ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Instituto de Ciencias Matemáticas

"Estudio Comparativo de Métodos de Estimación de parámetros para modelos lineales"

TESIS DE GRADO

Previa la obtención del Título:

INGENIERO EN ESTADÍSTICA – INFORMÁTICA

Presentada por:

FRANCISCO XAVIER VERA ALCÍVAR

Guayaquil – Ecuador

1999

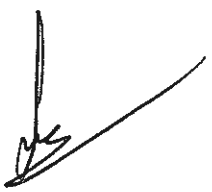
AGRADECIMIENTO

A Dios, Jehová.
A mis padres, José y Zenaida.
A mi Director de Tesis, Gaudencio Zurita

DEDICATORIA

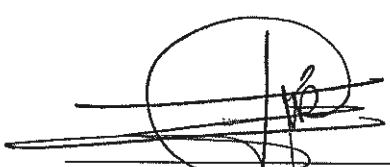
A la mujer que amo.
A mis hermanos.
A la Ciencia.

TRIBUNAL DE GRADUACIÓN

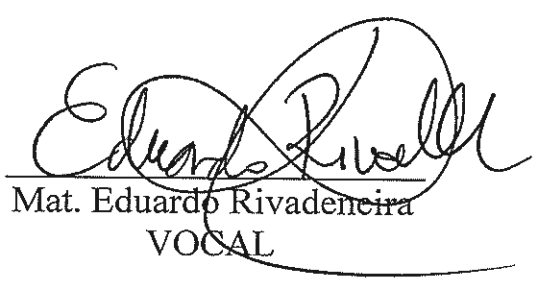


Ing. Félix Ramírez
DIRECTOR DEL ICM

Ing. Gaudencio Zurita
DIRECTOR DE TESIS



Mat. Jorge Medina
VOCAL

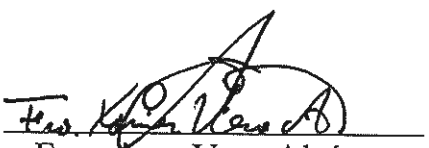


Mat. Eduardo Rivadeneira
VOCAL

DECLARACIÓN EXPRESA

"La responsabilidad del contenido de esta Tesis de Grado, me corresponden exclusivamente; y el patrimonio intelectual de la misma a la ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL"

(Reglamento de Graduación de la ESPOL).



Francisco Vera Alcívar

RESUMEN

En el presente trabajo, se realiza un estudio comparativo por simulación de diferentes métodos de estimación de parámetros de modelos lineales.

Principalmente, la intención es comparar un método robusto de estimación, contra otro que no es robusto, pero que presenta algunas características teóricas deseables.

A lo largo de este trabajo, se hace una presentación de lo que son los modelos lineales. Además se consideran tres métodos de estimación de parámetros, y se comparan dos de ellos.

La simulación, nos permite hacer una comparación experimental de los métodos, sin el costo que tienen los ensayos en la realidad. Además, cuando los tratados teóricos de un tema son complejos, la simulación puede ser una buena alternativa.

Uno de los métodos de estimación tratados en este trabajo, es relativamente nuevo. Los resultados de este trabajo, dan una calificación no satisfactoria a este método, sin embargo, el deseo y dedicación de explorar en un campo nuevo, debe estar en las mentes de los hombres que conforman la comunidad científica mundial, a pesar de los resultados que se puedan obtener.

TABLA DE CONTENIDO

1 INTRODUCCIÓN

1.1 PRESENTACIÓN DEL PROBLEMA

El principal objetivo de este trabajo es realizar un estudio comparativo, por simulación, de diferentes métodos de estimación de parámetros para los modelos lineales. La idea central es verificar bajo qué condiciones debe preferirse uno u otro estimador.

Uno de los métodos más usados para estimar parámetros de modelos lineales, y uno de los más estudiados, es el de mínimos cuadrados (MC). Un estimador de mínimos cuadrados cumple con muchas características teóricas deseables que debe tener todo "buen" estimador.

Para hallar el estimador de mínimos cuadrados de los parámetros de un modelo, se procede de la siguiente manera: dada una muestra aleatoria $y_1, y_2, \cdots, y_n$, de

la variable aleatoria Y, se trata de encontrar el valor de los parámetros del modelo, de tal forma que se minimice la suma cuadrática del error que está dada por $\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$, donde $\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_n$ representan la estimación de la variable aleatoria Y, para ciertos valores de una o más variables independientes. La diferencia $e_i=y_i-\hat{y}_i, i=1,\cdots,n$ representa la estimación del error de cada observación.

Esta suma mide en qué grado la estimación de los datos se aleja de los datos observados. La idea del método de mínimos cuadrados es intuitivamente buena, puesto que trata de minimizar una función que mide el error de estimación de los datos.

Estos estimadores presentan ciertas propiedades teóricas deseables en cualquier estimador: son insesgados, de mínima varianza, etc. Estas características teóricas son muy importantes, y se estudiarán en detalle próximamente.

A pesar que un estimador determinado con este método presenta algunos características deseables, es muy sensible a los valores aberrantes. Cuando los datos se contaminan con información no deseada, los estimadores MC pueden también perder sus propiedades, y podría resultar una estimación incorrecta de los parámetros poblacionales.

Existen otros métodos para estimar los parámetros de un modelo, como los procedimientos no paramétricos. Este procedimiento no hace suposiciones

específicas sobre la distribución del error, mas bien hace supuestos generales, como que la distribución del error debe ser simétrica. Este procedimiento exige menos restricciones sobre los datos que el de los mínimos cuadrados, pero no presenta algunos características teóricas del método MC. Es usado cuando la información disponible no cumple ciertos requisitos, como que el error de estimación siga una ley de distribución normal.

