



**ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL**

**Instituto de Ciencias Matemáticas**

"CONSTRUCCIÓN DE SOFTWARE PARA REGRESIÓN. EL CASO DE  
REGRESIÓN RIDGE Y ROBUSTA"

**INFORME DE LA MATERIA DE GRADUACIÓN**

Previa a la obtención del Título de:

**INGENIERO EN ESTADÍSTICA INFORMÁTICA**

Presentado por:

ESTEFANY MELISSA MINALLA ALAVA

MARIO DAVID SOLÓRZANO CARVAJAL

Guayaquil – Ecuador

2011

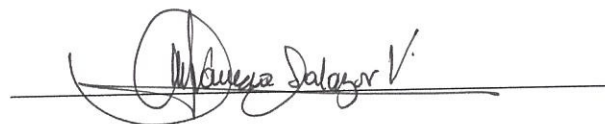
## **DEDICATORIA**

*A nuestros padres por su apoyo constante e incondicional. Al Máster Gaudencio Zurita por todas sus enseñanzas en nuestra formación profesional.*

# TRIBUNAL GRADUACIÓN



Máster Gaudencio Zurita Herrera  
PROFESOR DE LA MATERIA DE  
GRADUACIÓN



Ing. Vanessa Salazar Villalva  
DELEGADO DEL ICM

## DECLARACIÓN EXPRESA

"La responsabilidad del contenido de este Informe de Materia de Graduación, nos corresponde exclusivamente; y el patrimonio intelectual de la misma a la Escuela Superior Politécnica del Litoral".



Estefany Melissa Minalla Alava



Mario David Solórzano Carvajal



CIB-ESPOL

## 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 y la Regresión de Componentes Principales, Sin embargo, la Regresión Ridge es la que hace atractivo su uso.

En los modelos de Regresión Lineal cuando la correlación entre las variables de explicación causa que la matriz sea casi singular al estimar los parámetros por los 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.

# ÍNDICE GENERAL

INTRODUCCIÓN .....	1
CAPÍTULO I.....	8
1. ESTIMACIÓN.....	8
1.1. Introducción .....	8
1.2. Características de los estimadores.....	10
1.3. Teorema de Rao y Cramér .....	12
1.4. Robustez de un estimador.....	13
1.4.1. Contaminación de la Muestra.....	14
CAPÍTULO II.....	17
2. REGRESIÓN LINEAL .....	17
2.1. Introducción .....	17
2.2. Modelo de Regresión Lineal Simple .....	19
2.3. Estimación de Parámetros.....	23
2.3.1. Criterio de Máxima Verosimilitud .....	24
2.3.2. Criterio de Mínimos Cuadrados .....	26
2.4. Modelo Polinómico .....	32
2.5. Modelo de Regresión Lineal Múltiple.....	33
2.6. Estimación de los parámetros del Modelo .....	35
2.7. Tabla de Análisis de Varianza (ANOVA) .....	39

2.8. Intervalos de Confianza .....	44
2.9. Contraste de Hipótesis .....	44
2.9.1. Teorema de Cochran.....	45
CAPÍTULO III.....	52
3. REGRESIÓN RIDGE .....	52
3.1. Introducción .....	52
3.2. Multicolinealidad .....	53
3.3. Definición .....	56
3.3.1. Varianza y Valor Esperado de un Estimador Ridge.....	59
3.3.2. Media de Estimadores Ridge.....	60
3.3.3. Varianza de Estimadores de Ridge .....	60
3.4. Métodos para seleccionar k.....	61
3.4.1. Traza de Ridge .....	62
3.4.2. Método Analítico .....	64
CAPÍTULO IV.....	70
4. REGRESIÓN ROBUSTA.....	70
4.1. Introducción .....	70
4.2. Definición .....	71
4.3. Regresión de Mínimos Cuadrados Ponderados .....	75
4.3.1. Función de Pesos.....	77
4.3.2. Valores Iniciales .....	78
4.3.3. Escala Residual.....	78



4.3.4. Número de Iteraciones .....	79
CAPÍTULO V.....	86
5. PROGRAMACIÓN Y VALIDACIÓN .....	86
5.1. Introducción.....	86
5.2. Modulo de Regresión Ridge .....	87
5.3. Modulo de Regresión Robusta .....	90
5.4. Validación.....	93
5.4.1. Regresión Ridge .....	93
5.4.2. Regresión Robusta .....	96
COCLUSIONES.....	101
BIBLIOGRAFÍA.....	103
ANEXO .....	105

## SIMBOLOGÍA

$\beta$	Parámetro del modelo de regresión
$\sigma^2$	Varianza del modelo
$\mu$	Valor esperado de la variable a ser explicada y dado x
<b>H</b>	Matriz Hat
$\hat{\beta}$	Estimador de Mínimos Cuadrados
$R^2$	Coefficiente de Determinación
$\hat{\sigma}_{\beta_i}$	Varianza del estimador
$\hat{\beta}_R$	Estimador por Regresión Ridge
<b>W</b>	Vector de Pesos
$\hat{\beta}_{RB}$	Estimador por Regresión Robusta
<b>T</b>	Estadístico de Prueba t de student
<b>F</b>	Estadístico de Prueba F de Fisher
$\alpha$	Nivel de significancia de la Prueba

## ÍNDICE DE TABLAS

Tabla 2.1. Datos de Dosis e Incremento Porcentual.....	28
Tabla 2.2. Tabla de los residuos de los Datos de Dosis e Incremento Porcentual.....	31
Tabla 2.3. Datos de Calidad de Producto .....	38
Tabla 2.4. Tabla de Análisis de Varianza.....	40
Tabla 2.5. Tabla de Análisis de Varianza de la Calidad de un Producto.....	50
Tabla 2.6. Estadísticos de prueba para cada $\beta$ de Calidad de un producto.	51
Tabla 3.1. Datos de efecto de humedad, temperatura y producción de arroz .....	65
Tabla 3.2. Matriz de Correlación .....	66
Tabla 3.3. Estimación por el Método de Mínimos Cuadrados.....	66
Tabla 3.4. Estimación por Regresión Ridge.....	67
Tabla 4.1. Datos de la medición de gases en la atmosfera .....	81
Tabla 4.2. Iteraciones Huber con Mínimos Cuadrados Ponderados de medición de gases de la atmosfera .....	84
Tabla 4.3. Estimación por Regresión Robusto.....	84
Tabla 5.1. Datos del modelo .....	94
Tabla 5.2. Comparación de las Estimaciones con Mínimos Cuadrados vs Regresión Ridge .....	95

Tabla 5.3. Comparación de Media y Varianza de Mínimos Cuadrados vs Regresión Ridge .....	96
Tabla 5.4. Comparación de la Muestra Contaminada vs No contaminada por Mínimos Cuadrados y Regresión Robusta .....	99
Tabla 5.5. Estimaciones con Método de Mínimos Cuadrados y Regresión Robusta .....	99

## ÍNDICE DE GRÁFICOS

Gráfico 1.1. Sesgo y Eficiencia de un estimador.....	12
Gráfico 1.2. Ejemplo de Sesgo y Eficiencia de un estimador.....	13
Gráfico 1.3. Ejemplo de Sesgo y Eficiencia de un estimador con valor aberrante .....	14
Gráfico 2.1. Diagrama de Dispersión .....	20
Gráfico 2.2. Residuos vs Variable de explicación (a) Homocedástico (b) Heterocedástico .....	23
Gráfico 2.3. Diagrama de Dispersión de Dosis e Incremento Porcentual ...	29
Gráfico 2.4. Recta de Regresión Estimada de Dosis e Incremento Porcentual.....	30
Gráfico 2.5. Modelo de Regresión Polinómico .....	32
Gráfico 3.1. Distribución de Estimadores de Ridge y Mínimos Cuadrados..	61
Gráfico 3.2. Traza de Ridge.....	63
Gráfico 3.3. Traza de Ridge de datos del ejemplo .....	68
Gráficos4.1. (a) Conjunto de cinco puntos que definen correctamente la línea (b) Mismo conjunto de puntos presentado en (a) pero con un punto aberrante .....	74
Gráficos4.2. Función de Huber .....	78

Gráficos4.3. Gráfico de Dispersión .....	82
Gráficos4.4. Estimación Robusta.....	85
Gráfico 5.1. Estimación Robusta y de Mínimos Cuadrados (a) Muestra No Contaminada (b) Muestra Contaminada .....	98

# 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.

El software ERLA utiliza funciones gráficas y numéricas escritas en la plataforma Matlab\* para el desarrollo de los cálculos y Visual Basic\*\* para la creación de la interfaz gráfica. Para muestra del caso, ERLA contiene módulos específicos en donde se citan modelos de Regresión no usuales como Regresión Ridge, Regresión Robusta, Regresión Logística entre otros.

Visual Basic<sup>1</sup> pertenece a Microsoft y es por esta razón que el entorno grafico de ERLA es similar al del Sistema Operativo de Windows. Visual Basic 2008 es un sistema de desarrollo que puede ser utilizado para construir software de aplicación. Usando Visual Basic 2008 se pueden crear aplicaciones bajo el Sistema operativo Windows, la web y en muchos otros ambientes.

---

\***MATLAB** es un producto de Cleve Moler

\*\***VISUAL BASIC** es un producto de Microsoft

<sup>1</sup> [5]HALVORSON, M., (2008), "Microsoft Visual Basic 2008 Step by Step", Microsoft Press, Washington, EEUU



MATLAB<sup>2</sup> es un programa de computadora cuyo propósito es optimizar los cálculos científicos. En sus inicios era un programa diseñado para el desarrollo de la resolución de matrices, pero a través de los años se fue convirtiendo en un sistema de computación flexible capaz de resolver esencialmente cualquier problema del ámbito técnico. En los últimos tiempos Matlab ha implementado un subsistema llamada Matlab Builder que permite crear componentes COM, que es una acrónimo de *Component Object Model* los cuales son accesibles para Visual Basic o cualquier otro lenguaje que soporte COM, el cual es un lenguaje estándar binario para interoperabilidad de objetos, también conocido como Middleware. Cada objeto COM muestra una o más clases para el ambiente de programación de Visual Basic. Cada clase contiene un grupo de funciones llamados *Métodos*, correspondiente a las funciones originales de Matlab.

Para crear un componente, se debe proveer adicionalmente nombres a una o más clases. El nombre del componente representa el nombre del archivo \*.dll a crear.

Usando Matlab Builder para crear un componente COM es un proceso que requiere de una secuencia de cuatro pasos:

---

<sup>2</sup> [1]CHAPMAN, S., (2002), "MATLAB Programming for Engineers", Brooks-Cole, Canada.

- Crear un proyecto
- Administrar archivos M-Files
- Construir el proyecto
- Empaquetamiento y Distribución del componente

## Crear un Proyecto

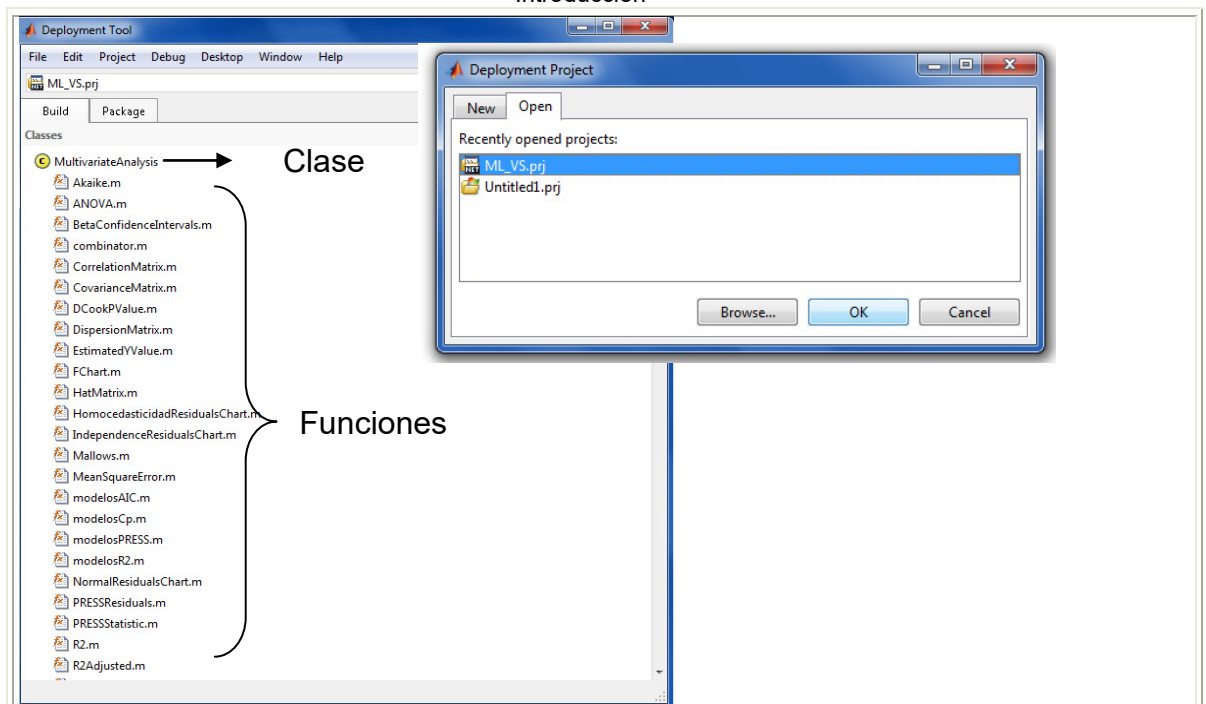
Para la creación del proyecto ingresamos en Matlab la línea de comando

```
>>deploytool
```

La ventana que aparece en Matlab es:

**Gráfico 1 Ventana de Deployment Tool**

Tesis de Grado  
Tema: Regresión Lineal Avanzada  
Introducción



Previamente hemos creado el proyecto ML\_VS.prj el cual contiene las funciones utilizadas en ERLA.

## Administrar M-Files

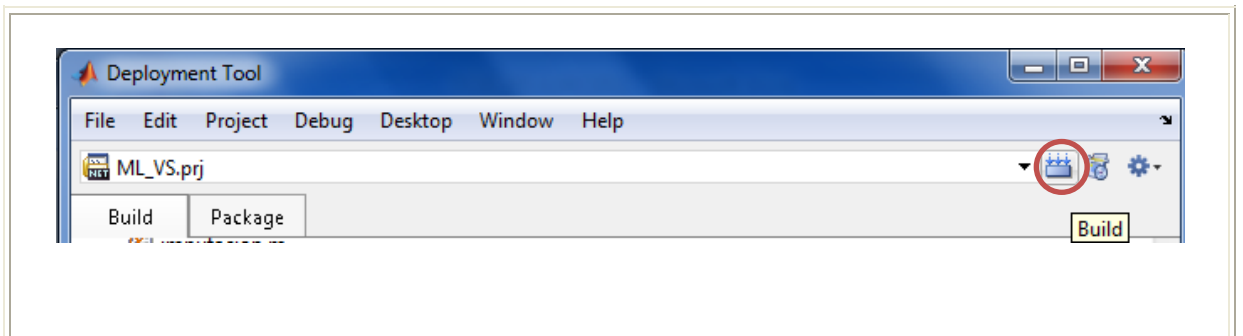
Añadimos una función de Matlab a alguna clase creada pulsando en el hipervínculo *AddFile* que aparece al final de la ventana mostrada en la grafica anterior

## Construcción del Proyecto

Después de definir las propiedades del proyecto y añadir las funciones de Matlab creadas, se construye la librería *\*.dll.* , al pulsar el botón *Build* se invoca al compilador de Matlab.

Gráfico 2 Botón Builder

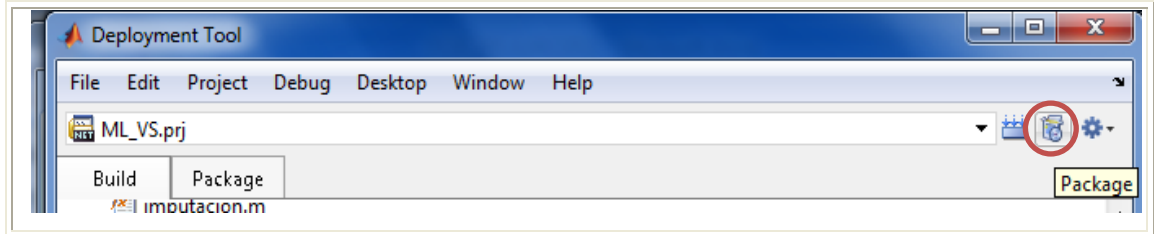
Tesis de Grado  
Tema: Regresión Lineal Avanzada  
Introducción



## Empaquetamiento y Distribución del Componente

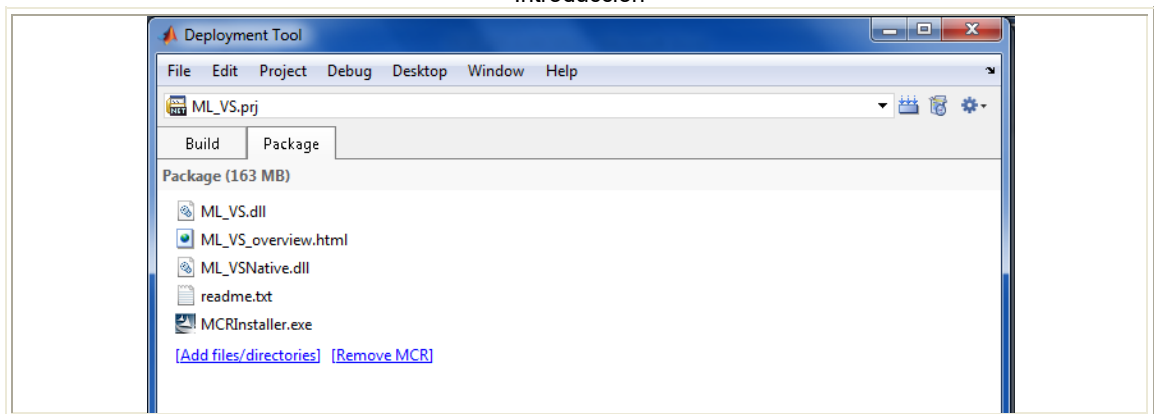
Una vez que se ha compilado satisfactoriamente los modelos y probado el objeto COM, procedemos a “empaquetar” el componente para la distribución a los usuarios finales. Pulsando la opción *Package*.

**Gráfico 3 Botón Package**  
 Tesis de Grado  
 Tema: Regresión Lineal Avanzada  
 Introducción



Este procedimiento extrae algunos archivos ejecutables, entre estos archivos se encuentra el `MCRInstaller.exe` el cual instala el MCR (Matlab Component Runtime), componente que necesita ser instalado en la máquina para el uso de ERLA.

**Gráfico 4 Ventaja con Archivos Ejecutables**  
 Tesis de Grado  
 Tema: Regresión Lineal Avanzada  
 Introducción



Visual Basic tiene acceso a estas componentes creadas en el Matlab Builder. Para crear la Interfaz Grafica del Usuario (GUI) utilizando las funciones creadas en Matlab, desde Visual Basic debemos hacer referencia a la librería `MW_Array.dll` la cual se crea al instalar el componente MCR; además de

hacer referencia a la librería que hemos creado desde el Matlab Builder, en nuestro caso la librería es ML\_VS.dll. De esta manera las funciones de la librería ML\_VS.dll creadas en Matlab podrán ser utilizados como funciones en Visual Basic.

