Construcción de Software para Regresión
El caso de Regresión Ridge y Robusta

[(1)] Estefany Melissa Minalla Alava [(2)] Mario David Solórzano Carvajal [(3)] Msc. Gaudencio Zurita Herrera
Instituto de Ciencias Matemáticas
Campus Gustavo Galindo, Km 30.5 vía Perimetral
Apartado 09-01-5863. Guayaquil-Ecuador
Email(s) [(1)]eminalla@espol.edu.ec [(2)]mdsolorz@espol.edu.ec [(3)]gzurita@espol.edu.ec

Resumen

En el contexto de la Regresión Lineal, se han propuesto diversos métodos para afrontar la multicolinealidad. Los principales de ellos constituyen estimadores sesgados de los coeficientes. Entre estos métodos, se encuentran la Regresión Ridge.
En los modelos de Regresión Lineal cuando la correlación entre las variables de explicación causa que la matriz de diseño sea casi singular al estimar los parámetros por Mínimos Cuadrados estos van a ser inestables, es decir su varianza será alta. La Regresión Ridge busca estimar nuevos parámetros del modelo minimizando la varianza de los mismos, estos estimadores de los parámetros a diferencia de los estimadores por mínimos cuadrados son sesgados.
Cuando en un modelo de Regresión Lineal las observaciones siguen una distribución no-Normal particularmente aquellas que poseen colas más alargadas o gruesas, el Método de Mínimos Cuadrados puede que no sea el apropiado. Las distribuciones con "colas gruesas" usualmente son generadas debido a la presencia de valores aberrantes, estos valores pueden influenciar mucho en las estimaciones por Mínimos Cuadrados. Los procedimientos de Regresión Robusta están diseñados para disminuir la influencia de los valores aberrantes obteniendo estimaciones más eficientes que las realizadas por Mínimos Cuadrados.

Palabras Claves: *Ridge, Robustez, Valores Aberrantes, Estimadores Segados*

Abstract

In the linear regression context, it have been proposed several methods in order to confront multicollinearity. The main of them are biased estimates of the coefficients. Among this methods, is the Ridge Regression.
In linear regression models when the correlation between explanations variables cause the design matrix is almost singular, the estimations of parameters by least squares will be unstable, i.e. its variance is high. Ridge Regression seeks to estimate new model parameters minimizing the variance of the same, these estimates of the parameters unlike the least squares estimators are biased.

When in a Linear Regression model the observations follow a non-normal distribution, particularly those that have elongated tails or fat, the least squares method may not be appropriate. Distributions with "fat tails" are usually generated due to the presence of outliers, these values can greatly influence the estimates by least squares.Robust Regression procedures are designed to reduce the influence of outliers getting more efficient estimates than those made by least squares.

Key words: *Ridge, Robustness, outliers, Biased Estimates*

1. Introducción

El presente trabajo es un proyecto previo a la obtención del título de Ingeniero en Estadística Informática, de la materia Regresión Avanzada, la cual se encuentra dirigida a la carrera de Ingeniería en Estadística Informática de la ESPOL, dictada en el I Término del 2010 por el profesor Gaudencio Zurita Herrera. Durante el curso se desarrolló ERLA (Estadística de Regresión Lineal Avanzada), un Software especializado en Análisis de Regresión.

El Análisis de Regresión es una técnica estadística que sirve para explicar valores de una o más variables de respuesta en términos de un grupo de variables predictoras o de explicación.

Para poder aplicar esta metodología, se postula una relación funcional entre las variables.

Debido a su simplicidad analítica, la forma funcional que más se utiliza en la práctica es la relación lineal.

En esta sección, en primer lugar se mostrará el modelo de Regresión Lineal Simple el cual considera una sola variable de explicación posteriormente se menciona la Regresión Lineal Múltiple en el cual se involucran dos o más variables de explicación además del uso de interacciones y términos polinómicos en el modelo. También se citarán modelos de Regresión no usuales como Regresión Ridge, Regresión Robusta, entre otros.

Un antecedente fundamental para estudiar Regresión Ridge y Regresión Robusta es la Regresión Lineal.

2. Estimación

2.1 Introducción

Para estimar parámetros de una población se utiliza información obtenida a partir de los datos que contiene una muestra. En este capítulo se discutirá algunas cualidades deseables de los estimadores que serán útiles para el desarrollo de los capítulos posteriores.

Supongamos que tenemos una variable aleatoria X que define una población, sea esta discreta o continua, definimos θ, una característica denominada parámetro poblacional, donde θ es una constante generalmente desconocida la cual deseamos estimar. Puede suceder que nos interese más de una característica de interés por ejemplo, su media, varianza y mediana poblacional, en ese caso tendríamos un vector en R^3 que contiene a los parámetros; en general, definiremos un vector Θ en R^p donde p es el número de parámetros que deseamos estimar.

Llamaremos estimadores puntuales a aquellos estimadores que asignen directamente al parámetro el valor obtenido. Nótese que, para un parámetro de una población puede existir más de un estimador de dicho parámetro, es decir, supongamos que se desea estimar la media μ de una población, un estimador puntual de μ podría ser la media muestral como también lo puede ser la mediana o la moda muestral, en la siguiente sección de este capítulo se analizarán características de los estimadores que nos permitan discernir entre un estimador y otro en base a las circunstancias.

