}^{n}\left[(y_t^*-\mu_t^*)-\phi\frac{\partial\mu_t^*}{\partial\phi}\right]\frac{1}{g'(\mu_t)}x_{ti}\right] \\ &= \sum_{t=1}^{n}\left[E\left[y_t^*-\mu_t^*\right]-\phi\frac{\partial\mu_t^*}{\partial\phi}\right]\frac{1}{g'(\mu_t)}x_{ti} \\ E\left(\frac{\partial^2 l(\beta,\phi)}{\partial\phi\partial\beta_i}\right) &= -\sum c_t\frac{1}{g'(\mu_t)}x_{ti}, \end{aligned}$$

donde $c_t = \phi\left[\psi'(\mu_t,\phi)\mu_t - \psi'\left((1-\mu_t)\phi\right)(1-\mu_t)\right]$. En notación matricial se puede escribir como

$$E\left(\frac{\partial^2 l(\beta,\phi)}{\partial\beta\partial\phi}\right) = -X^{tp}Tc, \tag{2.16}$$

con $c = (c_1,\ldots,c_n)^{tp}$.

Y por último, se halla la derivada de (2.12) respecto de ϕ

$$\begin{aligned} \frac{\partial^2 l(\beta,\phi)}{\partial^2\phi} &= \sum_{t=1}^{n}-\mu_t[\psi'(\mu_t\phi)\mu_t-\psi'((1-\mu_t)\phi)(1-\mu_t)]+\psi'(\phi)-\psi'((1-\mu_t)\phi)(1-\mu_t) \\ &= \sum_{t=1}^{n}-\mu_t^2\psi'(\mu_t\phi)-\psi'((1-\mu_t)\phi)((1-\mu_t)^2)+\psi'(\phi), \\ E\left(\frac{\partial^2 l(\beta,\phi)}{\partial^2\phi}\right) &= -\sum_{t=1}^{n}d_t, \end{aligned}$$

con $d_t = \mu_t^2\psi'(\mu_t\phi)+\psi'((1-\mu_t)\phi)(1-\mu_t)^2-\psi'(\phi)$, y en su forma matricial se obtiene,

$$E\left(\frac{\partial^2 l(\beta,\phi)}{\partial^2\phi}\right) = -\operatorname{Tr}(\operatorname{Diag}(d_i,...,d_n)) = -\operatorname{Tr}(D).$$

con Tr(.) que denota la traza de la matriz que está dentro del paréntesis, y $D = \operatorname{Diag}(d_i,...,d_n)$.

De esta manera, la matriz de información de Fisher puede ser escrita como

$$\mathfrak{J} = \mathfrak{J}(\beta,\phi) = \begin{pmatrix}\mathfrak{J}_{\beta\beta} & \mathfrak{J}_{\beta\phi}\\ \mathfrak{J}_{\phi\beta} & \mathfrak{J}_{\phi\phi}\end{pmatrix}, \tag{2.17}$$

donde:

$$\begin{aligned} \mathfrak{J}_{\beta\beta} &= \phi X^{tp}WX, \\ \mathfrak{J}_{\beta\phi} &= \mathfrak{J}_{\phi\beta}^{tp} = X^{tp}Tc, \\ \mathfrak{J}_{\phi\phi} &= \operatorname{Tr}(D). \end{aligned}$$

2.4. Estimador de máxima verosimilitud

La estimación de los parámetros β y ϕ mediante el método de máxima verosimilitud se realiza derivando la log-verosimilitud del modelo e igualando este término a cero. De esta manera, se construye un sistema de ecuaciones utilizando las funciones de score halladas en la anterior sección (2.11) y (2.12):

$$U_\beta(\beta,\phi) = 0 \quad \text{y} \quad U_\phi(\beta,\phi) = 0.$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Dado que este sistema de ecuaciones es no lineal, no es posible llegar a una solución con una forma cerrada.

Debido a este inconveniente, se recurre a un método de maximización numérica de la función de log-verosimilitud utilizando un algoritmo de optimización no lineal. El método de Fisher-Scoring (ver sección A.5 del apéndice) es un método numérico que repite el siguiente procedimiento hasta llegar a la convergencia de los estimadores de β y ϕ.

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \mathfrak{J}^{-1}(\hat{\theta}^{(k)}) \left.\frac{\partial l(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}^{(k)}}, \; k = 0, 1, \ldots,$$

donde $\hat{\theta}$, $\hat{\theta}^{(0)}$, $\hat{\theta}^{(k)}$ y $\mathfrak{J}^{-1}$ son el vector de parámetros, el valor inicial, los estimadores de los parámetros en la k-ésima iteración, y la inversa de la matriz de información de Fisher, respectivamente.

Al reemplazar la matriz de información de Fisher (2.18) y la función de score (2.11 y 2.12) en este problema en particular se tiene que,

$$\begin{pmatrix} \hat{\beta}^{(k+1)} \\ \hat{\phi}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \hat{\beta}^{(k)} \\ \hat{\phi}^{(k)} \end{pmatrix} + \begin{pmatrix} \mathfrak{J}_{\hat{\beta}\hat{\beta}} & \mathfrak{J}_{\hat{\beta}\hat{\phi}} \\ \mathfrak{J}_{\hat{\phi}\hat{\beta}} & \mathfrak{J}_{\hat{\phi}\hat{\phi}} \end{pmatrix}^{-1} \left.\begin{pmatrix} \frac{\partial l(\beta,\phi)}{\partial \beta} \\ \frac{\partial l(\beta,\phi)}{\partial \phi} \end{pmatrix}\right|_{\beta=\hat{\beta}^{(k)},\phi=\hat{\phi}^{(k)}}, \tag{2.18}$$

donde $\hat{\beta}$ y $\hat{\phi}$ son los estimadores de β y ϕ respectivamente.

Para ello, se requiere especificar un valor inicial para las iteraciones, Ferrari & Cribari-Neto [5] sugieren el estimador de mínimos cuadrados ordinarios de un modelo de regresión de las respuestas transformadas $g(y_1), \ldots, g(y_n)$ en X, es decir $(X^{tp}X)^{-1}X^{tp}z$ como valor inicial para β, donde $z = (g(y_1), \ldots, g(y_n))^{tp}$.

Para la construcción del valor inicial de ϕ, se comienza analizando la varianza de una distribución beta, $V(y_t) = \dfrac{\mu_t(1-\mu_t)}{1+\phi}$. Al despejar ϕ se obtiene $\phi = \dfrac{\mu_t(1-\mu_t)}{V(y_t)} - 1$.

Tomando en cuenta las expansiones de Taylor de primer grado para $g(y_t)$ sobre el valor μ_t, se tiene

$$\begin{aligned} g(y_t) &\approx g(\mu_t) + (y_t - \mu_t)g'(\mu_t), \\ V(g(y_t)) &\approx V\left[g(\mu_t) + (y_t - \mu_t)g'(\mu_t)\right] \\ &= V(y_t)g'(\mu_t)^2, \end{aligned}$$

es decir,

$$\begin{aligned} V(y_t) &\approx V\{g(y_t)\}\{g'(\mu_t)\}^{-2}, \\ \tfrac{\check{\mu}(1-\check{\mu})}{1+\phi} &\approx V\{g(y_t)\}\{g'(\mu_t)\}^{-2}, \\ \frac{\check{\mu}(1-\check{\mu})}{V\{g(y_t)\}\{g'(\mu_t)\}^{-2}} &\approx 1 + \phi. \end{aligned}$$

Por lo tanto, el valor inicial para ϕ puede ser considerado como:

$$\phi = \frac{1}{n}\sum_{t=1}^{n} \frac{\check{\mu}_t(1-\check{\mu}_t)}{\check{\sigma}_t^2} - 1,$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

donde $\check{\sigma}_t^2 = \dfrac{\check{e}^{tp}\check{e}}{(n-k)\{g'(\check{\mu}_t)\}^2}$, con $\check{e} = z - X(X^{tp}X)^{-1}X^{tp}z$ el vector de residuos de mínimos cuadrados ordinarios de la regresión lineal que emplea la respuesta transformada; y $\check{\mu}_t$ se obtiene aplicando la inversa de la función g a la función de enlace, es decir:

$$g(\mu_t) = x_t^{tp}\beta \qquad \rightarrow \qquad \check{\mu}_t = g^{-1}(x_t^{tp}(X^{tp}X)^{-1}X^{tp}z).$$

Para obtener los estimadores de β y ϕ se calculan los valores de $\hat{\beta}^{(k)}$ y $\hat{\phi}^{(k)}$ iterativamente hasta llegar a una convergencia de los mismos. Este algoritmo se implementa en el software estadístico R (ver apéndice B).

Por otro lado, dado $\theta = (\beta, \phi)^{tp}$, la matriz de información de Fisher para una muestra de tamaño n estará denotado por $\mathfrak{J}_n(\theta)$. De esta manera para demostrar que los estimadores $\hat{\theta} = (\hat{\beta}, \hat{\phi})$ son asintóticamente normales se deben realizar algunos supuestos:

- Sea θ_0 el parámetro real, entonces θ_0 es un punto interior del espacio paramétrico Θ ($\theta_0 \in int\Theta$).
- Se asume que $H(\theta) = \lim\limits_{n\to\infty} \dfrac{\mathfrak{J}_n(\theta)}{n}$ existe y es no singular.

Asumiendo estos dos supuestos, Rydlewski & Mielczarek [19] demostraron que los estimadores de máxima verosimilitud de un modelo de regresión beta son asintóticamente consistentes, además de ser normales. Específicamente,

$$\begin{pmatrix}\hat{\beta}\\ \hat{\phi}\end{pmatrix} \sim N\left(\begin{pmatrix}\beta\\ \phi\end{pmatrix}, \mathfrak{J}^{-1}\right), \text{ cuando } n \to \infty, \tag{2.19}$$

donde $\mathfrak{J}^{-1}$ es la inversa de la matriz de información de Fisher.

2.5. Estimación para ϕ conocido

Para llegar a una forma cerrada del estimador de β, se debe suponer que el parámetro de precisión ϕ es conocido. Utilizando la matriz de información de Fisher y la función de score para el estimador $\hat{\beta}$.

$$\begin{aligned}
\hat{\beta}^* &= \hat{\beta} + \tfrac{1}{\phi}(X^{tp}WX)^{-1}\phi X^{tp}T(y^* - \mu^*)\\
&= \hat{\beta} + (X^{tp}WX)^{-1}X^{tp}T(y^* - \mu^*)\\
&= (X^{tp}WX)^{-1}X^{tp}W\left(X\hat{\beta} + W^{-1}T(Y^* - \mu^*)\right)\\
\hat{\beta}^* &= (X^{tp}WX)^{-1}X^{tp}Wz,
\end{aligned}$$

con $z = (g(y_1), \ldots, g(y_n))^{tp}$.

Este resultado fue propuesto por Qasim, Mansson & Kibria [18], quienes coincidieron con el resultado hallado por Abonazel & Taha [1]. Observe que este estimador tiene la forma de un estimador de mínimos cuadrados iterativamente ponderados con la diferencia de que z no corresponde a la variable dependiente original, sino a la respuesta transformada por la función de enlace.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Capítulo 3

Estimadores de contracción

Los métodos de contracción nacen como solución a dos problemas que suelen presentarse en un modelo de regresión. El primero está relacionado con la existencia de multicolinealidad en el modelo, es decir, en la existencia de colinealidad entre las variables explicativas provocando que la varianza del modelo sea grande. Esto a su vez produce estimadores no muy confiables por su alta variabilidad.

El segundo problema viene dado por la alta dimensionalidad en un modelo de regresión. En la actualidad, la carencia de datos para estudiar un determinado fenómeno ha dejado de ser un problema. Es más, ahora se pueden encontrar bases de datos con una gran cantidad de variables, que en algunos casos, supera el número de observaciones, provocando de esta manera que la matriz $X^{tp}X$ no sea invertible, condición que es necesaria para hallar el estimador de mínimos cuadrados de forma única.

Por tales motivos, los estimadores de contracción realizan, según sea el caso, la eliminación de variables explicativas o una penalización a los estimadores del modelo de regresión. Ridge y lasso son dos métodos que tienen el objetivo de contraer los estimadores. Ambos métodos consisten en realizar una minimización de los errores al cuadrado, sujetos a una restricción; la restricción $\sum_j \beta_j^2 \leq t$ en el caso del método ridge, y $\sum_j |\beta_j| \leq t$ en el caso del método lasso.

El parámetro de contracción es constante y penaliza a todos los estimadores de β_i por igual, sin embargo, las variables explicativas podrían ser de distintas magnitudes, además de tener diferentes unidades de medida, lo que provocaría que el parámetro de contracción no penalice a los estimadores con el mismo peso. Por lo tanto, sin pérdida de generalidad, el proceso de estimación se realiza con las variables explicativas estandarizadas, es decir:

$$\sum_{i=1}^{n}(x_j)_i = 0 \quad \text{y} \quad \frac{1}{n}x_j^{tp}x_j = 1 \qquad \forall j = 1, \dots p.$$

En cuanto a la variable dependiente, no será estandarizada, debido a que está acotada en el intervalo $(0, 1)$ y al estandarizarla se elimina esta característica, perdiendo la esencia del modelo de regresión beta. Por este motivo, al realizar la estimación de β también se obtendrá un estimador para el intercepto.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

3.1. Estimador ridge

En esta sección se calculan dos estimadores del tipo ridge para el modelo de regresión beta: primero se hace uso del método ridge propuesto por Hoerl & Kennard [11] para el caso normal, que consiste en minimizar la suma ponderada de errores al cuadrado sujeta a una restricción cuadrática; la segunda es mediante el método de máxima verosimilitud penalizada tomando en cuenta la log-verosimilitud de la distribución beta.

Solo se está interesado en estimar el vector de parámetros β, por tanto, se toma el parámetro de precisión ϕ como fijo y conocido reemplazándolo por su estimador hallado en un modelo de regresión beta sin penalizaciones.

3.1.1. Estimador mínimos cuadrados ponderados

Para hallar el estimador ridge, se analiza primero la suma de los errores al cuadrado hallado con cualquier estimador de β ($\hat{\mathcal{B}}$). Esta metodología fue aplicada por Hoerl & Kennard [11] al modelo de regresión clásico. Tomando como base este análisis, Qasim, Mansson & Kibria [18] aplicaron esta metodología a la diferencia $(Y - X\hat{\mathcal{B}})$, cuyos resultados se detallan a continuación.

Suponga que $\hat{\mathcal{B}}$ es un estimador del vector de parámetros β y que el estimador de máxima verosimilitud (suponiendo que ϕ es conocido) está dado por $\hat{\beta} = (X^{tp}WX)^{-1}X^{tp}Wz$, entonces la suma ponderada de las diferencias entre Y y $X\hat{\mathcal{B}}$ al cuadrado está dada por:

$$
\begin{aligned}
\Theta &= (Y - X\hat{\mathcal{B}})^{tp}W(Y - X\hat{\mathcal{B}}) \\
&= Y^{tp}WY - Y^{tp}WX\hat{\mathcal{B}} - \hat{\mathcal{B}}^{tp}X^{tp}WY + \hat{\mathcal{B}}^{tp}X^{tp}WX\hat{\mathcal{B}} \\
&= Y^{tp}WY - \hat{\beta}^{tp}X^{tp}WX\hat{\mathcal{B}} - \hat{\mathcal{B}}^{tp}X^{tp}WY + \hat{\mathcal{B}}^{tp}X^{tp}WX\hat{\mathcal{B}} - Y^{tp}WX(X^{tp}WX)^{-1}X^{tp}WY \\
&\quad -Y^{tp}WX[(X^{tp}WX)^{-1}]^{tp}X^{tp}WY + 2\hat{\beta}^{tp}X^{tp}WX\hat{\beta} \\
&= Y^{tp}WY - Y^{tp}WX\hat{\beta} - \hat{\beta}^{tp}X^{tp}WY + \hat{\beta}^{tp}X^{tp}WX\hat{\beta} \\
&\quad +\hat{\mathcal{B}}^{tp}X^{tp}WX\hat{\mathcal{B}} - \hat{\mathcal{B}}^{tp}X^{tp}WY - \hat{\beta}^{tp}X^{tp}WX\hat{\mathcal{B}} + \hat{\beta}^{tp}X^{tp}WX\hat{\beta} \\
&= (Y - X\hat{\beta})^{tp}W(Y - X\hat{\beta}) + (\hat{\mathcal{B}} - \hat{\beta})^{tp}X^{tp}WX(\hat{\mathcal{B}} - \hat{\beta}) \\
\Theta &= \Theta_{min} + \Theta(\hat{\mathcal{B}}).
\end{aligned}
$$

Θ es la suma ponderada de los errores al cuadrado la cual se divide en dos partes, la primera es un término cuadrático $\Theta_{\text{mín}}$ que coincide con la mínima suma ponderada de errores al cuadrado para el estimador de máxima verosimilitud $\hat{\beta}$. El segundo término es otra función cuadrática $\Theta(\hat{\mathcal{B}}) > 0$ que corresponde al incremento fijo que aumenta la suma ponderada de los errores al cuadrado cuando el estimador de máxima verosimilitud $\hat{\beta}$ es sustituido por algún otro estimador $\hat{\mathcal{B}}$.

Si $\hat{\mathcal{B}}$ coincide con el estimador de máxima verosimilitud, es decir, si $\hat{\mathcal{B}} = \hat{\beta}$, entonces se anula el segundo término y se queda con la mínima suma ponderada de errores al cuadrado para un estimador insesgado. Si se extienden las posibles soluciones a estimadores sesgados, entonces para hacer mínima la suma ponderada de errores al cuadrado con $\hat{\mathcal{B}}$, se debe minimizar el segundo término $\Theta(\hat{\mathcal{B}}) = (\hat{\mathcal{B}} - \hat{\beta})^{tp}X^{tp}WX(\hat{\mathcal{B}} - \hat{\beta})$. A este término que minimiza la suma ponderada se la denota como Θ_0.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Por otra parte, el estimador de regresión beta ridge se obtiene minimizando la longitud de $\hat{\mathcal{B}}^{tp}\hat{\mathcal{B}}$ sujeto a la restricción $(\hat{\mathcal{B}}-\hat{\beta})^{tp}X^{tp}WX(\hat{\mathcal{B}}-\hat{\beta})=\Theta_0$, es decir:

$$\mathcal{Q}=\hat{\mathcal{B}}^{tp}\hat{\mathcal{B}} \qquad \text{Sujeto a } (\hat{\mathcal{B}}-\hat{\beta})^{tp}X^{tp}WX(\hat{\mathcal{B}}-\hat{\beta})-\Theta_0.$$

Esta ecuación puede ser escrita como una ecuación de lagrange de la siguiente manera:

$$\begin{aligned}
l(\mathcal{Q},k) &= \hat{\mathcal{B}}^{tp}\hat{\mathcal{B}}+\left(\frac{1}{k}\right)\{(\hat{\mathcal{B}}-\hat{\beta})^{tp}X^{tp}WX(\hat{\mathcal{B}}-\hat{\beta})-\Theta_0\}\\
l(\mathcal{Q},k) &= \hat{\mathcal{B}}^{tp}\hat{\mathcal{B}}+\left(\frac{1}{k}\right)\{\hat{\mathcal{B}}^{tp}X^{tp}WX\hat{\mathcal{B}}-\hat{\mathcal{B}}^{tp}X^{tp}WX\hat{\beta}-\hat{\beta}^{tp}X^{tp}WX\hat{\mathcal{B}}+\hat{\beta}^{tp}X^{tp}WX\hat{\beta}-\Theta_0\},\\
\frac{\partial l(\mathcal{Q},k)}{\partial\hat{\mathcal{B}}} &= 2\hat{\mathcal{B}}+\left(\frac{1}{k}\right)\{2X^{tp}WX\hat{\mathcal{B}}-X^{tp}WX\hat{\beta}-X^{tp}WX\hat{\beta}\}.
\end{aligned}$$

Igualando a cero se tiene

$$\begin{aligned}
2\hat{\mathcal{B}}+\left(\frac{1}{k}\right)\{2X^{tp}WX\hat{\mathcal{B}}-2X^{tp}WX\hat{\beta}\} &= 0,\\
k\hat{\mathcal{B}} &= -X^{tp}WX\hat{\mathcal{B}}+X^{tp}WX\hat{\beta},\\
X^{tp}WX\hat{\mathcal{B}}+k\hat{\mathcal{B}} &= X^{tp}WX\hat{\beta},\\
(X^{tp}WX+kI)\,\hat{\mathcal{B}} &= X^{tp}WX\hat{\beta},\\
\hat{\mathcal{B}} &= (X^{tp}WX+kI)^{-1}X^{tp}WX\hat{\beta}.
\end{aligned}$$

De esta manera, denotando por $\hat{\beta}_{ridge}$ al estimador ridge de β en el modelo de regresión beta,

$$\hat{\beta}_{ridge}=\left(X^{tp}WX+kI\right)^{-1}X^{tp}WX\hat{\beta}, \tag{3.1}$$

donde k es el parámetro de contracción e I es una matriz identidad de dimensión $p+1$.