Otro método de estimación de parámetros disponible es el de máxima verosimilitud. El estimador obtenido por este método para modelos lineales, coincide con el estimador de mínimos cuadrados. Este método conduce a las mismas ecuaciones que el de mínimos cuadrados.

Otro método, es el de la mínima mediana de los cuadrados (MMC). Este método presenta ciertas propiedades: es insesgado y robusto. No presenta la característica de mínima varianza, que es una propiedad deseable. Sin embargo, cuando los datos están contaminados, lo cual sucede con mucha frecuencia, este método podría ser preferible bajo ciertas condiciones.

La idea de la comparación que se va a realizar en este trabajo, es la de simular varios modelos de regresión. En cada uno de estos modelos se aplicarán los distintos métodos de estimación que se han mencionado, y se verificará como se comporta cada método en las distintas situaciones.

1.2 MODELOS LINEALES

En muchas situaciones, es necesario describir la relación existente entre dos o más variables. Los modelos lineales constituyen un instrumento teórico sumamente útil en este tipo de situaciones.

La idea principal es poder predecir ciertas variables dependientes en términos de variables independientes.

Todos los modelos lineales se pueden llevar al *modelo lineal general*:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \textbf{1.2-1}$$

La matriz columna $\mathbf{Y} \in \mathrm{M}_{n \times 1}$ representa un vector aleatorio de las observaciones de la variable a ser explicada.

La matriz $\mathbf{X} \in \mathrm{M}_{n \times p}$ denominada *matriz de diseño*, representa una matriz con los valores de las variables de explicación. Estas variables no son aleatorias.

La matriz columna $\beta \in \mathrm{M}_{p \times 1}$, donde p es el número de parámetros del modelo, representa un vector de parámetros desconocidos, que debemos estimar.

La matriz columna ε representa un vector aleatorio del ruido del modelo, tal que $E[\varepsilon] = \mathbf{0} \in \mathrm{R}^{n}$ y $\mathrm{var}(\varepsilon) = \boldsymbol{\Sigma} \in M_{n \times n}$. Σ es una matriz real simétrica, y por tanto diagonalizable ortogonalmente.

Existen otros modelos lineales que se desprenden del modelo lineal general. Los modelos de regresión, de análisis de varianza y otros, son casos particulares del modelo lineal general.

Ejemplo 1.2-1: Modelo de análisis de varianza

En este modelo, las variables de explicación son cualitativas y se denominan factor. Cada factor tiene varios niveles o tratamientos. La idea es medir el efecto que tiene cada factor en la variable explicada *Y*, que es cuantitativa. Este modelo es ampliamente usado en el diseño de experimentos.

Un modelo, denominado de dos vías, se expresa de la siguiente manera:

$$\begin{aligned} &y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}; \quad \sum_i \tau_i = \sum_j \beta_j = 0 \\ &i = 1,\cdots,a; \quad j = 1,\cdots,b; \quad \varepsilon_{ij} \sim N\left(0,\sigma^2\right) \end{aligned} \qquad \textbf{1.2-2}$$

Donde μ es la media del sistema, τ_i el efecto del nivel *i* del primer factor, β_j el efecto del nivel *j* del segundo factor, y ε_{ij} el error aleatorio de la observación.

Supongamos que el primer factor tienen dos niveles, es decir a = 2; y que el segundo factor tiene 3 niveles, esto es b = 3. La forma matricial del modelo sería:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \textbf{1.2-3}$$

donde

$$\mathbf{Y} = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

$$\beta = \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix}$$

Este modelo es lineal general con p = 6 parámetros a ser estimados. Nótese que los elementos de la matriz de diseño son solamente unos y ceros.

Ejemplo 1.2-2: Modelo de regresión múltiple

En este modelo, tanto las variables de explicación (o de predicción) como la variable explicada son cuantitativas. Este modelo tiene muchas aplicaciones en ingeniería, economía, medicina, etc. Es uno de los modelos lineales más utilizados.

El modelo de regresión es el siguiente:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i\,p-1}\,;\ i = 1, \ldots, n \qquad \textbf{1.2-4}$$

donde p es el número de parámetros a estimar.

La forma matricial de este modelo sería:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1\,p-1} \\ 1 & x_{21} & & x_{2\,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & & x_{np-1} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

1.3 REGRESIÓN

En el caso de la regresión, el supuesto que se hace es que el vector aleatorio de ruidos ε sigue una distribución normal multivariada con vector media **0** y matriz de varianzas y covarianzas σ^2 **I**, donde **I** es la matriz identidad de orden n y σ^2 es una constante real positiva.

Se deduce de lo anterior que los elementos del vector de ruido no están correlacionados. Se supone que los elementos del vector de ruidos son independientes. Además la esperanza del ruido es cero. A este tipo de ruido se lo conoce como *ruido blanco*, y se empleará este término en lo sucesivo.