Acerca del módulo correspondiente a nuestro tema Regresión Ridge y Regresión Robusta, un antecedente fundamental para estudiar es La Regresión Lineal. A éste dedicamos un capítulo de este trabajo. El segundo y tercer capítulo está orientado a examinar la necesidad de utilizar Regresión Ridge y Robusta como enfoque específico al Análisis de Regresión. Finalmente, en el cuarto capítulo se considera el desarrollo del software, el método de creación y la conexión entre los programas utilizados Matlab y Visual Basic además, la validación del Software ERLA.

# CAPÍTULO I

## 1. ESTIMACIÓN

### 1.1. 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  $\theta$ , una característica denominada parámetro poblacional, donde  $\theta$  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  $\theta$  en  $R^p$  donde  $p$  es el número de parámetros que deseamos estimar.

Sea  $X_1, X_2, \dots, X_n$  una muestra aleatoria de tamaño  $n$  tomada de la población  $X$  con vector de parámetros  $\theta \in R^p$ , el vector de estimadores  $\hat{\theta}$  de  $\theta$ , es una función  $\hat{\theta} : R^n \rightarrow R^p$ , definida en términos de los valores de la muestra y que, en su definición no incluye al vector de parámetros  $\theta$ .

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  $\mu$  de una población, un estimador puntual de  $\mu$  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.

## 1.2. 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  $\theta$  se lo considera insesgado si y solo sí, su esperanza, es decir  $E[\hat{\theta}] = \theta$ .

En este sentido vamos a citar el siguiente ejemplo; Supongamos que una población sigue la distribución normal  $N(\mu, 1)$  y que de esta población tomamos una muestra de tamaño  $n$  para estimar el parámetro desconocido  $\mu$ , vamos a utilizar como estimadores de la media los siguientes estadísticos:

$$\hat{\theta}_1 = \bar{x} ; \hat{\theta}_2 = \tilde{x} ; \hat{\theta}_3 = \frac{X_1 + X_2}{2} ; \hat{\theta}_4 = 5 ; \hat{\theta}_5 = \max_{0 \leq i \leq n} X_i \quad (1.1)$$

El valor esperado de los tres primeros estimadores es igual al valor del parámetro. El esperado de  $\hat{\theta}_4$  es 5 ya que es constante y únicamente sería insesgado en caso de que  $\mu = 5$ , por lo tanto tiene sesgo en general. En el caso de  $\hat{\theta}_5$  su valor esperado no puede ser igual a la media poblacional, puesto que el máximo de una muestra siempre estará por encima del valor promedio.



En este ejemplo, siguiendo el criterio de estimador insesgado, los tres primeros estadísticos son buenos estimadores de  $\mu$ , entonces podría resultar más cómodo estimar  $\mu$  utilizando  $\hat{\theta}_3$  que calcula el promedio de las dos primeras observaciones, a diferencia de  $\bar{x}$  que calcula el promedio de las  $n$  observaciones de la muestra, sin embargo se estaría dejando de lado un criterio fundamental en los estimadores como es la eficiencia del estimador.

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 si:

$$\frac{\text{Var}(\hat{\theta}_1)}{\text{Var}(\hat{\theta}_2)} < 1 \quad (1.2)$$

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

En el caso del ejemplo anterior si bien  $\hat{\theta}_1, \hat{\theta}_2$  y  $\hat{\theta}_3$  son insesgados,

$$\text{Var}(\hat{\theta}_1) = \text{Var}(\bar{x}) = \frac{1}{n}$$

$$\text{Var}(\hat{\theta}_2) = \text{Var}(\hat{x}) = 1$$

$$\text{Var}(\hat{\theta}_3) = \text{Var}\left(\frac{X_1 + X_2}{2}\right) = \frac{1}{2}$$

Por lo tanto la media muestral es un estimador más eficiente que  $\hat{\theta}_2$  y  $\hat{\theta}_3$ , en forma general el siguiente gráfico ilustra en forma analógica el concepto de sesgo y eficiencia de un estimador.

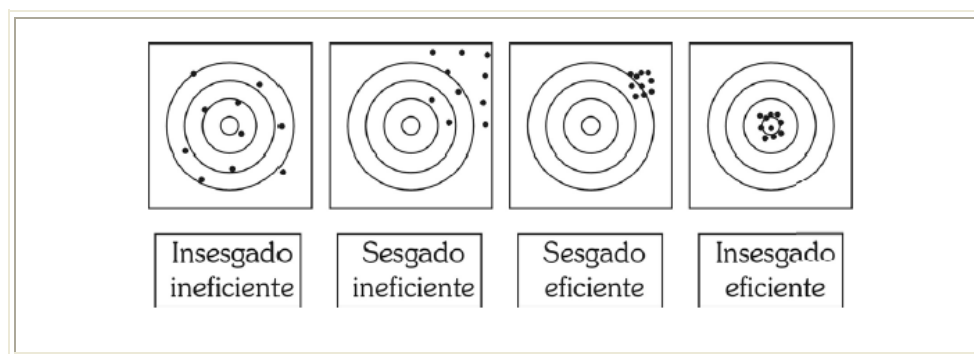


Gráfico 1.1 Sesgo y Eficiencia de un estimador

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.

### 1.3. 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  $f_\theta(x)$ ,  $\theta \in \Theta$  siendo  $\Theta = \{\theta | \alpha < \theta < \beta\}$ ;  $\alpha$  y  $\beta$  son conocidos; sea  $\hat{\theta}$  un estimador insesgado de  $\theta$ , bajo estas condiciones es verdad que:

$$Var(\hat{\theta}) \geq \frac{1}{nE\left[\left(\frac{\partial}{\partial \theta} \log_e f_\theta(x)\right)^2\right]} \quad (1.3)$$

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.

#### 1.4. 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.

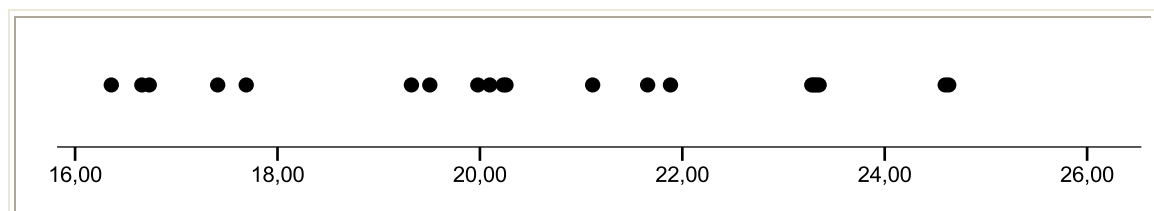


Gráfico 1.2 Ejemplo de Sesgo y Eficiencia de un estimador

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

### 1.4.1. Contaminación de la muestra

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

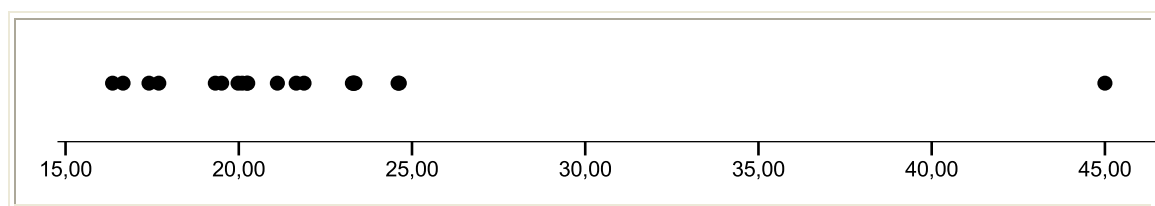


Gráfico 1.3 Ejemplo de Sesgo y Eficiencia de un estimador con valor aberrante

En el Gráfico 1.3 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,

$$MAD(\mathbf{X}) = MAD(X_1, X_2, \dots, X_n) = \text{Med}[|\mathbf{X} - \text{Med}(\mathbf{X})|] \quad (1.4)$$

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  $F_\theta$ , pero una pequeña proporción  $\varepsilon$  de las observaciones son valores aberrantes generados por otra variable aleatoria, Tukey [14]<sup>3</sup> propone la familia de contaminación  $F_\varepsilon$ , definida por

---

<sup>3</sup> [14] TUKEY, J. (1960), "A survey of sampling from contaminated distributions", Contributions to Probability and Statistics, CA: Stanford University Press, Stanford

$$F_\varepsilon = \{(1 - \varepsilon)F_\theta + \varepsilon H; \theta \in \Theta\} \quad (1.5)$$

Donde  $\varepsilon$  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.

# CAPÍTULO II

## 2. REGRESIÓN LINEAL

### 2.1. 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 a ser explicada o dependiente, que denominaremos  $Y$  con una o más variables de explicación o

independientes  $X_1, X_2, \dots, X_{p-1}$ . Formalmente, dado un conjunto de datos  $(\mathbf{X}_i, Y_i)$ , para  $i = 1, \dots, n$ , donde  $\mathbf{X}_i \in \mathbb{R}^{p-1}$  y  $Y_i$  es el valor de salida correspondiente al vector  $\mathbf{X}_i$ , dada una función  $f(\mathbf{X}_i, \boldsymbol{\beta})$ , se requiere encontrar el vector de parámetros  $\boldsymbol{\beta}$ , tal que

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\beta}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon_i \quad (2.1)$$

$$\begin{array}{ll} \mathbf{X}_i \in \mathbb{R}^{p-1} & \boldsymbol{\beta} \in \mathbb{R}^p \\ p: \# \text{ de parámetros del modelo} & \end{array}$$

Donde,  $Y_i$  se lee fijando el valor de  $\mathbf{X}_i$ . En general, la *Ecuación (2.1)* 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  $X$ .

La Regresión Lineal es la más conocida y utilizada entre las técnicas de regresión. Pudiera ser que existan más de dos variables que expliquen a  $Y$  el peso de una persona explicado, por ejemplo, a través de su estatura y edad.



Lo más frecuente es que  $f$  sea una función lineal debido a que el vector de variables de explicación  $x_i$  es una combinación lineal del vector de parámetros  $\beta$ . En caso de no serlo, se puede *linealizar* el modelo mediante transformaciones de las variables.

## 2.2. 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 *Gráfico 2.1*.

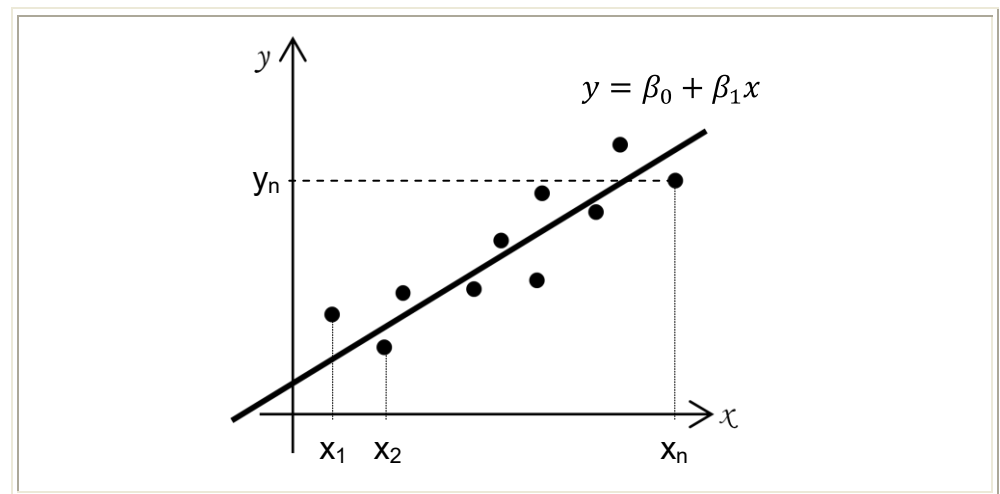


Gráfico 2.1 Diagrama de Dispersión

Como podemos observar en la *Gráfico 2.1* 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  $\varepsilon_i$ . Entonces para cada observación se plantea el modelo lineal siguiente:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (2.2)$$

con las condiciones

$$E[\varepsilon_i] = 0 \quad \text{VAR}(\varepsilon_i) = \sigma^2 \quad \text{COV}(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$$

Debido a  $\varepsilon_i$ ,  $Y$  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  $\beta_0$ ,  $\beta_1$  son constantes no conocidas, donde  $\beta_0$  representa la intersección de la recta con el eje  $Y$  y  $\beta_1$  la pendiente de la misma. En el modelo consideramos que  $\beta_0 + \beta_1 x_i$  es su parte determinística,  $\varepsilon_i$  es una variable aleatoria, es decir, la parte estocástica del modelo; y, sujeta a los supuestos mostrados en (2.2).

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

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

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

$$E[Y_i|X = x_i] = \beta_0 + \beta_1 x_i \quad (2.3)$$

El hecho de que la varianza  $\sigma^2$  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  $\beta_0$  y  $\beta_1$  que permitan la creación de una ecuación lineal para la estimación de  $Y$ :

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad (2.4)$$

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  $\varepsilon_i$  es decir:

$$\hat{\varepsilon}_i = e_i = \hat{Y}_i - Y_i \quad (2.5)$$

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 *Gráfico (2.2) (a)* y si en cambio existe una tendencia en los puntos como en el *Gráfico (2.2) (b)*, decimos que existe **Heterocedasticidad**.

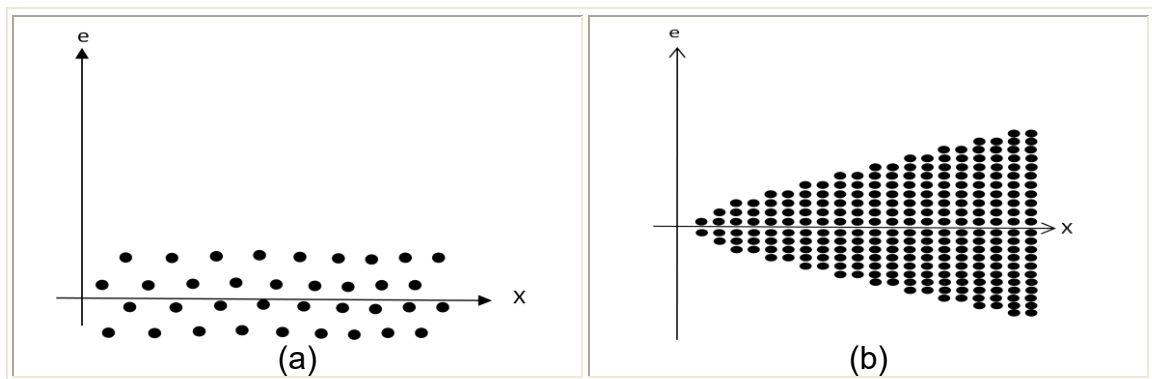


Gráfico 2.2 Residuos vs Variable de explicación (a) Homocedástico (b) Heterocedástico

Dado que  $\varepsilon_i$  es una variable aleatoria con media 0 y varianza  $\sigma^2$ ,  $Y_i$  es una variable aleatoria cuya media viene dada por la expresión (2.3) y  $VAR(Y_i) = \sigma^2$ . Suponiendo  $\varepsilon_i \sim N(0, \sigma^2)$  esto implica que,  $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ .

### 2.3. Estimación de Parámetros

Para estimar los parámetros del modelo de Regresión Lineal Simple nos basamos en la información obtenida de los datos observados condicionado a que se cumplan los supuestos. A continuación explicaremos los dos criterios de estimación más conocidos: Criterio de Máxima Verosimilitud el cual trabaja con la función de verosimilitud de los parámetros y el Criterio de Mínimos Cuadrados al que su nombre es debido a que minimiza la SCE.

### 2.3.1. Criterio de Máxima Verosimilitud

Dado el modelo de Regresión Lineal Simple

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

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$E[\varepsilon_i] = 0 \quad \text{VAR}(\varepsilon_i) = \sigma^2 \quad \text{COV}(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$$

Donde las variables  $Y_1, Y_2, \dots, Y_n$  tienen una distribución conjunta  $f$  que dependen de los parámetros  $\beta_0, \beta_1, \dots, \beta_{p-1}, \sigma^2$ , debido a la independencia entre  $y_1, y_2, \dots, y_n$  la distribución conjunta es igual al producto de sus marginales.

$$f(y_1, y_2, \dots, y_n | \beta_0, \beta_1, \sigma^2) = f_1(y_1 | \beta_0, \beta_1, \sigma^2) \dots f_n(y_n | \beta_0, \beta_1, \sigma^2) \quad (2.6)$$

$$f(y_1, y_2, \dots, y_n | \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^n f_i(y_i | \beta_0, \beta_1, \sigma^2)$$

A partir de esto definimos la Función de Verosimilitud en términos de los parámetros

$$L(\beta_0, \beta_1, \sigma^2 | y_1, y_2, \dots, y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right] \quad (2.7)$$

Donde la *Ecuación (2.7)* es la distribución conjunta de la muestra en términos de los parámetros. Algebraicamente la función de verosimilitud y la distribución conjunta es la

misma expresión, excepto que la densidad está en función de los  $y_i$ . En la función de verosimilitud se fijan los valores de  $y_i$ , y es una función de  $\beta_0, \beta_1, \dots, \beta_{p-1}, \sigma^2$ ; en la densidad conjunta  $\beta_0, \beta_1, \dots, \beta_{p-1}, \sigma^2$  están fijos y los valores de  $y_i$  pueden variar, mientras que en la función de verosimilitud los valores de  $y_i$  ya fueron observados y varían los parámetros.

El criterio de Máxima Verosimilitud estima los parámetros de tal manera que la Función de Verosimilitud de dichos parámetros se maximiza. De esta manera

$$\hat{\theta} = \arg\{max L(\theta)\} \quad (2.8)$$

Donde  $\hat{\theta}$  es el argumento que maximiza la función de máxima verosimilitud. Maximizar la función  $L(\beta_0, \beta_1, \sigma^2)$  es equivalente a maximizar  $\ln [L(\beta_0, \beta_1, \sigma^2)]$ , debido a que la función logaritmo natural es monótona creciente, el valor que maximice esta función es igual al máximo de la función  $L(\beta_0, \beta_1, \sigma^2)$ .

$$\ln[L(\beta_0, \beta_1, \sigma^2)] = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (2.9)$$

Maximizando la función de verosimilitud con respecto a  $\beta_0$  y  $\beta_1$

$$\frac{\partial(\ln(L(\beta_0, \beta_1, \sigma^2|x_i)))}{\partial \beta_0} = 0 \quad (2.10)$$

$$\frac{\partial(\ln(L(\beta_0, \beta_1, \sigma^2|x_i)))}{\partial \beta_1} = 0 \quad (2.11)$$

Igualando a cero las *Ecuaciones* (2.10) y (2.11) y verificando mediante la segunda derivada los estimadores de Máxima Verosimilitud

$$\frac{\partial(\ln(L(\beta_0, \beta_1, \sigma^2|x_i)))}{\partial \beta_0} = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) \quad (2.12)$$

$$\frac{\partial(\ln(L(\beta_0, \beta_1, \sigma^2|x_i)))}{\partial \beta_1} = \frac{1}{\sigma^2} \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) \quad (2.13)$$

### 2.3.2. Criterio de Mínimos Cuadrados

Tomando el modelo para Regresión Lineal Simple  $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ,  $i = 1, 2, \dots, n$  la estimación por Mínimos Cuadrados, minimiza la suma cuadrática de la diferencia entre los valores observados  $Y_i$  y los valores estimados  $\hat{Y}_i$  por el modelo. Entonces buscamos los valores de  $\hat{\beta}_0$  y  $\hat{\beta}_1$  que minimicen la función:

$$r_i = |y_i - \hat{y}_i|$$



$$SCE = \sum_{i=1}^n (r_i)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.14)$$

La *Expresión (2.14)* se denomina Suma Cuadrática del Error, de ahí el nombre de Mínimos Cuadrados ya que el objetivo es minimizar la Suma Cuadrática del Error.