A continuación se hacen algunas observaciones acerca del estimador de regresión beta ridge $\hat{\beta}_{ridge}$.

- El estimador $\hat{\beta}_{ridge}$ tiene la misma forma que Hoerl y Kennard hallan para el caso normal con la diferencia que este estimador esta ponderado por W. Al comparar con el estimador del tipo ridge de un modelo de regresión logístico, se observa que estos son idénticos tomando en cuenta que también existe una matriz de ponderadores W, aunque para el caso logístico, esta matriz es construida de manera distinta.
- Al igual que en el caso normal, este estimador ridge suma al término que está dentro de la inversa, una matriz diagonal kI.
- Al iniciar esta sección se mencionó que el parámetro ϕ era conocido, por tanto el estimador $\hat{\beta}_{ridge}$ está construido bajo esta suposición, ya que W depende del parámetro ϕ $\left(w_t=\phi\{\psi'(\mu_t\phi)+\psi'((1-\mu_t)\phi)\}\dfrac{1}{(g'(\mu_t))^2}\right)$.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

3.1.2. Estimador de máxima verosimilitud penalizada

La log-verosimilitud $l(\beta)$ penalizada por el término $k^*\|\beta\|^2$ está dada por:

$$l^{k^*}(\beta) = l(\beta) - k^*\|\beta\|^2,$$

donde $\|\beta\| = (\sum \beta_j^2)^{\frac{1}{2}}$ es la norma $L2$ del vector de parámetros β. El parámetro de ridge k controla la cantidad de contracción de la norma $L2$ de β. En el caso de que $k^* = 0$, este coincide con el estimador de máxima verosimilitud.

Al igual que la estimación de un modelo de regresión sin penalizar, el procedimiento de maximización de los estimadores se inicia calculando la primera (función de score) y segunda derivada de la función de log-verosimilitud. Calculando primero la función de score

$$U^{k^*}(\beta) = U(\beta) - 2k^*\beta, \tag{3.2}$$

con $U(\beta)$ la función de score de los estimadores de β hallados a partir de la verosimilitud logarítmica sin penalizar.

La segunda derivada de la función de log-verosimilitud penalizada esta denotado por:

$$\Omega^{k^*}(\beta) = \Omega(\beta) - 2k^*I, \tag{3.3}$$

donde $\Omega(\beta)$ es la segunda derivada de la función de log-verosimilitud sin penalizar.

El siguiente paso es aproximar la primera derivada de la log-verosimilitud, mediante el método de Newton Raphson (ver sección A.4 del apéndice), entonces, aplicando las expansiones de las series de Taylor de primer grado para $U^{k^*}(\hat{\beta}_{k^*})$ sobre el valor $\hat{\beta}_0$:

$$U^{k^*}(\hat{\beta}_{k^*}) = U^{k^*}(\hat{\beta}_0) + (\hat{\beta}_{k^*} - \hat{\beta}_0)'\Omega^{k^*}(\hat{\beta}_0) + 0(\|\hat{\beta}_{k^*} - \hat{\beta}_0\|).$$

Usando las ecuaciones (3.2), (3.3) e igualando $U^{k^*}(\beta_{k^*})$ a 0,

$$\begin{aligned}
0 &= U(\hat{\beta}_0) - 2k^*\hat{\beta}_0^{tp} + (\hat{\beta}_{k^*} - \hat{\beta}_0)^{tp}(\Omega(\hat{\beta}_0) - 2k^*I), \\
-(\hat{\beta}_{k^*} - \hat{\beta}_0)^{tp}(\Omega(\hat{\beta}_0) - 2k^*I) &= U(\hat{\beta}_0) - 2k^*\hat{\beta}_0^{tp}, \\
\hat{\beta}_{k^*} - \hat{\beta}_0 &= -(\Omega(\hat{\beta}_0) - 2k^*I)^{-1}(U(\hat{\beta}_0) - 2k^*\hat{\beta}_0^{tp})^{tp}, \\
\hat{\beta}_{k^*} &= \hat{\beta}_0 - \{\Omega(\hat{\beta}_0) - 2k^*I\}^{-1}\{U(\hat{\beta}_0) - 2k^*\hat{\beta}_0\} \\
&= \{\Omega(\hat{\beta}_0) - 2k^*I\}^{-1}\{\Omega(\hat{\beta}_0)\hat{\beta}_0 - 2k^*\hat{\beta}_0 - U(\hat{\beta}_0) + 2k^*\hat{\beta}_0\} \\
&= \{\Omega(\hat{\beta}_0) - 2k^*I\}^{-1}\{-U(\hat{\beta}_0) + \Omega(\hat{\beta}_0)\hat{\beta}_0\}.
\end{aligned}$$

Por otra parte, se puede demostrar que el estimador de máxima verosimilitud $(\hat{\beta})$, también puede ser aproximado mediante el método de Newton Raphson, obteniendo $\hat{\beta} = \hat{\beta}_0 + \Omega^{-1}(\hat{\beta}_0)U(\hat{\beta}_0)$. Asimismo, $\hat{\beta}_0 = \hat{\beta} + \Omega^{-1}(\hat{\beta}_0)U(\hat{\beta}_0)$. Entonces,

$$\begin{aligned}
\hat{\beta}_{k^*} &= \{\Omega(\hat{\beta}_0) - 2k^*I\}^{-1}\{-U(\hat{\beta}_0) + \Omega(\hat{\beta}_0)(\hat{\beta} + \Omega^{-1}(\hat{\beta}_0)U(\hat{\beta}_0))\} \\
&= \{\Omega(\hat{\beta}_0) - 2k^*I\}^{-1}\{-U(\hat{\beta}_0) + \Omega(\hat{\beta}_0)\hat{\beta} + \Omega(\hat{\beta}_0)\Omega^{-1}(\hat{\beta}_0)U(\hat{\beta}_0)\} \\
\hat{\beta}_{k^*} &= \{\Omega(\hat{\beta}_0) - 2k^*I\}^{-1}\{\Omega(\hat{\beta}_0)\hat{\beta}\}.
\end{aligned}$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Para obtener el estimador de tipo ridge se aproxima $\Omega(\hat{\beta}_0)$ con la matriz de información de Fisher que llega a ser el negativo de la esperanza de $\Omega(\hat{\beta}_0)$,

$$\mathfrak{J}_{\beta\beta} = E(-\Omega(\hat{\beta}_0)) = E\left(-\frac{\partial^2 l(\beta,\phi)}{\partial\beta\partial\beta^{tp}}\right).$$

De esta manera:

$$\hat{\beta}_{k^*} = \{\mathfrak{J}_{\beta\beta} + 2k^* I\}^{-1}\{\mathfrak{J}_{\beta\beta}\hat{\beta}\}. \tag{3.4}$$

Sustituyendo $\hat{\beta}_0$ por $\hat{\beta}$, y con ayuda de la ecuación (2.15) se tiene $\mathfrak{J}_{\beta\beta} = \phi X^{tp} W X$, que al reemplazar en la anterior ecuación se tiene que,

$$\begin{aligned}\hat{\beta}_{k^*} &= \left(\phi X^{tp} W X + 2k^* I\right)^{-1} \phi X^{tp} W X \hat{\beta} \\ &= \left(X^{tp} W X + 2\tfrac{k^*}{\phi} I\right)^{-1} \phi^{-1}\phi X^{tp} W X \hat{\beta}.\end{aligned}$$

Entonces,

$$\hat{\beta}^*_{ridge} = \left(X^{tp} W X + \frac{2k^*}{\phi} I\right)^{-1} X^{tp} W X \hat{\beta}, \tag{3.5}$$

donde $\hat{\beta}^*_{ridge}$ es el estimador del tipo ridge hallado mediante el método de máxima verosimilitud penalizada. Los estimadores $\hat{\beta}_{ridge}$ y $\hat{\beta}^*_{ridge}$ son bastante parecidos, la única diferencia se encuentra en el término que multiplica al parámetro de contracción k^*, ya que en el estimador $\hat{\beta}^*_{ridge}$ esta multiplicado por $\frac{2}{\phi}$ y en el estimador $\hat{\beta}_{ridge}$ por 1.

Sin embargo, el parámetro de precisión ϕ se supone conocido. Por tal motivo, se puede realizar el siguiente cambio de variables $k = \dfrac{2k^*}{\phi}$, lo que da como resultado el estimador ridge de un modelo de regresión beta hallado en la ecuación (3.1).

$$\hat{\beta}^*_{ridge} = \left(X^{tp} W X + kI\right)^{-1} X^{tp} W X \hat{\beta} = \hat{\beta}_{ridge},$$

demostrando que el estimador de β para un modelo de regresión beta, hallado mediante la minimización ponderada de los errores al cuadrado restringido y la maximización de la log-verosimilitud penalizada son el mismo.

Los dos estimadores beta ridge, suponen que el parámetro ϕ es conocido. Qasim, Mansson y Kibria [18] toman como su estimador al hallado con el método de máxima verosimilitud sin penalizar mostrado en la sección 2, el cual se mantiene fijo para todo k. Sin embargo, en el presente trabajo se hallará el estimador de ϕ específico para los estimadores beta ridge. La manera en que es hallado es mediante el método descenso coordinado explicado en la sección (A.2) del apéndice. Entonces, en primera instancia se hallan los estimadores de β suponiendo que ϕ es conocido, el cual toma el valor estimado del método de máxima verosimilitud sin penalizar. Posteriormente, con los estimadores ridge de β hallados se halla el estimador de ϕ; para posteriormente fijarlo y con este término volver a hallar los estimadores ridge de β, y así sucesivamente hasta alcanzar la convergencia. Este procedimiento se encuentra programado en la función **beta_ridge** de la sección (B.1.3) del apéndice y es utilizada en el capítulo de simulación y aplicación.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

3.1.3. Propiedades del estimador ridge

Para mostrar el rendimiento de los estimadores hallados en el presente trabajo, se utiliza el escalar del error cuadrático medio (ECM), que está definido como la esperanza de la diferencia entre el parámetro y el estimador al cuadrado,

$$ECM(\hat{\theta}) = E\left[(\hat{\theta}-\theta)^{tp}(\hat{\theta}-\theta)\right],$$

el cual puede ser escrito como la suma de la traza de la varianza y la suma del sesgo al cuadrado del estimador,

$$ECM(\hat{\theta}) = \text{Tr}\left[V(\hat{\theta})\right] + \left[\text{sesgo}(\hat{\theta})\right]^{tp}\left[\text{sesgo}(\hat{\theta})\right]. \tag{3.6}$$

Se comienza calculando el ECM del estimador de máxima verosimilitud $\hat{\beta}$, para compararlo posteriormente con el estimador ridge. Sin embargo, se debe resaltar que estos resultados son asintóticos. Entonces, de la ecuación (2.19) se tiene que $\hat{\beta}$ es un estimador asintóticamente insesgado. De esta manera, denotando con $V(\hat{\beta})$ a la varianza asintótica y haciendo uso de la ecuación (2.15), se tiene

$$\begin{aligned} ECM(\hat{\beta}) &= \text{Tr}\left[V(\hat{\beta})\right] + \left[\text{sesgo}(\hat{\beta})\right]^{tp}\left[\text{sesgo}(\hat{\beta})\right], \\ &= \text{Tr}\left[\tfrac{1}{\phi}(X^{tp}WX)^{-1}\right] + 0. \end{aligned}$$

Suponga que λ_i y v_i son el valor propio y el vector propio de la matriz $X^{tp}WX$ respectivamente con $i = 1, \ldots, p$, entonces, su ecuación característica está dada por $(X^{tp}WX)v_i = \lambda_i v_i$. Multiplicando ambos lados por $\frac{1}{\lambda_i}(X^{tp}WX)^{-1}$ se obtiene: $(X^{tp}WX)^{-1}v_i = \frac{1}{\lambda_i}v_i$, de esta manera, se tiene que el valor propio de la matriz $(X^{tp}WX)^{-1}$ es $\frac{1}{\lambda_i}$. Luego por propiedades de la traza, el error cuadrático medio del estimador $\hat{\beta}$ es:

$$\delta(k=0) = ECM(\hat{\beta}) = \tfrac{1}{\phi}\sum_{i=1}^{p}\tfrac{1}{\lambda_i}.$$

donde $\delta(k)$ es el error cuadrático medio para un parámetro de contracción k dado.

Luego, con el fin de comparar este resultado, se halla también el ECM del estimador ridge $\hat{\beta}_{ridge}$, calculando de manera separada el sesgo y la varianza del estimador.

Varianza del estimador ridge

Utilizando nuevamente la ecuación (2.15), se calcula la varianza:

$$\begin{aligned} V(\hat{\beta}_{ridge}) &= V\left[(X^{tp}WX+kI)^{-1}X^{tp}WX\hat{\beta}\right] \\ &= (X^{tp}WX+kI)^{-1}X^{tp}WXV(\hat{\beta})XWX^{tp}(X^{tp}WX+kI)^{-1} \\ &= \tfrac{1}{\phi}(X^{tp}WX+kI)^{-1}X^{tp}WX(X^{tp}WX)^{-1}XWX^{tp}(X^{tp}WX+kI)^{-1} \\ &= \tfrac{1}{\phi}(X^{tp}WX+kI)^{-1}XWX^{tp}(X^{tp}WX+kI)^{-1}. \end{aligned}$$

Partiendo nuevamente de la ecuación característica $(X^{tp}WX)v_i = \lambda_i v_i$ y sumando kv_i se obtiene el valor propio de $(X^{tp}WX+kI)$, que está dado por $\lambda_i + k$. Continuando de esta manera, dada la ecuación característica

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

$$\begin{aligned}
(X^{tp}WX + kI)^{-1}v_i &= \left(\frac{1}{\lambda_i + k}\right) v_i, \\
(X^{tp}WX + kI)^{-1}(\lambda_i v_i) &= \left(\frac{\lambda_i}{\lambda_i + k}\right) v_i, \\
(X^{tp}WX + kI)^{-1}(X^{tp}WX)v_i &= \left(\frac{\lambda_i}{\lambda_i + k}\right) v_i, \\
(X^{tp}WX + kI)^{-1}(X^{tp}WX)\left(\frac{1}{\lambda_i + k}\right) v_i &= \frac{\lambda_i}{(\lambda_i + k)^2} v_i, \\
(X^{tp}WX + kI)^{-1}(X^{tp}WX)(X^{tp}WX + kI)^{-1}v_i &= \frac{\lambda_i}{(\lambda_i + k)^2} v_i.
\end{aligned}$$

es decir, $\dfrac{\lambda_i}{(\lambda_i + k)^2}$ es el valor propio de $(X^{tp}WX + kI)^{-1}(X^{tp}WX)(X^{tp}WX + kI)^{-1}$.

Por lo tanto, denotando como $\delta_1(k)$ a la traza de la varianza del estimador ridge, se tiene que

$$\delta_1(k) = \mathrm{Tr}(V(\hat{\beta}_{ridge})) = \sum_{i=1}^{p} \frac{1}{\phi} \frac{\lambda_i}{(\lambda_i + k)^2}.$$

Observe que $\delta_1(k) \to 0$, cuando $k \to \infty$, mostrando que el estimador ridge disminuye la varianza del estimador cuando k se incrementa. Sin embargo, el incrementar k para disminuir la varianza, afecta al sesgo, propiedad que se analiza a continuación.

Sesgo del estimador ridge

El sesgo esta dado por:

$$\begin{aligned}
E(\hat{\beta}_{ridge}) - \beta &= E((X^{tp}WX + kI)^{-1}X^{tp}WX\hat{\beta}_{ML}) - \beta \\
&= (X^{tp}WX + kI)^{-1}X^{tp}WXE(\hat{\beta}_{ML}) - \beta \\
&= (X^{tp}WX + kI)^{-1}X^{tp}WX\beta - \beta \\
&= ((X^{tp}WX + kI)^{-1}X^{tp}WX - I)\,\beta \\
&= (X^{tp}WX + kI)^{-1}\,(X^{tp}WX - (X^{tp}WX + kI))\,\beta \\
&= (X^{tp}WX + kI)^{-1}\,(-kI)\,\beta \\
&= -k(X^{tp}WX + kI)^{-1}\beta,
\end{aligned}$$

y denotando como $\delta_2(k)$ al sesgo al cuadrado se tiene que,

$$\delta_2(k) = (E(\hat{\beta}_{ridge}) - \beta)^{tp}(E(\hat{\beta}_{ridge}) - \beta) = k^2\beta^{tp}(X^{tp}WX + kI)^{-2}\beta.$$

Ahora, suponga que Λ es la matriz de valores propios de $X^{tp}WX$ y P es una matriz ortogonal cuyas columnas son los vectores propios de $X^{tp}WX$, entonces $X^{tp}WX = P^{tp}\Lambda P$, luego

$$\begin{aligned}
\delta_2(k) = (E(\hat{\beta}_{ridge}) - \beta)^{tp}(E(\hat{\beta}_{ridge}) - \beta) &= k^2\beta^{tp}(P^{tp}(\Lambda + kI)P)^{-2}\beta \\
&= k^2\beta^{tp}P^{tp}(\Lambda + kI)^{-2}P\beta \\
&= k^2\gamma^{tp}(\Lambda + kI)^{-2}\gamma \\
&= k^2\sum_{i=1}^{p}\frac{\gamma_i^2}{(\lambda_i + k)^2},
\end{aligned}$$

donde $\gamma_i = P\beta$. Este término también puede ser escrito como,

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

$$\delta_2(k) = \sum_{i=1}^{p} \frac{\gamma_i^2}{\left(1+\frac{\lambda_i}{k}\right)^2}.$$

Observe que si el parámetro de contracción k crece, el sesgo al cuadrado también lo hará, esto sucede debido a que λ_i y k son positivos, entonces $\frac{\lambda_i}{k}$ es positivo y es monótonamente decreciente para incrementos en k, por lo cual, el sesgo al cuadrado es monótonamente creciente para incrementos de k. Es decir, que $\delta_2(k) \to \sum_{i=1}^{p} \gamma_i^2$, cuando $k \to \infty$.

Error cuadrático medio del estimador ridge

Tomando en cuenta los resultados hallados para el sesgo al cuadrado y la varianza, el error cuadrático medio del estimador de regresión ridge esta dado por:

$$\delta(k) = ECM(\hat{\beta}_{ridge}, k) = \sum_{i=1}^{p} \frac{1}{\phi} \frac{\lambda_i}{(\lambda_i + k)^2} + \sum_{i=1}^{p} \frac{k^2 \gamma_i^2}{(\lambda_i + k)^2}. \tag{3.7}$$

Observe que si $k = 0$, entonces el error cuadrático medio coincide con el del estimador máximo verosímil.

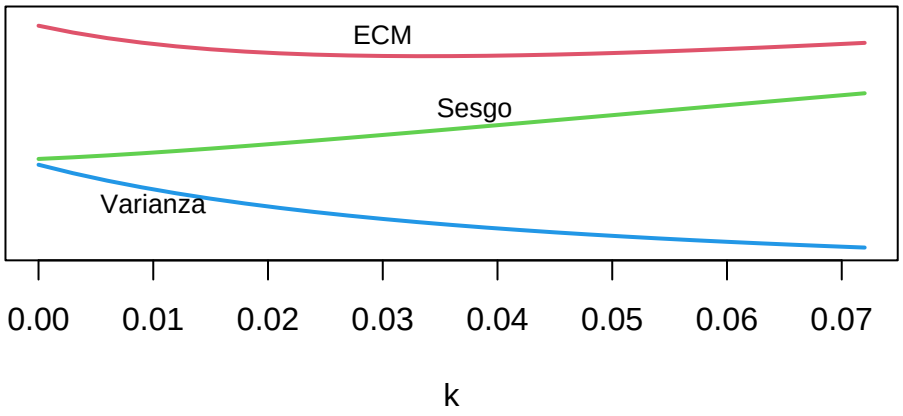


Figura 3.1: Relación entre el parámetro de contracción k y el $ECM(\hat{\beta}_{ridge})$

La figura (3.1) fue construida con datos simulados, con un tamaño de muestra de $n = 100$ observaciones y $p = 6$ variables explicativas que presentan una alta colinealidad[1], se realizan $S = 10000$ simulaciones con estas características, dando como resultado 10000 estimadores ridge con los cuales se halla el error cuadrático medio, el sesgo y la varianza. La figura muestra que mientras más crece el parámetro de contracción k, el sesgo al cuadrado se incrementa, mientras que la varianza disminuye, al sumar ambas funciones, se obtiene el error cuadrático medio que depende de k.