Ejemplo 1.3-1: Simulación de un modelo de regresión

Supongamos que la variable *Y* está relacionada con la variable *X* de la siguiente manera:

$$Y = 5 + 3X \qquad \textbf{1.3-1}$$

Debido a la imprecisión de las mediciones, esta relación no se cumple del todo, sino que se introduce un error aleatorio ε, con distribución normal con media 0 y varianza σ^2. Esto hace que la variable Y sea una variable aleatoria con distribución normal. Por tanto la relación de las variables sería la siguiente:

$$y = 5 + 3x + \varepsilon \qquad \textbf{1.3-2}$$

Supongamos que X es una variable que podemos controlar. Simularemos 10 observaciones del modelo y supondremos que el ruido sigue una distribución normal con media 0 y varianza 4. Entonces el modelo de regresión sería el siguiente:

$$y_i = 5 + 3x_i + \varepsilon_i ;\ \mathrm{i} = 1, \ldots, 10 \qquad \textbf{1.3-3}$$

donde

$$\varepsilon_i \sim N(0,4);\ \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ para } \mathrm{i} \neq \mathrm{j} \qquad \textbf{1.3-4}$$

Para este caso, $\beta_0 = 5$ y $\beta_1 = 3$.

Luego de la simulación se obtuvieron los siguientes datos:

i	y_i	$\hat{y}_i$	y_i teórico	x_i	ε_i	e_i
1	12.454	11.832	11	2	1.454	0.622
2	17.693	17.734	17	4	0.693	-0.041
3	26.172	23.636	23	6	3.172	2.536
4	28.681	29.538	29	8	-0.319	-0.857
5	32.474	35.44	35	10	-2.526	-2.966
6	41.273	41.342	41	12	0.273	-0.069
7	47.598	47.244	47	14	0.598	0.354
8	51.590	53.146	53	16	-1.410	-1.556
9	58.161	59.048	59	18	-0.839	-0.887
10	67.761	64.95	65	20	2.761	2.811

Tabla 1.3-1

Los parámetros originales del modelo, $\beta_0 = 5$ y $\beta_1 = 3$, son los que determinan los y_i "teóricos". En la realidad, estos son desconocidos.

Luego de realizar la estimación de los parámetros del modelo por el método de mínimos cuadrados, se obtuvo que:

$$\hat{\beta}_0 = 5.664 \text{ y } \hat{\beta}_1 = 2.944 \qquad \textbf{1.3-5}$$

Estos valores determinan la recta $\hat{y} = 5.664 + 2.944x$.

Los y_i representan los valores calculados del modelo para $X = x_1 = 2, X = x_2 = 4, \cdots, X = x_n = 20$.

Los ε_i son iguales a la diferencia de los y_i y los y_i teóricos, esto es $\varepsilon_i = y_i - y_{i\text{teórico}}$.

Los e_i son los estimadores de los ε_i. Esto es $\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$

Si hacemos la gráfica de los puntos correspondientes a los valores observados y la recta correspondiente a los valores estimados y teóricos obtenemos lo siguiente:

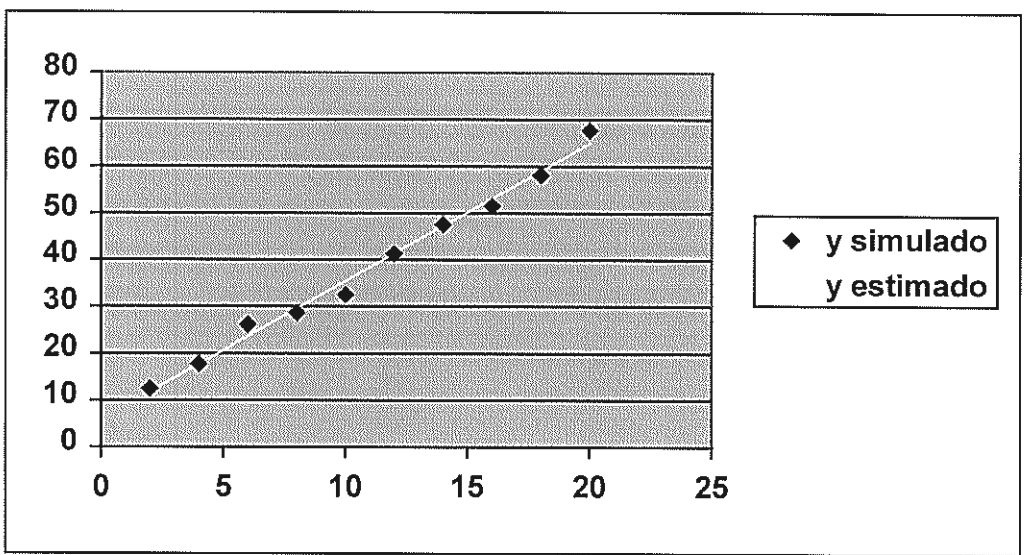


Gráfico 1.3-1: Recta de regresión lineal ajustada a los datos

Como puede apreciarse, los puntos simulados no coinciden exactamente con los puntos de una recta. Esto es debido a la presencia del ruido. Esto es debido a que en la una se utilizan los parámetros reales, y en la otra los estimadores de los parámetros. Recuérdese que β es un vector constante y $\hat{\beta}$ es un vector aleatorio.

1.4 CONCEPTOS GENERALES

Definición 1.4-1: Espacio muestral

Sea Ω el conjunto de todos los eventos elementales de un experimento. Sea S el mínimo campo de Borel de Ω. El par (Ω, S) se denomina *Espacio Muestral.*

En la definición anterior tenemos algunos elementos. Un *evento elemental* es un conjunto con un posible resultado de algún experimento. Un *campo de Borel* es aquel que contiene todas las uniones contables de los subconjuntos de un conjunto dado. El *mínimo campo de Borel* de un conjunto es la intersección de todos los campos de Borel de dicho conjunto.