2.2 Características de los estimadores

Existen ciertas características de los estimadores, que son deseables al momento de realizar inferencias estadísticas, el criterio más importante quizás es el del sesgo.

Un estimador $\hat{\theta}$ del parámetro θ se lo considera insesgado si y solo sí, su esperanza, es decir $E[\hat{\theta}] = \theta$.

Cuando tenemos dos estimadores insesgados $\hat{\theta}_1$ y $\hat{\theta}_2$ de un mismo parámetro, para diferenciarlos se recurre a la varianza de estos estimadores, diremos que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ si y solo sí:

$$\frac{Var(\hat{\theta}_1)}{Var(\hat{\theta}_2)} < 1$$

Es decir si existen dos estimadores insesgados de un mismo parámetro, se postula como más eficiente el de menor varianza.

En forma general el siguiente gráfico ilustra en forma analógica el concepto de sesgo y eficiencia de un estimador.

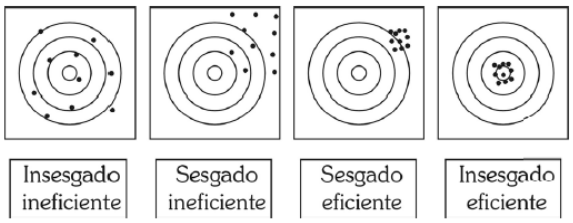


Figura 1 Sesgo y Eficiencia de un estimador

De lo expresado anteriormente, un estimador insesgado es más eficiente que otro en tanto su varianza sea menor, y mediante la cota de Rao y Cramér se obtiene el mínimo valor que puede tomar la varianza de un estimador insesgado.

2.3 Teorema de Rao y Cramér

Sea $X_1, X_2, \dots, X_n$ una muestra aleatoria de tamaño n tomada de una población X con densidad $f_\theta(x)$, $\theta \epsilon \Theta$ siendo $\Theta = \{\theta | \alpha < \theta < \beta\}$; α y β son conocidos; sea $\hat{\theta}$ un estimador insesgado de θ, bajo estas condiciones es verdad que:

$$Var(\hat{\theta}) \geq \frac{1}{nE\left[\left(\frac{\partial}{\partial\theta}\log_e f_\theta(x)\right)^2\right]}$$

Un estimador, cuya varianza alcance la cota de Rao y Cramér es un estimador eficiente y muy interesante, puesto que al hecho de ser insesgado se le habrá de sumar el hecho de tener la menor de las varianzas posibles.

2.4 Robustez de un estimador

Un parámetro poblacional se lo estima en base a la información que proporciona una muestra aleatoria tomada de dicha población, sin embargo en la práctica ocurre que muchas veces esta muestra se ve afectada por errores u observaciones atípicas denominados valores aberrantes, pues su comportamiento es diferente al resto de observaciones.

Para ilustrar lo mencionado tenemos la siguiente muestra de tamaño 20 tomada de una población normal con media $\mu = 20$ y varianza $\sigma^2 = 9$, realizamos el diagrama de puntos de la muestra.

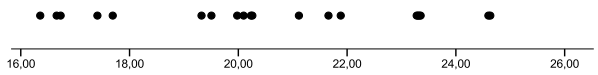


Figura 2 Ejemplo de Sesgo y Eficiencia de un estimador

Suponiendo que no conocemos el valor de la media poblacional y se la desea estimar, utilizaremos dos estimadores la media y la mediana muestral, así el valor de $\bar{x} = 20.56$ y el valor de la mediana $\tilde{x} = 20.24$ valores "cercanos" al valor de la media poblacional μ.

Contaminación de la muestra

Ahora que ocurriría si intercambiamos una observación de la muestra por un dato atípico o valor aberrante, siendo el grafico de puntos el siguiente:

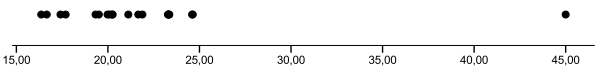


Figura 3 Ejemplo de Sesgo y Eficiencia de un estimador con valor aberrante

En la Figura 3 se puede observar un valor aberrante cercano a 45, realizando la estimación de la media poblacional nuevamente, nos queda, $\bar{x} = 21.98$ y el valor de la mediana $\tilde{x} = 20.68$, de este resultado se puede apreciar el cambio que ha sufrido la media aritmética, en relación al cambio de la mediana de los datos debido a la presencia de un valor aberrante en la muestra, esto nos da una idea de cómo la mediana es menos sensible a la presencia de valores aberrantes, lo cual ilustra la mayor robustez que la media aritmética muestral como estimador de la media poblacional.

De igual manera la desviación estándar se ve afectada por la presencia de estos valores, por lo que una alternativa robusta para la desviación estándar es la desviación absoluta de la mediana MAD, definida por,

$$MAD(\boldsymbol{X}) = MAD(X_1, X_2, \dots, X_n)$$
$$= Med[|\boldsymbol{X} - Med(\boldsymbol{X})|]$$

Este estimador utiliza la mediana dos veces la primera para obtener una estimación de los datos centrados alrededor de la mediana o residuales absolutos alrededor de la mediana, y la segunda, que obtiene la mediana de estos residuales absolutos.