[1]La colinealidad de las variables explicativas es determinada por el vector $a = c(0{,}8 \quad 0{,}75 \quad 0{,}85 \quad 0{,}9 \quad 0{,}95 \quad 0{,}85)$, donde a es un vector de $p \times 1$ que es definido en el siguiente capítulo.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

A continuación, se demuestra que existe un k, de manera que $\delta(k) = ECM(\hat{\beta}_{ridge}) < ECM(\hat{\beta}) = \delta(0)$, para ello, se halla la derivada del ECM, respecto de k

$$\begin{aligned} \delta'(k) = \frac{d\delta(k)}{dk} &= \delta_1'(k) + \delta_2'(k) \\ &= -\frac{2}{\phi}\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3} + 2k\sum_{i=1}^{p}\frac{\gamma_i^2\lambda_i}{(\lambda_i+k)^3}. \end{aligned} \tag{3.8}$$

Como ya se demostró en las anteriores dos propiedades de β_{ridge}, $\delta_1(k)$ y $\delta_2(k)$ son funciones monótonamente decrecientes y crecientes respectivamente, ya que sus primeras derivadas son no positivas y no negativas respectivamente. De esta manera, ya que $\delta'(0) = \frac{dECM(\hat{\beta})}{dk} = 0$, es suficiente mostrar que existe un k de tal manera que $\delta'(k) < 0$. De la ecuación (3.8) y tomando el escenario menos favorable con un γ_i más grande, ($\gamma_{\text{máx}}$) se tiene:

$$\begin{aligned} -\frac{2}{\phi}\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3} + 2k\sum_{i=1}^{p}\frac{\gamma_{\text{máx}}^2\lambda_i}{(\lambda_i+k)^3} &< 0, \\ 2k\gamma_{\text{máx}}^2\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3} &< \frac{2}{\phi}\sum_{i=1}^{p}\frac{\lambda_i}{(\lambda_i+k)^3}, \\ k &< \frac{1}{\gamma_{\text{máx}}^2\phi}, \end{aligned} \tag{3.9}$$

lo que demuestra que existe un k tal que $ECM(\hat{\beta}_{ridge}) < ECM(\hat{\beta})$.

3.2. Estimador lasso

El método lasso "least absolute shrinkage and selection operator", cumple con dos funciones: la primera que al igual que la regresión ridge, penaliza los coeficientes de regresión grandes y los contrae hacia cero; la segunda es que lasso produce coeficientes que son exactamente iguales a cero, eliminando de esta manera esos predictores del modelo, es decir, que realiza una selección de variables.

En un modelo de regresión clásico, donde se supone que la variable dependiente tiene una función de distribución normal, el método lasso consiste en minimizar el error al cuadrado más una penalización que depende del parámetro β (ver la sección A.3 del apéndice). En el caso de un modelo de regresión beta la función a minimizar es modificada por el negativo de la log-verosimilitud, estrategia que Noah et. al. [17] aplican para hallar el estimador lasso de un modelo de riesgos proporcionales de Cox. De esta manera, el problema de optimización en términos de lagrange está dado por:

$$\text{mín}\left(-\frac{1}{n}l(\beta) + k\|\beta\|\right), \tag{3.10}$$

donde $l(\beta)$ es la función de log-verosimilitud de una función de densidad beta dada por la ecuación (2.2) y $\|\beta\|_1 = \sum_j |\beta_j|$ es la norma $L1$ de los coeficientes de regresión.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Sea $\eta = X^{tp}\beta$ la función de enlace, $l'(\beta)$, $l''(\beta)$, $l'(\eta)$ y $l''(\eta)$ son el gradiente y el hessiano de la log-verosimilitud respecto de β y η respectivamente. Entonces, la log-verosimilitud $l(\beta)$ será aproximada mediante la expansión de las series de Taylor de segundo orden centrado en $\tilde{\beta}$.

$$l(\beta) \simeq l(\tilde{\beta}) + (\beta - \tilde{\beta})^{tp} l'(\tilde{\beta}) + \frac{(\beta - \tilde{\beta})^{tp} l''(\tilde{\beta})(\beta - \tilde{\beta})}{2}. \tag{3.11}$$

De las ecuaciones (2.11) y (2.15) se tiene que

$$l'(\beta) = \phi X^{tp} T(y^* - \mu^*) \qquad \text{y} \qquad l''(\beta) = -\phi X^{tp} W X,$$

respectivamente. El siguiente paso es calcular $l'(\eta)$ y $l''(\eta)$ de la misma manera, que en el anterior capítulo. Entonces, se inicia con la log-verosimilitud del modelo de regresión beta como se mostró en la ecuación (2.6).

Derivando respecto de $x_i\beta_i$ se tiene que:

$$\frac{\partial l(\beta,\phi)}{\partial x_i\beta_i} = \sum_{t=1}^{n} \frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t} \frac{\partial \mu_t}{\partial \eta_t} \frac{\partial \eta_t}{\partial x_i\beta_i},$$

de la ecuación (2.9),

$$\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t} = \phi(y_t^* - \mu_t^*),$$

con $y_t^* = \log\left(\frac{y_t}{1-y_t}\right)$ y $\mu_t^* = \psi(\mu_t\phi) - \psi((1-\mu_t)\phi)$, además $\frac{\partial \mu_t}{\partial \eta_t} = \frac{1}{g'(\mu_t)}$ y $\frac{\partial \eta_t}{\partial x_i\beta_i} = 1$, entonces:

$$l'(x_i\beta_i) = \frac{\partial l(\beta,\phi)}{\partial x_i\beta_i} = \sum_{t=1}^{n} \phi(y_t^* - \mu_t^*) \frac{1}{g'(\mu_t)},$$

y su forma matricial está dada por

$$l'(x\beta) = \phi T(y^* - \mu^*). \tag{3.12}$$

Para hallar el hessiano de la log-verosimilitud respecto de $x_i\beta_i$

$$\begin{aligned} l''(x_i\beta_i) &= \sum_{t=1}^{n} \frac{\partial}{\partial x_i\beta_i}\left(\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\frac{\partial \mu_t}{\partial \eta_t}\frac{\partial \eta_t}{\partial x_i\beta_i}\right) \\ &= \sum_{t=1}^{n} \frac{\partial}{\partial \mu_t}\left(\frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\frac{\partial \mu_t}{\partial \eta_t}\right)\frac{\partial \mu_t}{\partial \eta_t} \\ &= \sum_{t=1}^{n} \left(\frac{\partial^2 l_t(\mu_t,\phi)}{\partial^2 \mu_t}\frac{\partial \mu_t}{\partial \eta_t} + \frac{\partial l_t(\mu_t,\phi)}{\partial \mu_t}\frac{\partial^2 \mu_t}{\partial \mu_t \eta_t}\right)\frac{\partial \mu_t}{\partial \eta_t}. \end{aligned}$$

Aplicando esperanza a ambos lados, de la ecuación (2.13) el segundo término es anulado, en cuanto al primero se tiene la ecuación (2.14) donde se muestra que $\frac{\partial^2 l_t(\mu_t,\phi)}{\partial^2 \mu_t} = -\phi^2\{\psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)\}$ y además $\frac{\partial \mu_t}{\partial \eta_t} = \frac{1}{g'(\mu_t)}$, por tanto

$$l''(x_i\beta_i) = \sum_{t=1}^{n} -\phi w_t,$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

que en su forma matricial se tiene

$$l''(x\beta) = -\phi W, \tag{3.13}$$

donde $W = \text{Diag}(w_1, \ldots, , w_n)$ y $w_t = \phi(\psi'(\mu_t\phi) - \psi'((1-\mu_t)\phi))\dfrac{1}{(g'(\mu_t))^2}$.

Ahora, se busca una relación entre $l'(x\beta)$ y $l'(\beta)$, así,

$$\begin{aligned}(\beta-\tilde{\beta})^{tp}l'(\tilde{\beta}) &= (\beta-\tilde{\beta})^{tp}\phi X^{tp}T(y^*-\mu^*)\\ &= (X\beta-X\tilde{\beta})^{tp}\phi T(y^*-\mu^*)\\ &= (X\beta-X\tilde{\beta})^{tp}l'(\tilde{X\beta}),\end{aligned}$$

de la misma manera, se tiene que,

$$\begin{aligned}(\beta-\tilde{\beta})^{tp}l''(\tilde{\beta})(\beta-\tilde{\beta}) &= -(\beta-\tilde{\beta})^{tp}\phi X^{tp}WX(\beta-\tilde{\beta})\\ &= -(X\beta-X\tilde{\beta})^{tp}\phi W(X\beta-X\tilde{\beta})\\ &= (X\beta-X\tilde{\beta})^{tp}l''(\tilde{X\beta})(X\beta-X\tilde{\beta}).\end{aligned}$$

Luego, aplicando las últimas dos igualdades a la ecuación (3.11) y realizando algunas operaciones algebraicas se llega a una expresión que posteriormente será derivada para hallar el estimador de β.

$$\begin{aligned}l(\beta) &\simeq l(\tilde{\beta}) + (\beta-\tilde{\beta})^{tp}l'(\tilde{\beta}) + \tfrac{(\beta-\tilde{\beta})^{tp}l''(\tilde{\beta})(\beta-\tilde{\beta})}{2}\\ &= l(\tilde{\beta}) + (x\beta-x\tilde{\beta})^{tp}l'(x\tilde{\beta}) + \frac{(x\beta-x\tilde{\beta})^{tp}l''(x\tilde{\beta})(x\beta-x\tilde{\beta})}{2}\\ &= l(\tilde{\beta}) + \tfrac{1}{2}\Big[(x\tilde{\beta})^{tp}l''(x\tilde{\beta})(x\tilde{\beta}) - (x\tilde{\beta})^{tp}l'(x\tilde{\beta}) - (x\tilde{\beta})^{tp}l''(x\tilde{\beta})(x\beta) - l'(x\tilde{\beta})^{tp}(x\tilde{\beta})\\ &\quad + l'(x\tilde{\beta})l''(x\tilde{\beta})^{-1}l'(x\tilde{\beta}) + l'(x\tilde{\beta})^{tp}(x\beta) - (x\beta)^{tp}l''(x\tilde{\beta})(x\tilde{\beta}) + (x\beta)^{tp}l'(x\tilde{\beta})\\ &\quad + (x\beta)^{tp}l''(x\tilde{\beta})(x\beta)\Big] - \tfrac{1}{2}l'(x\tilde{\beta})l''(x\tilde{\beta})^{-1}l'(x\tilde{\beta})\\ &= l(\tilde{\beta}) + \tfrac{1}{2}\Big[(x\tilde{\beta})^{tp} - l''(x\tilde{\beta})^{-1}l'(x\tilde{\beta}) - x\beta\Big]\, l''(x\tilde{\beta})\, \Big[(x\tilde{\beta})^{tp} - l''(x\tilde{\beta})^{-1}l'(x\tilde{\beta}) - x\beta\Big]\\ &\quad - \tfrac{1}{2}l'(x\tilde{\beta})l''(x\tilde{\beta})^{-1}l'(x\tilde{\beta})\\ l(\beta) &= \tfrac{1}{2}\Big[z(x\tilde{\beta}) - x\beta\Big]\, l''(x\tilde{\beta})\, \Big[z(x\tilde{\beta}) - x\beta\Big] + C(x\tilde{\beta}),\end{aligned}$$

con $z(x\tilde{\beta}) = (x\tilde{\beta})^{tp} - l''(x\tilde{\beta})^{-1}l'(x\tilde{\beta})$ y $C(x\tilde{\beta}) = l(\tilde{\beta}) - \frac{1}{2}l'(x\tilde{\beta})l''(x\tilde{\beta})^{-1}l'(x\tilde{\beta})$.

De esta manera, mediante el método descenso coordinado se halla el estimador lasso suponiendo que todos los $\beta_i\ \forall i \neq j$ y el parámetro ϕ son conocidos, y β_j es el parámetro a estimar; entonces denotando con x_{-j} a la matriz de diseño después de eliminar la j-ésima columna y $\tilde{\beta}_{-j}$ al vector de estimadores después de eliminar el j-ésimo componente, la log-verosimilitud puede ser escrita como:

$$\begin{aligned}l(\beta) &= \tfrac{1}{2}\Big[z(x\tilde{\beta}) - x_{-j}\tilde{\beta}_{-j} - x_j\beta_j\Big]\, l''(x\tilde{\beta})\, \Big[z(x\tilde{\beta}) - x_{-j}\tilde{\beta}_{-j} - x_j\beta_j\Big] + C(x\tilde{\beta}),\\ \frac{\partial l(\beta)}{\partial \beta_j} &= -l''(x\tilde{\beta})x_j\Big[z(x\tilde{\beta}) - x_{-j}\tilde{\beta}_{-j} - x_j\beta_j\Big].\end{aligned}$$

El estimador lasso será hallado minimizando la siguiente expresión.

$$\beta_{lasso} = argmin\{-\tfrac{1}{n}l(\beta) + k|\beta_j| + k\sum_{i\neq j}|\beta_i|\}.$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Derivando la expresión anterior e igualando a cero

$$\begin{aligned}
\tfrac{1}{n}l''(x\tilde{\beta})x_j\left[z(x\tilde{\beta})-x_{-j}\tilde{\beta}_{-j}-x_j\beta_j\right]+k*\text{signo}(\beta_j) &= 0,\\
\tfrac{1}{n}l''(x\tilde{\beta})x_j\left[z(x\tilde{\beta})-x_{-j}\tilde{\beta}_{-j}\right]-\tfrac{1}{n}l''(x\tilde{\beta})x_jx_j\beta_j &= -k*\text{signo}(\beta_j),\\
\tfrac{1}{n}l''(x\tilde{\beta})x_jx_j\beta_j &= \tfrac{1}{n}l''(x\tilde{\beta})x_j\left[z(x\tilde{\beta})-x_{-j}\tilde{\beta}_{-j}\right]+\\
&\quad k*\text{signo}(\beta_j).
\end{aligned}$$

Aplicando el operador soft-thresholding (ver la sección A.1 del apéndice) y cambiando la matriz de segundas derivadas respecto de $x\tilde{\beta}$ por su valor esperado se tiene que:

$$\begin{aligned}
\frac{1}{n}\Omega(x\tilde{\beta})x_jx_j\beta_j &= S_k\left(\frac{1}{n}\Omega(x\tilde{\beta})x_j\left[z(x\tilde{\beta})-x_{-j}\tilde{\beta}_{-j}\right]\right),\\
\beta_j &= \frac{S_k\left(\frac{1}{n}\Omega(x\tilde{\beta})x_j\left[z(x\tilde{\beta})-x_{-j}\tilde{\beta}_{-j}\right]\right)}{\frac{1}{n}\Omega(x\tilde{\beta})x_jx_j},
\end{aligned}\tag{3.14}$$

con $z(x\tilde{\beta})=(x\tilde{\beta})^{tp}-\Omega(x\tilde{\beta})^{-1}l'(x\tilde{\beta})$ y $\Omega(x\tilde{\beta})=E(l''(x\tilde{\beta}))$.

De esta manera, reemplazando el vector de primeras derivadas (3.12) y la esperanza de la matriz de segundas derivadas (3.13) se tiene el estimador beta lasso

$$\hat{\beta}_{lasso(j)}=\frac{S_k\left(-\frac{\phi}{n}Wx_j\left[z(x\tilde{\beta})-x_{-j}\tilde{\beta}_{-j}\right]\right)}{-\frac{\phi}{n}Wx_jx_j},\tag{3.15}$$

donde $\hat{\beta}_{lasso(j)}$ es el estimador lasso y $z(x\tilde{\beta})=(x\tilde{\beta})^{tp}-l''(x\tilde{\beta})^{-1}l'(x\tilde{\beta})=(x\tilde{\beta})^{tp}-W^{-1}T(y^*-\mu^*)$. El valor inicial de $\hat{\beta}_{lasso}$ será el estimador de máxima verosimilitud beta $\hat{\beta}$.

En cuanto a la estimación del parámetro de precisión ϕ, también se realiza mediante el método descenso coordinado. Es decir, se mantiene fijo a los estimadores $\tilde{\beta}_j$ y se estima ϕ. En resumen, el proceso que se utiliza para la estimación mediante el método lasso es el siguiente:

- Se hallan los valores iniciales para $\hat{\beta}^{(0)}_{lasso}=\hat{\beta}$.
- Se debe seguir el siguiente algoritmo:
  - **repeat**
  - **for** $j=1,2,\ldots,p$
  - Calcular $W(\hat{\beta}^{(s-1)}_{lasso(j)})$ y $z(x\hat{\beta}^{(s-1)}_{lasso(j)})$
  - Hallar $\hat{\beta}^{(s)}_{lasso(j)}$ con la fórmula (3.15)
  - Calcular $dif er=\hat{\beta}^{(s)}_{lasso}-\hat{\beta}^{(s-1)}_{lasso}$
  - **Fin**
  - Se minimiza $-\frac{1}{n}l(\hat{\beta}^{(s)}_{lasso(j)})+k|\hat{\beta}^{(s)}_{lasso(j)}|$ respecto de $\hat{\phi}^{(s)}_{lasso}$
  - **Hasta** conseguir la convergencia,

donde s es el número de iteración en el que se encuentra el algoritmo. Al igual que el estimador ridge, el procedimiento para hallar los estimadores lasso de β y ϕ fue programado

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

en el software R, su código se encuentra en la función **beta_lasso** de la sección (B.1.4) del apéndice.

Debido a que el estimador lasso, no tiene una forma cerrada para hallar los estimadores, no se puede estudiar analíticamente las propiedades del estimador, como su sesgo, varianza y error cuadrático medio. En su lugar, se estudia vía simulación estas tres propiedades.

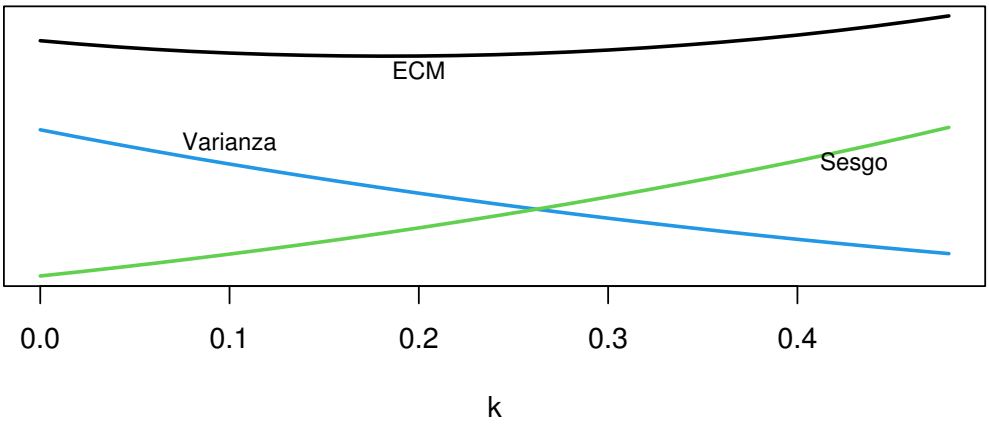


Figura 3.2: Relación entre el parámetro de contracción k y el $ECM(\hat{\beta}_{lasso})$

La figura (3.2) también es construida con datos simulados, donde el tamaño de la muestra es $n = 100$, el número de variables explicativas es de $p = 6$ los cuales presentan multicolinealidad[2], y se realizan $S = 10000$ simulaciones con las que se calcula el error cuadrático medio, el sesgo y la varianza. La figura muestra la construcción del ECM con la varianza y el sesgo del estimador. Al igual que el estimador ridge, la varianza del estimador lasso va disminuyendo según k va creciendo, por el contrario, el sesgo tiene una relación positiva con k, es decir, que mientras k va creciendo, el sesgo también lo hace.

3.3. Parámetro de contracción k

Hasta el momento, se ha realizado la estimación de los parámetros β sin tener conocimiento del valor que toma el parámetro de contracción, y como se vió en las figuras (3.1) y (3.2) depende de este parámetro el obtener el modelo más apropiado para los datos.

La estimación de este parámetro depende del método de estimación: en el caso de ridge, se proponen cuatro estimaciones de k comenzando con la propuesta de Hoerl & Kennard. En el caso de lasso se toma el método de validación cruzada que consiste en dividir el conjunto de datos en dos partes, denominados conjunto de prueba y conjunto de entrenamiento.

[2]El grado de colinealidad de las variables explicativas está determinado por $a = (0{,}85 \quad 0{,}79 \quad 0{,}75 \quad 0{,}85 \quad 0{,}9 \quad 0{,}85)$, el cual se define en el siguiente capítulo.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

3.3.1. Métodos de estimación de k para el estimador ridge

En el presente trabajo se describen cuatro estimadores de k, no obstante, se pueden encontrar otras alternativas en el artículo de Qasim, Mansson & Kibria [18].

De la sección 3.1.3, se tiene que $\hat{\gamma} = (\hat{\gamma}_1, \dots \hat{\gamma}_p)^{tp} = P\hat{\beta}$ donde P es una matriz ortogonal cuyas columnas son los vectores propios de la matriz $X^{tp}WX$, $\lambda_{\text{máx}}$ y $\lambda_{\text{mín}}$ son el valor propio máximo y mínimo de la matriz $X^{tp}WX$, entonces:

- El primer estimador de este grupo viene del resultado de comparar el ECM de los estimadores ridge y el estimador emv, por lo cual, de la desigualdad (3.8),

  $$k_{(HK)} = \frac{1}{\hat{\phi}\hat{\gamma}^2_{\text{máx}}},$$

  el cual fue propuesto por Hoerl & Kennard [11] para el caso del modelo de regresión clásico.