Definición 1.4-2: Probabilidad

Sea P una función definida sobre un espacio muestral (Ω, S) y cuyo imagen es R. Se dice que P es una *medida de probabilidad* si y solo sí cumple los siguientes axiomas:

1. $P(E) \geq 0, \forall E \in S$ **1.4-1**

2. $\bigcup_{i=1}^{\infty} E_i = \Omega; E_i \cap E_j = \phi, \forall i \neq j \Rightarrow \sum_{i=1}^{\infty} P(E_i) = 1$ **1.4-2**

De esta definición podemos deducir los siguientes resultados:

- $0 \leq P(E) \leq 1, \forall E \in S$
- $P(\phi) = 0$

- $P(\Omega) = 1$

- $P(\bigcup E_i) = \sum P(E_i)$ para cualquier unión contable de conjuntos disjuntos en S cuya unión también pertenezca a S.

La terna (Ω, S, *P*) es llamada *espacio probabilidad* o *sistema de probabilidad.*

Ahora estamos listos para definir una de los más importantes conceptos en estadística: variable aleatoria.

Definición 1.4-3: Variable Aleatoria

Sea (Ω, S, *P*) un espacio probabilidad y sea R un conjunto de números. Sea *X* una función de Ω en R. Sea B el mínimo campo de Borel de R. Se dice que *X* es una *variable aleatoria* si y solo sí las imágenes inversas de todos los elementos de B, son eventos. Es decir:

$$X^{-1}(B) = \{\omega : X(\omega) \in B\} \in s \qquad \textbf{1.4-3}$$

En la definición anterior, ω es evento elemental, esto es $\omega \in \Omega$. Nótese que no se dice que *X* tiene que ser invertible; la notación $X^{-1}(B)$ se refiere a la imagen inversa de B.

R puede ser el conjunto de los números reales, complejos, binarios, etc. Para efectos de este trabajo, supondremos que R es el conjunto de los número reales R.

Con toda variable aleatoria, se encuentra asociada una función denominada *distribución*. Esta se define a continuación:

Definición 1.4-4: Función distribución

Sea X una variable aleatoria definida en un espacio probabilidad (Ω, S, P). Sea F una función de variable real. Se dice que F es una *función distribución* de X, si y solo sí:

$$F(x) = P\{\omega : X(\omega) \leq x\} = P\left(X^{-1}(-\infty, x]\right) \qquad \textbf{1.4-4}$$

Puede demostrarse que F cumple las siguientes propiedades:

- $F(-\infty) = 0$
- $F(\infty) = 1$
- F es creciente.
- F es continua por la derecha.

La función distribución caracteriza a una variable aleatoria.

Se dice que X es una *variable aleatoria discreta* si y solo si el conjunto de los valores que puede tomar X es contable. Se dice que X es una *variable aleatoria continua* si y solo si el conjunto de los valores que puede tomar X es continuo. Una variable aleatoria puede ser discreta en ciertos intervalos y continua en otros.

Se puede demostrar que si la función de distribución es diferenciable en todo el intervalo real, la variable aleatoria es continua.

Definición 1.4-5: Función de densidad

Sea X una variable aleatoria continua con distribución F, sea f la derivada de F. Se dice que f es la *función de densidad* de la variable aleatoria X.

Nótese que la funciones de densidad solo existen para variables aleatorias continuas. Se puede demostrar que la función de densidad es única, y por tanto se puede hallar la función de distribución. Esto quiere decir que la función de densidad caracteriza a una variable aleatoria.

Definición 1.4-6: Esperanza Matemática

Sea X una variable aleatoria y sea u una función de variable real. Se define el *valor esperado* de $u(X)$ como una transformación E sobre $u(X)$, tal que:

$$E[u(X)] = \int_{R} u(x)dF(x) \qquad \textbf{1.4-5}$$

El integral al que se refiere esta definición es un integral de Stieltjes. Puede demostrarse que el valor esperado, cuando existe, es una transformación lineal sobre la variable aleatoria X.

Dos valores esperados muy importantes en estadística son la *media* y la *varianza*. La *media* se define como $E[X]$ y se denota por μ. La *varianza* se define como $E[(X-\mu)^2]$ y se denota por σ^2.

La media es una *medida de tendencia central*. Otras medidas de tendencia central son la *mediana* y la *moda*.

La *mediana* se denota por $\tilde{\mu}$ y se define como el valor en el que $F(\tilde{\mu}) = \frac{1}{2}$.

La *moda* se denota por μ_m y se define como el valor que maximiza la función de densidad, en el caso de variables aleatorias continuas; y como el valor cuya imagen inversa en Ω tiene mayor probabilidad, en el caso de variables aleatorias discretas.

Cuando se toman dos o más variables aleatorias del mismo espacio probabilidad, entonces se dice que son *variables aleatorias conjuntas*.

Definición 1.4-7: Covarianza

Sean X y Y dos variables aleatorias *conjuntas*, con media μ_X y μ_Y respectivamente. Se define la *covarianza* entre X y Y de la siguiente manera:

$$\operatorname{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] \qquad \textbf{1.4-6}$$

Nótese que la covarianza de una variable aleatoria consigo mismo, es la varianza de la variable aleatoria. Puede demostrarse que la covarianza es un producto interno Real.