Para modelar la situación en la que la mayoría de las observaciones provienen de una distribución F_θ, pero una pequeña proporción ε de las observaciones son valores aberrantes generados por otra variable aleatoria, Tukey[14] propone la familia de contaminación F_ε, definida por

$$F_\varepsilon = \{(1-\varepsilon)F_\theta + \varepsilon H;\ \theta \epsilon \Theta\}$$

Donde ε es el porcentaje de contaminación. Se espera que los estimadores robustos cumplan dos requerimientos, eficiencia y estabilidad, se llama eficiencia al hecho de que el estimador robusto se comporte "bien" sin la presencia de valores aberrantes, es decir, se lo pueda comparar con el estimador de máxima verosimilitud. Un estimador es estable cuando su comportamiento no varía ante la presencia de valores atípicos o aberrantes en la muestra.

3. Regresión Lineal

3.1. Introducción

En este Capítulo desarrollaremos el marco teórico y las aplicaciones para el problema de Regresión, utilizando varios modelos de Regresión. Entre ellos el Modelo de Regresión Lineal Simple, Polinómico y con interacciones, llevándolos hasta su forma matricial. Además discutiremos los métodos de estimación de los parámetros de estos modelos, utilizando la Tabla de Análisis de Varianza ANOVA, propondremos Contrates de Hipótesis basados en la partición de una forma cuadrática denominada Suma Cuadrática Total.

El problema general de regresión, consiste en encontrar la relación de una variable dependiente con un conjunto de variables independientes. Formalmente, dado un conjunto de datos (x_i, y_i), para $i = 1, \dots, n$, donde $x_i \epsilon R^p$ y y_i es el valor de salida correspondiente al vector x_i, dada una función $f(x_i, \beta)$, se requiere encontrar el vector de parámetros β, tal que

$$y_i = f(x_i, \beta) \text{ para } i = 1, \dots, n$$
$$\boldsymbol{X}_i \epsilon R^{p-1} \qquad \boldsymbol{\beta} \epsilon R^p$$
$$p: \#\ de\ parámetros\ del\ modelo$$

Donde, Y_i se lee fijando el valor de $\boldsymbol{X}_i$. En general, la expresión anterior es una aproximación, debido a la variabilidad de los datos del mundo real, una infinidad de factores que se reflejan de cada dato y la incertidumbre de muchas mediciones.

Con este tipo de datos trabajaremos en el Capítulo, para encontrar la relación funcional *f*, que explique a Y en términos de $\boldsymbol{X}$.

3.2 Modelo de Regresión Lineal Simple

El modelo más sencillo de Regresión Lineal, es aquel en el que explicamos la variable dependiente Y, en función de una sola variable independiente X,

conocido como modelo de Regresión Lineal Simple. Experimentalmente fijamos n valores para X y leemos Y, con lo que tendríamos n pares ordenados $(x_1\ y_1); (x_2\ y_2); \dots; (x_n\ y_n)$. En base a estos n pares encontraremos la relación funcional f que explique a Y en términos de X. Suponiendo que la relación existente entre X y Y es lineal, es decir, el gráfico de dispersión de los datos seguirán un patrón rectilíneo. Véase Figura 4.

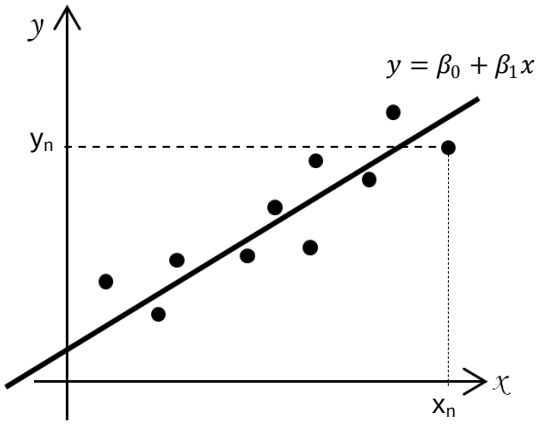


Figura 4 Diagrama de Dispersión

Como podemos observar en la *Figura 4* cada valor observado de Y no siempre determina un punto que pertenece a la recta, esto se debe a que al hacer la lectura de Y suponiendo que el modelo lineal $y = \beta_0 + \beta_1 x$ es válido, y fijando el valor de X, se comete un error aleatorio ε_i. Entonces para cada observación se plantea el modelo lineal siguiente:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i\ ,\ i = 1,2,\dots,n$$

con las condiciones

$$E[\varepsilon_i] = 0 \qquad VAR(\varepsilon_i) = \sigma^2$$

$$COV(\varepsilon_i, \varepsilon_j) = 0\ \ i \neq j$$

Debido a ε_i, Y es una variable aleatoria donde se espera que el valor de $Y_i = \beta_0 + \beta_1 X_i$, mientras X es una variable observable que condiciona a la variable Y, y es fijada según la necesidad del investigador. Los valores β_0, β_1 son constantes no conocidas, donde β_0 representa la intersección de la recta con el eje Y y β_1 la pendiente de la misma. En el modelo consideramos que $\beta_0 + \beta_1 x_i$ es su parte determinística, ε_i es una variable aleatoria, es decir, la parte estocástica del modelo; y, sujeta a los supuestos.