- Muniz & Kibria [16] propuso aplicar la mediana al término $\sqrt{\frac{1}{\phi\gamma_j^2}}$, de esta manera se tiene,

  $$k_{med} = \text{mediana}\left(\sqrt{\frac{1}{\hat{\phi}\hat{\gamma}_j^2}}\right).$$

- Por último, Abonazel & Taha [1] sugirieron los estimadores:

  $$k_{\text{máx}} = \frac{\lambda_{\text{máx}}}{\hat{\phi}\hat{\gamma}^2_{\text{máx}}} \text{ y } k_{\text{mín}} = \frac{\lambda_{\text{mín}}}{\hat{\phi}\hat{\gamma}^2_{\text{mín}}}.$$

3.3.2. Validación cruzada

Se explicarán dos tipos de validación cruzada: la primera que consiste en dejar una observación de los datos fuera del modelo; y la segunda que consiste en particionar el conjunto de datos en h grupos.

El primer método de validación cruzada extrae una observación de la muestra como dato de prueba, y el resto de las observaciones formaría el conjunto de entrenamiento. Es decir, suponiendo que se cuenta con n observaciones, se aplica el modelo con $n-1$ observaciones; con el dato sobrante se calcula el error de predicción con los estimadores del modelo construido con las $n-1$ observaciones.

Este procedimiento se realiza para cada una de las n observaciones de la muestra y la suma de estos errores da como resultado el error de predicción de las n observaciones.

$$EP = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

donde y_i es el valor de prueba y $\hat{y}_i$ es el valor estimado con los $n-1$ valores restantes.

Cada una de las n observaciones es analizado por este método. Luego son comparados y se selecciona el que tiene menor error de predicción. El problema que presenta este tipo

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

de estimación de k es el hecho de tener que evaluar el modelo para cada una de las n observaciones, lo que implica un mayor tiempo de análisis.

El segundo tipo de validación cruzada es bastante similar al anterior, la diferencia radica en que el dato de prueba, ahora es un conjunto de observaciones. El conjunto de datos se divide en h subconjuntos disjuntos, uno de estos será el conjunto de prueba y los $h-1$ subconjuntos restantes serán los datos de entrenamiento.

Entonces, con los $h-1$ subconjuntos se aplica el modelo de regresión, y con sus estimadores se calcula el error de predicción con el subconjunto restante (el conjunto de prueba). Por tal motivo, se obtiene un error de predicción asociado a cada conjunto de prueba. De esta manera, el error de predicción esta dado por:

$$EP_h = \frac{1}{H}\sum_{i=1}^{H}(y_{ih} - \hat{y}_{ih})^2,$$

donde y_{ih} corresponde al grupo de prueba h y $\hat{y}_{ih}$ es la estimación realizada con todos los datos a excepción del grupo h siendo h la cantidad de particiones o grupos realizados y H es el número de observaciones en el grupo h.

Por último, el estimador de k esta dado por:

$$k = \ \arg\min \ EP_h = \ \arg\min \ \frac{1}{H}\sum_{i=1}^{H}(y_{ih} - \hat{y}_{ih})^2. \tag{3.16}$$

En cuanto a los valores de k con los que se realiza este ejercicio, se debe construir una grilla, comenzando de un k lo suficientemente grande como para establecer que $\hat{\beta}_{lasso} = 0$, y se disminuye k hasta llegar a una solución óptima. Este procedimiento fue propuesto por Friedman, Hastie & Tibshirani [7].

Considere la ecuación (3.15). Entonces denotando por $w_i(0)$ al i-ésimo elemento de la diagonal de la matriz W evaluada en $\hat{\beta}_{lasso} = 0$; y $z(0)_i$ al i-ésimo elemento del vector z evaluada en $\hat{\beta}_{lasso} = 0$, luego se tiene que si $-\frac{\phi}{n}w_i(0)x_{ij}z(0)_i < k$ para todo j, entonces $\tilde{\beta} = 0$. Por lo tanto, se establece que el primer k está dado por:

$$k_{\text{máx}} = \max_j -\frac{\phi}{n}\sum_{i=1}^{n} w_i(0)x_{ij}z(0)_i,$$

de manera análoga, se establece $k_{\text{mín}}$ y se calculan las soluciones sobre una grilla de m valores entre $k_{\text{mín}}$ y $k_{\text{máx}}$. Para este caso, se trabaja con un $m = 50$.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Capítulo 4

Simulación y aplicación

4.1. Simulación

En esta sección se realizan simulaciones para evaluar las propiedades de los estimadores en muestras finitas. Para ello se usan distintas combinaciones de los parámetros y las variables que intervienen en los modelos de regresión beta. La programación se realiza en el software estadístico R en su versión 4.1.1.

Para iniciar se detalla el programa que calcula el error cuadrático medio y su relación con la colinealidad de las variables explicativas, posteriormente realiza una comparación entre los estimadores ridge, estimador lasso y el estimador máximo verosímil en función de su error cuadrático medio.

4.1.1. Programa para relacionar la correlación y el ECM

Con el objetivo de mostrar el rendimiento que tienen los estimadores presentados en este trabajo, se analiza el comportamiento de sus ECM en función del incremento de la colinealidad presentada en las variables explicativas mediante simulaciones.

Las características que se toman en cuenta para la simulación son las siguientes: el modelo de regresión beta está definido aplicando la función de enlace logit, el número de observaciones n, el número de parámetros a estimar p, la matriz de diseño X cuyas variables explicativas están estandarizadas, y presentan un grado de colinealidad, la cual es determinada mediante la siguiente fórmula propuesta por McDonald & Galarneau [15]:

$$x_{ij} = (1 - \alpha_j^2)^{\frac{1}{2}} z_{ij} + \alpha_j z_{ip} \qquad j = 1, \dots p-1 \quad : \quad i = 1, \dots, n, \tag{4.1}$$

donde z_{ij} tiene una distribución uniforme y α_j es el j-ésimo elemento de α, siendo α un vector de $p \times 1$ que especifica el grado de colinealidad entre las variables explicativas. Asimismo, dado que la matriz de diseño X está estandarizada, entonces $X^{tp}X$ tiene la forma de una matriz de correlaciones la cual dependerá de α.

Tomando en cuenta esta metodología, se crea una función que genera la matriz de diseño X (dicha función se encuentra en la sección B.1.1 con el nombre ***genera_x***), tomando como valores iniciales el número de observaciones **n**, el número de variables **p**, el vector $\boldsymbol{\alpha}$ y una semilla que genera la variable seudo aleatoria z_{ij} (**semilla**), con estos

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

valores se crea la matriz de diseño con la fórmula descrita en la ecuación (4.1).

Por otro lado, se crea otra función (***genera_y***) que genera **S** simulaciones, donde cada simulación crea una variable aleatoria y_i de n observaciones, con $i = 1, \ldots, S$ la cual representa a la i-ésima simulación, la creación de cada variable aleatoria se la realiza con la misma matriz de diseño (X) y el mismo vector de parámetros β (**b**), la salida de esta función es una matriz con **S** filas donde cada fila corresponde a una simulación de la variable aleatoria y_i con **n** observaciones. El procedimiento para la creación de y_i consiste en hallar la media μ del modelo mediante la función de enlace logit

$$\mu_i = \frac{\exp(X_i^{tp}\beta)}{1 + \exp(X_i^{tp}\beta)} \ , \ i = 1, 2, \ldots n,$$

donde $\mu = (\mu_1 \quad \ldots \quad \mu_n)$ es un vector de dimensión n. Posteriormente dado un parámetro de precisión ϕ (**phi**) se genera un vector y seudo aleatorio con distribución beta de parámetros $\mu\phi$ y $(1-\mu)\phi$. Cabe señalar que el vector y es creado con una semilla que se iguala con el número de simulación en el que se encuentra.

Esta función puede ser encontrada en la sección B.1.1, la cual tiene como valores de entrada: el número de simulaciones **S** de la variable y que se desean crear, el número de observaciones **n**, la matriz de diseño **X**, el vector de parámetros **b**, y el parámetro de precisión **phi** con el que se crea la variable aleatoria y_i. La salida de esta función es una matriz construida con las S simulaciones de la variable dependiente y con n observaciones, entonces Y es una matriz de dimensión $S \times n$.

Sin embargo, para crear la variable aleatoria y se necesitan los parámetros β del modelo, que estarán denotados como **b** en las funciones desarrolladas. Los parámetros del modelo se calculan con la suma de los vectores propios de la matriz $X^{tp}X$ divididos por la raíz cuadrada del número de parámetros, es decir:

$$\beta_j = \frac{\sum_{i=1}^{p} v_{ij}}{\sqrt{p}}, \tag{4.2}$$

donde v_{ij} es el i-ésimo componente del j-ésimo vector propio (v_j) de la matriz $X^{tp}X$.

De esta manera, se tiene un modelo de regresión beta simulado donde sus parámetros son conocidos, el siguiente paso será realizar la estimación de estos parámetros para diferentes niveles de colinealidad de las variables explicativas que son determinadas por el vector α, mediante distintos métodos de estimación: estimación sin penalizar, estimación ridge y estimación lasso. En el apéndice se presenta una función específica para cada uno de estos métodos de estimación, secciones B.1.2, B.1.3 y B.1.4 respectivamente.

Por último, el error cuadrático medio es calculado con la función ***calcula_ECM*** (sección B.1.5), donde se unen las funciones explicadas anteriormente, siguiendo el procedimiento que se describe a continuación:

- Se crea la matriz de diseño X de dimensión $n \times p$ con la función ***genera_x***.
- Con la matriz X creada, se halla el parámetro β (b) de dimensión $p \times 1$ con la ecuación (4.2).

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

- Con la matriz X y los parámetros β se hallan S simulaciones de la variable dependiente y con la función ***genera_y***.

- Se estandariza la matriz de diseño X y se crea una matriz **beta** de $S \times (p+1)$ donde se almacenan los estimadores de β y ϕ.

- Se crea un bucle de tamaño **S**, en cada ciclo se estiman los parámetros con la matriz de diseño simulada **X** y la i-ésima simulación de la variable dependiente (y_i), que se encuentra en la i-ésima fila de la matriz **Y**, recuerde que la matriz **Y** es de dimensión $S \times n$, entonces se tendrán **S** estimaciones de β, una por cada simulación de y, creando de esta manera una matriz **beta** de dimensión $S \times (p+1)$.

  Los S estimadores de la matriz **beta** son calculados con el mismo método de estimación, el cual será determinado por la variable de tipo cadena **regresion**, los valores que registra esta variable son **emv**, **ridge** y **lasso**:

  - **emv**, si desea estimar los parámetros con el método de máxima verosimilitud.
  - **ridge**, para estimar los parámetros con el método ridge.
  - **lasso**, si desea estimar los parámetros con el método lasso.

  Posteriormente a la estimación de los parámetros, en cada ciclo se aplica una estandarización inversa para convertir cada fila de la matriz **beta** a estimadores que se relacionen con las unidades de medida originales de las variables explicativas. Este procedimiento se realiza de la siguiente manera:

$$\tilde{\beta}_i = \frac{\tilde{\beta}_i^*}{\sigma_{x_i}},$$

  donde,

  $\tilde{\beta}_i$: es un estimador de β_i de las variables explicativas originales.
  $\tilde{\beta}_i^*$: es un estimador de β_i de las variables explicativas estandarizadas.
  σ_{x_i} : es la desviación estándar de la i-ésima variable explicativa original.

  Y el intercepto del modelo se calcula con:

$$\tilde{\beta}_0 = \tilde{\beta}_0^* - \sum_{i=1}^{p} \frac{\tilde{\beta}_i^*}{\sigma_{x_i}} \bar{x}_i,$$

  donde,

  $\tilde{\beta}_0$: es el intercepto del modelo de regresión.
  $\tilde{\beta}_0^*$: es el intercepto del modelo construido con las variables explicativas estandarizadas.
  $\tilde{\beta}_i^*$: es un estimador de β_i de las variables explicativas estandarizadas.
  σ_{x_i} : es la desviación estándar de la i-ésima variable explicativa original.
  $\bar{x}_i$: es el promedio de la i-ésima variable explicativa original.

- Hasta el momento se tiene una matriz de estimadores de dimensión $S \times (p+1)$, a partir de esta matriz se halla la varianza, el sesgo y el error cuadrático medio de los p estimadores de β. Estos últimos datos son los que se compararán con la colinealidad de las variables explicativas determinadas por el vector α.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

4.1.2. Comparación de los estimadores ridge con distintos parámetros de contracción

Para este ejercicio, se toma en cuenta la función **beta_ridge** que se encuentra en la sección B.1.2 del apéndice. Además, los estimadores de k con los que se estima β son aquellos que se describieron en la sección (3.3).

La función **estim_k** que se encuentra en la sección (B.1.6), calcula los estimadores de k mediante los cuatro métodos explicados, los parámetros de entrada son: la variable dependiente y, la matriz de diseño x y los estimadores de máxima verosimilitud.

En cuanto a la estimación de β, la función **beta_ridge** tiene como valores de entrada la matriz de diseño X denotada por **x**, la variable dependiente **y**, el valor del parámetro de contracción k (**k_cont**), la mínima diferencia para considerar que el estimador presenta convergencia denominado por **difer** y el máximo de iteraciones permitidas para buscar la convergencia **max_iter**.

Finalmente, con la función ***calcula_ECM*** se calcula el error cuadrático medio de los estimadores ridge calculados con los cuatro parámetros de contracción.

- Validación cruzada (k_{vc}).
- El estimador propuesto por Hoerl & Kennard [11] (k_{hk}).
- El estimador propuesto por Muniz & Kibria [16] (k_{med}).
- Y los estimadores propuestos por Abonazel & Taha [1] (k_{max}), (k_{min}).

La estimación es realizada con distintos tamaños de muestra: $n = 30, 50, 100, 200$; se generaron $S = 10000$ simulaciones de la variable dependiente y (mediante la función de enlace logit), por lo cual, se generaron también $S = 10000$ estimadores del vector de parámetros β (uno por cada simulación), el número de parámetros a estimar es de 4 y existe correlación entre las variables explicativas, tomando 3 grados de colinealidad mostrados a continuación:

- $baja = \begin{pmatrix} 0{,}12 & 0{,}21 & 0{,}32 & 0{,}22 \end{pmatrix}$ colinealidad baja.
- $moderada = \begin{pmatrix} 0{,}42 & 0{,}56 & 0{,}48 & 0{,}51 \end{pmatrix}$ colinealidad moderada.
- $alta = \begin{pmatrix} 0{,}9 & 0{,}96 & 0{,}88 & 0{,}89 \end{pmatrix}$ colinealidad alta.

Los resultados de las simulaciones se muestran en la tabla (4.1), los cuales fueron calculados tomando un parámetro de precisión igual a dos.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Tabla 4.1: Error cuadrático medio de los estimadores de máxima verosimilitud y ridge para distintos estimadores de k, simulaciones de $S = 10000$ repeticiones, para un tamaño de muestra de $n = 30, 50, 100, 200$, $\phi = 2$, $p = 4$ y distintos niveles de colinealidad

Colinealidad	n	*emv*	Ridge			
		$k = 0$	k_{hk}	k_{med}	k_{max}	k_{min}
Baja	30	3,701	3,3715	2,9189	1,2694	2,4177
	50	1,5464	1,4628	1,344	0,535	0,4227
	100	0,8081	0,7823	0,747	0,2795	0,2604
	200	0,2977	0,2955	0,292	0,1924	0,2066
Moderada	30	3,9711	3,7155	3,1726	1,8443	2,531
	50	1,7617	1,6615	1,4878	0,4861	0,4884
	100	0,8598	0,8329	0,7966	0,2443	0,1008
	200	0,3139	0,3116	0,3075	0,1783	0,0515
Alta	30	7,7282	7,0258	4,8047	1,6316	4,9935
	50	3,8445	3,5166	2,6888	0,4146	1,2832
	100	1,8354	1,0399	1,5233	0,0456	1,1492
	200	0,6894	0,6767	0,6462	0,134	0,1882

De la misma manera, la tabla (4.2) se construye con la diferencia que en este caso, el parámetro de precisión es igual a 4.

Tabla 4.2: Error cuadrático medio de los estimadores de máxima verosimilitud y ridge para distintos estimadores de k, simulaciones de $S = 10000$ repeticiones, para un tamaño de muestra de $n = 30, 50, 100, 200$, $\phi = 4$, $p = 4$ y distintos niveles de colinealidad

Colinealidad	n	*emv*	Ridge			
		$k = 0$	k_{hk}	k_{med}	k_{max}	k_{min}
Baja	30	2,489	2,3899	2,1885	1,6433	2,1708
	50	1,0327	1,0071	0,9593	0,6735	0,6115
	100	0,5911	0,5833	0,5664	0,3968	0,3897
	200	0,2977	0,2955	0,292	0,1924	0,2066
Moderada	30	2,7731	2,6925	2,4465	2,0855	2,3748
	50	1,3041	1,2725	1,1867	0,7664	0,7209
	100	0,64	0,6317	0,615	0,3884	0,2099
	200	0,3139	0,3116	0,3075	0,1783	0,0515
Alta	30	5,3851	5,1556	4,0327	2,6005	4,4885
	50	2,7984	2,6933	2,2663	0,9046	1,6541
	100	1,2812	0,9158	1,1583	0,0473	1,0995
	200	0,6894	0,6767	0,6462	0,134	0,1882

De la tabla (4.1) se observa que en todos los casos, los estimadores ridge (para cualquier estimador del parámetro de precisión) presentan un menor error cuadrático medio que el estimador de máxima verosimilitud (emv). Centrándose en los distintos estimadores de k, los que generan menor error cuadrático medio son los propuestos por Abonazel & Taha, k_{min} y k_{max}.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

En cuanto a la colinealidad, este afecta en el error cuadrático medio de los estimadores de manera positiva, es decir, mientras mayor es el grado de colinealidad, mayor es el error cuadrático medio. Respecto al tamaño de la muestra, cuando $n = 30$ el error cuadrático medio es más grande en comparación a los hallados para un tamaño de muestra más grande.

Por último, analizando las tablas (4.1) y (4.2), en donde la primera se calcula con un parámetro de precisión $\phi = 2$ y la segunda es calculada con un $\phi = 4$, y tomando en cuenta que el parámetro de precisión juega un papel inverso a la varianza, entonces, se muestra que mientras mayor es el parámetro de precisión ($\phi = 4$) menor es el error cuadrático medio.

4.1.3. Comparación de los estimadores de máxima verosimilitud, ridge y lasso

Para la comparación de los estimadores de máxima verosimilitud, estimador ridge y estimador lasso se toma en cuenta las ecuaciones (2.18), (3.1) y (3.15) respectivamente. Las funciones que hallan estos estimadores están en el apéndice en las secciones B.1.2, B.1.3 y B.1.4, con los nombres ***reg_beta***, ***beta_ridge*** y ***beta_lasso*** respectivamente.

Las tres funciones tienen como valores de entrada la matriz de diseño X denotada por **x** y la variable dependiente **y**, sin embargo, los estimadores ridge y lasso requieren también del parámetro de contracción k denotándose como **k_cont**, la mínima diferencia para considerar que el estimador presenta convergencia denominado por **difer** (con un valor preasignado igual a 0,0001); y el máximo de iteraciones permitidas para buscar la convergencia **max_iter** (con un valor preasignado de 200). La salida de las funciones es un vector donde se encuentran los estimadores de β y ϕ.

En cuanto al parámetro de contracción k (**k_cont**), en el caso de ridge se utiliza el estimador $k_{\text{mín}}$; y en el caso de lasso, se estima mediante el método de validación cruzada cuyo procedimiento se encuentra programado en la función **valid_cruz** de la sección (B.1.6), con $h = 10$ subconjuntos.

La función ***calcula_ECM*** calcula el error cuadrático medio con los tres métodos de estimación analizados en función al grado de colinealidad. Los resultados son mostrados en las tablas (4.3), (4.4) y (4.5) realizadas con distintos tamaños de muestra: $n = 30, 50, 100$ respectivamente; se generaron $S = 10000$ simulaciones de la variable dependiente y (mediante la función de enlace logit), con sus respectivos estimadores, el número de parámetros a estimar es de 4 y presentan un grado de colinealidad baja, moderada y alta definidos en la anterior sección.