Definición 1.4-8: Vector aleatorio

Se dice que el Vector **X** es un *vector aleatorio* definido sobre un espacio probabilidad (Ω, S, P), si y solo sí sus componentes son variables aleatorias

definida sobre el mismo espacio (*conjuntas*). Es decir $\mathbf{X}^T = [X_1 \quad X_2 \quad \cdots \quad X_n]$ es la transpuesta de un vector aleatorio si X_i es una variable aleatoria para i = 1, ..., n

La esperanza de un vector aleatorio se define como el vector en cuyas coordenadas se hallan las esperanzas de cada uno de los elementos del vector. De esta manera de define la media como:

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_n] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \qquad \textbf{1.4-7}$$

Definición 1.4-9: Función de distribución Multivariada

Sea $\mathbf{X}^T = [X_1 \quad X_2 \quad \cdots \quad X_n]$ un vector aleatorio definido sobre un espacio probabilidad (Ω, S, P). Sea F una función de R^n en R. Se dice que F es una *función de distribución multivariada* si y solo si:

$$F(x_1, x_2, \cdots, x_n) = P\{\omega : X_1(\omega) \le x_1, X_2(\omega) \le x_2, \cdots, X_n(\omega) \le x_n\} \qquad \textbf{1.4-8}$$

Si la función de distribución de un vector aleatorio es diferenciable en todos sus puntos, entonces el vector aleatorio tiene *función de densidad conjunta*, y está determinada por

$$f(x_1, x_2, \cdots, x_n) = \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F(x_1, x_2, \cdots, x_n) \qquad \textbf{1.4-9}$$

Existen ciertas distribuciones especiales, conocidas por su aplicación en la teoría de la estadística. Para este trabajo, se empleará principalmente la distribución normal multivariada, que es una de las más aplicadas en estadística.

Definición 1.4-10: Distribución normal multivariada

Sea $\mathbf{X}^T = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}$ un vector aleatorio, sea $\mathbf{x}^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_p \end{bmatrix}$ un vector variable en R^p, sea $\boldsymbol{\mu}^T = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_p \end{bmatrix}$ un vector constante en R^p y sea Σ una matriz definida positiva en M_{pxp}. Se dice que el vector aleatorio $\mathbf{X}$ tiene una *distribución normal p - multivariada* si y solo si su función de densidad está dada por:

$$f(x_1, x_2, \cdots, x_p) = (2\pi)^{-p/2} (\det \Sigma)^{-1/2} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \qquad \textbf{1.4-10}$$

La expresión anterior está en términos matriciales.

Otra distribución que se emplea en este trabajo, es la distribución T^2 de Hotelling.

Definición 1.4-11: Distribución T^2

Sea $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ una sucesión de vectores aleatorios independientes, con distribución normal p-multivariada, con vector media μ y matriz de varianzas y covarianzas Σ. Sean $\bar{\mathbf{x}}$ y $\mathbf{S}$ los estimadores de máxima verosimilitud de μ y Σ respectivamente. Sea Y la variable aleatoria determinada por

$Y = n(\bar{\mathbf{x}} - \boldsymbol{\mu})^{T} \mathbf{S}^{-1} (\bar{\mathbf{x}} - \boldsymbol{\mu})$. Entonces, se dice que Y tiene una distribución T^2 con p y n - 1 grados de libertad.

La distribución T^2 es tal, que puede tomar valores negativos con probabilidad cero, y valores positivos con probabilidad uno. Esta distribución fue propuesta por Hotelling en 1947.

Se puede aproximar la distribución T^2 mediante la siguiente relación:

$$T^2_{\alpha,p,n-1} \approx p \frac{n-1}{n-p} F_{\alpha,p,n-p} \qquad \textbf{1.4-11}$$

donde F, es una variable aleatoria con distribución F. De esta manera, podemos hallar los valores críticos de T, a partir de los valores de los valores críticos de F.

Definición 1.4-12: Matriz de covarianzas

Sean **X** y **Y** dos vectores aleatorios, tal que $\mathbf{X} \in \mathrm{R}^m$ y $\mathbf{Y} \in \mathrm{R}^n$. Se define la *matriz de covarianzas* de **X** y **Y**, como:

$$\operatorname{cov}(\mathbf{X},\mathbf{Y}) = E[(\mathbf{X}-\boldsymbol{\mu}_X)(\mathbf{Y}-\boldsymbol{\mu}_Y)] = \begin{bmatrix} \operatorname{cov}(X_1,Y_1) & \operatorname{cov}(X_1,Y_2) & \cdots & \operatorname{cov}(X_1,Y_n) \\ \operatorname{cov}(X_2,Y_1) & \operatorname{cov}(X_2,Y_2) & \cdots & \operatorname{cov}(X_2,Y_n) \\ \vdots & \vdots & & \vdots \\ \operatorname{cov}(X_m,Y_1) & \operatorname{cov}(X_m,Y_2) & \cdots & \operatorname{cov}(X_m,Y_n) \end{bmatrix}$$

1.4-12

Se define la *matriz de varianzas y covarianzas* de un vector aleatorio, como la covarianza de un vector consigo mismo. . Se denota por $\Sigma = V(\mathbf{X})$ y sus posiciones por σ_{ij} donde $\sigma_{ij} = \text{cov}(X_i, X_j)$. Los elementos de la diagonal principal de esta matriz representan las varianzas de las variables aleatorias. Se puede probar que esta matriz es simétrica y definida positiva. La matriz Σ se aplica en el estudio de técnicas estadísticas multivariadas.

2 CRITERIOS PARA DETERMINAR CONDICIONES DESEABLES DE LOS ESTIMADORES

2.1 INTRODUCCIÓN

La estimación es uno de los conceptos básicos en Estadística. Intuitivamente, la estimación consiste en hacer una aproximación de ciertos parámetros poblacionales desconocidos, a través de la información que puedan proveer los datos obtenidos de muestras aleatorias.