Bajo los supuestos anteriores sobre el modelo condicional expresamos el valor esperado del mismo de la siguiente manera:

$$E[Y_i|X = x_i] = E[\beta_0 + \beta_1 x_i + \varepsilon_i]$$

$$E[Y_i|X = x_i] = E[\beta_0 + \beta_1 x_i] + 0$$

$$E[Y_i|X = x_i] = \beta_0 + \beta_1 x_i$$

El hecho de que la varianza σ^2 del error sea constante durante todo el proceso, es un supuesto fuerte y cuando así lo hacemos, el modelo utilizado es calificado como **homocedástico**, esto es, de variabilidad constante.

Podemos comprobar que el modelo es Homocedástico cuando el grupo de puntos es cercano a cero, es decir si $E[\varepsilon_i] = 0$.

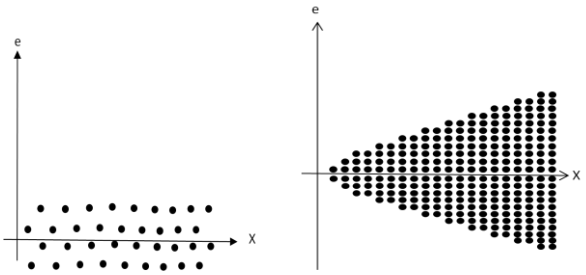
Estadísticamente se plantea estimar los valores de β_0 y β_1 que permitan la creación de una ecuación lineal para la estimación de Y:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Donde Y_i y su estimador $\hat{Y}_i$ no son necesariamente iguales, la diferencia $(\hat{Y}_i - Y_i)$ estima el valor del Error ε_i es decir:

$$\hat{\varepsilon}_i = e_i = \hat{Y}_i - Y_i$$

Donde e_i es conocido como los residuos del modelo. El grafico de los residuos de un modelo vs la variable de explicación x_i, refleja el comportamiento de la varianza del error, se observa la tendencia que siguen los puntos, si no poseen tendencia alguna y el promedio de los residuos es cero, podemos decir que el modelo es homocedástico como en el *Figura 5(a)* y si en cambio existe una tendencia en los puntos como en el *Figura 5 (b)*, decimos que existe **Heterocedasticidad**.



(a) (b)

Figura 5 Residuos vs Variable de explicación (a) Homocedástico (b) Heterocedástico

Dado que ε_i es una variable aleatoria con media 0 y varianza σ^2, Y_i es una variable aleatoria cuya media $E[Y_i|X = x_i] = \beta_0 + \beta_1 x_i$ y $VAR(Y_i) = \sigma^2$.

Suponiendo $\varepsilon_i \sim N(0, \sigma^2)$ esto implica que, $Y_i \sim N(\beta_0 + \beta_1 x_i\ , \sigma^2)$.

3.3 Modelo de Regresión Lineal Múltiple

No siempre una característica Y puede ser explicada en términos de una sola variable, es frecuente que exista *más de una variable de explicación*. Cuando se t iene esto, hablaremos de *Regresión Múltiple*.

En esta Sección se considera el Modelo de Regresión Lineal con *p* parámetros a estimar quedando la ecuación:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i\ ,\ i = 1,2,\dots,n$$

donde $X_{i,j}$ es el i-ésimo valor de la variable de explicación X_j.

En este caso la variable aleatoria Y se lee fijando los valores de las variables de explicación $X_1, X_2, X_3 \dots, X_{p-1}$. De esta forma el esperado de Y_i en términos de las variables $X_{i,j}$ es

$$E[Y_i|X_1 = x_{i1};\ X_2 = x_{i2}; \dots; X_{p-1} = x_{p-1}] = E[\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i]$$

$$= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1}$$

Esta expresión representa la parte determinística del modelo de Regresión Lineal Múltiple.

Usando notación matricial el modelo para n observaciones queda de la forma:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2p-1} \\ 1 & x_{31} & x_{32} & \dots & x_{3p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n\,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{y} \in R^n \qquad \boldsymbol{X} \in M_{nxp} \qquad \boldsymbol{\beta} \in R^p \qquad \boldsymbol{\varepsilon} \in R^n$$

Es decir $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, donde, $\boldsymbol{X} \in M_{nxp}$ se la conoce como Matriz de Diseño, cuyos vectores columnas (a excepción de la primera) representan a cada una de las variables de explicación, es decir $\boldsymbol{X}$ es una matriz de rango *p*, o rango completo. Con esta notación los supuestos del Error se pueden escribir en forma matricial de tal forma;

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

$$E[\boldsymbol{\varepsilon}] = \mathbf{0}$$

$$V(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I} = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

Además se introduce un término de **interacción** cuando se cree que una variable x_i influye sobre y en la relación entre otra variable x_j independiente

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i \ , \quad i = 1,2,\dots,n$$