Los resultados de las tablas son el ECM de cada estimador de β y ϕ, la última columna corresponde al error cuadrático medio total dado por $ECM = \sum_{i=0}^{p} ECM_i$ donde ECM_i es el error cuadrático medio del estimador de β_i. Cabe señalar que ϕ no se incluye en el error cuadrático medio total.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Tabla 4.3: Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 30$, $\phi = 2$ y $p = 4$

Colinealidad	Método	ECM_i						ECM
		β_0	β_1	β_2	β_3	β_4	ϕ	*total*
Baja	emv	1,2143	0,5535	0,4224	0,8229	0,6878	0,992	3,701
	ridge	0,7663	0,3962	0,3018	0,5196	0,4339	0,8914	2,4177
	lasso	0,3112	0,0086	0,0001	0,0154	0,0496	0,5386	0,3848
Moderada	emv	1,1571	0,5686	0,4426	0,8573	0,9456	1,1296	3,9711
	ridge	0,789	0,3872	0,2769	0,5111	0,5667	0,9961	2,531
	lasso	1,5393	0,0027	0,0033	0,0058	0,0182	1,1343	1,5693
Alta	emv	0,6054	1,8955	1,2099	1,9216	2,0958	1,0953	7,7282
	ridge	0,5043	1,1319	0,7797	1,1654	1,4122	0,9928	4,9935
	lasso	1,1939	0,0142	0	0,0015	0,0013	1,0366	1,2108

Tabla 4.4: Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 50$, $\phi = 2$, $p = 4$

Colinealidad	Método	ECM_i						ECM
		β_0	β_1	β_2	β_3	β_4	ϕ	*total*
Baja	emv	0,3283	0,3078	0,3425	0,2526	0,3153	0,3331	1,5464
	ridge	0,1333	0,074	0,0707	0,0635	0,0812	0,2526	0,4227
	lasso	0,2933	0,0002	0,0003	0,0178	0,04	0,5968	0,3516
Moderada	emv	0,2968	0,3449	0,4585	0,2964	0,3651	0,3911	1,7617
	ridge	0,1679	0,0938	0,0779	0,068	0,0807	0,3031	0,4884
	lasso	2,022	0,4926	0,0018	0,015	0,0049	0,4107	2,5363
Alta	emv	0,202	1,0182	0,9295	0,8144	0,8803	0,3878	3,8445
	ridge	0,1752	0,3113	0,2918	0,229	0,276	0,3214	1,2832
	lasso	1,5092	0,013	0,0001	0,0001	0,0006	1,183	1,523

Tabla 4.5: Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 100$, $\phi = 2$, $p = 4$

Colinealidad	Método	ECM_i						ECM
		β_0	β_1	β_2	β_3	β_4	ϕ	*total*
Baja	emv	0,2071	0,1372	0,1443	0,1465	0,173	0,124	0,8081
	ridge	0,0796	0,0389	0,0407	0,0399	0,0613	0,1071	0,2604
	lasso	1,5915	0,0024	0,003	0,0003	0,0593	1,2283	1,6564
Moderada	emv	0,1818	0,1492	0,1687	0,1629	0,1971	0,1284	0,8598
	ridge	0,0521	0,008	0,0097	0,0081	0,0229	0,1012	0,1008
	lasso	1,791	0,0012	0,0052	0,0006	0,0236	1,2911	1,8216
Alta	emv	0,1109	0,4706	0,3648	0,4129	0,4762	0,115	1,8354
	ridge	0,1048	0,2759	0,2305	0,2486	0,2896	0,1094	1,1492
	lasso	0,5771	0,1135	0,0003	0	0	0,6271	0,6909

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Observe que, en la mayoría de los casos, el estimador ridge presenta un menor ECM, que los otros dos estimadores. En cuanto al estimador lasso, no se destaca frente a los otros dos estimadores con algunas excepciones, (cuando $n = 30$). Otro aspecto importante que se muestra en las tablas, es el comportamiento del ECM respecto al tamaño de muestra; para $n = 30$, el ECM es más grande en comparación a los demás, cuanto más crece el tamaño de muestra, mayor es la disminución del ECM. Sin embargo, esta característica no es muy clara en el estimador lasso.

Al igual que el ejercicio de la anterior sección, ahora se comparan los estimadores con las mismas características con la diferencia de que ahora, el parámetro de precisión es $\phi = 10$. De esta manera, al comparar las tres tablas anteriores con las tablas (4.6), (4.7) y (4.8) se ratifica que el error cuadrático medio y ϕ tienen una relación negativa, es decir que mientras más grande es el parámetro de precisión, menor es el ECM incluyendo al estimador lasso.

Tabla 4.6: Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 30$, $\phi = 10$, $p = 4$

Colinealidad	Método	ECM_i						ECM
		β_0	β_1	β_2	β_3	β_4	ϕ	*total*
Baja	emv	0,3893	0,1713	0,1331	0,259	0,2191	24,3447	1,1718
	ridge	0,3776	0,1679	0,1303	0,2513	0,2126	24,3079	1,1396
	lasso	0,0393	0,5003	0,0007	0,0157	0,0498	10,9343	0,6058
Moderada	emv	0,3967	0,199	0,1562	0,2979	0,3465	24,3405	1,3963
	ridge	0,3832	0,1927	0,1498	0,2842	0,3308	24,2705	1,3408
	lasso	0,0925	0,0015	0,0033	0,0058	0,0182	59,5522	0,1213
Alta	emv	0,2058	0,6797	0,4151	0,6607	0,7513	24,3387	2,7126
	ridge	0,1989	0,6205	0,3842	0,6029	0,7001	24,2276	2,5065
	lasso	0,0545	0,0143	0	0,0015	0,0013	56,7486	0,0716

Tabla 4.7: Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 50$, $\phi = 10$, $p = 4$

Colinealidad	Método	ECM_i						ECM
		β_0	β_1	β_2	β_3	β_4	ϕ	*total*
Baja	emv	0,1051	0,1004	0,1141	0,0837	0,1056	8,4719	0,5088
	ridge	0,0933	0,0845	0,0931	0,0705	0,0882	8,3705	0,4297
	lasso	0,0235	0,0068	0,0003	0,0178	0,04	41,8347	0,0884
Moderada	emv	0,1042	0,1302	0,1813	0,1131	0,1401	8,8059	0,669
	ridge	0,0974	0,1101	0,1389	0,0921	0,114	8,669	0,5525
	lasso	0,1597	0,3209	0,0017	0,0151	0,0049	33,1071	0,5024
Alta	emv	0,0677	0,3867	0,3467	0,3027	0,3217	8,6404	1,4255
	ridge	0,0661	0,3142	0,2842	0,242	0,2636	8,5267	1,1701
	lasso	0,0922	0,0561	0,0001	0,0001	0,0006	53,7335	0,1491

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Tabla 4.8: Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para $n = 100$, $\phi = 10$, $p = 4$

Colinealidad	Método	ECM_i						ECM
		β_0	β_1	β_2	β_3	β_4	ϕ	*total*
Baja	emv	0,0768	0,0521	0,0543	0,0564	0,0677	2,9381	0,3075
	ridge	0,0714	0,0477	0,0496	0,0514	0,0624	2,9243	0,2825
	lasso	0,1114	0,2939	0,003	0,0003	0,0593	35,3727	0,4678
Moderada	emv	0,0688	0,0577	0,0668	0,0646	0,0763	2,9534	0,3342
	ridge	0,059	0,0427	0,0465	0,0463	0,0569	2,9003	0,2514
	lasso	0,1331	0,1534	0,0052	0,0006	0,0236	47,2756	0,3159
Alta	emv	0,0369	0,1607	0,1276	0,1462	0,165	2,8874	0,6364
	ridge	0,0367	0,1547	0,1236	0,1411	0,1594	2,8847	0,6155
	lasso	0,008	0,686	0,0003	0	0	8,5761	0,6943

Como se pudo apreciar, en la mayoría de los casos, cuando existe colinealidad entre las variables explicativas y tomando como medida de rendimiento al ECM, el estimador ridge es superior a lasso y emv. Sin embargo, estudiando otras características del modelo, lasso presentaría un mejor rendimiento, una de estas características es la selección de variables, la cual, es una de las ventajas que lasso presenta frente a ridge.

En este sentido, las tablas (4.9), (4.10) y (4.11) son generadas con distintos tamaños de muestra: $n = 30, 50, 100$ respectivamente, además se realizaron $S = 10000$ simulaciones, el número de parámetros a estimar es de 6 de los cuales 2 son iguales a 0, es decir $\beta_5 = \beta_6 = 0$; y el grado de colinealidad presentado en el resto de las variables explicativas esta dado por:

- $baja = \begin{pmatrix} 0{,}2 & 0{,}25 & 0{,}27 & 0{,}3 & - & - \end{pmatrix}$ colinealidad baja.
- $moderada = \begin{pmatrix} 0{,}52 & 0{,}61 & 0{,}45 & 0{,}49 & - & - \end{pmatrix}$ colinealidad moderada.
- $alta = \begin{pmatrix} 0{,}86 & 0{,}95 & 0{,}94 & 0{,}88 & - & - \end{pmatrix}$ colinealidad alta.

Tabla 4.9: Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para un tamaño de muestra de $n = 30$, $\phi = 2$, $p = 6$ con 2 parámetros iguales a 0 ($\beta_5 = \beta_6 = 0$)

Colinealidad	Método	ECM_i								ECM
		β_0	β_1	β_2	β_3	β_4	β_5	β_6	ϕ	*total*
Baja	emv	1	0,43	0,5	0,58	0,41	0,57	0,6	6,41	4,1
	ridge	0,95	0,41	0,46	0,55	0,38	0,53	0,57	6,33	3,85
	lasso	0,15	0,12	0	0	0,02	0	0	2,69	0,29
Moderada	emv	0,59	0,48	0,72	0,63	0,42	0,72	0,81	6,24	4,37
	ridge	0,58	0,46	0,68	0,61	0,4	0,68	0,78	6,19	4,19
	lasso	0,1	0,14	0	0	0,02	0	0	2,43	0,25
Alta	emv	0,32	1,49	4,55	3,24	1,52	2,16	3,39	6,75	16,68
	ridge	0,31	1,46	4,32	3,14	1,47	2,09	3,29	6,71	16,08
	lasso	0,35	0,63	0	0	0	0	0	1,44	0,99

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Tabla 4.10: Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para un tamaño de muestra de $n = 50$, $\phi = 2$, $p = 6$ con 2 parámetros iguales a 0 ($\beta_5 = \beta_6 = 0$)

Colinealidad	Método	ECM_i								ECM
		β_0	β_1	β_2	β_3	β_4	β_5	β_6	ϕ	*total*
Baja	emv	0,62	0,33	0,39	0,37	0,36	0,37	0,38	0,49	2,81
	ridge	0,4	0,18	0,2	0,19	0,21	0,2	0,21	0,41	1,59
	lasso	0,68	0	0	0	0	0	0	0,8	0,69
Moderada	emv	0,51	0,41	0,51	0,42	0,42	0,53	0,41	0,51	3,22
	ridge	0,34	0,15	0,16	0,15	0,17	0,18	0,16	0,39	1,3
	lasso	1,23	0	0	0	0,02	0	0	1,05	1,24
Alta	emv	0,29	1,16	2,93	2,27	1,38	1,48	1,28	0,52	10,78
	ridge	0,28	0,96	2,02	1,66	1,14	1,22	1,08	0,49	8,36
	lasso	1,31	0,01	0	0	0	0	0	1,03	1,32

Tabla 4.11: Error cuadrático medio de los estimadores emv, ridge y lasso mediante simulaciones de $S = 10000$ repeticiones, para un tamaño de muestra de $n = 100$, $\phi = 2$, $p = 6$ con 2 parámetros iguales a 0 ($\beta_5 = \beta_6 = 0$)

Colinealidad	Método	ECM_i								ECM
		β_0	β_1	β_2	β_3	β_4	β_5	β_6	ϕ	*total*
Baja	emv	0,27	0,15	0,15	0,14	0,17	0,16	0,17	0,14	1,21
	ridge	0,06	0,02	0,02	0,02	0,02	0,02	0,02	0,1	0,17
	lasso	0,64	0,01	0	0	0	0	0	0,78	0,65
Moderada	emv	0,19	0,19	0,21	0,17	0,19	0,21	0,22	0,14	1,37
	ridge	0,06	0,01	0,01	0,01	0,01	0,01	0,01	0,1	0,13
	lasso	0,82	0,01	0	0	0	0	0	0,83	0,83
Alta	emv	0,11	0,51	1,23	0,99	0,62	0,63	0,82	0,15	4,91
	ridge	0,1	0,34	0,46	0,4	0,36	0,36	0,44	0,13	2,45
	lasso	1,22	0,01	0	0	0	0	0	1	1,22

En estas últimas 3 tablas, el estimador lasso presenta un mejor rendimiento que los otros estimadores, esto debido a que existen variables explicativas que no aportan al modelo ($\beta_5 = \beta_6 = 0$), los cuales son detectados por el método lasso, por tal motivo, cuando se quiere realizar una selección de variables, el estimador lasso se ajusta de mejor manera al modelo de regresión beta que los estimadores emv y ridge. En cuanto a la relación del ECM y el tamaño de la muestra, no se observa algo claro.

4.2. Aplicación

En esta sección se toma los datos de comunas que pertenecen a tres regiones de Chile: Ñuble, Biobío y La Araucanía, información que fue recolectada en un proyecto de tesis cuyos autores son: Fuentes & Ortiz [9].

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

La base de datos está compuesta por 86 observaciones (comunas), y se tomarán 11 variables para este análisis, las cuales se describen a continuación:

- **Proporción de pobreza por ingreso (y).-** construida con el ingreso del hogar que se define como la suma del ingreso autónomo del hogar, las transferencias monetarias que recibe el hogar por parte del Estado, y una imputación por concepto de arriendo de la vivienda.
- **Tasa de plantación forestal (x_1).-** Representa la superficie comunal cubierta por plantaciones forestales, donde 0 representa ausencia de plantaciones y 1 corresponde a la comuna con la mayor proporción de estos.
- **Porcentaje de personas que trabajan en el sector primario (x_2).-** El sector primario corresponde a las actividades extractivas, que contemplan las áreas de agricultura, pesca, ganadería y minería
- **Porcentaje de personas que trabajan en el sector secundario (x_3).-** Este sector considera las empresas, instituciones o actividades por cuenta propia que realizan principalmente actividades industriales y manufactureras.
- **Porcentaje de personas que trabajan en el sector terciario (x_4).-** Este sector contempla las empresas, instituciones o actividades por cuenta propia que pertenecen al área de servicios.
- **Años de estudio del jefe del hogar (x_5).-** Corresponde al promedio de años de estudio formales reconocidos de las personas que se declaran jefe o jefa de hogar.
- **Índice de complejidad económica (ICE) (x_6).-** Capacidad de que un país genere un valor determinado en base a la diversificación de sus exportaciones.
- **Proporción de hogares sin agua (x_7).-** Hogares que no tienen acceso al agua de fuentes como: red pública con medidor propio, con medidor compartido, sin medidor, o de alguna fuente no potable
- **Proporción de personas pertenecientes al área rural (x_8).-** Conjunto de viviendas concentradas o dispersas con población menor o igual a 1.000 habitantes, o entre 1.001 y 2.000 habitantes donde menos del 50 % de la población económicamente activa se dedica a actividades secundarias o terciarias.
- **Proporción de personas pertenecientes a pueblos originarios (x_9).-** Personas censadas que se consideran pertenecientes a un grupo indígena u originario.
- **Olas de calor y precipitaciones (x_{10}).-** Indica las áreas que han sido más afectadas por los cambios de temperatura y precipitación.

La variable dependiente es la proporción de pobreza por ingreso que será modelada con 10 variables explicativas anteriormente definidas. Debido a que la variable dependiente es una proporción, no puede ser modelada con un modelo de regresión clásico, ya que y esta acotado entre 0 y 1. Por lo tanto, se aplicará el modelo de regresión beta.

De esta manera, se analiza el conjunto de datos comenzando con las variables explicativas, a continuación se muestra la matriz de correlaciones de la matriz de diseño X:

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x_1	1	-0,06	0,19	-0,05	0	-0,05	-0,23	-0,26	-0,46	0,46
x_2	-0,06	1	-0,33	-0,81	-0,8	-0,53	0,59	0,67	0,28	0,1
x_3	0,19	-0,33	1	0,21	0,3	0,34	-0,35	-0,51	-0,34	0,17
x_4	-0,05	-0,81	0,21	1	0,86	0,69	-0,5	-0,67	-0,13	-0,23
x_5	0	-0,8	0,3	0,86	1	0,8	-0,52	-0,75	-0,31	0,01
x_6	-0,05	-0,53	0,34	0,69	0,8	1	-0,39	-0,58	-0,17	0,07
x_7	-0,23	0,59	-0,35	-0,5	-0,52	-0,39	1	0,65	0,46	-0,07
x_8	-0,26	0,67	-0,51	-0,67	-0,75	-0,58	0,65	1	0,43	-0,03
x_9	-0,46	0,28	-0,34	-0,13	-0,31	-0,17	0,46	0,43	1	-0,41
x_{10}	0,46	0,1	0,17	-0,23	0,01	0,07	-0,07	-0,03	-0,41	1

Observe que existen correlaciones entre algunas de las variables explicativas siendo la más alta entre x_4 y x_5 $(0,86)$, el porcentaje de personas que trabajan en el sector terciario y los años de estudio.

Asimismo, para evidenciar si existe colinealidad entre las variables explicativas, y no necesariamente entre un par de variables, se recurre al estadístico número de condición,

$$h(X) = \sqrt{\frac{\lambda_{\text{máx}}}{\lambda_{\text{mín}}}},$$

donde $\lambda_{\text{máx}}$ y $\lambda_{\text{mín}}$ es el valor propio más grande y más pequeño de la matriz $X^{tp}X$ respectivamente. Si el número de condición está entre 20 y 30, se tiene un problema grave de multicolinealidad, si está por encima de 30, entonces el problema de multicolinealidad es manifiestamente grave.

El número de condición para la base de datos en estudio se muestra a continuación:

$$h(X) = \sqrt{\frac{6678,825}{0,02346}} = 504,49,$$

indicando que las variables explicativas del conjunto de datos presentan un grave problema de multicolinealidad, por lo tanto, se recurre a realizar la estimación mediante los métodos ridge y lasso, sin embargo, para fines comparativos, también se realiza la estimación de máxima verosimilitud sin penalizar.

Antes de realizar la estimación de β, se estima el parámetro de contracción k para los métodos ridge y lasso, en el caso del método ridge se toma el estimador $k_{\text{máx}}$, obteniendo un valor de 0,095 y para lasso se halla el estimador mediante validación cruzada obteniendo un valor de 0,01. De esta manera, la tabla (4.12) muestra los resultados de los estimadores.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Tabla 4.12: Estimadores de máxima verosimilitud, ridge y lasso para el conjunto de datos

Parámetro	Descripción	Estimadores		
		emv	*ridge*	*lasso*
β_0	Intercepto	1,15215	0,52465	0,1935
		(0,434)	(0,307)	(0,154)
β_1	Tasa de plantación	-0,09043	-0,06962	-0,0729
		(-1,883)	(-1,401)	(-1,738)
β_2	Sector primario	-3,20691	-2,10997	-1,27639
		(-2,056)	(-2,671)	(-1,321)
β_3	Sector secundario	-6,07222	-3,19347	-2,28458
		(-0,668)	(-1,107)	(-0,529)
β_4	Sector terciario	-2,16916	-1,75968	-1,55845
		(-1,133)	(-1,251)	(-1,214)
β_5	Años de estudio	-0,20148	-0,16641	-0,14365
		(-12,657)	(-10,386)	(-8,425)
β_6	*ICE*	0,08063	0,04203	0,01109
		(4,473)	(2,271)	(0,578)
β_7	Proporción sin agua	0,447	0,39838	0,33885
		(8,425)	(8,062)	(7,23)
β_8	Proporción área rural	0,01172	0,13141	0,12213
		(0,094)	(1,121)	(1,213)
β_9	Proporción pueblos	0,66034	0,6799	0,68445
		(13,441)	(13,988)	(13,63)
β_{10}	Olas y precipitación	-0,03233	-0,11742	-0,13427
		(-0,195)	(-0,756)	(-1,064)
ϕ		75,70416	75,70416	75,70416
ECM		15,713	7,247	8,23

Los valores en paréntesis son los valores t hallados con el método de remuestreo o bootstrap

Estos resultados fueron obtenidos mediante la función de enlace logit, por lo cual, de la ecuación (2.5) se tiene:

$$\mu_i = \frac{1}{1+\exp(-X_i^{tp}\tilde{\beta})} \ , \ i = 1, \dots 87,$$

donde

$$X^{tp}\tilde{\beta} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{\beta}_3 x_3 + \tilde{\beta}_4 x_4 + \tilde{\beta}_5 x_5 + \tilde{\beta}_6 x_6 + \tilde{\beta}_7 x_7 + \tilde{\beta}_8 x_8 + \tilde{\beta}_9 x_9 + \tilde{\beta}_{10} x_{10},$$

con $\tilde{\beta}$ un estimador de β. Entonces, si un estimador $\tilde{\beta}_i$ es positivo, un incremento en la variable explicativa x_i supone un incremento en μ; en el caso de que β_i sea negativo, un incremento en la variable explicativa x_i, supondría un decremento en μ.

De esta manera, tomando los estimadores de máxima verosimilitud, los coeficientes más relevantes en el modelo se interpretan a continuación: las variables x_2, x_3 y x_4 corresponden a las proporciones de personas que trabajan en el sector primario, secundario y terciario respectivamente; en los tres casos, sus coeficientes tienen signo negativo, por lo

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

que, un incremento en la proporción de personas que trabajan en alguno de estos sectores económicos supone una disminución en la proporción de pobreza por ingreso (y), sin embargo, el sector secundario tiene un mayor efecto en la disminución de la proporción de pobreza, seguido por el sector primario.