Definición 2.1-1: Estimador

Sea $X_1, X_2, \cdots, X_n$ una sucesión de variables aleatorias independientes e idénticamente distribuidas. Sea $\theta \in \mathrm{R}^p$ un vector de parámetros de la población. Sea $\hat{\theta}$ una función de $X_1, X_2, \cdots, X_n$, cuyo conjunto de llegada es R^p. Entonces se dice que el estadístico $\hat{\theta}$ es un *estimador* del vector de parámetros θ.

En la definición anterior, θ es un vector de la forma

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} \qquad \textbf{2.1-1}$$

mientras que el estimador $\hat{\boldsymbol{\theta}}$ es una función cuyo dominio es R^n, y cuyo imagen es un subconjunto de R^p. Esta función asigna un vector p - dimensional a cada sucesión de variables aleatorias. Esta asignación representa la aproximación que se menciona al principio de esta sección.

Ejemplo 2.1-1: Estimadores

Supongamos que se toma una muestra aleatoria de tamaño n, de un población con media μ y varianza σ^2. Entonces

$$\boldsymbol{\theta} = \left\{ (\mu, \sigma^2) \middle| \mu, \sigma^2 \in \mathrm{R} \quad , \sigma^2 > 0 \right\} \qquad \textbf{2.1-2}$$

Es el vector de parámetros de esta población. Luego, el estimador correspondiente a este vector de parámetros es

$$\hat{\boldsymbol{\theta}} = \left\{ (\bar{x}, s^2) \middle| \bar{x}, s^2 \in \mathrm{R} \quad , s^2 > 0 \right\} \qquad \textbf{2.1-3}$$

donde

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad\qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad \textbf{2.1-4}$$

Existen ciertas propiedades atribuibles a los estimadores y que determinan las condiciones deseables para un estimador. Algunas de estas propiedades se utilizarán más adelante para la realización del estudio comparativo. Estas propiedades se estudian a continuación.

2.2 CONVERGENCIA DE SUCESIONES DE VARIABLES ALEATORIAS

Teóricamente, algunas propiedades de las variables aleatorias se cumplen cuando *n* el número de datos tiende a infinito. A estas propiedades se les conoce como *propiedades asintóticas*. En la práctica, algunas propiedades asintóticas se cumplen con un número suficientemente grande de datos, y este número depende de la propiedad que se estudia, y de la población a la cual pertenecen los datos.

Definición 2.2-1: Convergencia en probabilidad (convergencia débil)

Sea $X_1, X_2, \cdots, X_n$ una sucesión de variables aleatorias. Sea $c \in \mathrm{R}$, una constante. Se dice que la sucesión $\{ X_n \}$ *converge en probabilidad* (débilmente) a la constante *c*, o simplemente $X_n \xrightarrow{p} c$, si y solo si

$$\forall \varepsilon > 0, \lim_{n\to\infty} P\left(|X_n - c| > \varepsilon\right) = 0 \qquad \textbf{2.2-1}$$

Dicha en palabras, esta definición es como sigue: Para toda constante positiva ε, el límite de la probabilidad de que la sucesión difiera de la constante *c* por una valor mayor que ε, es igual a cero.

La convergencia en probabilidad es una propiedad asintótica deseable en un estimador.

Definición 2.2-2: Convergencia casi segura (convergencia fuerte)

Sea $X_1, X_2, \cdots, X_n$ una sucesión de variables aleatorias. Sea $c \in \mathrm{R}$, una constante. Se dice que la sucesión $\{ X_n \}$ *converge casi seguramente* (fuertemente) a la constante c, o simplemente $X_n \xrightarrow{C.S.} c$, si y solo si

$$P\left(\lim_{n\to\infty} X_n = c\right) = 1 \qquad \textbf{2.2-2}$$

Otra forma equivalente de esta expresión es

$$\lim_{N\to\infty} P\left(\sup_{n\geq N} |X_n - c| > \varepsilon\right) = 0 \qquad \textbf{2.2-3}$$

Esta expresión es cercana a la de convergencia débil, excepto que en esta última se toma el supremo, para $n \geq N$, de las diferencias de la sucesión y la constante. Además el límite no es sobre n, sino sobre N.

Esta definición es mucho más fuerte, matemáticamente hablando, puesto que la probabilidad se refiere a un conjunto en R^{∞} donde cada uno de sus puntos (sucesiones) cumpla con la condición deseada. Esta condición es menos fácil de probar que la anterior.

Definición 2.2-3: Convergencia en media cuadrática

Sea $X_1, X_2, \cdots, X_n$ una sucesión de variables aleatorias. Sea $c \in \mathbb{R}$, una constante. Se dice que la sucesión $\{ X_n \}$ *converge en media cuadrática* a la constante c, o simplemente $X_n \overset{M.C.}{\to} c$, si y solo si

$$\lim_{n\to\infty} E\left[(X_n - c)^2\right] = 0 \qquad \textbf{2.2-4}$$

Esta convergencia es "más fuerte" que la convergencia en probabilidad. Se puede probar que todas las sucesiones que convergen en media cuadrática convergen en probabilidad

A continuación se enumeran tres resultados que establecen relaciones entre los tres tipos de convergencia vistos hasta aquí:

- $X_n \overset{M.C.}{\to} c \Rightarrow X_n \overset{P}{\to} c$
- $X_n \overset{C.S.}{\to} c \Rightarrow X_n \overset{P}{\to} c$
- $X_n \overset{M.C.}{\to} c \wedge \sum_{n=1}^{\infty} E\left[(X_n - c)^2\right] < \infty \Rightarrow X_n \overset{C.S.}{\to} c$

Los tres tipos de convergencia vistos hasta aquí son de una sucesión de variables aleatorias hacia una constante. Ahora veremos la convergencia de una sucesión de variables aleatorias hacia una variable aleatoria específica.