3.4 Estimación de los parámetros del Modelo

Para la estimación del modelo consideramos el Método de Mínimos Cuadrados, en el cual minimizamos $\sum_{i=1}^{n} \varepsilon_i^2$ con respecto a $\boldsymbol{\beta}$, es decir,

$$\boldsymbol{\mu} = E[\boldsymbol{y}|\boldsymbol{\beta}] = E[\boldsymbol{X\beta} + \boldsymbol{\varepsilon}] = \boldsymbol{X\beta}$$

$$SCE = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{\mu})'(\boldsymbol{y} - \boldsymbol{\mu})$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2p-1} \\ 1 & x_{31} & x_{32} & \dots & x_{3p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n\,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

Los elementos del vector $\boldsymbol{\mu}$ vienen dado por:

$$\mu_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \ , \qquad i = 1, \dots, n$$

Las derivadas parciales con respecto a los parámetros quedan,

$$\frac{\partial (SCE)}{\partial \beta_j} = 2 \sum_{i=1}^{n} \varepsilon_i \frac{\partial \varepsilon_i}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ij} \varepsilon_i \ , \ j = 0,1,2,..,p$$

En forma vectorial,

$$\boldsymbol{x'}_j(\boldsymbol{y} - \boldsymbol{\mu}) = 0 \ , j = 0,1,2,\dots,p$$

El sistema de p ecuaciones nos queda,

$$\boldsymbol{X}'\boldsymbol{X\beta} = \boldsymbol{X'y}$$

Los valores de $\widehat{\boldsymbol{\beta}}$ que minimizan la SCE viene dada por:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$\widehat{\boldsymbol{\beta}}$ es el vector que contiene los estimadores de mínimos cuadrados de los parámetros, además son insesgados, es decir,

$$E\left[\widehat{\boldsymbol{\beta}}\ \right] = E[(\boldsymbol{X'X})^{-1}\boldsymbol{X'y}]$$

$$E\left[\widehat{\boldsymbol{\beta}}\ \right] = (\boldsymbol{X'X})^{-1}\boldsymbol{X'}E[\boldsymbol{y}]$$

$$E\left[\widehat{\boldsymbol{\beta}}\ \right] = (\boldsymbol{X'X})^{-1}\boldsymbol{X'X\beta}$$

$$E\left[\widehat{\boldsymbol{\beta}}\ \right] = \boldsymbol{I\beta} = \boldsymbol{\beta}$$

De aquí que, reemplazando esta expresión en el modelo de la ecuación *(2.29)*.

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

$$\widehat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

$$\widehat{\mathbf{Y}} = \boldsymbol{H}\boldsymbol{y}$$

La notación de **H** denominada "*Matriz Hat*", es

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$$

Donde $\boldsymbol{H} \epsilon M_{nxn}$ es simétrica e idempotente, esto significa que **HH** = $\mathbf{H}^2$ = **H**.

De igual forma como lo hicimos en Regresión Lineal

4. Regresión Ridge

4.2 Definición

La Regresión Ridge busca estimar nuevos parámetros del modelo minimizando la varianza de los mismos; los estimadores ridge a diferencia de los estimadores por mínimos cuadrados son sesgados. La Regresión Ridge es la que presenta mayor regularidad en el proceso de estimación y debido a esto es más atractivo su uso.

Suponiendo que se puede determinar un estimador sesgado de β denotado por $\hat{\beta}_r$ (Estimador Ridge), que tenga menor varianza que el estimador insesgado $\hat{\beta}$. El error cuadrático medio del estimador $\hat{\beta}_r$ se denota:

$$ECM(\widehat{\boldsymbol{\beta}}_{\mathbf{R}}) = \mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\mathbf{R}}) + \left[\mathrm{E}(\widehat{\boldsymbol{\beta}}_{\mathbf{R}}) - \boldsymbol{\beta}\right]^2$$

es decir

$$ECM(\widehat{\boldsymbol{\beta}}_{\mathbf{R}}) = \mathrm{Var}(\widehat{\boldsymbol{\beta}}_{\mathbf{R}}) + (\text{sesgo de } \widehat{\boldsymbol{\beta}}_{\mathbf{R}})^2$$

El Error Cuadrático Medio (ECM) no es más que la distancia esperada, de $\hat{\beta}_R$ a β, elevada al cuadrado. Nótese que si se permite una pequeña cantidad de sesgo en $\hat{\beta}_R$, la varianza de $\hat{\beta}_R$ se puede hacer pequeña, de tal modo que el error cuadrático medio de $\hat{\beta}_R$ sea menor que la varianza del estimador insesgado $\hat{\beta}$. En la Figura 6 se presenta el caso en el que la varianza del estimador sesgado es bastante menor que la del estimador insesgado.

Denotamos al estimador de Mínimos Cuadrados en forma matricial.