Tomando de la *Expresión (2.14)*

$$Q = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \quad (2.15)$$

Derivando con respecto a  $\beta_0$  y  $\beta_1$  e igualando a cero la función anterior

$$\frac{\partial Q}{\partial \beta_0} = \frac{\partial Q}{\partial \beta_1} = 0 \quad (2.16)$$

nos queda un sistema de dos ecuaciones como sigue:

$$\frac{\partial Q}{\partial \beta_0} = n\beta_0 + (\sum_{i=1}^n x_i)\beta_1 = \sum_{i=1}^n y_i \quad (2.17)$$

$$\frac{\partial Q}{\partial \beta_1} = (\sum_{i=1}^n x_i)\beta_0 + (\sum_{i=1}^n x_i^2)\beta_1 = \sum_{i=1}^n x_i y_i \quad (2.18)$$

Las *Ecuaciones (2.17)* y *(2.18)* se las conoce como *Ecuaciones Normales*. Dejando en función de  $\hat{\beta}_0$  y  $\hat{\beta}_1$  ambas expresiones, la solución viene dada por:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \quad (2.19)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \text{ donde } \bar{y} = \frac{\sum y_i}{n} \text{ y } \bar{x} = \frac{\sum x_i}{n}$$

Ilustrando estos conceptos utilizaremos un ejercicio tomado de *Zurita*[15]<sup>4</sup>.

**Ejemplo:** A fin de evitar la retención de líquidos que puede ser peligrosa para pacientes con presión sanguínea, un investigador se propone estudiar el incremento porcentual Y, en la cantidad de orina que una droga provoca al ser administrada a diez personas. Se prescriben cinco dosis X, en miligramos, y se replican la dosis cada dos personas. Los resultados (x y) son los que se anotan en la *Tabla 1.1*.

X	1	1	2	2	3	3	4	4	5	5
Y	1.93	2.18	2.12	5.01	11.1	12.7	15.2	18.7	26.2	28.9

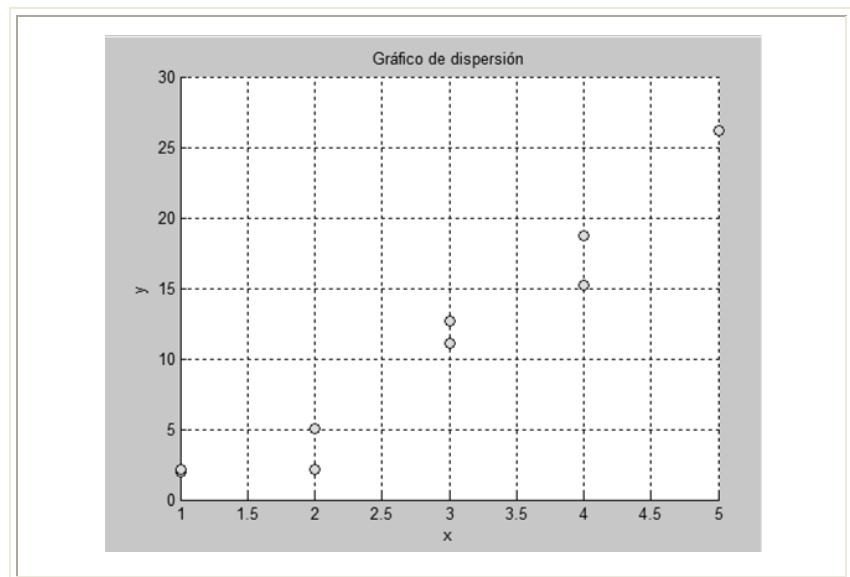
**Tabla 2.1 Datos de Dosis e Incremento Porcentual**

*Fuente: G. Zurita H.*

---

<sup>4</sup> [15]Zurita, G. (2010), "Probabilidad y Estadística, Fundamentos y Aplicaciones", Ediciones del Instituto de Ciencias Matemáticas, pág. 584-585, Guayaquil, Ecuador

Haciendo uso del software de Estadística en Regresión Lineal Avanzada ERLA realizamos el Gráfico de Dispersión de los datos contenidos en la *Tabla 2.1*.



**Gráfico 2.3** Diagrama de Dispersión de Dosis e Incremento Porcentual

En el *Gráfico 2.3* podemos observar la relación existente entre los dos grupos de variables, nótese que su tendencia es lineal, por lo que una estimación de los parámetros, utilizando un modelo de Regresión Lineal Simple parece ser adecuado.

Planteamos el modelo

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, 10$$

Utilizando el criterio de mínimos cuadrados

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 3 \quad \bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = 12.404$$

De aquí que,

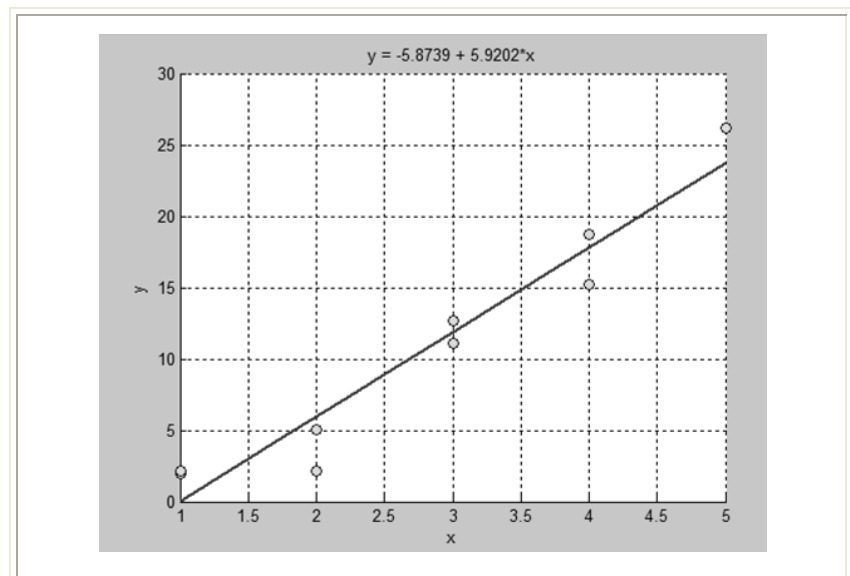
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} (x_i - 3)(y_i - 12.404)}{\sum_{i=1}^{10} (x_i - 3)^2} = 6.438$$

$$\hat{\beta}_0 = 12.404 - (6.438)3 = -6.909$$

La estimación del modelo bajo el criterio de mínimos cuadrados queda:

$$\hat{y}_i = -6.909 + 6.438x_i, \quad i = 1, 2, \dots, 10$$

Graficando la recta de regresión estimada en el diagrama de dispersión,



**Gráfico 2.4 Recta de Regresión Estimada de Dosis e Incremento Porcentual**

Calculando los valores estimados para cada valor de x, y los residuos del modelo,

$x_i$	$y_i$	$\hat{y}_i$	$e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i$
1	1,93	-0,471	2,401
1	2,18	-0,471	2,651
2	2,12	5,967	-3,847
2	5,01	5,967	-0,957
3	11,1	12,404	-1,304
3	12,7	12,404	0,296
4	15,2	18,842	-3,642
4	18,7	18,842	-0,142
5	26,2	25,279	0,921
5	28,9	25,279	3,621

**Tabla 2.2** Tabla de los residuos de los Datos de Dosis e Incremento Porcentual

*Autores: Minalla y Solórzano*

Sumando los residuos,

$$\sum_{i=1}^{10} e_i = 0$$

Dado que la suma de los residuos del modelo es cero, más adelante estudiaremos el método a utilizar para verificar que los valores estimados del modelo se encuentran dentro del intervalo de confianza, mediante el uso del análisis de varianza.

## 2.4. Modelo Polinómico

Puede ocurrir que la relación entre la variable  $Y$  y  $X$  no sea rectilínea como se presenta en los diagramas de dispersión del *Gráfico 2.5*, donde se observa que la relación entre las variables no es rectilínea, más bien es de forma polinómica; por lo que un modelo  $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$  podría no ser suficiente para predecir  $Y$ , en vez de esto consideramos un modelo **Polinómico** del tipo:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (2.20)$$

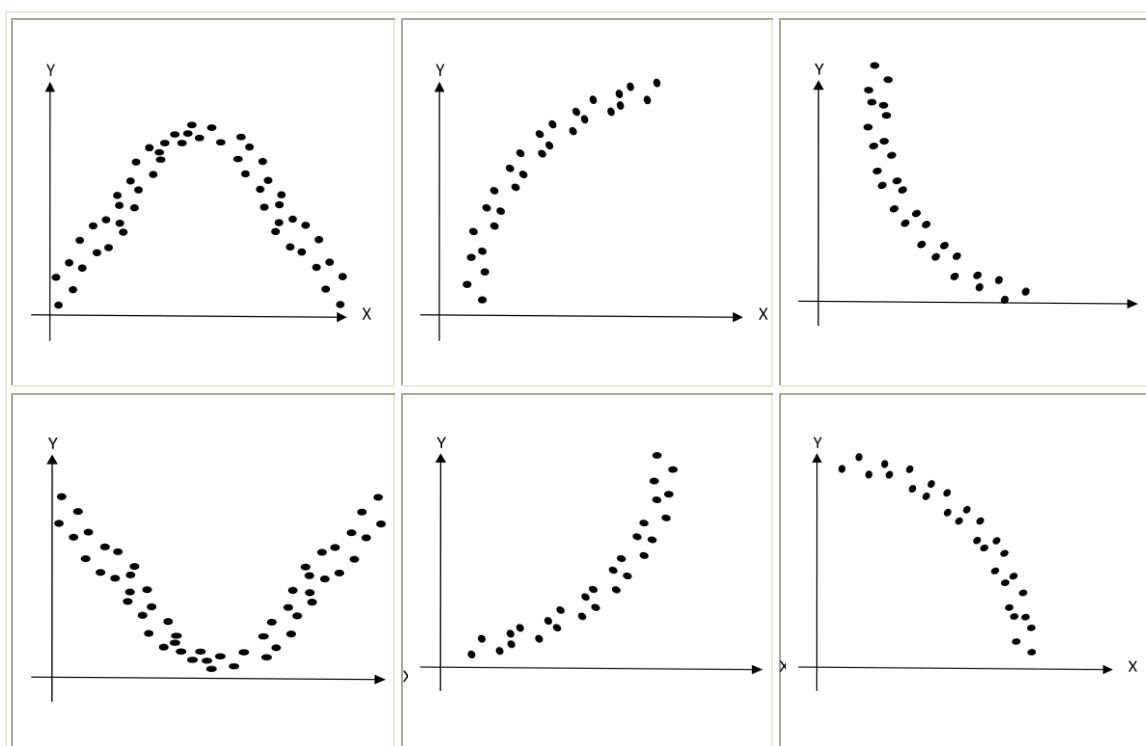


Gráfico 2.5 Modelo de Regresión Polinómico

Nótese que en este modelo estimamos tres coeficientes donde  $\beta_2$  es el coeficiente correspondiente a la variable de explicación

al cuadrado, cabe resaltar que tenemos una sola variable de explicación en el modelo ( $x_i$ ).

Se puede expresar la *Ecuación 2.20* en forma matricial:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12}^2 \\ 1 & x_{21} & x_{22}^2 \\ 1 & x_{31} & x_{32}^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad (2.21)$$

Que puede quedar representada en la siguiente expresión:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (2.22)$$

$$\mathbf{y} \in R^n \quad \mathbf{X} \in M_{n \times p} \quad \boldsymbol{\beta} \in R^p \quad \boldsymbol{\varepsilon} \in R^n$$

## 2.5. 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 tiene esto, hablaremos de *Regresión Múltiple*.

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (2.23)$$

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

$$\begin{aligned} E[Y_i | X_1 = x_{i1}; X_2 = x_{i2}; \dots; X_{p-1} = x_{i,p-1}] &= E[\beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i,p-1} + \varepsilon_i] \\ &= \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i,p-1} \quad \mathbf{(2.24)} \end{aligned}$$

La expresión (2.24) representa la parte determinística del modelo de Regresión Lineal Múltiple.

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

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2,p-1} \\ 1 & x_{31} & x_{32} & \dots & x_{3,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \mathbf{(2.25)}$$

$$\mathbf{y} \in R^n \quad \mathbf{X} \in M_{n \times p} \quad \mathbf{\beta} \in R^p \quad \mathbf{\varepsilon} \in R^n$$

Es decir  $\mathbf{y} = \mathbf{X}\mathbf{\beta} + \mathbf{\varepsilon}$ , donde,  $\mathbf{X} \in M_{n \times p}$  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  $\mathbf{X}$  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;

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (2.26)$$

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

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

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

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

## 2.6. 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,

$$\boldsymbol{\mu} = E[\mathbf{y}|\boldsymbol{\beta}] = E[\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}] = \mathbf{X}\boldsymbol{\beta} \quad (2.27)$$

$$SCE = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu}) \quad (2.28)$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p-1} \\ 1 & x_{21} & x_{22} & \dots & x_{2p-1} \\ 1 & x_{31} & x_{32} & \dots & x_{3p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$$

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

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{ip-1}, \quad i = 1, \dots, n \quad (2.29)$$

Las derivadas parciales con respecto a los parámetros quedan,

$$\frac{\partial(SCE)}{\partial \beta_j} = 2 \sum_{i=1}^n \varepsilon_i \frac{\partial \varepsilon_i}{\partial \beta_j} = -2 \sum_{i=1}^n x_{ij} \varepsilon_i, \quad j = 0, 1, 2, \dots, p \quad (2.30)$$

En forma vectorial,

$$\mathbf{x}'_j (\mathbf{y} - \boldsymbol{\mu}) = 0, \quad j = 0, 1, 2, \dots, p \quad (2.31)$$

El sistema de  $p$  ecuaciones de la *Expresión (2.31)* nos queda,

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \quad (2.32)$$

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

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad (2.33)$$

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

$$E[\hat{\boldsymbol{\beta}}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}]$$

$$E[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{y}]$$

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

$$E[\widehat{\beta}] = I\beta = \beta \quad (2.34)$$

De aquí que, reemplazando la ecuación (2.27) en el modelo de la ecuación (2.29).

$$\widehat{Y} = X\widehat{\beta} \quad (2.35)$$

$$\widehat{Y} = X(X'X)^{-1}X'y \quad (2.36)$$

$$\widehat{Y} = Hy \quad (2.37)$$

La notación de  $H$  denominada “*Matriz Hat*”, es

$$H = X(X'X)^{-1}X' \quad (2.38)$$

Donde  $H \in M_{n \times n}$  es simétrica e idempotente, esto significa que  $HH = H^2 = H$ .

De igual forma como lo hicimos en Regresión Lineal Simple utilizaremos un ejercicio tomado de *Zurita* [15]<sup>5</sup> para ilustrar Regresión Lineal Múltiple.

**Ejemplo:** Se plantea dos variables de producción  $x_1$  y  $x_2$ ; y  $y$  una característica cuantitativa que refleja mejora en la calidad de un producto que se está fabricando para consumo público.

---

<sup>5</sup> [15]Zurita, G. (2010), “*Probabilidad y Estadística, Fundamentos y Aplicaciones*”, Ediciones del Instituto de Ciencias Matemáticas, pág. 584-585, Guayaquil, Ecuador

Nueve observaciones efectuadas en la línea de producción son los que se muestran en la *Tabla 2.3*.

$x_{1i}$	0.38	0.42	0.51	0.63	0.64	0.72	0.75	0.85	0.85
$x_{2i}$	2.00	2.05	2.10	2.15	2.30	2.35	2.40	2.50	2.60
$y_i$	3.20	3.15	3.70	3.76	3.87	4.67	5.73	6.83	7.43

**Tabla 2.3 Datos de Calidad de Producto**

*Fuente: G. Zurita*

De aquí se plantea el modelo para medir la calidad del producto que se fabrica, en base a las dos variables de producción, el modelo propuesto es:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, \quad i = 1, 2, \dots, 9$$

Expresado en forma matricial:

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

Es decir,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_9 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ \vdots & \vdots & \vdots \\ 1 & x_{91} & x_{92} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_9 \end{pmatrix}$$

$$\begin{pmatrix} 3.20 \\ 3.15 \\ \vdots \\ 7.43 \end{pmatrix} = \begin{pmatrix} 1 & 0.38 & 2.00 \\ 1 & 0.42 & 2.05 \\ 1 & 0.51 & 2.10 \\ \vdots & \vdots & \vdots \\ 1 & 0.85 & 2.60 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_9 \end{pmatrix}$$

Utilizando la estimación de Mínimos Cuadrados en forma matricial,

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

Al hacer el cálculo matricial,

$$\hat{\beta} = \begin{pmatrix} -12.97 \\ -1.105 \\ 8.089 \end{pmatrix}$$

El modelo de regresión estimado queda como sigue:

$$y_i = -12.97 - 1.105x_{1i} + 8.089x_{2i} + \varepsilon_i, \quad i = 1, 2, \dots, 9$$

Si deseamos estimar la calidad del producto cuando las variables de producción son  $x_{13}$  y  $x_{23}$ , queda,

$$\hat{y}_i = -12.97 - 1.105(0.51) + 8.089(2.10)$$

$$\hat{y}_i = 3.453$$

ERLA nos permite hacer el cálculo de las estimaciones Mínimos Cuadráticas en la opción de Regresión Múltiple, para una descripción más detallada de esta opción consulte el Manual de Usuario de ERLA.

## 2.7. Tabla Análisis de Varianza (ANOVA)

La *Tabla de Análisis de Varianza*, utilizada en Regresión para analizar estadísticamente algunas características del modelo  $y_i = \beta_0 + \beta_1x_i + \varepsilon_i$  y los supuestos alrededor del mismo, consiste en una tabla de cinco columnas y con un mínimo de tres

filas, en las filas se encuentran las tres fuentes de variación del modelo de Regresión, Error y Total; sus grados de libertad; las sumas y medias cuadráticas; y el estadístico de prueba F, como se muestra en la *Tabla 2.4*.

Fuentes de Variación	Grados de Libertad	Sumas Cuadráticas	Medias Cuadráticas	Estadístico de Prueba F	Valor p
Regresión	p-1	$SCR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	$MCR = \frac{SCR}{p-1}$	$F = \frac{MCR}{MCE}$	$P(F > \frac{MCR}{MCE})$
Error	n-p	$SCE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$MCE = \frac{SCE}{n-p}$		
Total	n-1	$SCT = \sum_{i=1}^n (y_i - \bar{y})^2$			

**Tabla 2.4** Tabla de Análisis de Varianza

*Fuente: G. Zurita*

La Suma Cuadrática Total definida por SCT se explica en función de la Suma Cuadrática de Regresión y la Suma Cuadrática del Error, es decir la  $SCT=SCR+SCE$ , para verificar esta ecuación partimos de la definición de la Suma Cuadrática Total.

$$SCT = \sum_{i=1}^n (y_i - \bar{y})^2$$

Podemos descomponer el término de la sumatoria de la siguiente manera

$$SCT = \sum_{i=1}^n [(\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)]^2$$

$$SCT = \sum_{i=1}^n [(\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 + 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i)]$$

Se puede probar que

$$\sum_{i=1}^n 2(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$$

Finalmente queda

$$SCT = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

La expresión  $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  se denomina a la Suma Cuadrática de Regresión y  $\sum_{i=1}^n (y_i - \hat{y}_i)^2$  a la Suma Cuadrática del Error.