Definición 2.2-4: Convergencia en ley (en distribución)

Sea $X_1, X_2, \cdots, X_n$ una sucesión de variables aleatorias. Sea $F_1, F_2, \cdots, F_n$ la sucesión de funciones de distribución correspondientes a las sucesión de variables aleatorias. Sea X una variable aleatoria con función de distribución F. Se dice que la sucesión $\{ X_n \}$ *converge en ley* (en distribución) a la variable aleatoria X, o simplemente $X_n \xrightarrow{L} X$, si y solo si en todos los puntos de continuidad x de F se cumple que:

$$\lim_{n\to\infty} F_n(x) = F(x) \qquad \textbf{2.2-5}$$

La convergencia en distribución nos ayuda a explicar el comportamiento asintótico de una sucesión de variables aleatorias. Con un número suficientemente grande de observaciones, su puede suponer que la distribución de X_n es aproximadamente la misma que la distribución de X. Una aplicación de este concepto es el teorema del límite central, que explica el comportamiento de la media aritmética de una muestra de cualquier población, a través de una distribución normal.

Ejemplo 2.2-1: Convergencia en ley

Sea $\{ X_n \}$ una sucesión de variables aleatorias, tal que la sucesión de funciones de distribución correspondiente está dada por:

$$F_n(x_n)=\begin{cases} 0 & ; \quad x_n<0 \\ \dfrac{n}{2n-1}x_n & ; \quad 0\le x_n\le \dfrac{2n-1}{n} \\ 1 & ; \quad x_n>\dfrac{2n-1}{n} \end{cases} \qquad \textbf{2.2-6}$$

Sea X una variable aleatoria con función de distribución F dada por

$$F(x)=\begin{cases} 0 & ; \quad x<0 \\ \dfrac{1}{2}x & ; \quad 0\le x\le 2 \\ 1 & ; \quad x>2 \end{cases} \qquad \textbf{2.2-7}$$

Entonces

$$\lim_{n\to\infty} F_n(x)=\lim_{n\to\infty}\frac{n}{2n-1}x=\lim_{n\to\infty}\frac{1}{2-\frac{1}{n}}x=\frac{1}{2}x=F(x), \quad x\in(0,2) \qquad \textbf{2.2-8}$$

es decir la sucesión X_n converge en ley a la variable aleatoria X ($X_n \xrightarrow{L} X$).

Podemos observar gráficamente la convergencia en distribución de { F_n } hacia F.

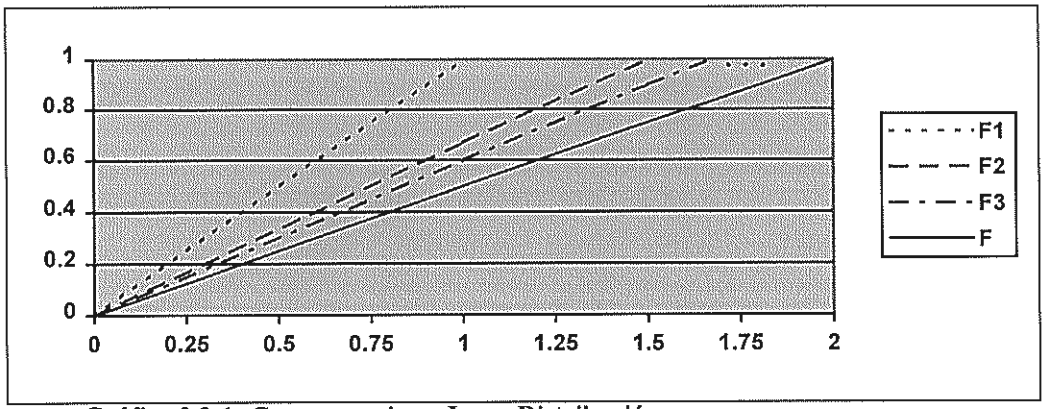


Gráfico 2.2-1: Convergencia en Ley o Distribución

2.3 *CRITERIOS DE SESGO Y EFICIENCIA.*

Las propiedades de convergencia vistas en la sección anterior, son generales para las variables aleatorias. Los estimadores son variables aleatorias, y es deseable que converjan al parámetro que estiman.

Existen otras propiedades, que son más direccionadas para estimadores. En esta sección discutiremos estas propiedades.

Definición 2.3-1: Estimador Insesgado

Sea $X_1, X_2, \cdots, X_n$ una sucesión de variables aleatorias independientes e idénticamente distribuidas tomadas de una población con vector de parámetros $\theta \in R^p$. Sea $\hat{\theta}$ un estimador del vector de parámetros θ. Se dice que $\hat{\theta}$ es un *estimador insesgado* de θ si y solo si

$$E\left[\hat{\theta}\right] = \theta \qquad \textbf{2.3-1}$$

Es decir,

$$E\left[\hat{\boldsymbol{\theta}}\right] = E\begin{bmatrix}\hat{\theta}_1\\ \hat{\theta}_2\\ \vdots\\ \hat{\theta}_p\end{bmatrix} = \begin{bmatrix}E[\hat{\theta}_1]\\ E[\hat{\theta}_2]\\ \vdots\\ E[\hat{\theta}_p]\end{bmatrix} = \begin{bmatrix}\theta_1\\ \theta_2\\ \vdots\\ \theta_p\end{bmatrix} = \boldsymbol{\theta} \qquad \textbf{2.3-2}$$

El *sesgo* de un estimador se define como $B = E\left[\hat{\boldsymbol{\theta}}\right] - \boldsymbol{\theta}$. Si B es el vector cero, obviamente que $\hat{\boldsymbol{\theta}}$ es un estimador insesgado. En el caso de $p = 1$, es decir si $\theta \in R$, entonces podemos hablar del signo de B. Si B es positivo, se dice que el estimador es sesgado a la derecha; por el contrario, si B es negativo se dice que el estimador es sesgado a la izquierda.