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Se han desarrollado varios procedimientos para obtener estimadores sesgados de coeficientes de regresión. Uno de esos procedimientos es la Regresión Ridge (o de la cresta), propuesta originalmente por

Hoerl y Kennard (1960a, b) [7],[8]. En forma específica, el estimador de ridge $\hat{\beta}_R$ se define como la solución de

$$(\mathbf{X}'\mathbf{X} + k\mathbf{I})\hat{\boldsymbol{\beta}}_{\mathbf{R}} = \mathbf{X}'\mathbf{y}$$

que es

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

donde *k* es una constante positiva, nótese que cuando k=0 el Estimador Ridge es igual al Estimador por Mínimos Cuadrados.

4.2.1 Varianza y Valor Esperado de un Estimador Ridge

La Regresión Ridge es un método de estimación válido cuando las variables de explicación no son independientes. El uso de estimadores de Ridge ayudan a disminuir la varianza de los estimadores considerando el sesgo de estimación. A continuación desarrollaremos el valor esperado y la Varianza de los Estimadores Ridge.

Media de Estimadores de Ridge

$$\begin{aligned} \mathrm{E}[\hat{\boldsymbol{\beta}}_{\mathbf{R}}] &= \mathrm{E}[(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})\mathrm{E}[\hat{\boldsymbol{\beta}}] \\ &= \mathrm{Z}_k\hat{\boldsymbol{\beta}} \end{aligned}$$

Donde $\mathrm{Z}_k = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})$ es una función que depende de los datos y de la constante *k*.

Varianza de Estimadores de Ridge

$$\begin{aligned} &\mathrm{Var}[\hat{\boldsymbol{\beta}}_{\mathbf{R}}] = \mathrm{Var}[(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}] \\ &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathrm{Var}[\mathbf{y}]\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \end{aligned}$$

$\hat{\theta}_1$ estimador insesgado de θ

$\hat{\theta}_2$ estimador sesgado de θ

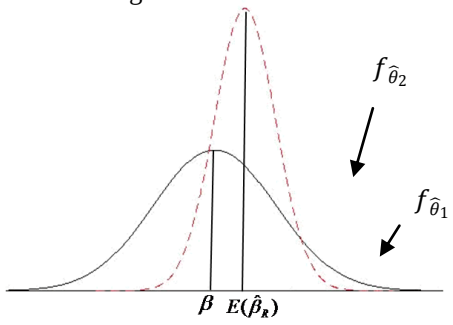


Figura 6 Distribución de Estimadores de Ridge y Mínimos Cuadrados

4.2.1 Métodos para seleccionar k

Al usar Regresión Ridge sería bueno escoger un valor de *k*, tal que la reducción en el término de varianza sea mayor que el aumento en el sesgo al cuadrado. De aquí que el Error Cuadrático Medio del estimador ridge $\hat{\beta}_R$ será menor que la varianza del estimador $\hat{\beta}$, por Mínimos Cuadrados. Hoerl y Kennard demostraron que existe un valor de *k* distinto de cero para el cual el Error Cuadrático Medio de $\hat{\beta}_R$ es menor que la varianza del estimador $\hat{\beta}$ por mínimos cuadrados, siempre y cuando el escalar $\boldsymbol{\beta}^T\boldsymbol{\beta}$ sea acotado.

A aumentar *k*, algunos de los Estimados Ridge variarán considerablemente. En cierto valor de *k* se estabilizarán los estimados ridge $\hat{\boldsymbol{\beta}}_{\mathbf{Ri}}$. El objetivo es seleccionar un valor de *k* razonablemente pequeño, en el cual los estimados ridge de $\hat{\boldsymbol{\beta}}_{\mathbf{Ri}}$ sean estables. Es posible que así se produzca un conjunto de estimados con el Error Cuadrático Medio menor que los estimados por mínimos cuadrados.

4.2.3 Traza de Ridge

Un primer método a utilizar es la Traza de Ridge el cual es un grafico en función de *k* y los valores de k en un intervalo entre [0,1]. En este método calculamos los estimadores de los $\hat{\boldsymbol{\beta}}_{\mathbf{Ri}}(k)$ para diferentes valores de *k* graficando para cada $\hat{\boldsymbol{\beta}}_{\mathbf{Ri}}$ la curva de coeficientes estimados. *Véase Figura 7.*

A aumentar *k*, algunos de los Estimados Ridge variarán en forma dramática. En cierto valor de *k* se estabilizarán los estimadores ridge $\hat{\boldsymbol{\beta}}_{\mathbf{Ri}}$. El objetivo es seleccionar un valor de *k* "razonablemente pequeño", en el cual los estimadores ridge de $\hat{\boldsymbol{\beta}}_{\mathbf{Ri}}$ tienen menor varianza que los otros estimadores. Es posible que así se produzca un conjunto de estimados con el Error Cuadrático Medio menor que los estimados por mínimos cuadrados.

Para elegir *k* hay que considerar los siguientes aspectos:

1. Que los valores de los coeficientes de regresión se estabilicen.
2. Que los coeficientes de regresión que tenían un valor demasiado grande comiencen a tener valores cercanos al valor real del coeficiente.
3. Que los coeficientes de regresión que inicialmente pudieran tener el signo equivocado cambien de signo.