Lo deseable es que la SCE sea lo más “pequeña” posible con respecto a la SCT; el modelo sería idealmente ajustado si la SCT es igual a la SCR; se daría la doble condición, primero, no se ha cometido error al medir y luego, el modelo propuesto es válido para tales datos es decir  $\hat{y}_i = y_i$ . Por cierto que la variación total contenida en los datos estaría explicada por el Modelo de Regresión.

Se pueden representar las Sumas Cuadráticas en forma matricial en función de la “Matriz Hat” representada por  $\mathbf{H}$ , como se muestra en las ecuaciones (2.39), (2.40) y (2.41).

$$SCT = \mathbf{Y}^T \left[ \mathbf{I} - \left( \frac{1}{n} \right) \mathbf{J} \right] \mathbf{Y} \quad (2.39)$$

$$SCE = \mathbf{Y}^T [\mathbf{I} - \mathbf{H}] \mathbf{Y} \quad (2.40)$$

$$SCT = \mathbf{Y}^T \left[ \mathbf{H} - \left( \frac{1}{n} \right) \mathbf{J} \right] \mathbf{Y} \quad (2.41)$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (2.42)$$

$$\mathbf{H} \in M_{n \times n}$$

$$\mathbf{J} = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & 1 & \dots & 1 \end{bmatrix}; \quad \mathbf{J} \in M_{n \times n} \quad (2.43)$$

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}; \quad \mathbf{I} \in M_{n \times n} \quad (2.44)$$

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix}; \quad \mathbf{Y} \in M_{n \times 1} \quad (2.45)$$

De la relación entre la Suma Cuadrática y los grados de libertad, tanto del Error como el de Regresión se obtienen las Medias Cuadráticas para cada una de las fuentes de variación. Se puede



probar que la Media Cuadrática del Error es un estimador insesgado de la varianza del modelo, es decir

$$E[MCE] = \sigma^2 \quad (2.46)$$

El Coeficiente de Determinación  $R^2$  tratando de medir la calidad del modelo de Regresión estimado, se lo define como la división entre la Suma Cuadrática de Regresión y la Suma Cuadrática Total.

$$R^2 = \frac{SCR}{SCT} \quad (2.47)$$

Este coeficiente nos indica la proporción del total de la variación de la variable  $y$  explicada por el modelo de Regresión, es decir si el coeficiente de determinación es uno el modelo explicara el 100% de la variación de la variable aleatoria  $Y$ .

Evidentemente que este coeficiente es no negativo y su valor se encuentra entre cero y uno. La Potencia de Explicación del Modelo es definida como el porcentaje:

$$R^2 100\% \quad (2.48)$$

## 2.8. Intervalos de Confianza

Suponiendo que los errores del modelo de regresión lineal se distribuye en forma normal con media cero y varianza constante  $\sigma^2$  y que además son independientes entre sí, se puede establecer que un intervalo de confianza del  $100(1 - \alpha)\%$  para el coeficiente de regresión lineal  $\beta_i$  es

$$\hat{\beta}_i - t_{(n-p-1, \alpha/2)} \frac{\sigma_{\hat{\beta}_i}}{\sqrt{n}} \leq \beta_i \leq \hat{\beta}_i + t_{(n-p-1, 1-\alpha/2)} \frac{\sigma_{\hat{\beta}_i}}{\sqrt{n}} \quad (2.49)$$

Donde  $t_{(n-p-1, \alpha/2)}$  y  $t_{(n-p-1, 1-\alpha/2)}$  representa a los percentiles del  $(\alpha/2)100\%$  y de  $(1 - \alpha/2)100\%$  de una distribución *t de student* con  $(n - p - 1)$  grados de libertad, siendo  $p$  el número de variables de explicación en el modelo.

## 2.9. Contraste de Hipótesis

Supongamos que tenemos el modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i \quad (2.50)$$

con las condiciones

$$E[\varepsilon_i] = 0 \quad VAR(\varepsilon_i) = \sigma^2 \quad COV(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$$

definimos un contraste de hipótesis para comprobar estadísticamente que los valores de los  $\beta_i$  son diferentes de cero,

de tal manera que las variables de explicación sean significativas para el modelo. El contraste de hipótesis queda como sigue

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0 \quad (2.51)$$

vs

$$H_1: \text{al menos un } \beta_i \neq 0$$

Se utiliza el estadístico  $F_0 = \frac{MCR}{MCE}$ , como estadístico de prueba,

$F_0$  tiene distribución  $F_{(p-1, n-p)}$  por lo que con  $(1 - \alpha)\%$  de confianza rechazamos la  $H_0$  en favor de  $H_1$  si y solo si

$$F > F_{\alpha(p-1, n-p)} \quad (2.52)$$

Se puede demostrar que  $F$  sigue una distribución  $F$  de Fisher utilizando el *Teorema de Cochran*.

### 2.9.1. Teorema de Cochran

Sean  $x_1, x_2, \dots, x_n$   $n$  variables aleatorias independientes e idénticamente distribuidas con distribución  $N(\mu, \sigma^2)$ . Sea

$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$  con distribución  $\bar{x} \sim N(\mu, \frac{\sigma^2}{n})$ , entonces

$$\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\sigma^2} \sim \chi_{(n-1)}^2 \quad (2.53)$$

**Demostración:**

$$\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\sigma^2} = \frac{1}{\sigma^2} \mathbf{X}^T \left( \mathbf{I} - \frac{1}{n} \mathbf{J} \mathbf{J}^T \right) \mathbf{X},$$

$$\mathbf{J} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}; \in M_{n \times 1} \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}; \mathbf{I} \in M_{n \times n} \quad \mathbf{X} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}; \mathbf{X} \in M_{n \times 1}$$

Sea  $\mathbf{A} = \mathbf{I} - \frac{1}{n} \mathbf{J} \mathbf{J}^T$ , que al ser simétrica ( $\mathbf{A} = \mathbf{A}^T$ ), es diagonalizable ortogonalmente, por lo tanto se puede expresar de la siguiente forma,

$$\mathbf{A} = \mathbf{P} \mathbf{\Lambda} \mathbf{P}^T \quad (2.54)$$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix}; \mathbf{\Lambda} \in M_{n \times n}$$

$$\mathbf{P} = [\mathbf{p}_{.1} \quad \mathbf{p}_{.2} \quad \mathbf{p}_{.3} \quad \cdots \quad \mathbf{p}_{.n}]; \mathbf{P} \in M_{p \times n}$$

Donde  $\mathbf{\Lambda}$  contiene los valores propios de  $\mathbf{A}$  cuyos correspondientes vectores propios ortogonales son  $\mathbf{p}_{.1}, \mathbf{p}_{.2}, \dots, \mathbf{p}_{.n}$ . La matriz  $\mathbf{A}$  es idempotente ( $\mathbf{A} = \mathbf{A}^2$ ) por lo que sus valores propios son ceros o unos.

Definimos la

$$\begin{aligned} \text{tr}(\mathbf{A}) &= \lambda_1 + \lambda_2 + \dots + \lambda_n \\ \text{tr}(\mathbf{A}) &= \text{tr}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\mathbf{J}^T\right) = n - \frac{1}{n}n = n - 1 \quad (2.55) \end{aligned}$$

Debido a que los valores propios de  $\mathbf{A}$  son ceros o unos y su traza es  $(n-1)$ , entonces  $\lambda_1 = \lambda_2 = \dots = \lambda_{n-1} = 1 \wedge \lambda_n = 0$ . Ya que el último valor propio es cero podemos prescindir de él, de tal forma,

$$\mathbf{\Lambda}_- = \begin{bmatrix} \lambda_1 & 0 & 0 & \dots & 0 \\ 0 & \lambda_2 & 0 & \dots & 0 \\ 0 & 0 & \lambda_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_{n-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} = \mathbf{I};$$

$$\mathbf{I} \in M_{(n-1) \times (n-1)}$$

$$\mathbf{P}_- = [\mathbf{p}_{.1} \quad \mathbf{p}_{.2} \quad \mathbf{p}_{.3} \quad \dots \quad \mathbf{p}_{.n-1}]; \quad \mathbf{P}_- \in M_{px(n-1)} \quad (2.56)$$

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T = \mathbf{P}_-\mathbf{I}\mathbf{P}_-^T = \mathbf{P}_-\mathbf{P}_-^T \quad (2.57)$$

Denotamos

$$S = \frac{1}{\sigma^2} \mathbf{X}^T \left( \mathbf{I} - \frac{1}{n} \mathbf{J}\mathbf{J}^T \right) \mathbf{X} = \frac{1}{\sigma^2} \mathbf{X}^T (\mathbf{A}) \mathbf{X} = \frac{1}{\sigma^2} \mathbf{X}^T (\mathbf{P}_- \mathbf{P}_-^T) \mathbf{X}$$

De la expresión  $S$ , definimos,  $\mathbf{Z} = \mathbf{P}_-^T \mathbf{X}$ ;  $\mathbf{Z} \in M_{(n-1) \times 1}$

Por lo tanto,  $S = \frac{1}{\sigma^2} \mathbf{Z}^T \mathbf{Z}$ ;  $\mathbf{Z}^T \mathbf{Z}$  es una forma cuadrática que representa la suma de (n-1) variables aleatorias al cuadrado en este caso normales. Se puede demostrar que  $E[\mathbf{Z}] = \mathbf{0}$  y que  $VAR(\mathbf{Z}) = \sigma^2 \mathbf{I}$ , es decir,

$$S = \frac{1}{\sigma^2} (Z_1^2 + Z_2^2 + \dots + Z_{n-1}^2); Z_i \sim N(0, \sigma^2) \quad (2.58)$$

De aquí que,

$$S = \left[ \left( \frac{Z_1}{\sigma} \right)^2 + \left( \frac{Z_2}{\sigma} \right)^2 + \dots + \left( \frac{Z_{n-1}}{\sigma} \right)^2 \right]; \frac{Z_i}{\sigma} \sim N(0,1)$$

Nótese que S es la suma de n-1 variables normales estándar al cuadrado por lo que se concluye que,

$$S = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\sigma^2} \sim \chi_{(n-1)}^2 \quad (2.59)$$

A partir de este teorema,  $\frac{SCT}{\sigma^2} = \sum_{i=1}^n \frac{(y_i - \bar{y})^2}{\sigma^2}$  tiene una distribución  $\chi_{(n-1)}^2$ , además  $\frac{SCE}{\sigma^2} = \frac{(n-p)MCE}{\sigma^2}$  sigue una distribución  $\chi_{(n-p)}^2$ , siendo la  $SCT = SCR + SCE$ , entonces,

$$\frac{SCR}{\sigma^2} = \frac{SCT}{\sigma^2} - \frac{SCE}{\sigma^2} \quad (2.60)$$

La distribución de  $\frac{SCR}{\sigma^2}$  es igual a la resta de  $\chi^2_{(n-1)} - \chi^2_{(n-p)}$ , teniendo en cuenta que  $\chi^2_{(n)} = \sum_{i=1}^n \chi^2_{(1)}$ , entonces  $\frac{SCR}{\sigma^2} \sim \chi^2_{(p-1)}$ .

Siendo,

$$F = \frac{MCR}{MCE} = \frac{\frac{SCR}{(p-1)\sigma^2}}{\frac{SCE}{(n-p)\sigma^2}} \quad (2.61)$$

Nos queda la división de una variable  $\chi^2_{(p-1)}$  para sus grados de libertad sobre una variable  $\chi^2_{(n-p)}$  para sus grados de libertad, esto permite afirmar que F es una variable aleatoria F de Fisher con (p-1) gl. en el numerador y (n-p) gl. en el denominador.

Lo que se espera es Rechazar  $H_0$ . Una vez que existe evidencia estadística para Rechazar  $H_0$  mediante otro contraste de Hipótesis para cada uno de los  $\beta_i$  buscamos cual es significativo para el modelo. El contraste de Hipótesis para cada  $\beta_i$  es el siguiente

$$H_0: \beta_i = 0, i = 1, 2, \dots, p \quad (2.62)$$

vs.

$$H_1: \beta_i \neq 0$$

Sea  $T = \frac{\hat{\beta}_1 - \beta_1}{S_{\beta_1}}$ , con  $(1 - \alpha)\%$  de confianza rechazamos la  $H_0$  en favor de  $H_1$  si y solo si

$$|T| > t_{\frac{\alpha}{2}, (n-p)} \quad (2.63)$$

Con los datos de la *Tabla 2.3* relativos al Ejemplo a la Calidad de un Producto se procederá

$$SCR = \sum_{i=1}^n (\hat{y} - \bar{y})^2 = 18.222$$

$$SCE = \sum_{i=1}^n (y_i - \hat{y})^2 = 2.054$$

Fuentes de Variación	Grados de	Sumas Cuadráticas	Medias Cuadráticas	Estadístico de Prueba F	Valor p
Regresión	2	18.222	9.111	26.614	0.001
Error	6	2.054	0.342		
Total	8	20.276			

**Tabla 2.5** *Tabla de Análisis de Varianza de la Calidad de un Producto*  
Autores: Minalla y Solórzano

Coefficiente de Determinación  $R^2 = 0.8987$

Para comprobar la Hipótesis Nula  $H_0: \beta_1 = \beta_2 = 0$  vs  $H_1: \text{al menos un } \beta_i \neq 0$  utilizamos el valor p de la *Tabla 2.5* cuyo valor es 0.001 dado que es “pequeño” podemos concluir que existe evidencia estadística para rechazar la Hipótesis Nula, es decir al menos un  $\beta_i$  es significativo para el modelo. Ahora tendremos que realizar el



Contraste de Hipótesis para cada  $\beta_i$  y determinar cuáles de estos aportan al modelo planteado.

$$H_0: \beta_i = 0, i = 1, 2$$

vs.

$$H_1: \beta_i \neq 0$$

	T	Valor p
$\beta_1$	-0.232	0.821
$\beta_2$	2.09	0.081

**Tabla 2.6 Estadísticos de prueba para cada  $\beta$  de Calidad de un producto**  
**Autores: Minalla y Solórzano**

Los datos de la *Tabla 2.6* se obtuvieron utilizando el Software ERLA que nos calculo el estadístico de Prueba t y el valor p del Contraste de Hipótesis para cada uno de los  $\beta$ .

Observando los valores p de la *Tabla 2.6* para el coeficiente  $\beta_1$  el valor es de 0.821 el cual es cercano a 1 por lo que no existe evidencia estadística que nos permita Rechazar la Hipótesis Nula caso contrario con el el valor p de  $\beta_2$  el cual es menor a 0.10.

# CAPÍTULO III

## 3. REGRESIÓN RIDGE

### 3.1. Introducción

A fin de afrontar los diversos problemas de multicolinealidad; es decir que no existe independencia entre las variables de explicación, en este Capítulo se analiza con detalle la Regresión Ridge, la estimación de sus parámetros y las propiedades de sus estimadores.

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. A continuación se presentan técnicas que pueden solucionar los problemas que pueden plantearse en una regresión, o al menos diagnosticarlos con más precisión, entre estos métodos se encuentran la *Regresión Ridge* y la *Regresión de Componentes*

*Principales.* Sin embargo, la Regresión Ridge es la que presenta una mejor precisión en el proceso de estimación, en el ajuste de los estimadores, lo que hace más atractivo su uso.

### 3.2. Multicolinealidad

Uno de los supuestos del modelo de regresión lineal  $y = X\beta + \varepsilon$  establece que las variables de explicación son linealmente independientes, es decir la igualdad

$$\lambda_1 x_1 + \lambda_2 x_2 + \dots + \lambda_p x_p = 0 \quad (3.1)$$

Se cumple para  $\lambda_1 = \lambda_2 = \dots = \lambda_p = 0$ . Este supuesto asegura que la matriz  $X$  de orden  $n \times p$  tenga un rango igual a  $p$ , el número de parámetros del modelo. De aquí que la matriz  $X'X$  de orden  $p \times p$  también tenga rango igual a  $p$ . En definitiva el supuesto de ausencia de Multicolinealidad garantiza que el sistema de ecuaciones normales:

$$X'X\beta = X'y$$

Tenga solución única y está viene dada por el estimador de mínimos cuadrados:

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

Ahora bien, cuando las columnas de la matriz de diseño  $X$  son linealmente dependientes, entonces el rango de la matriz  $X$  será menor a  $p$ , la matriz  $X'X$  es singular y el estimador de mínimos cuadrados es indeterminado.

Consideremos el modelo de regresión lineal múltiple:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (3.2)$$

Y supongamos que  $X_{2i} = kX_{1i}$ , donde  $k$  es un número real conocido, esta ecuación presenta multicolinealidad exacta y no es posible obtener los estimadores de mínimos cuadrados de los coeficientes. Sin embargo sí es posible estimar combinaciones de los parámetros. Incorporando la igualdad  $X_{2i} = kX_{1i}$  en la ecuación de regresión:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 k X_{1i} + \varepsilon_i$$

Quedando,

$$Y_i = \beta_0 + (\beta_1 + \beta_2 k) X_{1i} + \varepsilon_i$$

Lo que nos da un modelo de regresión lineal simple de  $Y_i$  sobre  $X_{1i}$  que puede estimarse por mínimos cuadrados. Ahora bien, no es posible obtener estimaciones individuales de  $\beta_1$  y  $\beta_2$ . De aquí, la principal consecuencia de la multicolinealidad, es que no

podemos medir los efectos individuales de las variables de explicación. En otras palabras, no podemos estimar la respuesta de  $Y_i$  ante un cambio unitario de  $X_{1i}$ , porque al cambiar  $X_{1i}$  también lo hace  $X_{2i}$ . Solo podemos estimar la respuesta de  $Y_i$  ante un cambio unitario de  $X_{1i}$  y un cambio igual a  $k$  veces de  $X_{2i}$ , respuesta que es igual a  $\beta_1 + \beta_2 k$ . En este sentido la multicolinealidad no es un problema para la predicción de la variable  $Y_i$ .

Ahora bien, la precisión o acuracidad de un estimador insesgado se define en relación a la inversa de su varianza, cuanto menor sea la varianza de un estimador, mayor será la precisión del mismo. Siendo la matriz de varianzas y covarianzas de los estimadores,

$$\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \quad (3.3)$$

La varianza de  $\hat{\beta}_i$ , puede expresarse como,

$$\text{Var}[\hat{\beta}_i] = \text{Var}[\hat{b}_i] \frac{1}{1-R_i^2} \quad (3.4)$$

Donde  $\text{Var}[\hat{b}_i]$ , es la varianza de la estimación del modelo de regresión lineal simple de  $Y_i$  sobre  $X_i$ , y  $R_i^2$  es el coeficiente de determinación en el modelo de regresión múltiple de  $X_i$  sobre las restantes variables de explicación.