Por otro lado, un incremento en la proporción de personas sin agua (x_7) representa un aumento en la proporción de pobreza por ingreso. Finalmente, las variables proporción de personas que pertenecen al área rural (x_8) y proporción de personas que pertenecen a pueblos originarios (x_9) tienen coeficientes positivos, lo que significa que un incremento en estas variables causa un incremento en la proporción de pobreza por ingreso.

Realizando una comparación con los estimadores ridge y lasso, iniciando con las variables porcentaje de personas que trabajan en los sectores primario, secundario y terciario, se observa que sus coeficientes disminuyen, sin embargo, cabe notar que ahora el sector terciario presenta mayor importancia, dado que su estimador disminuye en menor proporción a los otros dos sectores, inclusive en el estimador lasso, supera al sector primario, que era el sector que más afectaba a la proporción de pobreza después del sector secundario.

En cuanto a la variable porcentaje de personas pertenecientes al área rural, en los modelos ridge y lasso, se observa que su coeficiente aumenta, es decir, que el incremento de la proporción de personas en el área rural dentro de una comuna incide en mayor magnitud a la pobreza en comparación al estimador de máxima verosimilitud.

Los valores t de los estimadores se muestran entre paréntesis en la tabla (4.12) ubicados debajo de cada coeficiente para ver su significancia. El valor t se calcula mediante el cociente entre el estimador y la desviación estándar, sin embargo, debido a que los estimadores estudiados son sesgados, la desviación estándar es sustituida por la raíz cuadrada del coeficiente de variación. Estos estadísticos son calculados mediante la técnica de remuestreo o bootstrap con 10000 repeticiones.

Tomando un nivel de confianza del 95 %, el estadístico t de tablas es igual a 1,99, entonces los valores t que superan en valor absoluto a 1,99 serán significativos. Por lo tanto, los coeficientes de las variables explicativas x_1, x_3, x_4, x_8 y x_{10} hallados con los métodos *emv*, ridge y lasso no son significativos ya que sus valores t en valor absoluto no superan al estadístico t de tablas.

Por otro lado, se muestra el error cuadrático medio total de los estimadores emv, ridge y lasso, que al igual que los valores t son calculados con el método de remuestreo. Se observa que los estimadores ridge y lasso presentan un menor error cuadrático medio total en comparación al estimador emv.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Capítulo 5

Conclusiones y trabajos futuros

5.1. Conclusiones

Se estimaron los parámetros del modelo de regresión beta mediante el método de máxima verosimilitud sin penalizar, el cual llega a un sistema de ecuaciones no lineal, que es resuelto con el método de Newton Rapshon, de esta manera, se hallan los estimadores de β y ϕ.

Posteriormente, se halla el estimador ridge $\hat{\beta}_{ridge}$, tomando como estimador de ϕ al estimador de máxima verosimilitud, es decir $\hat{\phi}$. La estimación se realiza desde dos perspectivas: la primera mediante el método de mínimos cuadrados ponderados penalizado por una función cuadrática, y la segunda mediante el método de máxima verosimilitud penalizada, los resultados de estos estimadores muestran que ambos métodos de estimación llegan al mismo resultado.

El último método de estimación estudiado es lasso, el cual recurre a las condiciones de Karush Khun Tucker (A.1) debido a que el modelo es no diferenciable, luego se estiman los coeficientes mediante el método descenso coordinado. Con este mismo método se halla $\hat{\phi}_{lasso}$ por lo cual, los estimadores de ϕ mediante lasso y emv son distintos, asimismo, se recurre al mismo método para hallar el estimador de ϕ en un modelo de regresión beta ridge.

Una vez se obtuvieron los estimadores emv, ridge y lasso, se halló el estimador de contracción k. Para el caso de lasso, se utilizó validación cruzada; y en el caso de ridge, la estimación fue realizada mediante cuatro métodos que fueron elaborados por distintos autores. Mediante simulaciones se comprobó que la estimación ridge para los distintos estimadores de contracción k presentaban un menor error cuadrático medio que el del estimador de máxima verosimilitud. Sin embargo, entre estos cuatro estimadores ridge, el que corresponde al estimador de contracción propuesto por Abonazel & Taha [1] tiene el menor error cuadrático medio, por lo cual se utiliza este estimador de contracción para comparar los resultados de ridge con lasso y emv.

Se realizaron simulaciones de los estimadores emv, ridge y lasso, las conclusiones de este ejercicio son las siguientes:

- Cuando las variables explicativas presentan multicolinealidad, en la mayoría de los casos, el estimador ridge presenta un menor error cuadrático medio, seguido por lasso.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

- Si la variable dependiente del modelo es construida con algún o algunos parámetros $\beta_i = 0$, entonces, el estimador lasso presenta un menor error cuadrático medio.
- Mientras mayor es el tamaño de muestra, menor es el error cuadrático medio.
- Mientras mayor es el parámetro de precisión, menor es el error cuadrático medio.

Por último, el modelo de regresión beta fue aplicado en una base de datos de comunas de tres regiones de Chile cuyas variables explicativas presentan multicolinealidad. Se observa que los estimadores ridge y lasso afectan en los coeficientes, en algunos casos disminuyendo el efecto que tienen algunas variables y en otros casos incrementando su efecto, es el caso de la variable proporción de personas del área rural, donde el estimador *emv* le da un peso relativamente pequeño, pero los estimadores ridge y lasso agrandan el efecto de esta variable sobre la proporción de pobreza. Por otro lado, con la técnica de remuestreo se observa que los estimadores ridge y lasso tienen un menor error cuadrático medio y presentan menor varianza que el estimador *emv*.

Por lo tanto, si en un modelo la variable dependiente está acotada y sus variables explicativas presentan multicolinealidad, los estimadores beta ridge y lasso son mejores que los estimadores *emv* tomando como medida de rendimiento al error cuadrático medio, es más, aplicando la técnica de remuestreo, se llega a la misma conclusión.

5.2. Trabajos futuros

En este proyecto se analizaron los estimadores ridge y lasso, métodos de contracción que tienen el objetivo de disminuir el error cuadrático medio de los estimadores cuando existe multicolinealidad en las variables explicativas, sin embargo, hay una gran variedad de estimadores de contracción que pueden ser estudiados.

Para comenzar, se sugiere analizar el estimador elastic net para un modelo de regresión beta, el cual realiza una combinación ponderada de regularización de los estimadores ridge y lasso. Para ello, se añade otro parámetro que determina cuanto peso le corresponde a la penalización del estimador lasso y cuanto a la penalización de ridge.

El siguiente estimador que se investigará, es el estimador beta bridge, que generaliza los estimadores ridge y lasso, en este caso, toma importancia el orden de la penalización del estimador, lasso tendría un orden 1 de penalización (norma L1), y ridge tomaría el orden 2 (norma L2), sin embargo, las penalizaciones pueden realizarse en un orden q (norma Lq). Este estimador ha sido estudiado para un modelo de regresión clásico, donde demuestra ser una opción sólida en diversas circunstancias en comparación con otros estimadores de contracción (ridge, lasso y elastic net).

Por último, se sugiere hallar los estimadores ridge y lasso, para un modelo que tiene una distribución beta generalizada, es decir, que la variable dependiente y esté acotada por un intervalo (a, b) en lugar del intervalo $(0, 1)$.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Apéndice A

Principios fundamentales

Para realizar la estimación de los parámetros del modelo mediante los métodos ridge y lasso, se hace uso de métodos numéricos ya que las ecuaciones a resolver no son lineales, entre los métodos a utilizar, están Newton Raphson, Fisher-Scoring y descenso coordinado, los cuales se explicarán en esta sección.

Asimismo, debido a que la función a optimizar, en la estimación de los parámetros del modelo mediante el método lasso, depende del valor absoluto de los estimadores, entonces no se puede aplicar la metodología para la estimación como en el método ridge, ya que la función valor absoluto no es diferenciable, por tal motivo, se extienden algunos cálculos básicos al caso de funciones no diferenciables.

A.1. Conceptos básicos para la estimación lasso

Operador umbral suave.

El operador umbral suave o soft-thresholding está definido como sigue:

$$\mathbb{S}_k(x) = \begin{cases} x+k & \text{si } x < -k \\ 0 & \text{si } |x| \leq k \\ x-k & \text{si } x > k \end{cases},$$

el cual se grafica en la figura A.1, donde la línea roja es el operador soft-thresholding para un $k = 1$.

Conjunto convexo.

Un conjunto $U \subset \mathbb{R}$ es convexo si $\forall x, y \in U$ y $\lambda \in [0,1]$ entonces $\lambda x + (1-\lambda)y \in U$.

Función convexa.

Sea $U \in \mathbb{R}^p$ y $f : U \to (-\infty, \infty]$ una función, entonces, f es una función convexa si su dominio es un conjunto convexo y se cumple que $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$, $\forall x, y, \in Dom(f)$ y $\lambda \in [0,1]$ con $Dom(f)$ el dominio de f.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

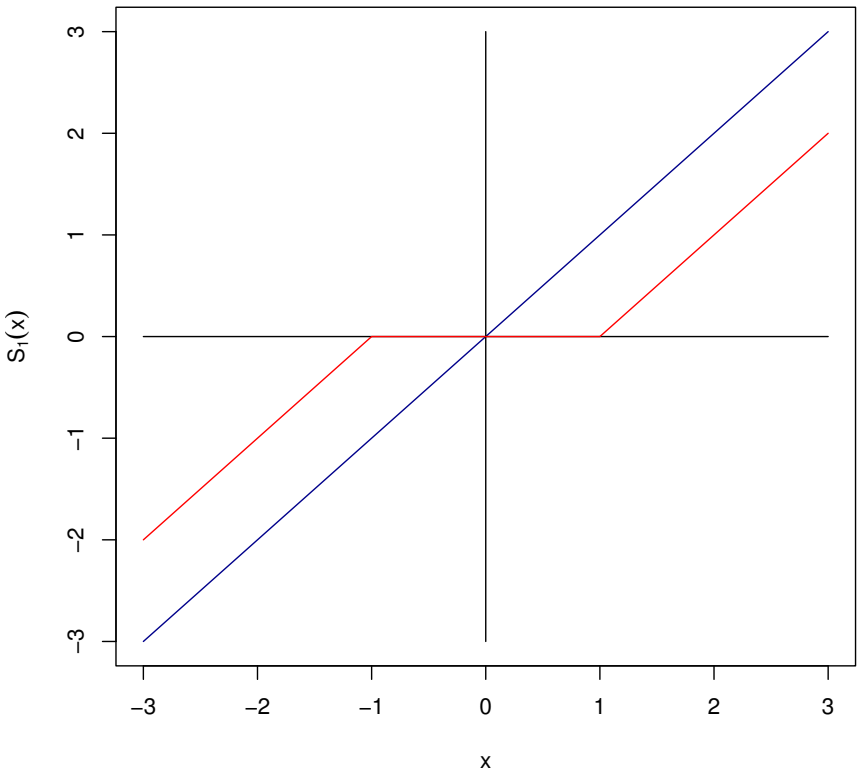


Figura A.1: Operador soft-thresholding

Sub-gradiente.

Sea f una función convexa, no necesariamente diferenciable, $g \in \mathbb{R}^p$ es el sub-gradiente de f si $f(y) \geq f(x) + g^{tp}(y - x)$, $\forall y$.

La función valor absoluto $f(x) = |x|$ no es diferenciable en el punto $x = 0$. Sin embargo, si tiene sub-gradiente, si $x > 0$, el sub-gradiente es $g = 1$, si $x < 0$ entonces el sub-gradiente es $g = -1$, pero si $x = 0$ entonces el sub-gradiente es cualquier elemento del conjunto $[-1, 1]$.

Sub-diferencial.

La sub-diferencial de la función f en el punto x $(\partial f(x))$ es el conjunto de todos los sub-gradientes de la función convexa f en x

$$\partial f(x) = \{g \in \mathbb{R}^p : g \text{ es un sub-gradiente de } f \text{ en } x\}$$

La sub-diferencial de la función valor absoluto es la función signo $\partial|x| = \text{signo}(x)$, donde la función signo está dada por:

$$\text{signo}(x) = \begin{cases} -1 & \text{si } x < 0 \\ [-1, 1] & \text{si } x = 0 \\ 1 & \text{si } x > 0 \end{cases}$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

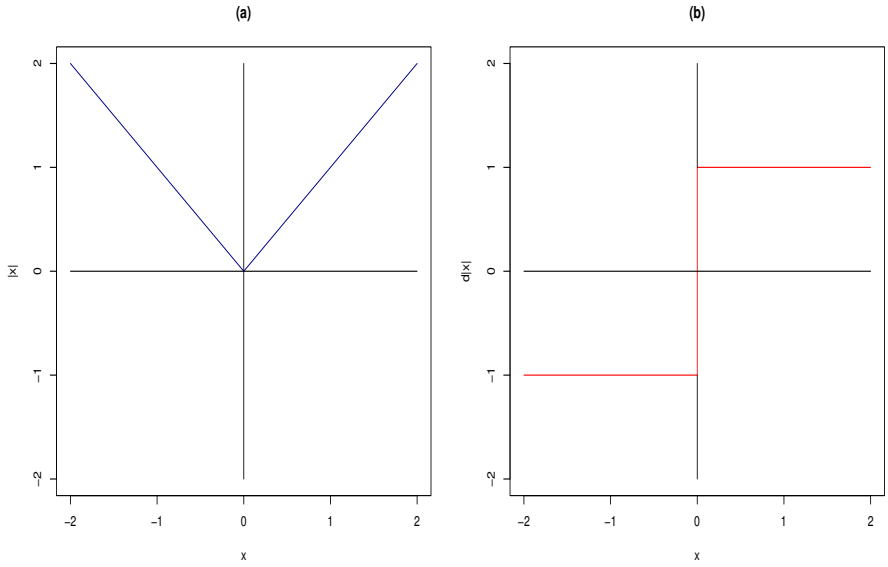


Figura A.2: Función valor absoluto y su derivada

En la figura A.2.a se observa la gráfica del valor absoluto, tomando a las pendientes como las derivadas de la función, se observa que para $x > 0$, la pendiente de la línea es 1, es decir que la derivada de $|x|$ para $x > 0$ es 1, si $x < 0$ entonces la pendiente y por tanto la derivada es -1, sin embargo, se pueden graficar varias pendientes para $x = 0$, como se observa en el gráfico, las pendientes para este punto pueden ir desde -1 a 1. Este mismo resultado puede ser visto en la figura A.2.b, la derivada para el punto $x = 0$ esta entre $[-1, 1]$.

Problemas de optimización convexa.

Un problema de optimización convexa tiene la siguiente forma:

$$\begin{aligned} \text{mín } f(x) \text{ sujeto a} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, r \\ & h_i(x) = 0, \quad i = 1, \ldots, s, \end{aligned} \tag{A.1}$$

donde x es la variable de optimización, f y g_i son funciones convexas y h_i son funciones del tipo $Ax + b$.

Dualidad de lagrange.

Dado un problema de optimización convexa de la forma de la ecuación (A.2), el lagrangiano del problema de optimización está definido como la función objetivo aumentado con una suma ponderada de las restricciones.

$$L(x, u, v) = f(x) + \sum_{i=1}^{r} u_i g_i(x) + \sum_{i=1}^{s} v_i h_i(x) \tag{A.2}$$

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

donde $u_i \leq 0 \forall i$ y los vectores $u_i \in \mathbb{R}^r$ y $v_i \in \mathbb{R}^s$ son llamadas variables duales.

Condiciones de Karush-Kuhn-Tucker.

Sea $f(x)$ definido como un problema de optimización primal, $L(x, u, v)$ su lagrangiano, x^* una solución óptima primal y (u^*, v^*) una solución óptima dual entonces las condiciones de Karush-Kuhn-Tucker (KKT) están definidos por:

- Condición de estacionariedad: el sub-diferencial de $L(x, u, v)$ en el punto (x^*, u^*, v^*) debe contener el punto 0.

$$0 \in \partial f(x^*) + \sum_{i=1}^{r} u_i^* \partial g_i(x^*) + \sum_{i=1}^{s} v_i^* \partial h_i(x^*)$$

- Holgura complementaria

$$u_i^* g_i(x^*) = 0, \forall i = 1, \ldots, r$$

- Condición factible primal

$$\begin{aligned} g_i(x^*) &\leq 0, \quad \forall i = 1, \ldots, r \\ h_i(x^*) &= 0, \quad \forall i = 1, \ldots, s \end{aligned}$$

- Condición factible dual

$$u_i^* \geq 0, \forall i = 1, \ldots, r$$

A.2. Descenso coordinado.

El descenso coordinado es un proceso iterativo, que optimiza la función objetivo respecto a un solo parámetro a la vez suponiendo que los demás son constantes, recorriendo de esta manera por todos los parámetros, este procedimiento se irá repitiendo hasta alcanzar la convergencia.

Para aplicar este procedimiento, considere la función objetivo f convexa que puede fragmentarse en dos componentes, de las cuales una es no diferenciable, de esta manera:

$$f(\beta) = g(\beta) + \sum_{i=j}^{p} h_j(\beta_j)$$

donde g es una función diferenciable y convexa; y h es una función convexa no necesariamente diferenciable. Este algoritmo es bastante eficiente, ya que facilita su cálculo convirtiendo el problema no lineal a uno lineal.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

A.3. Método lasso para un modelo de regresión clásico

El procedimiento para hallar los estimadores de regresión lasso en un modelo beta es un tanto complejo, por lo cual, en la presente sección, se hallan los estimadores lasso considerando una sola variable explicativa y tomando el modelo de regresión clásico ya que ambos procedimientos son bastante similares.

Entonces, el método lasso para un modelo de regresión clásico puede ser escrito como un problema de optimización convexa de la siguiente manera:

$$\min_{\beta\in\mathbb{R}^p}\left(\frac{1}{2n}\|y-X\beta\|^2\right) \text{ sujeto a } \|\beta\|_1 \leq t, \tag{A.3}$$

donde t es el equivalente al parámetro de contracción. Ilustrando primero para una sola variable explicativa x_1, se tiene que el problema de optimización en términos de lagrange esta dado por:

$$\min\left(\frac{1}{2n}\|y-x_1\beta_1\|^2+k|\beta_1|\right).$$

Sea $\tilde{\beta}_1$ una solución al problema de optimización descrito, entonces de las condiciones de KKT (ver la sección A.1 del apéndice) la sub-diferencial debe contener el 0, entonces:

$$\begin{aligned}\frac{\partial\left(\frac{1}{2n}\|y-x_1\beta_1\|^2\right)}{\partial\beta_1}+k\frac{\partial|\beta_1|}{\partial\beta_1} &= 0,\\ -\tfrac{2}{2n}x_1^{tp}(y-x_1\tilde{\beta}_1)+k\,\text{signo}(\tilde{\beta}_1) &= 0,\\ \tfrac{1}{n}x_1^{tp}(y-x_1\tilde{\beta}_1) &= k\,\text{signo}(\tilde{\beta}_1).\end{aligned}$$

Dado que se está trabajando con una matriz de diseño estandarizada, se tiene que $\frac{1}{n}x_1^{tp}x=1$, entonces

$$\begin{aligned}\tfrac{1}{n}x_1^{tp}y-\tfrac{1}{n}x_1^{tp}x_1\tilde{\beta}_1 &= k\,\text{signo}(\tilde{\beta}_1),\\ \tilde{\beta}_1 &= \tfrac{1}{n}x_1^{tp}y-k\,\text{signo}(\tilde{\beta}_1),\end{aligned}$$

$$\tilde{\beta}_1 = \begin{cases}\frac{1}{n}x_1^{tp}y+k & \text{si } \frac{1}{n}x_1^{tp}y<-k\\ 0 & \text{si } \frac{1}{n}x_1^{tp}y\leq k\\ \frac{1}{n}x_1^{tp}y-k & \text{si } \frac{1}{n}x_1^{tp}y>k\end{cases}$$

el cual puede ser escrito como un operador soft-thresholding (ver la sección A.1 del apéndice), de esta manera:

$$\tilde{\beta}_1=\mathbb{S}_k\left(\frac{x_1^{tp}y}{n}\right),$$

es el estimador de β_1.

A.4. Método de Newton-Raphson

El método de Newton se desarrolló originalmente para encontrar raíces de ecuaciones no lineales, que consiste en realizar una aproximación cuadrática de la ecuación de manera iterativa.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Este método se basa en la expansión de Taylor para una función ($l(\theta)$) cuya raíz se quiere calcular partiendo de un punto lo suficientemente cercano a la raíz ($\theta^{(t)}$), entonces

$$l(\theta) = l(\theta^{(t)}) + dl(\theta^{(t)}) \left(\theta - \theta^{(t)}\right) + \frac{1}{2} \left(\theta - \theta^{(t)}\right)^{tp} d^2 l(\theta^{(t)})(\theta - \theta^{(t)})$$

y luego maximizar la aproximación cuadrática.