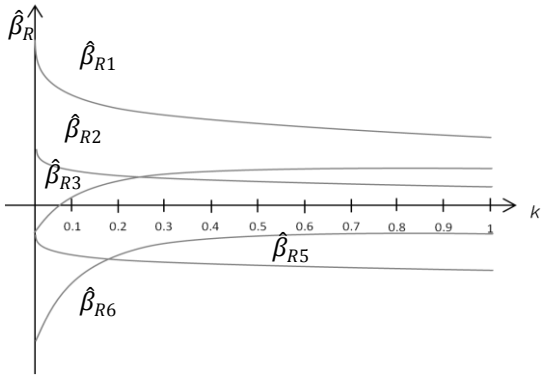


Figura 7 Traza de Ridge

4.2.4 Método Analítico

Algunos autores sugieren otros procedimientos para elegir *k*, entre ellos *Hoerl, Kennard y Baldwin* [6], proponen que una elección adecuada de *k* es

$$k = \frac{p\hat{\sigma}^2}{\hat{\beta}^T\hat{\beta}}$$

donde $\hat{\beta}$ y $\hat{\sigma}^2$ se determinan con la solución por Mínimos Cuadrados del modelo $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, siendo *p* el número de parámetros del modelo.

5. Regresión Robusta

5.2 Definición

Las distribuciones con "colas gruesas" usualmente son generadas debido a la presencia de valores aberrantes, estos valores pueden influenciar mucho en las estimaciones por Mínimos Cuadrados.

Mínimos Cuadrados presenta una clara desventaja al utilizar los residuales cuadráticos, ya que si dentro del conjunto de datos conocido existe un elemento que se encuentre lo suficientemente alejado del modelo original, éste repercutirá en gran medida sobre el ajuste que Mínimos Cuadrados nos proporcione, debido a que el elemento contribuirá en gran medida al error que se trata de minimizar, generando un ajuste sesgado en dirección del elemento indeseado.

Este tipo de elementos se conocen en la literatura como valores aberrantes o extremos.

Los valores aberrantes son muy comunes al trabajar con datos reales. Estos provienen de diversas fuentes, por ejemplo errores de captura, mal posicionamiento de puntos decimales, errores de almacenamiento, etc. En la mayoría de los casos este tipo de datos pasan desapercibidos, debido a que resulta muy complicado realizar un análisis preliminar.

Un caso claro y radical para ponderar en regresión se da cuando se utiliza una variable ficticia, con el propósito de reducir el efecto que provoca los residuos. En base a este contexto se utiliza el método de Regresión Robusta IRLS (Iteratively Reweighted Least Squares) el cual utiliza el Método de Mínimos Cuadrados Ponderados para disminuir la influencia de los valores aberrantes.

5.2.1 Regresión de Mínimos Cuadrados Ponderados

Una de las aplicaciones de la ponderación analítica de las observaciones es eliminar el error producido por la presencia de heterocedasticidad en los datos. Recordando que para el Método de Mínimos Cuadrados la suma cuadrática del error es

$$SCE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

La idea básica en Mínimos Cuadrados Ponderados es calcular el estimador $\hat{\boldsymbol{\beta}}_{\mathbf{RB}}$ que minimiza la siguiente función

$$SCE = \sum_{i=1}^{n} w(e_i)(e_i)^2 = \sum_{i=1}^{n} w(y_i - \hat{y}_i)(y_i - \hat{y}_i)^2$$

donde $w(e_i)$ es una función de ponderación que se introduce para reducir e incluso eliminar el efecto de los residuos altos. Por tanto se definen los pesos $w(e_i)$ de forma que tomen valores pequeños en los residuos e_i "grandes". Para aplicar esta definición es necesario conocer los residuos e_i.

El modo de ponderación que dispone ERLA es mediante un algoritmo iterativo que se construye mediante un procedimiento que se explica a continuación.

Función de Pesos

Muchas funciones de pesos han sido propuestas para "amortiguar" la influencia de los valores aberrantes. La función de Huber usa pesos *w* tales que:

$$w = \begin{cases} 1 \\ \dfrac{1.345}{|u|} \end{cases} ; \begin{matrix} |u| \leq 1.345 \\ |u| > 1.345 \end{matrix}$$

La escala residual u_i es el siguiente

$$u_i = \frac{e_i}{MAD}$$

En la función de Huber, *w* denota el peso o ponderación, y u denota la escala residual que será explicada más adelante. La constante 1.345 en la función de peso Huber mostrada en la Figura 8 es conocida como *constante de ajuste*. Este valor fue escogido para trabajar con un 95% de confianza para los modelos cuyo error sigue una distribución normal.

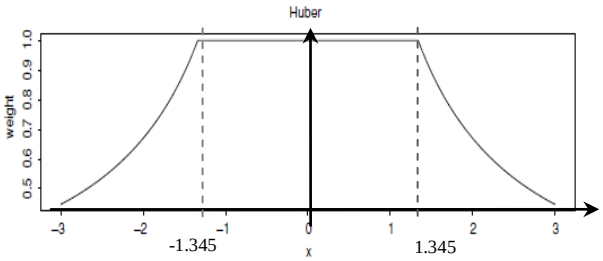


Figura 8 Función de Huber

Valores Iniciales

Algunas funciones de peso son muy sensibles a los valores iniciales; con otras, esto no es un problema. Cuando la función de peso Huber es empleada, los residuos iniciales pueden ser obtenidos de un ajuste de mínimos cuadrados.