La ecuación 3.4 Muestra que la varianza de  $\hat{\beta}_i$  aumenta a medida que aumenta  $R_i^2$ . Esta varianza toma el valor más pequeño  $\text{Var}[\hat{b}_i]$ , cuando las variables de explicación son ortogonales, es decir  $R_i^2 = 0$ , y el valor máximo cuando existe multicolinealidad exacta,  $R_i^2 = 1$ . Entre estos extremos, podemos graduar la multicolinealidad como alta, media o baja por medio del factor de inflación de la varianza que viene dada por la siguiente expresión,

$$\text{FIV}_i = \frac{1}{1-R_i^2} \quad (3.5)$$

La principal consecuencia de una multicolinealidad alta,  $R_i^2 \cong 1$ , es que las varianzas de las estimaciones asociadas a las variables colineales son muy altas. Ahora bien, aún siendo muy grandes, son las menores que puede alcanzar un estimador insesgado.

### 3.3. Definición

En los modelos de Regresión Lineal cuando la correlación entre las variables de explicación causa que la matriz  $\mathbf{X}'\mathbf{X}$  no sea inversible al estimar los parámetros por Mínimos Cuadrados estos estimadores tendrán varianza alta. El problema con el

Método de Mínimos Cuadrados es el requisito que  $\hat{\beta}$  sea un *estimador insesgado* de  $\beta$  de los cuales no hay garantía de que su varianza sea pequeña. Este caso se presenta en el *Gráfico 3.1(a)* donde se ve la distribución de muestreo de  $\hat{\beta}$ , el estimador insesgado de  $\beta$ . La varianza de  $\hat{\beta}$  es grande, y eso implica que los intervalos de confianza serán grandes, y el estimado de punto de  $\hat{\beta}$  sea muy inestable. Una forma de aliviar este problema es eliminar el requisito que el estimador de  $\beta$  sea insesgado.

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  $\beta$  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:

$$ECM(\hat{\beta}_R) = \text{Var}(\hat{\beta}_R) + [E(\hat{\beta}_R) - \beta]^2 \quad (3.6)$$

es decir

$$ECM(\hat{\beta}_R) = \text{Var}(\hat{\beta}_R) + (\text{sesgo de } \hat{\beta}_R)^2 \quad (3.7)$$

El Error Cuadrático Medio (ECM) no es más que la distancia esperada, de  $\hat{\beta}_R$  a  $\beta$ , 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 el *Gráfico 3.1* se presenta el caso en el que la varianza del estimador sesgado es bastante menor que la del estimador insesgado. En consecuencia, los intervalos de confianza de  $\beta_i$  serán mucho más angostos si se usa el estimador sesgado. La pequeña varianza del estimador sesgado implica también que  $\hat{\beta}_R$  es un estimador más estable de  $\beta$  que el estimador insesgado de  $\hat{\beta}$ .

Denotamos en la *Ecuación (3.3)* el estimador de Mínimos Cuadrados en forma matricial.

$$\hat{\beta} = (X'X)^{-1}X'y \quad (3.8)$$

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]. El Estimador Ridge se determina modificando las ecuaciones normales de (3.3). En forma específica, el estimador de ridge  $\hat{\beta}_R$  se define como la solución de

$$(\mathbf{X}'\mathbf{X} + k\mathbf{I})\hat{\beta}_R = \mathbf{X}'\mathbf{y} \quad (3.9)$$

que es

$$\hat{\beta}_R = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y} \quad (3.10)$$

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

### 3.3.1. 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.

### 3.3.2. Media de Estimadores de Ridge

$$E[\widehat{\beta}_R] = E[(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}] \quad (3.11)$$

$$\begin{aligned} &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}(\mathbf{X}'\mathbf{X})E[\widehat{\beta}] \\ &= Z_k\widehat{\beta} \end{aligned} \quad (3.12)$$

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

### 3.3.3. Varianza de Estimadores de Ridge

$$\text{Var}[\widehat{\beta}_R] = \text{Var}[(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}] \quad (3.13)$$

$$\begin{aligned} &= (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\text{Var}[\mathbf{y}]\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1} \end{aligned} \quad (3.14)$$

De el *Gráfico 3.2* la distribución muestral de los estimadores por mínimos cuadrados y de estimadores Ridge. Podemos observar cómo, si bien el valor esperado de los estimadores de Ridge no es igual al valor del parámetro, su varianza, hace que la probabilidad que, un valor estimado por regresión Ridge esté alejado del valor del parámetro, sea menor que la probabilidad por mínimos cuadrados, en el caso en el que exista colinealidad entre las variables independientes.

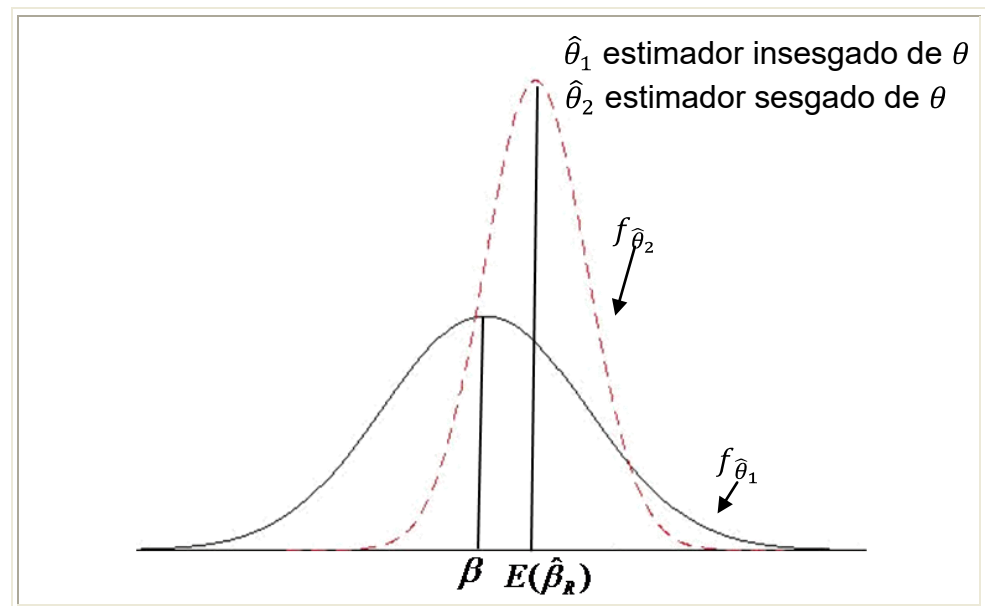


Gráfico 3.1 Distribución de Estimadores de Ridge y Mínimos Cuadrados

### 3.4. 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  $\beta^T \beta$  sea acotado.

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

### 3.4.1. Traza de Ridge

Un primer método a utilizar es la Traza de Ridge el cual es un grafico bidimensional de los elementos de  $\hat{\beta}_{Ri}(k)$  estimadores Ridge en función de  $k$  y los valores de  $k$  en un intervalo entre  $[0,1]$ . En este método calculamos los estimadores de los  $\hat{\beta}_{Ri}(k)$  para diferentes valores de  $k$  graficando para cada  $\hat{\beta}_{Ri}$  la curva de coeficientes estimados. Véase *Gráfico 3.2*.

A aumentar  $k$ , algunos de los Estimados Ridge variarán en forma dramática. En cierto valor de  $k$  se estabilizarán los estimadores ridge  $\hat{\beta}_{Ri}$ . El objetivo es seleccionar un valor de  $k$  “razonablemente pequeño”, en el cual los estimadores ridge de  $\hat{\beta}_{Ri}$  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.

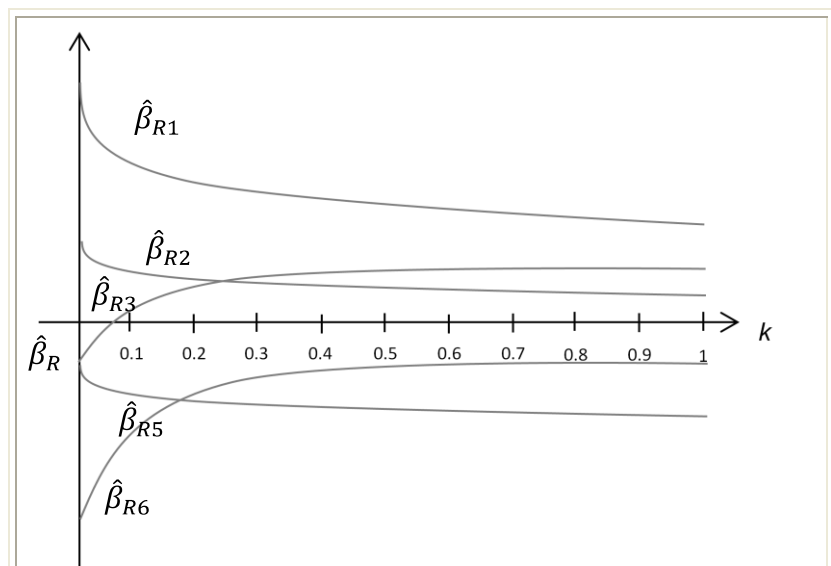


Gráfico 3.2 Traza de Ridge

### 3.4.2. 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

$$k = \frac{p\hat{\sigma}^2}{\hat{\beta}^T\hat{\beta}} \quad (3.15)$$

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

El Software ERLA (Estadística Regresión Lineal Avanzada) diseñado a través de nuestro proyecto de graduación consta de un modulo específico en donde se puede calcular el valor de  $k$  utilizando cualquiera de estos dos métodos: la Traza de Ridge y el Método Analítico. Para una mejor ilustración se pide al lector consultar el Anexo de este documento, en donde se presenta una explicación detallada de cómo usar este modulo.

Para ilustrar el uso de regresión Ridge tomaremos un ejemplo del libro “*Probabilidad y Estadística Fundamentos y*

*Aplicaciones*” de G. Zurita el cual se ajusta a las necesidades de nuestro problema:

**Ejemplo:** Los integrantes de una cooperativa arrocera están interesados en que se les ayude a conocer el efecto de la humedad, en cm/m<sup>2</sup> y la temperatura, en grados Celsius, sobre la producción de arroz  $y$ , en toneladas métricas por hectárea. Se presentan los siguientes datos:

<b>Temperatura</b>	25	27	29	31	32	33	34	35	36	30	24
<b>Humedad</b>	17	18	19	22	25	27	29	33	23	20	15
<b>Producción</b>	5.3	5.2	5.1	5.0	4.8	4.9	5.3	3.2	3.1	5.9	5.2

**Tabla 3.1 Datos de efecto de humedad, temperatura y producción de arroz**  
Fuente: G. Zurita

El modelo nos queda de la siguiente forma:

$$\text{Producción} = \beta_0 + \beta_1 \text{Temperatura} + \beta_2 \text{Humedad} + \varepsilon_i$$

Para determinar el modelo de Regresión a utilizar observamos la Matriz de Correlación de los datos que se presentan en la *Tabla 3.1*, podemos observar que todos los datos tienen una correlación mayor 0.5 y existe una

correlación de 0.8177 entre Temperatura y Humedad que son nuestras variables de explicación. Para ilustrar el problema de la alta correlación entre las variables de explicación realizaremos el análisis de Regresión tanto con el Método de Mínimos Cuadrados como con Regresión Ridge.

	Temperatura	Humedad	Producción
Temperatura	1	0.8177	-0.6746
Humedad	0.8177	1	-0.5235
Producción	-0.6746	-0.5235	1

**Tabla 3.2 Matriz de Correlación**  
*Autor: Solórzano*

El resultado de la estimación por el Método de Mínimos Cuadrados para cada uno de los  $\beta_i$  se muestra en la *Tabla 3.3* y por Regresión Ridge se presentan en la *Tabla 3.7*.

Coefficiente	Valor estimado	Error Estándar	T	Valor p
$\beta_{0MC}$	9	2.1639	4.2808	0.0027
Temperatura	-0.1547	0.1159	-13.347	0.2187
Humedad	0.0124	0.0836	0.1485	0.8856

**Tabla 3.3 Estimación por el Método de Mínimos Cuadrados**  
*Autores: Minalla y Solórzano*



<b>Coefficiente</b>	<b>Valor estimado</b>	<b>Error Estándar</b>	<b>T</b>	<b>Valor p</b>
$\beta_{OR}$	8	2.0505	3.8818	0.0047
<b>Temperatura</b>	-0.0927	0.1117	-0.8297	0.4308
<b>Humedad</b>	-0.0144	0.0838	-0.1721	0.8676

**Tabla 3.4 Estimación por Regresión Ridge**  
**Autores: Minalla y Solórzano**

De los modelos calculados procedemos a hacer la estimación de la observación  $y_1$  reemplazando en el modelo estimado los valores de  $x_1$  y  $x_2$  tanto para Mínimos cuadrados como para Ridge.

**Modelo estimado por Regresión Ridge**

$$\text{Producción} = 8 - 0.0927 \text{ Temperatura} - 0.0144 \text{ Humedad}$$

$$\text{Producción} = 8 - 0.0927 (25) - 0.0144 (17)$$

$$\text{Producción} = 5.3975$$

**Modelo estimado por Mínimos Cuadrados**

$$\text{Producción} = 9 - 0.1547 \text{ Temperatura} + 0.0124 \text{ Humedad}$$

$$\text{Producción} = 9 - 0.1547 (25) + 0.0124 (17)$$

$$\text{Producción} = 5.6064$$

El valor observado de  $y_1$  es igual a 5.3, el valor Ridge  $\hat{y}_{1R} = 5.3975$  y el valor por Mínimos Cuadrados es  $\hat{y}_{1MC} = 5.6064$ .

$$|y_1 - \hat{y}_{1R}| < |y_1 - \hat{y}_{1MC}|$$

$$|5.3 - 5.3675| < |5.3 - 5.6064|$$

$$0.0975 < 0.3064$$

La estimación de Regresión de Ridge es mucho más cercana al valor observado  $y_1$  que la estimación por el Método de Mínimos Cuadrados.

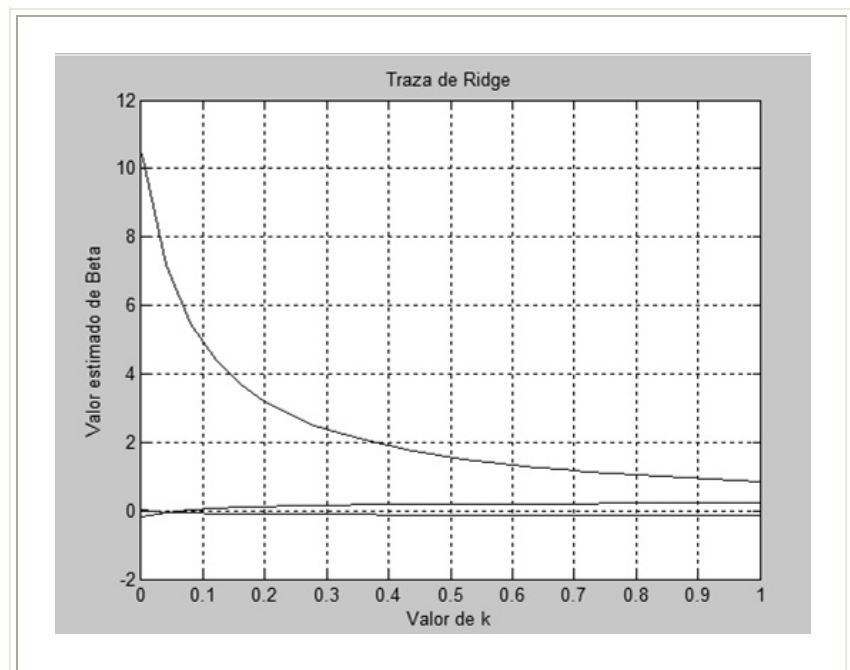


Gráfico 3.3 Traza de Ridge de datos del ejemplo

En el *Gráfico 3.3* muestra la Traza de Ridge para los datos de ejemplo sobre la producción de arroz donde se presenta las estimaciones para los  $\hat{\beta}_R$  con los diferentes valores de k. Como el resultado del algoritmo de estimación en el Software ERLA el mejor valor de k es 0.0159.

# CAPÍTULO IV

## 4. REGRESIÓN ROBUSTA

### 4.1. Introducción

En este Capítulo se analiza con detalle la Regresión Robusta, la estimación de sus parámetros y las propiedades de sus estimadores para cuando existe la presencia de valores aberrantes en la muestra.

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 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. Las suposiciones de los estimadores basados en mínimos cuadrados, sobre los

errores, no siempre se cumplen, lo cual producirá desviaciones importantes en los resultados obtenidos.

El limitado desempeño del estimador, basado en los mínimos cuadrados, es claramente visible ante la presencia de puntos cuyas mediciones hayan sido erróneas, las cuales arrastran al resto hacia un comportamiento alejado del patrón esperado. Por el contrario, los procedimientos de Regresión Robusta, tienen la capacidad de permanecer invariantes en presencia de errores “gruesos” y puntos desfavorables.

Finalmente, se comparan sus resultados con el estimador de Mínimos Cuadrados, los cuales confirman la superioridad del estimador por Regresión Robusta.

## 4.2. Definición

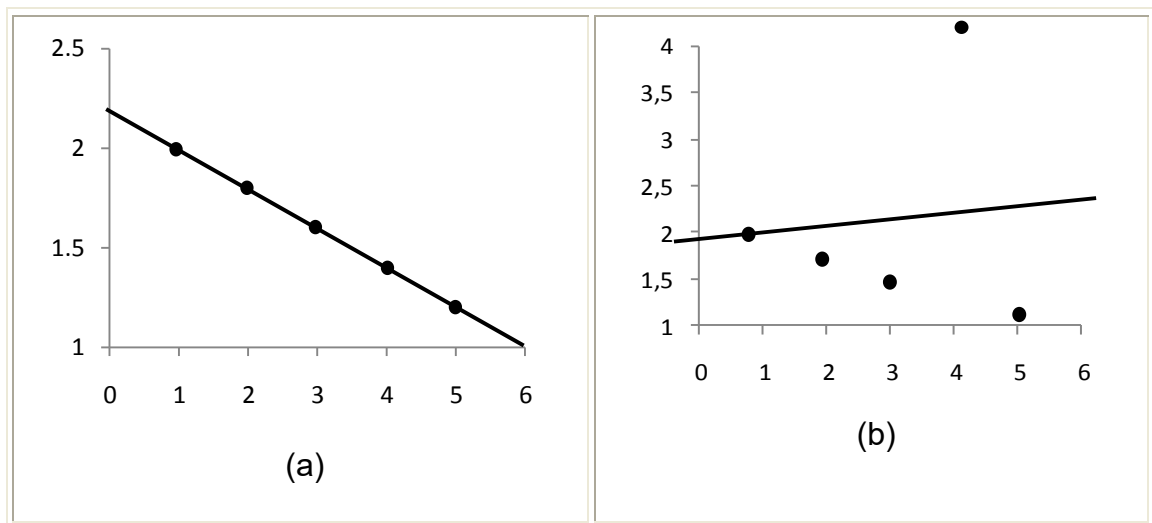
Para utilizar Regresión Lineal se debe determinar un conjunto de parámetros de un modelo que se ajusten a las variables de explicación. En general, un modelo lineal denotado por  $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$ ,  $i = 1, 2, \dots, n$ , donde,  $y_i$  es condicionado a los valores de  $x_i$  y que se lee fijando los valores de  $x_i$  y el error de varianza constante del modelo es  $\varepsilon_i$ . El

vector de estimación de los parámetros  $\hat{\beta}$  de dimensiones  $p \times 1$ , donde  $p$  es el número de parámetros del modelo. Una vez resuelto el problema de ajuste, el método entrega la estimación que mejor se “ajusta” al conjunto de datos dados. 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. Muchas veces los datos son procesados automáticamente y no se tiene acceso a ellos [10].