Ahora, suponga que $dl(\theta^{(t)})$ existe y que la diferencia $\theta - \theta^{(t)}$ es muy pequeña, otra manera de escribir el método de Newton-Raphson es despreciando el tercer término del anterior desarrollo, por lo que se tiene,

$$l(\theta) = l(\theta^{(t)}) + dl(\theta^{(t)}) \left(\theta - \theta^{(t)}\right).$$

A.5. Método de Fisher-Scoring

El método de Fisher-Scoring es un caso particular del método de Newton Raphson, donde se sustituye la derivada $d^2 l(\theta^{(t)})$ por el negativo de su valor esperado que estará denotado por $\mathfrak{J} = -E\left[d^2 l(\theta^{(t)})\right]$, denominada como la matriz de información de fisher.

Las iteraciones para este método numérico están definidas por:

$$\theta = \theta^{(t)} + \mathfrak{J}^{-1}(\theta^{(t)}) U(\theta^{(t)}).$$

donde $U(\theta^{(t)})$ es la función de score.

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Apéndice B

Códigos en R

Los códigos realizados en R se dividen en 3 partes, la primera son las funciones creadas que son utilizadas para generar las variables aleatorias y para la estimación de los parámetros; la segunda son los códigos con los que se crearon las tablas del capítulo de simulación; y la tercera es el código con el que se realizó la aplicación.

B.1. Funciones utilizadas

B.1.1. Simulación de las variables aleatorias

genera_x es una función que genera p variables explicativas con n observaciones los cuales presentan una correlación que está definida por el vector a, estos valores son generados a partir de una semilla. De esta manera, la salida que arroja esta función es una matriz de $n \times p$.

```
genera_x <- function(n,p,a,semilla=1001){
set.seed(1001)
  z <- matrix(runif(n*(p+1)),n,p+1)
  x <- matrix(NA, n,p)
  for (t in 1:n) {
  for (j in 1:p) {
      x[t,j] <- sqrt(1-a[j]^2) * z[t,j] + a[j] * z[t,(p+1)]
    }
  }
  x
}
```

En cuanto a la función ***genera_Y*** genera una matriz de $n \times S$, donde n es el número de observaciones y S es la cantidad de simulaciones que se desean simular, X es la matriz de diseño, b son los parámetros del modelo y phi es el parámetro de precisión del error que afecta a la función de enlace.

```
genera_Y <- function(S, n, X, b, phi){
  Y <- matrix(NA, S,n)
  mu <- exp(X %*% b)/(1+exp(X %*% b))
  a <- mu*phi
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
  b <- (1-mu)*phi
  for (i in 1:S){
  set.seed(i)
    Y[i,] <- rbeta(n,a,b)
  }
  Y
}
```

B.1.2. Código para el modelo de regresión beta

La función ***reg_beta*** es el código creado para el modelo de regresión beta sin penalizaciones hallado mediante el método de máxima verosimilitud con ayuda del método fisher-scoring, cuyos valores de entrada son la matriz de diseño X sin tomar en cuenta la columna del intercepto, y la variable dependiente y. La salida que arroja esta función, es una lista en donde se tiene los estimadores de los parámetros β y ϕ, los residuales de los estimadores, los valores ajustados, el número de observaciones y la varianza del modelo.

```
reg_beta <- function(x,y,k_cont=0){
  nobs <- NROW(x); p <- NCOL(x)
  xx <- as.matrix(cbind(1,x))
  control <- list(phi=T,method = "BFGS",maxit = 5000,hessian = F,
                  trace = F, start = NULL, fsmaxit = 200,
                  fstol = 1e-08,fnscale = -1, reltol = 9.02333e-14)
  control$phi <- control$method <- control$hessian <- NULL
  control$start <- control$fsmaxit <- control$fstol <- NULL
  z <- matrix(1, ncol = 1, nrow = nobs)
  colnames(z) <- "(Intercepto)"; rownames(z) <- rownames(x)

  fun_enlac <- function(mu) log(mu/(1-mu))
  inv_enlac <- function(y) exp(y)/(1+exp(y))
  dif_mu.eta <- function(eta) (exp(eta))/(1+exp(eta))^2
  phi_fun_enl <- function(x) x
  phi_fun_inv <- function(x) x
  phi_mu.eta <- function(x) 1
  ystar <- qlogis(y)
  inicio <- NULL

  auxreg <- lm.fit(xx, fun_enlac(y))
  beta <- auxreg$coefficients
  yhat <- inv_enlac(auxreg$fitted.values)
  der_enlac <- 1/dif_mu.eta(fun_enlac(yhat))
  res <- auxreg$residuals
  sigma2 <- sum(res^2)/((nobs - p) * (der_enlac)^2)
  phi_y <- yhat * (1 - yhat)/(nobs * sigma2) - 1/nobs
  phi <- sum(phi_y)
  phi <- rep.int(0, 1)
  phi[1L] <- suppressWarnings(phi_fun_enl(sum(phi_y)))
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
if (!isTRUE(phi_fun_inv(phi[1L]) > 0)) {
  warning("no se encontró un valor de inicio válido para phi
          se usa 1 en su lugar")
  phi[1L] <- 1
}
inicio <- list(media = beta, precision = phi)
inicio <- do.call("c", inicio)
ajuste <- function(par, deriv = 0L) {
  beta <- par[seq.int(length.out = p+1)]
  gamma <- par[seq.int(length.out = 1) + (p+1)]
  eta <- as.vector(xx %*% beta)
  phi_eta <- as.vector(z %*% gamma)
  mu <- inv_enlac(eta)
  phi <- phi_fun_inv(phi_eta)
  mustar <- if (deriv >= 1L)
  digamma(mu * phi) - digamma((1 - mu) * phi) else NULL
  psi1 <- if (deriv >= 2L)
  trigamma(mu * phi) else NULL
  psi2 <- if (deriv >= 2L)
  trigamma((1 - mu) * phi) else NULL
  list(beta = beta, gamma = gamma, eta = eta, phi_eta = phi_eta,
  mu = mu, phi = phi, mustar = mustar, psi1 = psi1,
  psi2 = psi2)
}
log_MV <- function(par, fit = NULL) {
  fit <- ajuste(par)
  alpha <- fit$mu * fit$phi
  beta <- (1 - fit$mu) * fit$phi
  ll <- suppressWarnings(dbeta(y, alpha, beta, log = TRUE))
  if (any(!is.finite(ll)))
  NaN else sum(ll)
}
gradfun <- function(par, sum = TRUE, fit = NULL) {
  fit <- ajuste(par, deriv = 1L)
  mu <- fit$mu
  phi <- fit$phi
  eta <- fit$eta
  phi_eta <- fit$phi_eta
  mustar <- fit$mustar
  rval <- cbind(phi * (ystar - mustar) * dif_mu.eta(eta) * xx,
                (digamma(phi) - digamma((1 - mu) * phi)
         + mu * (ystar - mustar)  + log(1 - y))*phi_mu.eta(phi_eta))
  if (sum)
    colSums(rval)
  else rval
}
hessfun <- function(par, inverse = FALSE, fit = NULL) {
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
  fit <- ajuste(par, deriv = 2L)
  mu <- fit$mu
  phi <- fit$phi
  eta <- fit$eta
  phi_eta <- fit$phi_eta
  mustar <- fit$mustar
  psi1 <- fit$psi1
  psi2 <- fit$psi2
  a <- psi1 + psi2
  wpp <- (psi1*mu^2 + psi2*(1 - mu)^2 - trigamma(phi))*phi_mu.eta(phi_eta)^2
  wbb <- phi^2 * a * dif_mu.eta(eta)^2
  wbp <- phi * (mu * a - psi2) * dif_mu.eta(eta) * phi_mu.eta(phi_eta)
  kbb <- if (p > 0L)
  crossprod(sqrt(wbb) * xx) else crossprod(xx)
  kpp <- if (1 > 0L)
  crossprod(sqrt(wpp)) else crossprod(z)
  kbp <- if (p > 0L & 1 > 0L)
  crossprod(wbp * xx, z) else crossprod(xx, z)
  K <- cbind(rbind(kbb, t(kbp)), rbind(kbp, kpp))
  if (!inverse)
    K else chol2inv(chol(K))
}
opt <- optim(par = inicio, fn = log_MV, gr = gradfun, method = "BFGS",
             hessian = F, control = control)
par <- opt$par

umbral = 1e-10; max_iter = 100
dif <- 10000; conteo <- 0
while(dif > umbral ){
  fit <- ajuste(par, deriv = 2L)
  scores <- gradfun(par, fit = fit)
  InfoInv <- try(hessfun(par, fit = fit, inverse = TRUE))
  par <- par + (InfoInv %*% scores)
  dif <- sum((InfoInv %*% scores)^2)
  conteo <- conteo + 1
  if(conteo > max_iter) {
    stop("El modelo no tiene estimadores convergentes.")
  }
}

fit <- ajuste(par, deriv = 2L)
beta <- fit$beta
gamma <- fit$gamma
mu <- fit$mu
names(beta) <- colnames(xx)
names(gamma) <- colnames(z)
yhat <- inv_enlac(structure(mu, .Names = names(y)))
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
  der_enlac <- 1/dif_mu.eta(fun_enlac(yhat))
  res <- y - mu
  sigma2 <- sum(sum(res^2)/((nobs - p) * (der_enlac)^2))
  rval <- list(coeficientes = list(media = beta, precision = gamma),
    residuales = y - mu, valor_ajustado = structure(mu, .Names = names(y)),
    n = nobs, gl.residual = nobs - p - 1, sigma2 = sigma2)
  rval
}
```

B.1.3. Código para el modelo de regresión beta ridge

La función ***beta_ridge*** realiza la estimación del modelo de regresión beta ridge mediante el método descenso coordinado, los valores de entrada son la matriz de diseño X sin tomar en cuenta la columna del intercepto, la variable dependiente y y el parámetro de contracción k. La salida de la función son los estimadores de los parámetros β y ϕ.

```
beta_ridge <- function(x, y, k_cont, difer = 0.0001, max_iter = 200){
  x <- as.matrix(x)
  xx <- as.matrix(cbind(1,x))
  p <- dim(x)[2]
  mod <- reg_beta(x,y)$coeficiente
  init_beta <- c(mod$media,mod$precision)
  iter <-1
  n <- nobs <- NROW(x)
  fun_enlac <- function(mu) log(mu/(1-mu))
  inv_enlac <- function(y) exp(y)/(1+exp(y))
  r_mu.eta <- function(eta) (exp(eta))/(1+exp(eta))^2
  while(iter < max_iter){
    beta_curr <- init_beta
    for (i in 1:(length(beta_curr)-1)) {
      beta <- init_beta[seq.int(length.out = (p+1))]
      gamma <- init_beta[seq.int(length.out = 1) + (p+1)]
      ml <- function(betai){
        beta[i] <- betai
        eta <- as.vector(xx %*% beta)
        phi <- as.vector(gamma)
        mu <- inv_enlac(eta)
        li1 <- log(1/beta(mu*phi,(1-mu)*phi))
        li2 <- (mu*phi-1)*log(y)+((1-mu)*phi-1)*log(1-y)
        -sum(li1+li2)+k_cont*sum(beta[-1]^2)
      }
      init_beta[i] <- optim(par = beta[i],fn = ml)$par
    }
    beta <- init_beta[seq.int(length.out = (p+1))]
    gamma <- init_beta[seq.int(length.out = 1) + (p+1)]
    ml2 <- function(phi){
      eta <- as.vector(xx %*% beta)
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
      mu <- inv_enlac(eta)
      li1 <- log(1/beta(mu*phi,(1-mu)*phi))
      li2 <- (mu*phi-1)*log(y)+((1-mu)*phi-1)*log(1-y)
      -sum(li1+li2)+k_cont*sum(beta[-1]^2)
    }
    init_beta[p+2] <- optim(par = as.vector(gamma),fn = ml2)$par
    iter <- iter +1
    if(sum(abs(init_beta - beta_curr)) < difer){
      iter <- max_iter
    }
  }
  return(init_beta)
}
```

B.1.4. Código para el modelo de regresión lasso

La función ***beta_lasso*** halla los estimadores del modelo de regresión beta lasso hallado mediante la minimización del negativo de la log-verosimilitud penalizada, para esto se utiliza la ecuación (3.15), posteriormente se busca la convergencia de los estimadores con el método descenso coordinado. Los valores de entrada son la matriz de diseño X sin tomar en cuenta la columna del intercepto, la variable dependiente y y el parámetro de contracción k. La salida de la función son los estimadores de los parámetros β y ϕ.

```
beta_lasso <- function(x, y, k_cont, difer = 0.0001, max_iter = 200){
  x <- as.matrix(x)
  xx <- as.matrix(cbind(1,x))
  p <- dim(x)[2]
  n <- dim(x)[1]
  iter <-1
  ystar <- qlogis(y)
  fun_enlac <- function(mu) log(mu/(1-mu))
  inv_enlac <- function(y) exp(y)/(1+exp(y))
  dif_mu.eta <- function(eta) (exp(eta))/(1+exp(eta))^2
  phi_fun_enl <- function(x) log(x)
  phi_fun_inv <- function(x) exp(x)
  ajuste <- function(par) {
    beta <- par[1:(p+1)]
    gamma <- par[p+2]
    eta <- as.vector(xx %*% beta)
    phi <- as.vector(1 %*% gamma)
    mu <- inv_enlac(eta)
    mustar <- digamma(mu * phi) - digamma((1 - mu) * phi)
    psi1 <- trigamma(mu * phi)
    psi2 <- trigamma((1 - mu) * phi)
    list(beta = beta, gamma = gamma, eta = eta,
    mu = mu, phi = phi, mustar = mustar, psi1 = psi1,
    psi2 = psi2)
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
}
gradfun <- function(par) {
  fit <- ajuste(par)
  mu <- fit$mu
  phi <- fit$phi
  eta <- fit$eta
  mustar <- fit$mustar
  rval <- phi * (ystar - mustar) * dif_mu.eta(eta)
  rval
}
hessfun <- function(par) {
  fit <- ajuste(par)
  mu <- fit$mu
  phi <- fit$phi
  eta <- fit$eta
  mustar <- fit$mustar
  psi1 <- fit$psi1
  psi2 <- fit$psi2
  w <- diag(phi * (psi1 + psi2) * dif_mu.eta(eta)^2)
  -phi * w
}
soft_thresh <- function(a, b) {
  salida <- rep(0, length(a))
  salida[a >  b] <- a[a > b] - b
  salida[a < -b] <- a[a < -b] + b
  salida
}
mod <- reg_beta(x,y)$coeficiente
init_beta <- c(mod$media,mod$precision)

while(iter < max_iter){
  beta_curr <- init_beta
  for (i in 1:(length(beta_curr)-1)) {
    beta <- init_beta[seq.int(length.out = (p+1))]
    phi <- init_beta[seq.int(length.out = 1) + (p+1)]
    deriv1 <- gradfun(init_beta)
    deriv2 <- hessfun(init_beta)
    z <- xx %*% beta - solve(deriv2)%*%deriv1
    xj <- xx[,-i] %*% beta[-i]
    init_beta[i] <- soft_thresh(sum(deriv2 %*% xx[,i] * (z - xj))/n,
                                k_cont)/(sum(deriv2 %*%xx[,i]*xx[,i])/n)
  }
  beta <- init_beta[seq.int(length.out = (p+1))]
  gamma <- init_beta[seq.int(length.out = 1) + (p+1)]
  ml2 <- function(phi){
    eta <- as.vector(xx %*% beta)
    mu <- inv_enlac(eta)
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
      li1 <- log(1/beta(mu*phi,(1-mu)*phi))
      li2 <- (mu*phi-1)*log(y)+((1-mu)*phi-1)*log(1-y)
      -1/n*sum(li1+li2)+k_cont*sum(abs(beta[-1]))
    }
    init_beta[p+2] <- optim(par = as.vector(gamma),fn = ml2)$par
    iter <- iter +1
    if(sum(abs(init_beta - beta_curr)) < difer){
      iter <- max_iter
    }
  }
  return(init_beta)
}
```

B.1.5. Código para calcular el ECM

Por último, la función ***calcula_ECM*** une las funciones anteriores para calcular el error cuadrático medio para un estimador en específico, los valores de entrada son el número de simulaciones S que se desean realizar, el número de observaciones n, el número de parámetros p, el vector que define el grado de correlación existente entre las variables explicativas a, el valor semilla con el que se generarán las variables aleatorias *semilla*, el parámetro de precisión *phi*, *regresion* que es la variable en la que se selecciona que modelo ajustar (ridge, lasso o el estimador beta sin penalizaciones) y finalmente el parámetro de contracción k_{cont}.

```
calcula_ECM <- function(S,n,p,a,semilla,phi, regresion, k_cont,nulo=F){
  x_orig <- genera_x(n=n, p=p, a=a, semilla=semilla)
  X_orig <- cbind(1,x_orig)
  q <- eigen(t(X_orig)%*%X_orig)$vectors
  b <- colSums(q)/sqrt(p)
  if(nulo == T) {b[c(p,p+1)] <- 0}
  Y <- genera_Y(S=S, n=n, X=X_orig, b=b, phi = phi)
  x <- as.matrix(scale(x_orig))
  X <- cbind(1,x)
  p <- dim(X)[2]-1
  beta <- matrix(NA, S, p+1)
  phi_est <- rep(NA, S)
  for (i in 1:S) {
    if(regresion == "emv") {
      rr <- reg_beta(x,Y[i,])
      r_NR <- rr$coeficientes
      b_i <- r_NR$media[-1]/apply(x_orig, 2, sd)
      b_0 <- r_NR$media[1] -
      sum(r_NR$media[-1]/apply(x_orig, 2, sd)*apply(x_orig, 2, mean))
      beta[i,] <- c(b_0,b_i)
      phi_est[i] <- r_NR$precision
    }
    if(regresion == "ridge") {
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
      rr <- beta_ridge(x,Y[i,], k_cont)
      b_i <- rr[2:(p+1)]/apply(x_orig, 2, sd)
      b_0 <- rr[1] -
      sum(rr[2:(p+1)]/apply(x_orig, 2, sd)*apply(x_orig, 2, mean))
      beta[i,] <- c(b_0,b_i)
      phi_est[i] <- rr[p+2]
    }
    if(regresion == "lasso") {
      rr <- beta_lasso(x,Y[i,], k_cont)
      b_i <- rr[1:(p)]/apply(x, 2, sd)
      b_0 <- mean(log(Y[i,]/(1-Y[i,]))) -
      sum(rr[1:(p)]/apply(x, 2, sd)*apply(x, 2, mean))
      beta[i,] <- c(b_0,b_i)
      phi_est[i] <- rr[p+2]
    }
  }
  var <- apply(beta, 2, var)
  sesgo <- (apply(beta, 2, mean) - b)^2
  MSE <- rep(0,p+1)
  MSE_phi <- 0
  for (i in 1:S) {
    MSE <- MSE + (beta[i,] - b)^2
    MSE_phi <- MSE_phi + (phi_est[i] - phi)^2
  }
  salida <- list(TMSE = MSE/S,sesgo = sesgo, var  = var,
              MSE_phi = MSE_phi/S,estimador = beta,phi = phi_est,Y = Y,
              X = x_orig, x_stand=x, parametro = b)
  salida
}
```

B.1.6. Código de funciones para estimar k

Función para estimar el parámetro de contracción k mediante validación cruzada.

```
valid_cruz <- function(x,y){
  vc_ecm <- function(x,y,k_cont){
    n <- dim(x)[1]
    m <- trunc(n/10)
    p <- dim(x)[2]
    set.seed(1)
    orden <- sample(c(1:n),n,replace = F)
    dif_mu.eta <- function(eta) (exp(eta))/(1+exp(eta))^2
    ECM <- numeric()
    for (i in 1:9) {
      x_train <- x[-orden[((i-1)*m+1):(i*m)],]
      y_train <- y[-orden[((i-1)*m+1):(i*m)]]
      beta <- beta_lasso(x_train,y_train,k_cont)
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
    b_i <- beta[2:(p+1)]/apply(x_train, 2, sd)
    b_0 <- beta[1] - sum(beta[2:(p+1)]/apply(x_train, 2, sd)
                *apply(x_train, 2, mean))
    beta <- c(b_0,b_i)
    x_test <- x[orden[((i-1)*m+1):(i*m)],]
    y_test <- y[orden[((i-1)*m+1):(i*m)]]
    eta <- cbind(1,x_test) %*% beta
    ECM[i] <- sum((dif_mu.eta(eta)-y_test)^2)/length(eta)
  }
  x_train <- x[-orden[(4*m+1):n],]
  y_train <- y[-orden[(4*m+1):n]]
  beta <- beta_lasso(x_train,y_train,k_cont)[-(p+2)]
  x_test <- x[orden[(4*m+1):n],]
  y_test <- y[orden[(4*m+1):n]]
  eta <- cbind(1,x_test) %*% beta
  ECM[10] <- sum((dif_mu.eta(eta)-y_test)^2)/length(eta)
  mean(ECM)
}
x <- as.matrix(scale(x))
xx <- as.matrix(cbind(1,x))
p <- dim(x)[2]
n <- dim(x)[1]
iter <-1
ystar <- qlogis(y)
fun_enlac <- function(mu) log(mu/(1-mu))
inv_enlac <- function(y) exp(y)/(1+exp(y))
dif_mu.eta <- function(eta) (exp(eta))/(1+exp(eta))^2
phi_fun_enl <- function(x) log(x)
phi_fun_inv <- function(x) exp(x)
ajuste <- function(par) {
  beta <- par[1:(p+1)]
  gamma <- par[p+2]
  eta <- as.vector(xx %*% beta)
  phi <- as.vector(1 %*% gamma)
  mu <- inv_enlac(eta)
  mustar <- digamma(mu * phi) - digamma((1 - mu) * phi)
  psi1 <- trigamma(mu * phi)
  psi2 <- trigamma((1 - mu) * phi)
  list(beta = beta, gamma = gamma, eta = eta,
       mu = mu, phi = phi, mustar = mustar, psi1 = psi1,
       psi2 = psi2)
}
gradfun <- function(par) {
  fit <- ajuste(par)
  mu <- fit$mu
  phi <- fit$phi
  eta <- fit$eta
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
    mustar <- fit$mustar
    rval <- phi * (ystar - mustar) * dif_mu.eta(eta)
    rval
  }
  hessfun <- function(par) {
    fit <- ajuste(par)
    mu <- fit$mu
    phi <- fit$phi
    eta <- fit$eta
    mustar <- fit$mustar
    psi1 <- fit$psi1
    psi2 <- fit$psi2
    w <- diag(phi * (psi1 + psi2) * dif_mu.eta(eta)^2)
    -phi * w
  }
  mod <- reg_beta(x,y)$coeficiente
  init_beta <- c(rep(0,(p+1)),mod$precision)

  beta_curr <- init_beta
  beta <- init_beta[seq.int(length.out = (p+1))]
  phi <- init_beta[seq.int(length.out = 1) + (p+1)]
  deriv1 <- gradfun(init_beta)
  deriv2 <- hessfun(init_beta)
  z <- xx %*% beta - solve(deriv2)%*%deriv1
  k_max <- NA
  for (i in 1:(p+1)) {
    k_max[i] <- sum(diag(deriv2) * xx[,i] * z/n)
  }
  ks <- seq(max(min(k_max),0.01),max(k_max),length.out = 25)
  ecm<- numeric()
  for (i in 1:25) {
    ecm[i] <- vc_ecm(x,y,ks[i])
  }
  pos <- which(ecm == min(ecm))[1]
  ks[pos]
}
```