Escala Residual

La función de peso de Huber está diseñada para ser usada con la escala residual definida por la expresión:

$$u_i = \frac{e_i}{MAD}$$

Sin embargo, en presencia de valores aberrantes, $\sqrt{MSE}$ no es un estimador sensible ante la presencia de dichos valores; la magnitud de $\sqrt{MSE}$ puede ser influenciada en gran medida por uno o muchos valores aberrantes. También, $\sqrt{MSE}$ no es un estimador robusto de σ cuando la distribución del error está lejos

de ser Normal. Sin embargo, la Desviación Absoluta de la Mediana (MAD) es el estimador a menudo empleado debido a su estabilidad como estimador robusto de la desviación estándar:

$$MAD = \frac{1}{0.6745} mediana(|e_i - mediana(e_i)|)$$

La constante 0.6745 provee una estimación insesgada de σ para observaciones independientes provenientes de una distribución normal. En este caso sirve para proveer un estimador que es aproximadamente insesgado.

Número de Iteraciones

Los procesos iterativos para obtener un nuevo ajuste, nuevos residuales y por ende nuevos pesos, se repite con los nuevos pesos hasta que el proceso converja, es decir los valores cambien poco.

$$\min\{\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{RB}}\right)' diag(\mathbf{W})\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{RB}}\right)\}$$

Descomponiendo las matrices nos queda

$$(e_1 \quad e_2 \quad e_3 \quad \cdots \quad e_n)\begin{pmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & \cdots & 0 \\ 0 & 0 & w_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_n \end{pmatrix}\begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{pmatrix}$$

Las Ecuaciones Normales del Método de Mínimos Cuadrados Ponderados son

$$\mathbf{X}'\mathbf{W}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{RB}}\right) = 0$$
$$\mathbf{X}'\mathbf{W}\mathbf{Y} - \mathbf{X}'\mathbf{W}\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{RB}} = 0$$
$$\mathbf{X}'\mathbf{W}\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathbf{RB}} = \mathbf{X}'\mathbf{W}\mathbf{Y}$$
$$\widehat{\boldsymbol{\beta}}_{\mathbf{RB}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y}$$

Con los nuevos estimadores $\widehat{\boldsymbol{\beta}}_{\mathbf{RB}}$ se obtienen unos nuevos residuos $e_i(1)$ y se repite este proceso hasta obtener la convergencia de las estimaciones.

6. Referencia

[1] ATKINSON, A. & RIANA, M. (2000), ‘‘*Robust Diagnostic Regresion Analysis’’*, Springer-Verlag, New York, EEUU

[2] BOVAS, A. & LEDOTTER, J. (2006), ‘‘*Introduction to Regression Modeling*’’, Thomson Brooks, Canada.

[3] CHAPMAN, S., (2002), *“MATLAB Programming for Engineers”,* Brooks-Cole, Canada.

[4] GYU SIM, D. & HONG PARK, R. (1998*), “Robust Reweighted MAP Motion Estimation”,* IEEE transactions on pattern analysis and machine intelligence, vol. 20, no. 4, Sogang University, Seoul, Korea.

[5] HALVORSON, M., (2008), *“Microsoft Visual Basic 2008 Step by Step”*, Microsoft Press, Washington, EEUU.

[6] HOERL, A. E., KENNARD, R. W. AND BALDWIN, K. F. (1975), *“Ridge regression: some simulations”*. *Communications in Statistics*, 4, 105-123.

[7] HOERL, A. E. & KENNARD, R. (1970)a, “Ridge Regression: Applications to Nonorthogonal Problems”, *Technometrics*; Vol. 12, No. 1., 55-67.

[8] HOERL, A. E. & KENNARD, R. (1970)b, “Ridge Regression: Biased Estimation for Nonorthogonal Problems*”, Technometrics*; Vol. 12, No. 1., 69-82.

[9] ITHAKA (2011), “*Jstor*”, http://www.jstor.org, Fecha de Ultima Visita: enero de 2011, Michigan, EEUU

[10] LÓPEZ, G. (2010), “*Ajuste robusto usando heurísticas”,* Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México, D.F.

[11] MARONNA, R., MARTIN, D. AND YOHAI, V., “*Robust Statistics: Theory and Methods”* , John Wiley &Sons, The Atrium, Southern Gate, Chichester, West Sussex, England

[12] MINITAB, INC. (2010), *“ MeetMinitab”,* Minitab Español, Versión 16.1.0., *EEUU*

[13] SEBER, G. & LEE, A. (2003), ‘‘Linear Regression Analysis’’, Segunda Edición, John Wiley & Sons, Inc, Hoboken, New Jersey.

[14] TUKEY, J. (1960), *“A survey of sampling from contaminated distributions”*, Contributions to Probability and Statistics, CA: Stanford University Press, Stanford

[15] ZURITA, G. (2010), *“Probabilidad y Estadistica, Fundamentos y Aplicaciones”*, Ediciones del Instituto de Ciencias Matematicas, Guayaquil, Ecuador