La teoría de estimación robusta parte de la premisa que en la práctica se puede plantear un modelo aproximado, al cual no necesariamente se ajustarán todos los datos, pero sí la mayoría de ellos. Aquellos puntos que no se ajustan al modelo utilizado son los denominados valores aberrantes, las poblaciones con la presencia de estos valores se la conoce como población contaminada, la influencia de la contaminación debe ser minimizada. El concepto de robustez se define como “la insensibilidad de los resultados a pequeñas desviaciones o contaminaciones con respecto al modelo”.

En el *Grafico 4.1* se puede observar como el Método de Mínimos Cuadrados proporciona un excelente ajuste cuando los datos no están contaminados o no tiene valores aberrantes. Sin embargo, al encontrarse con un punto que está lo suficientemente alejado de la línea que se desea ajustar, este repercute significativamente en el ajuste, por lo que el mismo se sesga en dirección del punto contaminado o valor aberrante.



**Gráfico 4.1 (a) Conjunto de cinco puntos que definen correctamente la línea (b) Mismo conjunto de puntos presentado en (a) pero con un punto aberrante**

Además del método de Mínimos Cuadrados, la regresión puede realizarse ponderando los datos con una determinada cantidad de tal forma que en la estimación del modelo algunas observaciones influyan más que otras.

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.



### 4.3. Regresión de Mínimos Cuadrados Ponderados

El propósito del ajuste robusto es realizar una buena aproximación de las ecuaciones que definen un modelo a partir de un conjunto de datos conocidos. Este método usa pesos que se basan en la distancia a la que se encuentra una observación del valor estimado del modelo, medido por el valor de los residuos de cada observación.

Una de las aplicaciones de la ponderación analítica de las observaciones es eliminar el error producido por la presencia de heterocedasticidad en los datos. Recuérdese que por este término se entiende como varianzas no constantes y que una de las consecuencias consiste en que el error estándar de los estimadores calculados por el método de mínimos cuadrados ordinarios sea sesgado.

Recordando que para el Método de Mínimos Cuadrados la suma cuadrática del error es

$$SCE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (4.1)$$

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

$$SCE = \sum_{i=1}^n w(e_i)(e_i)^2 = \sum_{i=1}^n w(y_i - \hat{y}_i)(y_i - \hat{y}_i)^2 \quad (4.2)$$

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

1. Elija la función de pesos para ponderar los casos
2. Obtenga pesos iniciales para todos los casos
3. Use los pesos iniciales en Mínimos Cuadrados Ponderados y obtenga los residuos de la función de regresión ajustada
4. Use los residuos del paso 3 para obtener los nuevos pesos
5. Continúe las iteraciones hasta obtener la convergencia

Nosotros ahora discutiremos cada uno de los pasos de la regresión robusta ponderada

### 4.3.1. 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:

$$w = \begin{cases} 1 & ; |u| \leq 1.345 \\ \frac{1.345}{|u|} & ; |u| > 1.345 \end{cases} \quad (4.3)$$

La escala residual  $u_i$  es el siguiente

$$u_i = \frac{e_i}{MAD} \quad (4.4)$$

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 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. El Gráfico 4.2 muestra la función de peso Huber. Nótese que, cada vez que la función de peso declina, la escala absoluta residual aumenta, y que la función de peso es simétrica alrededor de  $u=0$ . También en la función de Huber su peso no disminuye de 1 hasta que la escala absoluta del residuo exceda de 1.345, y además que los pesos son siempre positivos, sin importar que tan grande sea la escala residual absoluta.

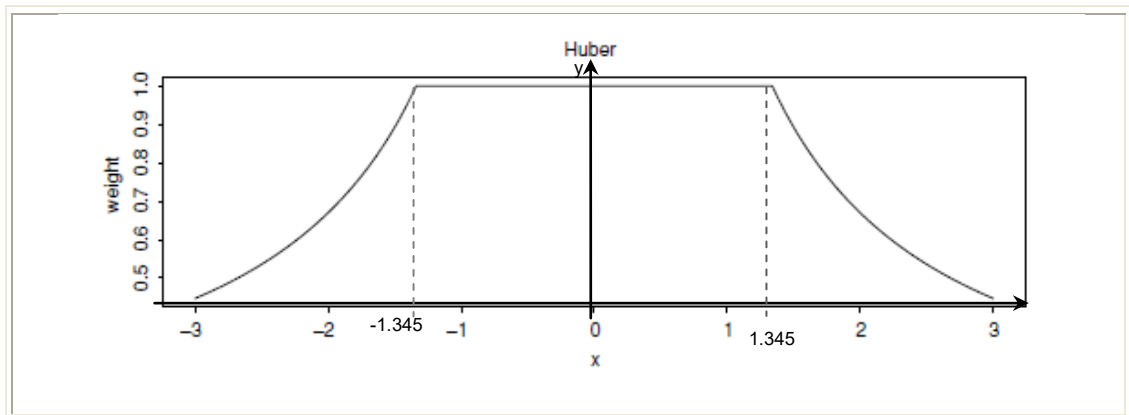


Gráfico 4.2 Función de Huber

### 4.3.2. 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.

### 4.3.3. Escala Residual

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

$$u_i = \frac{e_i}{MAD} \quad (4.5)$$

En presencia de valores aberrantes, la  $\sqrt{MSE}$  es un estimador sensible a los mismos; 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  $\sigma$  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:

$$MAD = \frac{1}{0.6745} \text{mediana}(|e_i - \text{mediana}(e_i)|) \quad (4.6)$$

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

#### 4.3.4. 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 converga. La convergencia puede ser medida observando si los pesos cambian relativamente poco, si los valores residuales cambian relativamente poco, si los coeficientes de regresión estimados cambian relativamente poco, o si los valores ajustados cambian relativamente poco.

En el modelo de Regresión Lineal minimizamos en (4.2)

$$\min\{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{RB}})' \text{diag}(\mathbf{W})(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{RB}})\} \quad (4.6)$$

Descomponiendo las matrices de (4.6) nos queda

$$(e_1 \ e_2 \ e_3 \ \dots \ e_n) \begin{pmatrix} w_1 & 0 & 0 & \dots & 0 \\ 0 & w_2 & 0 & \dots & 0 \\ 0 & 0 & w_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & w_n \end{pmatrix} \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{pmatrix} \quad (4.7)$$

Las Ecuaciones Normales del Método de Mínimos Cuadrados Ponderados son

$$\mathbf{X}'\mathbf{W}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{RB}}) = 0 \quad (4.8)$$

$$\mathbf{X}'\mathbf{W}\mathbf{Y} - \mathbf{X}'\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{RB}} = 0$$

$$\mathbf{X}'\mathbf{W}\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{RB}} = \mathbf{X}'\mathbf{W}\mathbf{Y}$$

$$\hat{\boldsymbol{\beta}}_{\text{RB}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{Y} \quad (4.9)$$

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

**Ejemplo:** Los grupos ambientalistas se preocupan de verificar la contaminación que el medio ambiente sufre por efectos de la combustión de gases e la atmosfera. En términos de tal actitud consiguen financiamiento para efectuar experimentos para medir presencia de gases contaminantes en la atmosfera. Uno de tales experimentos consiste en monitorear el efecto de la velocidad X de los vehículos en Kmts/h, y la cantidad del monóxido de carbono Y, en ppm/Km que se libera en la atmosfera; de tal experimento con diez vehículos es el siguiente:

<b>Velocidad</b>	45	50	55	60	65	70	75	80	85	90
<b>Cantidad</b>	110	118	129	159	166	171	153	159	166	169

**Tabla 4.1 Datos de la medición de gases en la atmosfera**

*Fuente: G. Zurita*

Para realizar el modelo que mejor explique la cantidad de Monóxido de Carbono de un vehículo en función de su velocidad. En el *Gráfico 4.3* se observan un grupo de valores alejados de la tendencia del resto de puntos, los que han sido señalados en el mismo como posibles valores aberrantes.

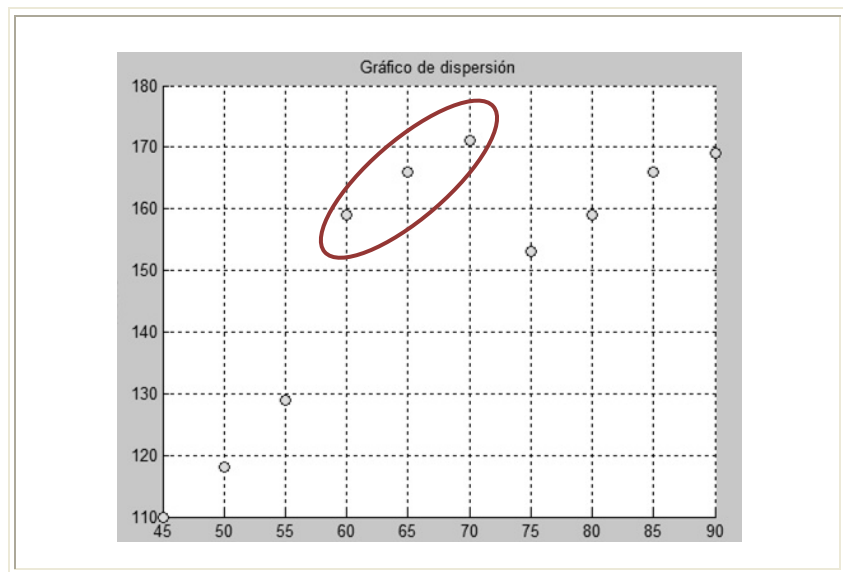


Gráfico 4.3 Gráfico de Dispersión

Dado que cerca del 30% de los datos al parecer no se ajustan al modelo, es decir son aberrantes, estimar los parámetro por Mínimos Cuadrados no sería la mejor opción debido a la sensibilidad que este método presenta a la presencia de valores aberrantes, por esta razón utilizamos la Regresión Robusta para estimar el modelo, al utilizar el método de Mínimos Cuadrados Ponderados se presenta los resultados en la *Tabla 4.2*.

En vista de la presencia de los valores aberrantes, ajustaremos una regresión lineal robusta, usando Regresión de Mínimos Cuadrados Ponderados y la



función de pesos Huber (4.3). Ilustraremos los casos para el primer par de observaciones.

El modelo bajo el criterio de mínimos cuadrados es:

$$\hat{y}_i = -15.247 + 0.5516x_i, \quad i = 1, 2, \dots, 10$$

El valor residual  $e_1 = 0.4341$ . Los valores residuales se muestran en la primera columna de la Tabla 4.2. El valor de la mediana de los diez valores residuales es  $mediana(e_1) = 0.1407$ , donde  $e_1 - mediana(e_1) = 0.4341 - 0.1407 = 0.2934$ . La mediana de los diez residuos absolutos es:

$$mediana\{e_1 - mediana(e_1)\} = 8.2450$$

Entonces el estimador MAD (4.5) es:

$$MAD = \frac{8.2450}{0.6745} = 12.2239$$

Donde, la escala residual (4.6) es:

$$u_1 = \frac{0.4341}{12.2239}$$

Los valores de las escalas residuales se muestran en la *Tabla 4.2*, segunda columna. Desde  $|u_1| = 0.4341 \leq$

1.345, el peso inicial Huber es  $w_1 = 1$ . Los pesos iniciales se muestran en la *Tabla 4.2*, columna 3

$i$	$e_i$	$u_i$	$w_i$
1	0,4341	0,0355	1
2	-0,1527	-0,0125	1
3	0,9154	0,0749	1
4	12,4648	1,0197	1
5	11,3264	0,9266	1
6	9,0846	0,7432	1
7	-5,8451	-0,4782	1
8	-7,5352	-0,6164	1
9	-8,6736	-0,7096	1
10	-12,0187	-0,9832	1

**Tabla 4.2 Iteraciones Huber con Mínimos Cuadrados Ponderados de medición de gases de la atmosfera**  
Autores: Minalla y Solórzano

Este ejemplo es considerado como un caso especial de pesos de mínimos cuadrados debido a que todos los valores de los pesos son iguales a 1, es decir que esta va a ser la única iteración que calcularemos.

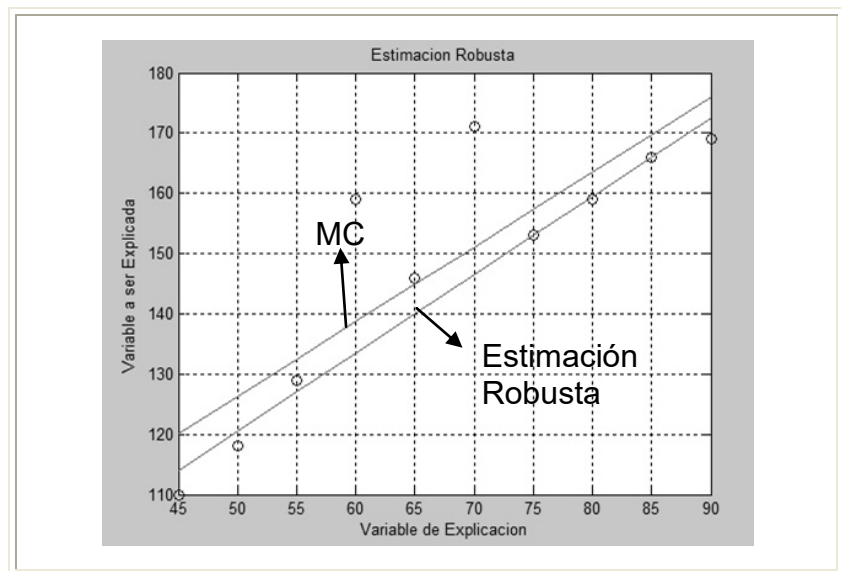
Coeficiente	Valor Estimado
$\hat{\beta}_{0RB}$	55.687
<b>Velocidad</b>	1.2975

**Tabla 4.3 Estimación por Regresión Robusta**  
Autores: Minalla y Solórzano

La estimación del modelo con Regresión Robusta queda:

$$\hat{y}_i = 55.687 + 1.2975x_i, \quad i = 1, 2, \dots, 10$$

El *Gráfico 4.4* corresponde al diagrama de dispersión de los datos con las rectas de Estimación Robusta y de Mínimos Cuadrados. La estimación robusta proporciona un mejor ajuste a la tendencia de las observaciones, en contraste con la de Mínimos Cuadrados que presenta una variación debido a los valores aberrantes.



**Gráfico 4.4 Estimación Robusta**

# CAPÍTULO V

## 5. PROGRAMACIÓN Y VALIDACIÓN

### 5.1. Introducción

Este Capítulo se trata específicamente el análisis de los modelos de estimación relatados en los Capítulos 3 y 4. Procedemos a validar los algoritmos creados en el Software ERLA analizando dos casos específicos, Regresión Ridge y Regresión Robusta.

En ambos casos se plantearán modelos en donde los parámetros de estimación son conocidos con el objetivo de observar qué tanto se asemejan los datos obtenidos por el Software de las estimaciones con los parámetros. Para Regresión Ridge tomaremos muestras y con cada una de ellas plantearemos el modelo de estimación de las que se obtendrán la distribución muestral de los parámetros estimados, analizaremos la varianza de los mismos.

En el caso de Regresión Robusta se tendrá una muestra la cual será contaminada con valores de una distribución de probabilidad, determinaremos los modelos tanto para el modelo de la muestra contaminada y no contaminada, los que serán estimados utilizando el Método de Mínimos Cuadrados Ordinarios y el Método Robusto.

## 5.2. Modulo de Regresión Ridge

Hemos creado una función en Matlab denominada `CoficienteRegresionRidge.m`, el cual recibe a “y” y “MX” donde Y representa el vector de la variable a ser explicada y MX es la matriz que contiene a la variables de explicación del modelo que permita el cálculo de los valores estimados de los Betas.

La programación de la función en Matlab es la siguiente:

```
function Bkr=CoficienteRegresionRidge(y,MX)
format long g;
d=size(MX);
n=d(1);
p=d(2)+1;
j=ones(n,1);
X=[j,MX];
I=eye(n);
J=ones(n);
A=inv(X'*X);
H=X*A*X';
SCE=y'*(I-H)*y;
MCE=SCE/(n-p);
b=A*X'*y;
SB=MCE*A;
```

```

Sb=diag(SB);
%u vectores propios
%v valores propios
[u,v]=eig(X'*X);
alpha=u'*b;
B=alpha'*alpha;
k=p*MCE/B;
Ip=eye(p);
Bk=inv(X'*X+k*Ip)*X'*y;
MCEK=(SCE+(Bk-b)'*(X'*X)*(Bk-b))/(n-p);
SBK=MCEK*inv(X'*X+k*Ip);
Sbk=diag(SBK);
Sbk;
Bkr=zeros(p,4);
Bkr(:,1)=Bk;

for i=1:p
    Bkr(i,2)=sqrt(Sbk(i));
    Bkr(i,3)=Bkr(i,1)/Bkr(i,2);
    Bkr(i,4)=abs(Bkr(i,3));
    Bkr(i,4)=tcdf(Bkr(i,4),n-p);
    Bkr(i,4)=(1-Bkr(i,4))*2;
End

```

Recordemos que Matlab interpreta a *%u vectores propios* y a *%v valores propios* como comentarios.

Usamos la siguiente subrutina de Visual Basic en donde primero se definen las Librerías de Visual entre ellas se encuentra ML\_VS que es la permite la conexión entre Visual Basic y Matlab. Entonces ejecutamos esta subrutina por medio de la GUI, el cual crea una instancia de la clase AnalysisMultivariate.