Función para estimar el parámetro de contracción mediante otros métodos.

```
estim_k <- function(x,y, estim){
  dif_mu.eta <- function(eta) (exp(eta))/(1+exp(eta))^2
  inv_enlac <- function(y) exp(y)/(1+exp(y))
  gen_w <- function(par) {
    beta <- par[seq.int(length.out = p+1)]
    gamma <- par[seq.int(length.out = 1) + (p+1)]
    eta <- as.vector(xx %*% beta)
    mu <- inv_enlac(eta)
    phi <- gamma
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
    psi1 <- trigamma(mu * phi)
    psi2 <- trigamma((1 - mu) * phi)
    w <- diag(phi * (psi1 + psi2) * dif_mu.eta(eta)^2)
    w
  }
  x <- scale(x)
  xx <- cbind(1,x)
  p <- dim(x)[2]
  W <- gen_w(estim)
  xwx <- t(cbind(1,x))%*% W  %*% cbind(1,x)
  valor <- eigen(xwx)$values
  vector <- eigen(xwx)$vectors %*% estim[-(p+2)]
  data.frame(metodo = c("hk","med","max","min"),
             k_cont = c(1/(max(vector)^2*estim[(p+2)]),
             median(sqrt(1/(vector^2*estim[(p+2)]))),
             max(valor)/(max(vector)^2*estim[(p+2)]),
             min(valor)/(min(vector)^2*estim[(p+2)])))
}
```

B.2. Creación de tablas de la sección simulación

El siguiente código genera la tabla (4.1), en cuanto a la tabla (4.2), únicamente se debe cambiar el parámetro de precisión ϕ (**phi**) de 2 a 4.

```
multi <- c(rep("Baja",4),rep("Moderada",4),rep("Alta",4))
muestra <- rep(c(30,50,100,200),3)
A <- data.frame(multi,muestra,emv = NA,hk= NA,med = NA,max = NA,min=NA)

n_muestra <- c(30,50,100,200); S <- 10000; p<- 4; phi<- 2; semilla = 100
corel <- matrix(c(0.12,0.21,0.32,0.22,
                  0.42,0.56,0.48,0.51,
                  0.9,0.8,0.87,0.85),4,3)
for(g in 1:4){
  n <- n_muestra[g]
  for(i in 1:3){
    a <- corr <- corel[,i]
    regresion <- "emv"; k_cont <- 0
    emv <- calcula_ECM(S=S, n = n, p = p, a = corr, semilla = semilla,
            phi = phi, regresion = regresion,k_cont = k_cont)
    x <- as.matrix(emv$X)
    k_ridge <- matrix(NA,S,4)
    for (j in 1:S) {
      y <- as.numeric(emv$Y[j,])
      estim <- c(emv$estimador[j,],emv$phi[j])
      k_cont <- estim_k(x,y,estim)
      k_ridge[j,] <- k_cont$k_cont
    }
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
    k_cont <- apply(k_ridge,2,median)
    regresion <- "ridge"
    emv_ridge_hk <- calcula_ECM(S=S,n=n,p= p,a=corr,semilla=semilla,
                        phi = phi, regresion = regresion,k_cont = k_cont[1])
    emv_ridge_med <- calcula_ECM(S=S, n =n,p=p,a=corr,semilla =semilla,
                        phi = phi, regresion = regresion,k_cont = k_cont[2])
    emv_ridge_max <- calcula_ECM(S=S,n=n,p= p,a=corr,semilla=semilla,
                        phi = phi, regresion = regresion,k_cont = k_cont[3])
    emv_ridge_min <- calcula_ECM(S=S,n=n,p= p,a=corr,semilla=semilla,
                        phi = phi, regresion = regresion,k_cont = k_cont[4])
    A[(i-1)*4+g,3] <- sum(emv$TMSE)
    A[(i-1)*4+g,4] <- sum(emv_ridge_hk$TMSE)
    A[(i-1)*4+g,5] <- sum(emv_ridge_med$TMSE)
    A[(i-1)*4+g,6] <- sum(emv_ridge_max$TMSE)
    A[(i-1)*4+g,7] <- sum(emv_ridge_min$TMSE)
  }
}
```

Las tablas (4.3) a (4.8) pueden ser generadas con el siguiente código, cambiando apropiadamente los valores de n y ϕ (phi).

```
n <- 30; S <- 10000; p <- 4; phi <- 2; semilla = 100
corel <- matrix(c(0.12,0.21,0.32,0.22,
0.42,0.56,0.48,0.51,
0.9,0.8,0.87,0.85),4,3)
A <- matrix(NA,9,(p+3))
for(i in 1:3){
  a <- corr <- corel[,i]
  regresion <- "emv"; k_cont <- 0
  emv <- calcula_ECM(S=S, n = n, p = p, a = corr, semilla = semilla,
                        phi = phi, regresion = regresion,k_cont = k_cont)
  x <- as.matrix(emv$X)
  k_ridge <- matrix(NA,S,4)
  for (j in 1:S) {
    y <- as.numeric(emv$Y[j,])
    estim <- c(emv$estimador[j,],emv$phi[j])
    k_cont <- estim_k(x,y,estim)
    k_ridge[j,] <- k_cont$k_cont
  }
  k_cont <- apply(k_ridge,2,median)[4]
  regresion <- "ridge"
  emv_ridge <- calcula_ECM(S=S, n = n, p = p, a = corr, semilla = semilla,
                        phi = phi, regresion = regresion,k_cont = k_cont)
  k_cont <- valid_cruz(x,emv$Y[1,])
  regresion <- "lasso"
  emv_lasso_dc <- calcula_ECM(S=S,n=n,p=p,a =corr,semilla=semilla,
                        phi = phi, regresion = regresion,k_cont = k_cont)
  A[(i-1)*3+1,] <- c(emv$TMSE,emv$MSE_phi,sum(emv$TMSE))
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
  A[(i-1)*3+2,] <- c(emv_ridge$TMSE,emv_ridge$MSE_phi,sum(emv_ridge$TMSE))
  A[(i-1)*3+3,] <- c(emv_lasso_dc$TMSE,emv_lasso_dc$MSE_phi,
                     sum(emv_lasso_dc$TMSE))
}
```

De igual manera, las tablas (4.9) a (4.11) pueden ser generadas con el siguiente código, cambiando los valores de n.

```
n <- 30; S <- 10; p <- 6; phi <- 4; semilla = 100
corel <- matrix(c(0.2, 0.25, 0.27, 0.30, 0.22, 0.1,
0.52, 0.61, 0.45, 0.49, 0.58, 0.5,
0.86, 0.95, 0.94, 0.88, 0.87, 0.9),p,3)
A <- matrix(NA,9,(p+3))
for(i in 1:3){
  a <- corr <- corel[,i]
  regresion <- "emv"; k_cont <- 0
  emv <- calcula_ECM(S=S, n = n, p = p, a = corr, semilla = semilla,
                     phi = phi, regresion = regresion,k_cont = k_cont,T)
  x <- as.matrix(emv$X)
  k_ridge <- matrix(NA,S,4)
  for (j in 1:S) {
    y <- as.numeric(emv$Y[j,])
    estim <- c(emv$estimador[j,],emv$phi[j])
    k_cont <- estim_k(x,y,estim)
    k_ridge[j,] <- k_cont$k_cont
  }
  k_cont <- apply(k_ridge,2,median)[4]
  regresion <- "ridge"
  emv_ridge <- calcula_ECM(S=S, n = n, p = p, a = corr, semilla = semilla,
                     phi = phi, regresion = regresion,k_cont = k_cont,T)

  k_cont <- valid_cruz(x,emv$Y[1,])
  regresion <- "lasso"
  emv_lasso_dc <- calcula_ECM(S=S,n=n,p = p, a = corr, semilla = semilla,
                     phi = phi, regresion = regresion,k_cont = k_cont,T)
  A[(i-1)*3+1,] <- c(emv$TMSE,emv$MSE_phi,sum(emv$TMSE))
  A[(i-1)*3+2,] <- c(emv_ridge$TMSE,emv_ridge$MSE_phi,sum(emv_ridge$TMSE))
  A[(i-1)*3+3,] <- c(emv_lasso_dc$TMSE,emv_lasso_dc$MSE_phi,
                              sum(emv_lasso_dc$TMSE))
}
```

B.3. Código utilizado para la aplicación

```
library(readxl)
region <- read_excel("Datos regiones.xlsx")
dim(region)
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
y <- region$pobreza_ing
x <- region[,c(-1,-2,-3,-4,-11)]
p <- dim(x)[2]
cor(x)
valores <- eigen(t(as.matrix(x))%*%as.matrix(x))$values
sqrt(max(valores)/min(valores))
xx <- x
xx$escol_jefe_hogar <- scale(xx$escol_jefe_hogar)
# estimación de máxima verosimilitud
mod <- reg_beta(xx,y)
estim <- mod$coeficientes
estim_b <- estim$media
b_i <- estim_b[-1]/apply(x, 2, sd)
estim_b[6] <- b_i[5]
b_0 <- estim_b[1] - (estim_b[-1]/apply(x, 2, sd)*apply(x, 2, mean))[5]
estim_emv <- c(b_0,estim_b[-1],estim$precision)
# estimador ridge
k_ridge <- estim_k(xx,y,estim_emv)
estim_r <- beta_ridge(xx,y,k_ridge[3,2])
estim_b <- estim_r[-(p+2)]
b_i <- estim_b[-1]/apply(x, 2, sd)
estim_b[6] <- b_i[5]
b_0 <- estim_b[1] - (estim_b[-1]/apply(x, 2, sd)*apply(x, 2, mean))[5]
estim_ridge <- c(b_0,estim_b[-1],estim$precision)
# estimador lasso
k_lasso <- valid_cruz(xx,y)
estim_l <- beta_lasso(xx,y,k_lasso)
estim_b <- estim_l[-(p+2)]
b_i <- estim_b[-1]/apply(x, 2, sd)
estim_b[6] <- b_i[5]
b_0 <- estim_b[1] - (estim_b[-1]/apply(x, 2, sd)*apply(x, 2, mean))[5]
estim_lasso <- c(b_0,estim_b[-1],estim$precision)
#  Generación de los resultados
aplic <- data.frame(nombre = names(estim_emv), emv = estim_emv,
                    ridge = estim_ridge, lasso = estim_lasso)
aplic[13:14,1] <- c("AIC", "BIC")
aplic[13:14,2] <- diagnostico(as.matrix(xx),y,c(estim$media,estim$precision))
aplic[13:14,3] <- diagnostico(as.matrix(xx),y,estim_r)
aplic[13:14,4] <- diagnostico(as.matrix(xx),y,estim_l)
aplic
# Lectura y preparación de la base de datos
n <- length(y)
p <- dim(x)[2]
boot <- 3
k_ridge <- k_ridge[3,2]
M_emv <- matrix(NA,boot,(p+2))
M_ridge <- matrix(NA,boot,(p+2))
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
M_lasso <- matrix(NA,boot,(p+2))
for (i in 1:boot) {
set.seed(i)
  remuestra <- sample(1:n,n,replace = T)
  x_boot <- x[c(remuestra),]
  y_boot <- y[c(remuestra)]
  x_boot$escol_jefe_hogar <- as.numeric(scale(x_boot$escol_jefe_hogar))
# estimación de máxima verosimilitud
  mod <- reg_beta(x_boot,y_boot)
  estim <- mod$coeficientes
  estim_b <- estim$media
  b_i <- estim_b[-1]/apply(x, 2, sd)
  estim_b[6] <- b_i[5]
  b_0 <- estim_b[1] - (estim_b[-1]/apply(x, 2, sd)*apply(x, 2, mean))[5]
  M_emv[i,] <- c(b_0,estim_b[-1],estim$precision)
# estimación ridge
  estim_r <- beta_ridge(x_boot,y_boot,k_ridge)
  estim_b <- estim_r[-(p+2)]
  b_i <- estim_b[-1]/apply(x, 2, sd)
  estim_b[6] <- b_i[5]
  b_0 <- estim_b[1] - (estim_b[-1]/apply(x, 2, sd)*apply(x, 2, mean))[5]
  M_ridge[i,] <- c(b_0,estim_b[-1],estim$precision)
# estimación lasso
  estim_l <- beta_lasso(x_boot,y_boot,k_lasso)
  estim_b <- estim_l[-(p+2)]
  b_i <- estim_b[-1]/apply(x, 2, sd)
  estim_b[6] <- b_i[5]
  b_0 <- estim_b[1] - (estim_b[-1]/apply(x, 2, sd)*apply(x, 2, mean))[5]
  M_lasso[i,] <- c(b_0,estim_b[-1],estim$precision)
}
var_emv <- apply(M_emv, 2, var)
sesgo_emv <- (apply(M_emv, 2, mean) - estim_emv)^2
MSE_emv <- rep(0,p+2)
for (i in 1:boot) {
  MSE_emv <- MSE_emv + (M_emv[i,] - estim_emv)^2
}
var_r <- apply(M_ridge, 2, var)
sesgo_r <- (apply(M_ridge, 2, mean) - estim_ridge)^2
MSE_r <- rep(0,p+2)
for (i in 1:boot) {
  MSE_r <- MSE_r + (M_ridge[i,] - estim_ridge)^2
}
var_l <- apply(M_lasso, 2, var)
sesgo_l <- (apply(M_lasso, 2, mean) - estim_lasso)^2
MSE_l <- rep(0,p+2)
for (i in 1:boot) {
  MSE_l <- MSE_l + (M_lasso[i,] - estim_lasso)^2
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile


```
}
sum(MSE_emv[-(p+2)])/boot
sum(MSE_r[-(p+2)])/boot
sum(MSE_l[-(p+2)])/boot
desv_est <- data.frame(nombre = names(estim_emv[-(p+2)]),
               emv = sqrt(var_emv[-(p+2)]), ridge = sqrt(var_r[-(p+2)]),
               lasso = sqrt(var_l[-(p+2)]))
desv_est[12,1] <- c("ECM")
desv_est[12,2] <- sum(MSE_emv[-(p+2)])/boot
desv_est[12,3] <- sum(MSE_r[-(p+2)])/boot
desv_est[12,4] <- sum(MSE_l[-(p+2)])/boot
desv_est
```

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

Bibliografía

[1] Abonazel, M. R. & Taha, I. M. , *Beta ridge regression estimators: simulation and application*. Communications in Statistics - Simulation and Computation, 2021 (pg. 1-13).

[2] Breiman, L., *Better subset selection using the non-negative garotte*. Technometrics, volumen 37, número 4, 1995 (pg. 373-384).

[3] Cribari-Neto, F. & Queiroz, M. P. F., *On testing inference in beta regressions*. Journal of Statistical Computation and Simulation, volumen 84, número 1, 2012 (pg. 186-203).

[4] Dobson, A. J. & Barnett, A. G. , *An Introduction to Generalized Linear Models*. Texts in Statistical Science Series, 3ra edición, 2008.

[5] Ferrari, S. & Cribari-Neto, F., *Beta Regression for Modelling Rates and Proportions*. Journal of Applied Statistics, volumen 31, número 7, 2004 (pg. 799-815).

[6] Firinguetti, L.; Kibria, G. & Araya, R., *Study of partial least squares and ridge regression methods*. Communications in Statistics - Simulation and Computation, volumen 46, número 8, 2017 (pg. 1-14).

[7] Friedman, J.; Hastie, T. & Tibshirani, R., *Regularization Paths for Generalized Linear Models via Coordinate Descent*. Journal of Statistical Software, volumen 33, número 1, 2010(pg. 1-22).

[8] Friedman, J.; Hastie, T.; Hofling, H. & Tibshirani, R., *Pathwise coordinate optimization*. The Annals of Applied Statistics, volumen 1, número 2, 2007 (pg. 302-332).

[9] Fuentes, O. G. A. & Ortiz, M. T. A., *Análisis de incidencia de indicadores comunales en la zona centrosur; mediante métodos de regresión y aglomeración*. Tesis de lic. Ing. Estadística - Universidad del Bío-Bío, 2022.

[10] Gauraha, N., *Introduction to the LASSO - A Convex Optimization Approach for High-dimensional Problems*. Resonance, volumen 23, 2018 (pg. 439-464).

[11] Hoerl, A. E. & Kennard, R. W., *Ridge Regression: Biased Estimation for Nonorthogonal Problems*. Journal technometrics, volumen 42, número 1, 2000 (pg. 80-86).

[12] Hoerl, A. E. & Kennard, R. W., *Ridge Regression: Applications to Nonorthogonal Problems*. Journal technometrics, volumen 12, número 1, 1970 (pg. 69-82).

Universidad del Bío-Bío. Sistema de Bibliotecas - Chile

[13] Inan, D. & Sancar, N., *Particle swarm optimization based ridge logistic estimator.* Communications in Statistics - Simulation and Computation, volumen 49, número 3, 2020 (pg. 669-683).

[14] Le Cessie, & Van Houwelingen, J. C. , *Ridge Estimators in Logistic Regression.* Journal of the Royal Statistical Society. Series C, volumen 41, número 1, 1992 (pg. 191-201).

[15] McDonald, G. C. & Galarneau, D. I., *A monte carlo evaluation of some ridge-type estimators.* Journal of the American Statistical Association, Volumen 70, número 350, 1975 (pg. 407-416).

[16] Muniz, G. & Kibria, B. M. G. , *On some ridge regression estimators: An empirical comparisons.* Communication in statistics-simulation and computation, volumen 38, número 3, 2009 (pg. 621-630).

[17] Noah, S.; Friedman, J.; Hastie, T. & Tibshirani, R., *Regularization paths for cox's proportional hazards model via coordinate descent.* Journal of Statistical Software, volumen 39, número 5, 2011 (pg. 1-13).

[18] Qasim, M.; Månsson, K. & Kibria, G., *On some beta ridge regression estimators: method, simulation and application.* Journal of Statistical Computation and Simulation, volumen 91, número 9, 2021 (pg. 1-14).

[19] Rydlewski, J. P. & Mielczarek, D., *On the maximum likelihood estimator in the generalized beta regression model.* Opuscula Mathematica, volumen 32, número 4, 2012 (pg. 761-774).

[20] Schaefer, R. L.; Roi, L. D. & Wolfe, R.A., *A ridge logistic estimator.* Communications in Statistics - Theory and Methods. Volumen 13, número 1, 1984 (pg. 99-113).

[21] Seber, G. A. F. & Lee, A. J. , *Linear Regression Analysis.* Wiley Series in probability and statistics, 2da edición, 2003.

[22] Tibshirani, R., *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society, volumen 58, número 1, 1996 (pg. 267-288).

[23] Van Wieringen, W. N. , *Lecture notes on ridge regression.* Version 0.31, 2020.

[24] Wenjiang, J. F., *Penalized Regressions: The Bridge Versus the Lasso.* Journal of Computational and Graphical Statistics, volumen 7, número 3, 1998 (pg. 397-416).

[25] Zou, H. & Hastie, T., *Regularization and variable selection via the elastic net.* J. R. Statist, volumen 67, número 2, 2005 (pg. 301-320).