Finalmente, este llama al Método CoeficienteRegresionRidge.m de Matlab.

```

Imports MathWorks.MATLAB.NET.Utility
Imports MathWorks.MATLAB.NET.Arrays
Imports ML_VS
Public Class frmRidge

    Private Sub btnAceptar_Click(ByVal sender As System.Object,
ByVal e As System.EventArgs) Handles btnAceptar.Click
        Dim rows, columns As Integer

        rows = oWS.dgvWorksheet.Rows.Count - 1
        columns = csXY.lstX.Items.Count

        Dim yvar As New MWNumericArray(New Double(rows - 1, 0) {})
        Dim xvar As New MWNumericArray(New Double(rows - 1, columns
- 1) {})

        Dim xcol, ycol As String
        Dim xval, yval As Double

        For i As Integer = 1 To rows
            ycol = csXY.lstY.Items(0).ToString
            yval = Convert.ToDouble(oWS.dgvWorksheet(ycol, i -
1).Value)
            yvar(i, 1) = yval
            For j As Integer = 1 To columns
                xcol = csXY.lstX.Items(j - 1).ToString
                xval = Convert.ToDouble(oWS.dgvWorksheet(xcol, i -
1).Value)
                xvar(i, j) = xval
            Next
        Next
        Dim constR As MWArray = Nothing
        Dim BetasR As MWArray = Nothing
        Dim mva As New MultivariateAnalysis

        constR = mva.RidgeRegressionConstant(yvar, xvar)
        BetasR = mva.RidgeRegressionCoefficient(yvar, xvar)
        Dim ar_r2 As Array = constR.ToArray
        Dim ar As Array = BetasR.ToArray

        If (csXY.lstX.Items.Count > 0 And csXY.lstY.Items.Count > 0)
Then
            Dim oRS As New clsRegStats
            csXY.CreateVars()
            'TABLA DE COEFICIENTES
            Dim tblrow1(csXY.lstX.Items.Count), tblcol1(3) As String
            tblcol1(0) = "COEFICIENTE"
            tblcol1(1) = "COEF.E.E."
            tblcol1(2) = "T"
            tblcol1(3) = "P"
            tblrow1(0) = "CONSTANTE"
            For i As Integer = 1 To csXY.lstX.Items.Count
                tblrow1(i) = csXY.lstX.Items(i - 1).ToString
            Next
        End If
    End Sub
End Class

```

```

Next
'Reporte publicado en el webbrowser control
oWR.InsertTitle("Regresion Ridge ")
oWR.InsertParagraph("Valor de la constante k:" &
Decimal.Round(ar_r2(0, 0), Precision).ToString)
oWR.InsertParagraph("Valores estimados de los parametros
para el modelo:")
oWR.InsertTable(oRS.VSGHTMLtable("Tabla de
Estimadores", oRS.VSRidgeRegressionCoefficients(csXY.yvar,
csXY.xvar), tblcoll, tblrow1).ToString)
oWR.InsertLine()
If RidgeT.Checked = True Then
Dim oRS1 As New clsRegStats
oRS1.VStrazaridge(csXY.yvar, csXY.xvar)
End If
Me.Close()
Else
MessageBox.Show("Seleccione las variables a utilizar.",
Application.ProductName, MessageBoxButtons.OK,
MessageBoxIcon.Exclamation)
End If
End Sub

```

Podremos observar con ejemplos el resultado de estas operaciones en el Manual de Usuario que se encuentra como Anexo en este documento.

### 5.3. Modulo de Regresión Robusta

De la misma manera que en Regresión Ridge, aquí se crea una función en Matlab denominada `CoeficienteRegresionRobusta.m`, en el que se ingresa a “y” y “MX”, y que presenta como resultado las estimaciones de los parámetros a través de la GUI creada en Visual Basic. El



siguiente algoritmo ilustra el procedimiento que se utilizó para realizar dichas estimaciones.

```
function Br=CoficienteRegresionRobusta (y,MX)
format long g;
d=size(MX);
n=d(1);
p=d(2)+1;
j=ones(n,1);
X=[j,MX];
I=eye(n);
J=ones(n);
A=inv(X'*X);
H=X*A*X';
SCE=y'*(I-H)*y;
MCE=SCE/(n-p);
coef0=A*X'*y;
coef=coef0;
for i=1:20
    res = X*coef0 - y;
    weights = exp(-20*abs(res)/max(abs(res)))';
    % compute the weighted estimate using these weights
    %coef1 = lscov(X,y,weights);
    w=diag(weights);
    coef1=inv(X'*w*X)*X'*w*y
    coef0=coef1;
end
Br=coef1;
yhat = X*coef1;
yhat1 = X*coef;
close
figure('Name','Gráfico de Estimacion
Robusta','NumberTitle','on','Color',[0.6824 0.8000 0.909])
plot(MX,y,'o',MX,yhat,'-'), title 'Outliers in the data',
hold on,
plot(MX,y,'o',MX,yhat1,'-'), grid on,
title('Estimacion Robusta')
xlabel('Variable de Explicacion')
ylabel('Variable a ser Explicada')
```

De igual manera como en la programación de Regresión Ridge primero se definen las librerías, luego hacemos referencia a la función en Matlab. En la siguiente parte tomamos los datos de la Hoja de Trabajo y los transformamos en vectores

almacenándolos en la variable `csXY`. Finalmente hacemos uso de la función en Matlab `CoeficienteRegresionRobusta.m`, el resultado de las estimaciones se muestra en `WebReport`. A continuación se muestra con detalle la programación en Visual Basic.

```
Imports MathWorks.MATLAB.NET.Utility
Imports MathWorks.MATLAB.NET.Arrays
Imports ML_VS

Public Class frmRobust

    Private Sub btnAceptar_Click(ByVal sender As System.Object,
ByVal e As System.EventArgs) Handles btnAceptar.Click

        If (csXY.lstX.Items.Count > 0 And csXY.lstY.Items.Count > 0)
Then
            Dim oRS As New clsRegStats
            csXY.CreateVars ()

            'TABLA DE COEFICIENTES
            Dim tblrow1 (csXY.lstX.Items.Count), tblcol1(0) As String
            tblcol1(0) = "COEFICIENTE"
            tblrow1(0) = "CONSTANTE"
            For i As Integer = 1 To csXY.lstX.Items.Count
                tblrow1(i) = csXY.lstX.Items(i - 1).ToString
            Next

            'Reporte publicado en el webbrowser control
            oWR.InsertTitle("Regresion Robusta ")
            oWR.InsertParagraph("Valores estimados de los parametros
para el modelo:")
            oWR.InsertTable(oRS.VSGTHTMLtable("Tabla de
Estimadores", oRS.VSRobustRegressionCoefficient(csXY.yvar,
csXY.xvar), tblcol1, tblrow1).ToString)
            oWR.InsertLine ()

            Me.Close ()
        Else
            MessageBox.Show("Seleccione las variables a utilizar.",
Application.ProductName, MessageBoxButtons.OK,
MessageBoxIcon.Exclamation)
        End If
    End Sub
End Class
```

Para ambos casos, en Visual Basic se crea un Formulario que es el que se presenta al usuario, el que contiene una serie de ListBox que permiten el ingreso de las variables. Este formulario se muestra en el Manual de Usuario que se encuentra en el Anexo.

## 5.4. Validación

### 5.4.1. Regresión Ridge

Para validar el Software ERLA planteamos un ejercicio en la cual definimos una función  $Y$  en términos de dos variables  $x_1$  y  $x_2$  respectivamente, tal que

$$y = 6 + 3x_1 + 4x_2$$

haciendo que la variable  $x_1$  tome valores enteros desde 1 hasta 20 y que la variable  $x_2$  sea igual a  $2(x_1) + 5$ , de tal manera que la variable  $x_2$  este correlacionada con  $x_1$  de esta forma a la variable  $y_1$  se le suma una variable  $\varepsilon \sim N(0,16)$ , está claro que no sesgamos los datos debido a que el valor esperado de la variable es cero. La varianza es alta en relación a las variables de explicación, debido a que los Estimadores Ridge son sesgados pero disminuyen la varianza del estimador; recordemos que la

Varianza del estimador por Mínimos Cuadrados es  $\sigma_{\hat{\beta}}^2 = (X'X)^{-1}\hat{\sigma}^2$ , de esta expresión podemos ver que aumentado la varianza del error en el modelo también se incrementa la varianza del estimador ya que su relación es directamente proporcional, la Tabla 5.1 queda como sigue:

$x_1$	$x_2$	$y = 6 + 3x_1 + 4x_2$	$\varepsilon \sim N(0, 16)$	$\hat{y}_i$
1	7	37	-0,67848186	36.3215181
2	9	48	-0,50604392	47.4939561
3	11	59	0,59926208	59.5992621
4	13	70	-0,1562746	69.8437254
5	15	81	-2,41690988	78.5830901
6	17	92	-0,42661982	91.5733802
7	19	103	0,07164939	103.071649
8	21	114	-1,84883613	112.151164
9	23	125	0,50815019	125.50815
10	25	136	-3,17414403	132.825856
11	27	147	-0,10690617	146.893094
12	29	158	-2,82717069	155.172829
13	31	169	0,41653535	169.416535
14	33	180	0,62529534	180.625295
15	35	191	1,28161221	192.281612
16	37	202	0,99564058	202.995641
17	39	213	2,31691447	215.316914
18	41	224	1,75748168	225.757482
19	43	235	1,70180427	236.701804
20	45	246	0,07856513	246.078565

**Tabla 5.1 Estimación de los datos del modelo con  $\varepsilon \sim N(0, 16)$**

**Autores: Minalla y Solórzano**

Es decir se tomó una muestra de tamaño 20, donde  $\hat{y}_i$  es el estimador de  $y$ . Se realizaron 15 muestras de igual tamaño generando valores aleatorios para la variable  $\varepsilon$  y estimando para cada una de las muestras el valor de los  $\beta_i$  por Mínimos Cuadrados y por Regresión Ridge. El resultado de la estimación para cada uno de los  $\beta_i$  se muestra en la siguiente tabla.

Muestra	Mínimos Cuadrados			Regresión Ridge		
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{R0}$	$\hat{\beta}_{R1}$	$\hat{\beta}_{R2}$
1	-1.2	17.47	3.18	0.11	2.15	4.63
2	11.2	17.23	3.51	-0.16	2.62	4.74
3	4.6	17.73	3.36	0.01	2.39	4.81
4	10.52	17.08	3.46	-0.14	2.58	4.72
5	10.74	17.03	3.5	-0.15	2.57	4.7
6	4.97	17.57	3.44	0.007	2.39	4.8
7	7.12	17.42	3.47	-0.04	2.46	4.77
8	8.008	17.48	3.49	-0.06	2.49	4.79
9	10.17	16.99	3.39	-0.13	2.54	4.67
10	8.91	17.33	3.53	-0.09	2.52	4.78
11	8.16	17.22	3.47	-0.06	2.49	4.76
12	9.775	17.26	3.49	-0.11	2.55	4.75
13	5.69	17.74	3.44	-0.008	2.43	4.83
14	3.73	17.88	3.4	0.03	2.37	4.83
15	6.08	17.72	3.528	-0.016	2.45	4.85

**Tabla 5.2 Comparación de las Estimaciones con Mínimos Cuadrados vs Regresión Ridge**

*Autores: Minalla y Solórzano*

Obteniendo el valor de la media y la varianza para cada uno de los  $\hat{\beta}_i$

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{R0}$	$\hat{\beta}_{R1}$	$\hat{\beta}_{R2}$
<b>Media</b>	7.231	17.410	3.443	-0.053	2.466	4.762
<b>Varianza</b>	11.240	0.077	0.007	0.005	0.013	0.003

**Tabla 5.3 Comparación de Media y Varianza de Mínimos Cuadrados vs Regresión Ridge**

*Autores: Minalla y Solórzano*

Observamos como las medias de los valores de los  $\hat{\beta}_i$  por Regresión de Ridge son mucho más cercanos a los valores de los parámetros  $\beta_i$  a diferencia de los  $\hat{\beta}_i$  estimados por Mínimos Cuadrados. Además la varianza de los  $\hat{\beta}_i$  por Regresión de Ridge es menor a la varianza de los  $\hat{\beta}_i$  estimados por Mínimos Cuadrados.

#### 5.4.2. Regresión Robusta

Para el caso de Regresión Robusta formularemos un modelo en donde los parámetros son conocidos, con una sola variable de explicación  $x_1$ , es decir un modelo de Regresión Lineal Simple.

Supongamos que conocemos la relación funcional que liga a  $y$  con  $x$  a través de la expresión  $y = 6 + 3x$ , es fácil saber el valor que toma el valor de  $y$  dado que  $x = 4$  es  $y = 6 + 3(4) = 18$  debido a que  $y$  es una variable determinística es decir no está sujeta a incertidumbre.

Si queremos plantear un modelo de Regresión Lineal a partir de la relación funcional anterior necesitamos agregar una variable de incertidumbre en el modelo es decir el error.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (5.1)$$

Al agregar el Error en el modelo, la variable  $y$  pasa de ser una variable determinística para convertirse en una variable aleatoria.

Para ilustrar nuestro caso específico hemos planteado que el error del modelo tenga una Distribución Normal con media cero y varianza cuatro, es decir  $\varepsilon \sim N(0,4)$ .

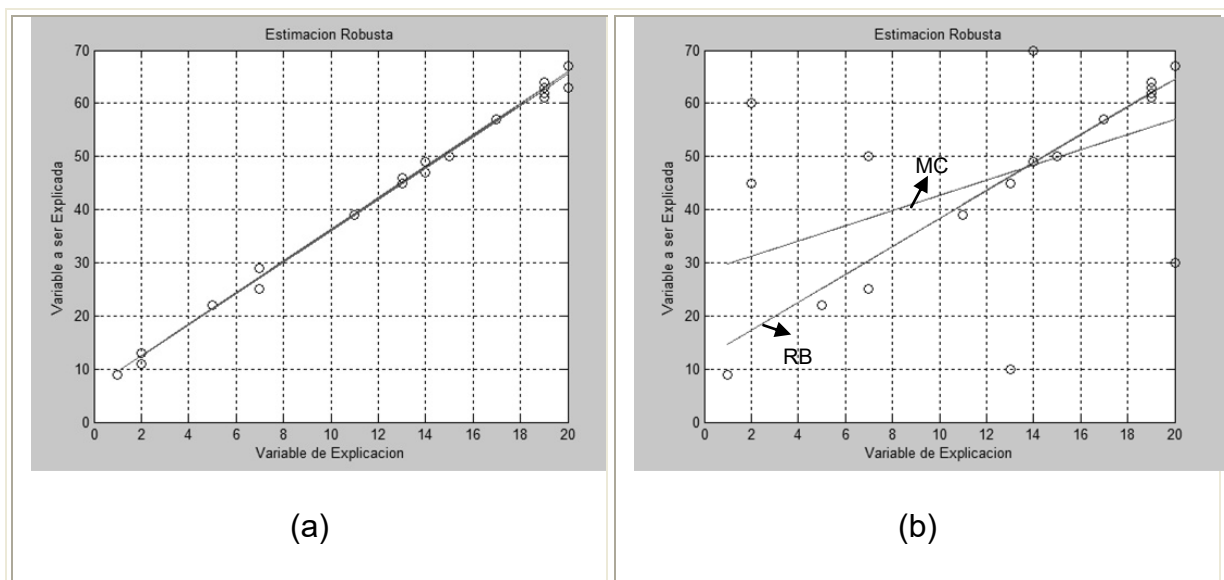
La variable  $x$  es una variable determinística, por tal razón no está sujeta a incertidumbre; a partir de esta situación

hemos tomado una muestra de tamaño 20 y fijamos a conveniencia que  $x$  tome valores comprendidos entre 1 y 20.

Al incluir estos valores a la expresión (5.1) nos da como resultado los valores correspondientes a la variable aleatoria  $y$ . Con los datos resultantes de las variables  $y$  y  $x$  procederemos a estimar los parámetros del modelo. Recordemos que los parámetros son conocidos, es decir

$$y_i = 6 + 3x_i, \quad i = 1, 2, \dots, 20 \quad (5.2)$$

obtenidos del modelo determinístico planteado.



**Gráfico 5.1** Estimación Robusta y de Mínimos Cuadrados (a) Muestra No Contaminada (b) Muestra Contaminada



Para efectos del estudio hemos contaminado la muestra con un 30% de valores aberrantes, hemos estimado los modelos de Regresión. Como podemos observar en el *Gráfico 5.2* el ajuste por el Método de Mínimos Cuadrados fue tan bueno como el Método de Regresión Robusta debido a que los puntos se ajustan muy bien a las rectas por ambos métodos.

Muestra	Mínimos Cuadrados		Regresión Robusta	
	CONSTANTE	$x$	CONSTANTE	$x$
No contaminada	6,4725	2,9544	6,4862	2,9759
Contaminada	28,4065	1,4298	11,9441	2,6359

**Tabla 5.4 Comparación de la Muestra Contaminada vs No contaminada por Mínimos Cuadrados y Regresión Robusta**

*Autores: Minalla y Solórzano*

La *Tabla 5.3* presenta los resultados de las estimaciones por el Método de Mínimos Cuadrados y Regresión Robusta de ambas muestras, a continuación en la *Tabla 5.4* se presentan estimaciones para dos valores de  $y$ , uno que es un valor aberrante y otro que no es valor aberrante.

	<b>Mínimos Cuadrados</b>	<b>Regresión Robusta</b>
Valor Aberrante = 2	$\hat{y}^* = 28,4065 + 1,4298(2)$ = 31,9308	$\hat{y}^* = 11,9441 + 2,6359(2)$ = 18,4413
Valor No Aberrante =5	$\hat{y} = 28,4065 + 1,4298(5)$ = 34,8977	$\hat{y} = 11,9441 + 2,6359(5)$ = 23,911

**Tabla 5.5 Estimaciones con Método de Mínimos Cuadrados y Regresión Robusta**

**Autores:** Minalla y Solórzano

Cuando  $x = 2$ , el valor  $y^* = 60$ , mientras  $y = 11$  en el caso de la muestra sin contaminar. La estimación realizada es mejor en el método de robustez ya que el valor estimado se encuentra más cercano al verdadero valor de  $y$ , de igual manera para el valor de  $x = 5$ .

## CONCLUSIONES

El desarrollo de ERLA se lo realizo durante la materia de graduación: “*Regresión lineal avanzada*”, para los modelos de regresión Ridge y regresión Robusta podemos concluir:

1. Los estimadores insesgados, no siempre son “buenos” estimadores de un parámetro, en ciertos casos como la regresión Ridge se prefieren estimadores sesgados pero de menor varianza.
2. Si bien la estimación de mínimos cuadrados para un modelo de regresión disminuye la varianza de los estimadores de los coeficientes de regresión; ante la presencia de multicolinealidad en los datos, se probó que los estimadores Ridge tienen un mejor “comportamiento”.
3. Se confirma la sensibilidad de estimadores de mínimos cuadrados para modelos de regresión lineal ante la presencia de valores aberrantes.

4. Se verifica que mediante el uso de regresión IRLS las estimaciones son eficientes tanto para poblaciones contaminadas como para poblaciones sin contaminar.

## REFERENCIAS BIBLIOGRAFICAS

- [1] ATKINSON, A. & RIANA, M. (2000), "*Robust Diagnostic Regression Analysis*", Springer-Verlag, New York, EEUU.
- [2] BOVAS, A. & LEDOTTER, J. (2006), "*Introduction to Regression Modeling*", Thomson Brooks, Ciudad-Pais.
- [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 Estadística, Fundamentos y Aplicaciones*", Ediciones del Instituto de Ciencias Matemáticas, Guayaquil, Ecuador.

# **ANEXO**

# ANEXO

## Manual de Usuario

---

### Objetivos

---

A través de este manual, usted aprenderá a:

- Instalar e iniciar la aplicación ERLA;
- Conocer los requerimientos necesarios para instalar ERLA;
- Conocer brevemente cuáles son los métodos y técnicas estadísticas que puede utilizar con ERLA.

### Revisión general

Este manual presenta las características utilizadas en ERLA y usted aprenderá a usar funciones, crear graficas y generar estadísticas. La mayoría de los análisis estadísticos requieren una serie de pasos, con frecuencia orientados por un conocimiento previo o por el área en cuestión que se investiga.

Al ser ERLA, un software especializado en la técnica de Regresión Lineal, es posible evaluar la calidad de los modelos obtenidos, realizar estimaciones de todos los modelos que se hayan generado y además seleccionar el mejor modelo considerando todas las variables que usted considere sean relevantes en el estudio.

A medida que vaya leyendo este manual usted encontrará el icono que aparece a la izquierda, el cual le permitirá obtener información adicional sobre el tema que se esté desarrollando.

---

### Acerca de ERLA

---

ERLA es el nombre abreviado de Estadística de Regresión Lineal Avanzada. Es una aplicación para realizar análisis estadísticos en un entorno gráfico, utilizando menús



descriptivos y cuadros de diálogo sencillos que realizan la mayor parte del trabajo.

El software ERLA utiliza funciones gráficas y numéricas escritas en la plataforma Matlab 2010 para el desarrollo de los cálculos y Visual Basic .NET 2008 para la creación de la interfaz gráfica.

Esta aplicación estadística fue desarrollada en el año 2010 completamente en el idioma Español. El sistema operativo requerido para ERLA puede ser Windows XP o Vista 7.

Los requisitos del equipo para la instalación de ERLA es tener un sistema operativo Microsoft .NET Framework 2.0 o superior y también instalado MATLAB ComponentRuntime 7.13 o superior.

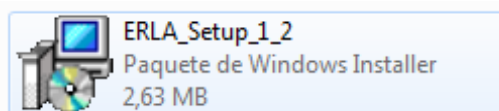
Las pantallas que aparecen en este manual corresponden al de un ordenador Windows XP. Pueden variar en función del Sistema Operativo.

## Instalación

Usted podrá obtener el instalador de ERLA a través del siguiente enlace:

<http://code.google.com/p/erla/>

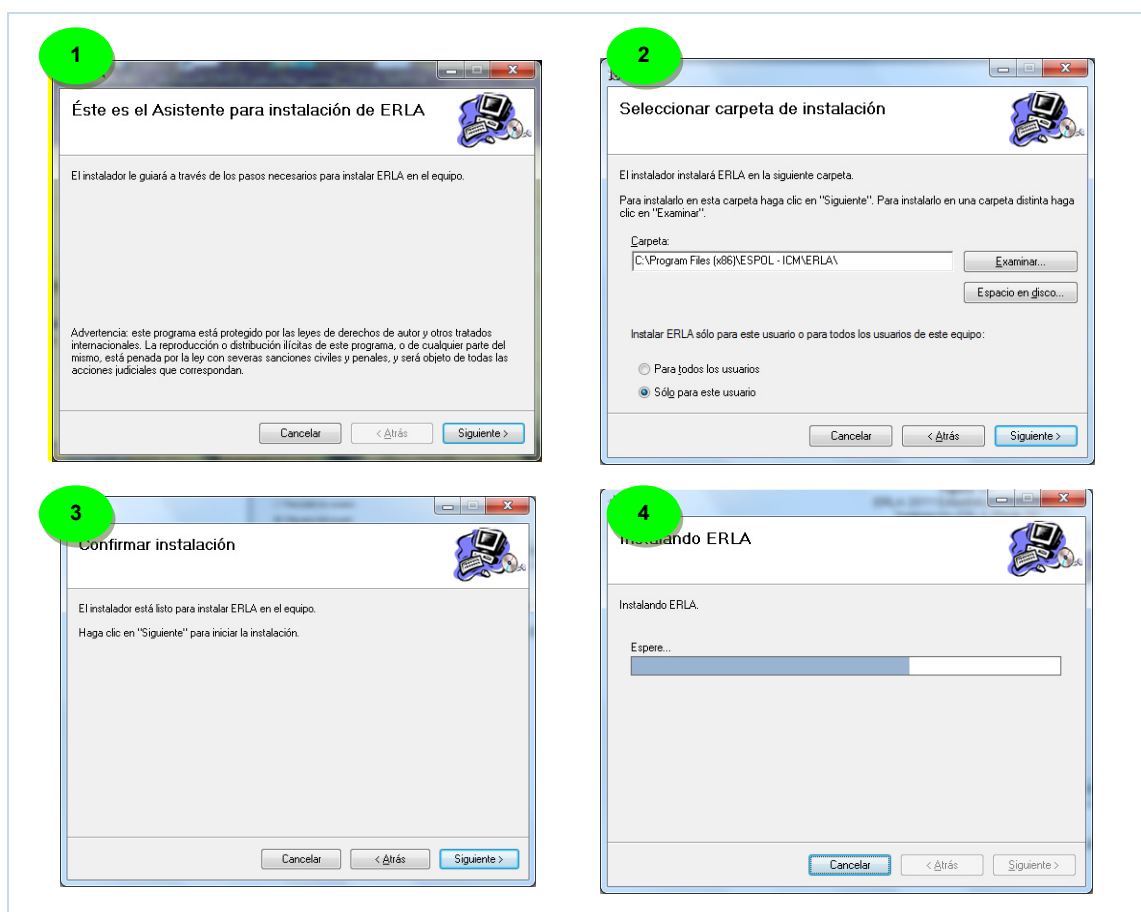
Una vez obtenido el instalador de ERLA, pulse dos veces el archivo **ERLA\_Setup\_[versión].msi**



1. Lea el cuadro de dialogo del asistente de instalación y luego pulse el botón **Siguiente**.
2. Por defecto se instalará en la carpeta **Archivos de programa** de la unidad de sistema (C:).
3. Para iniciar la instalación pulse el botón **Siguiente**.

4. Espere mientras se realiza la instalación.
5. Una vez que se haya completado la instalación pulse el botón **Cerrar**.
6. Una vez instalado ERLA, se crea un acceso directo en el escritorio y otro en el menú **Inicio** a través de los cuales puede empezar a utilizar el software.

Véase *Figura 1*



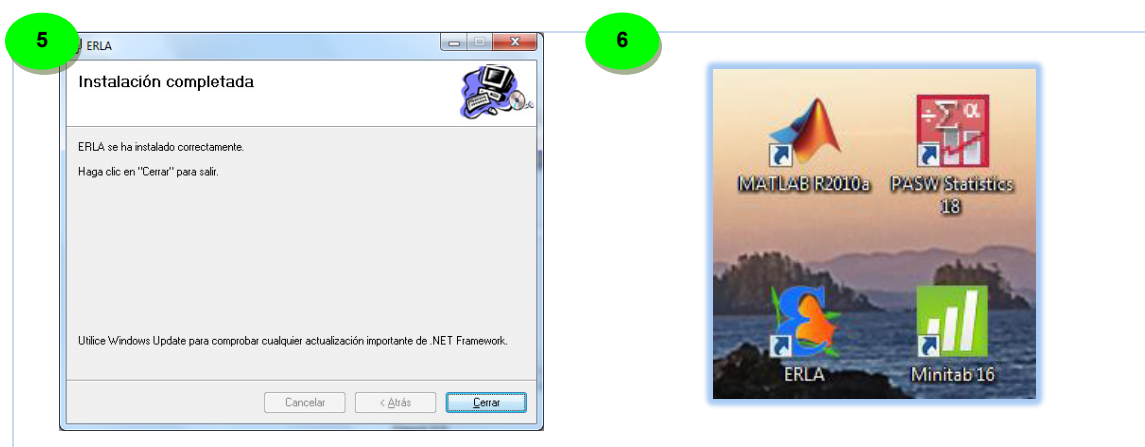


Figura 1 "Instalación de ERLA"

## Inicio de ERLA

Luego de instalar ERLA, abra la **ventana principal** pulsando dos veces el acceso directo del escritorio o:

En la barra de tareas de Windows, elija **Inicio > Todos los Programas > ERLA > Software Estadístico ERLA**.

Mientras se carga el programa usted podrá visualizar la Figura 2.



Figura 2 "Inicio ERLA"

La **ventana principal** de ERLA está formada por una **barra de título**, una **barra de menús**, una **barra de herramientas** y

además contiene dos ventanas:

- **Ventana de datos**, aquí se deben ingresar los datos correspondientes.
- **Ventana de reporte**, en la cual se presentan los resultados.

Por último en la parte inferior se encuentra la **barra de estado**. Véase Figura 3.

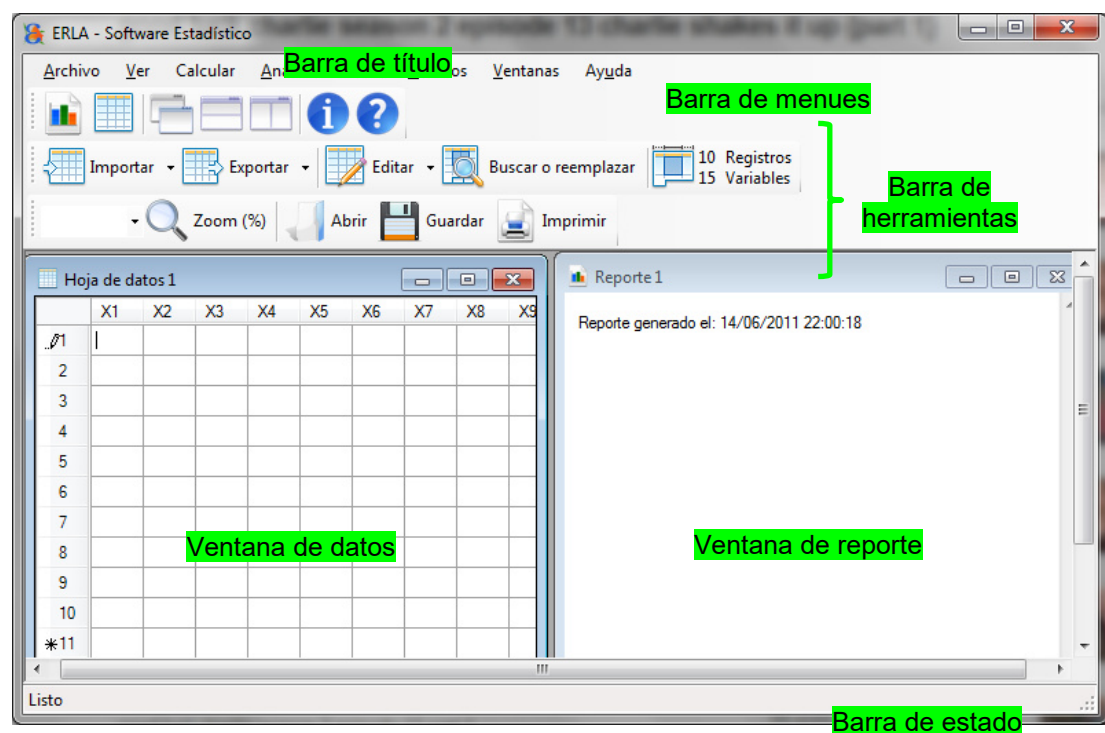


Figura 3 “Ventana principal de ERLA”

## Regresión Ridge

Para ilustrar el uso de regresión Ridge tomaremos un ejemplo del libro *“Probabilidad y Estadística Fundamentos y Aplicaciones”* de G. Zurita el cual se ajusta a las necesidades de nuestro problema:

Los integrantes de una cooperativa arrocera están interesados en que se les ayude a conocer el efecto de la humedad, en  $\text{cm/m}^2$  y la temperatura, en grados Celsius, sobre la producción de arroz (variable dependiente), en toneladas métricas por hectárea. Los datos se presentan en la Tabla 1.

**Tabla 1** “Datos para regresión Ridge”

<b>TEMPERATURA</b>	25	27	29	31	32	33	34	35	36	30	24
<b>HUMEDAD</b>	17	18	19	22	25	27	29	33	23	20	15
<b>PRODUCCIÓN</b>	5.3	5.2	5.1	5	4.8	4.9	5.3	3.2	3.1	5.9	5.2

Fuente: G. Zurita

Para realizar la regresión Ridge en ERLA siga la siguiente secuencia de pasos:

1. **Barra de menú**► **Análisis de datos**► **Regresión**► **Regresión Ridge**
2. Seleccione las variables que desee y luego pulse el botón **Aceptar**.

Los resultados se presentan en la **ventana de reporte** junto con el gráfico correspondiente. Véase Figura 4.

Salida en  
ventana de  
reporte

Salida en  
ventana de  
gráfico

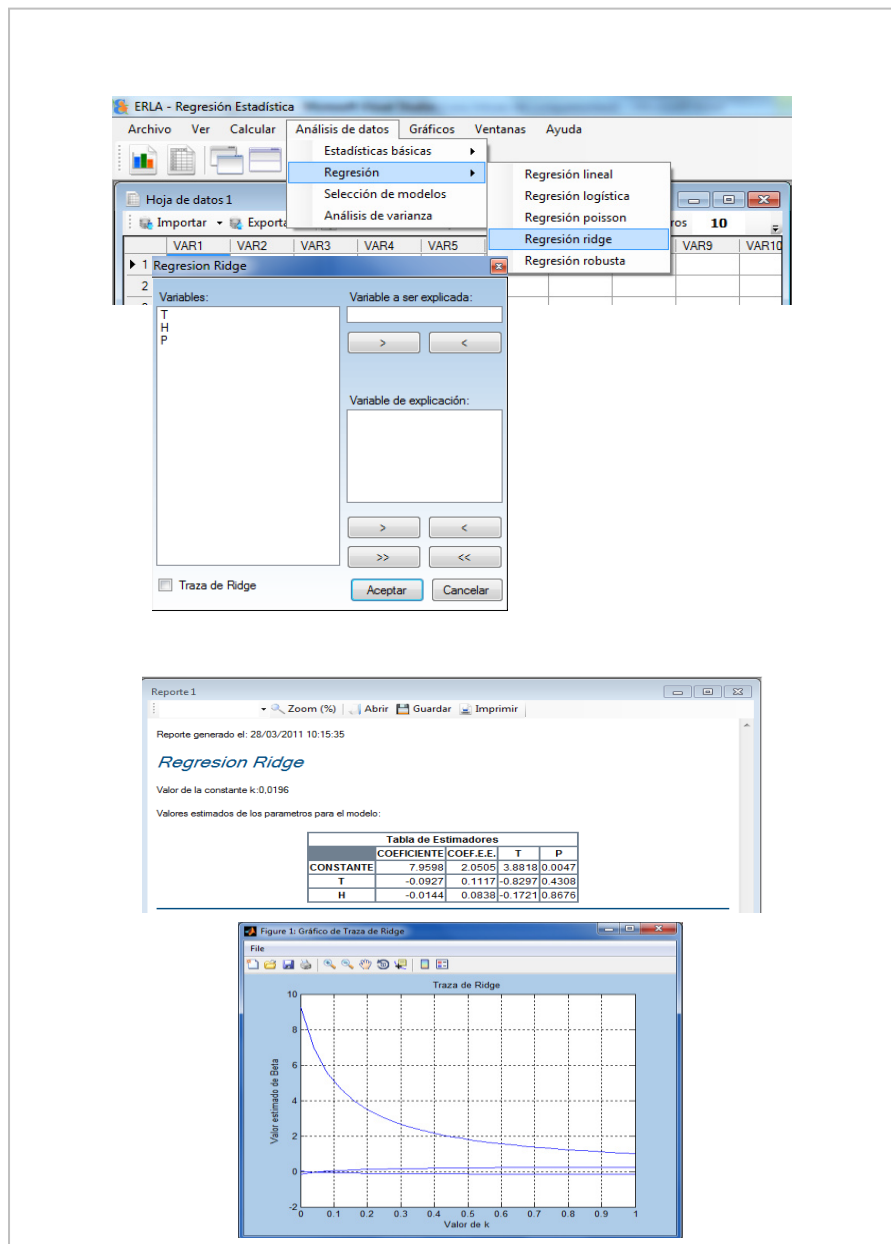


Figura 4 "Regresión Ridge"

Interpretación  
de resultados

Los resultados que se presentan en la **ventana de reporte** corresponden a:

- Valor de la Constante k es decir que del intervalo comprendido entre [0 1] el valor de 0.0196 hace que

los coeficientes de regresión que tenían un valor demasiado grande empiezan a tener valores razonables.

- Tabla que contiene los valores estimados de los parámetros para el modelo, aquí tenemos los coeficientes, los coeficientes del Error Estándar, el Estadístico de Prueba F y el valor p.

En la **ventana de gráfico** se muestra la Traza de Ridge para los diferentes valores de k con respecto a los diferentes valores de Beta.

## Regresión robusta

Ahora en el caso de regresión robusta se emplea el mismo ejemplo usado en regresión lineal simple, donde se trata de explicar la variable *P10(Identifico a los estudiantes de la ESPOL por su honestidad)* en función de la variable *P9(Identifico a los estudiantes de la ESPOL por su responsabilidad)*.

Para dar una mejor ilustración de nuestro problema hemos contaminado el 20% de la muestra. Ahora mediante la sentencia:

1. **Barra de menú>Análisis de datos>Regresión>Regresión Ridge**
2. Seleccione las variables que desee y luego pulse el botón **Aceptar**. Véase Figura 5.

Los resultados se presentan en la **ventana de reporte** junto con el gráfico correspondiente. Véase Figura 6.

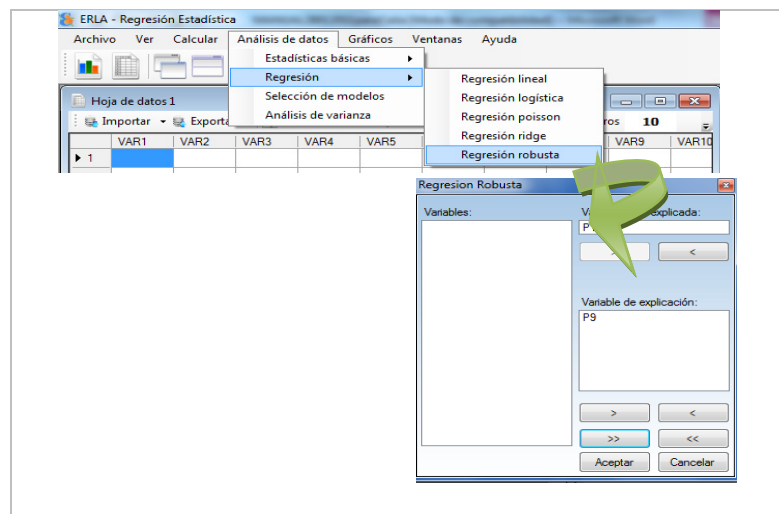


Figura 5 “Regresión robusta”

Salida en  
ventana de  
reporte



Salida en  
ventana de  
gráfico

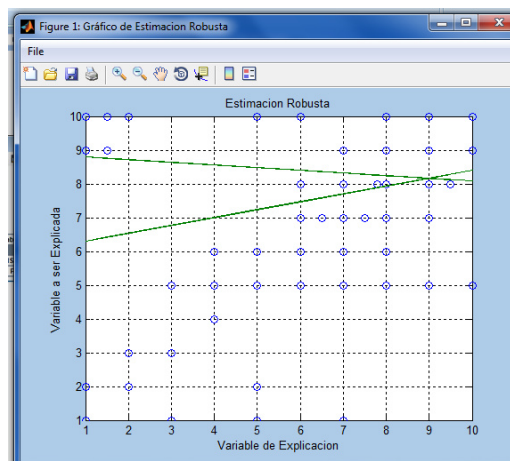


Figura 6 “Regresión robusta: Resultados”



**Interpretación  
de resultados**

En la **ventana de reporte** se presenta la tabla que contiene los coeficientes de los parámetros para el modelo.

En la **ventana de gráfico** se muestra la representación gráfica de la estimación robusta, que corresponde al diagrama de dispersión de los datos con las rectas de Estimación Robusta y de Mínimos Cuadrados. Observamos que la estimación robusta proporciona un mejor ajuste a la tendencia de las observaciones mientras que la recta de estimación por Mínimos Cuadrados no se ajusta, debido a la presencia de valores aberrantes.