Ejemplo 2.3-1: Estimador insesgado

Supongamos la sucesión de variables aleatorias $X_1, X_2, \cdots, X_n$ independientes e idénticamente distribuidas, tomadas de una población con media μ y varianza σ^2. Un estimador insesgado del par $\left[\mu, \sigma^2\right]$ es el par $\left[\bar{x}, s^2\right]$, donde las componentes del último par son:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad\qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2 \qquad \textbf{2.3-3}$$

Demostraremos que $E[\bar{x}] = \mu$ y $E\left[s^2\right] = \sigma^2$

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n\mu}{n} = \mu$$

$$E[s^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2\right] = \frac{1}{n-1}\sum_{i=1}^{n}E\left[((X_i - \mu) - (\bar{x} - \mu))^2\right]$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}E\left[(X_i - \mu)^2\right] - 2\sum_{i=1}^{n}E\left[(X_i - \mu)(\bar{x} - \mu)\right] + \sum_{i=1}^{n}E\left[(\bar{x} - \mu)^2\right]\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 - 2\frac{n\sigma^2}{n} + \frac{n\sigma^2}{n}\right) = \frac{1}{n-1}\left(n\sigma^2 - \sigma^2\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

En la demostración se utiliza el hecho que las variables aleatorias de la sucesión $X_1, X_2, \cdots, X_n$ con independientes e idénticamente distribuidas.

Estos estimadores tienen la propiedad de insesgados, puesto que sus esperanzas coinciden con los parámetros que estima.

No solo se busca que la esperanza del estimador sea el parámetro, sino que también se busca que el estimador tenga poca variabilidad.

Definición 2.3-2: Estimador eficiente

Sea $X_1, X_2, \cdots, X_n$ una sucesión de variables aleatorias tomadas de una población con vector de parámetros $\theta \in \mathrm{R}^p$. Sea $t(X_1, X_2, \cdots, X_n)$ un estimador de la función escalar $c(\theta_1, \theta_2, \cdots, \theta_p)$. Se dice que $t(X)$ es un *estimador eficiente* de $c(\theta)$ si y solo si:

- $E[t(X)] = c(\theta)$.
- $\mathrm{var}(t(X)) \leq \mathrm{var}(t^*(X))$, donde $t^*(X)$ es cualquier estimador insesgado de $c(\theta)$.

Nótese que un estimador eficiente es insesgado, y además es de mínima varianza.

Este estimador se lo conoce también como *Uniformemente Insesgado de Mínima Varianza*, o simplemente estimador UMVU (Uniformly Minimum Variance Unbiased).

Para verificar la propiedad de UMVU se utiliza el teorema de Rao - Blackwell. Este teorema establece bajo que condiciones un estimador es eficiente.

Si el número de parámetros es uno ($p = 1$), entonces se puede utilizar un caso particular del teorema de Rao – Blackwell, conocido como el teorema de Rao – Cramer.

No todos los estimadores son eficientes. Sin embargo, esta es una propiedad deseable para un estimador en general.

2.4 CRITERIOS DE CONSISTENCIA Y SUFICIENCIA

Todos los tipos de convergencia vistos en la sección 2.2 son propiedades asintóticas generales de la variables aleatorias. Los estimadores, como variables aleatorias, también pueden tener o no esas propiedades.

La convergencia en probabilidad es de especial interés para los estimadores, y ha sido clasificada dentro de las propiedades de los estimadores como la consistencia.

Definición 2.4-1: Estimador consistente

Sea $X_1, X_2, \cdots, X_n$ una sucesión de variables aleatorias independientes e idénticamente distribuidas, con vector de parámetros $\theta \in R^p$. Sea $t(X_1, X_2, \cdots, X_n)$ un estimador de la función escalar $c(\theta_1, \theta_2, \cdots, \theta_p)$. Se dice que $t(X)$ es un *estimador consistente* de $c(\theta)$ si y solo si:

$$t(X) \xrightarrow{p} c(\theta) \qquad \textbf{2.4-1}$$

Existen otros tipos de consistencia, como la consistencia en media cuadrática, que no es otra cosa que la convergencia a $c(\theta)$ en media cuadrática del estimador.

Ejemplo 2.4-1: Estimador consistente

Un estimador consistente de la media poblacional es la media aritmética. Supongamos la sucesión $X_1, X_2, \cdots, X_n$, de variables aleatorias independientes e idénticamente distribuidas, con media μ y varianza σ^2. Sea $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ la media aritmética de la sucesión. Demostraremos que $\bar{x}$ es un estimador consistente de μ.

$$P(|\bar{x} - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \Rightarrow \lim_{n\to\infty} P(|\bar{x} - \mu| > \varepsilon) = 0 \qquad \textbf{2.4-2}$$

La parte izquierda de la expresión anterior es consecuencia directa de la desigualdad de Tchebysheff. La parte derecha coincide con la definición de convergencia en probabilidad. Por tanto queda demostrado que $\bar{x} \xrightarrow{p} \mu$.

Este ejemplo, es conocido como la *versión débil de la ley de los números grandes.*

Definición 2.4-2: Estadístico suficiente

Sea $X_1, X_2, \cdots, X_n$ una sucesión de variables aleatorias tomadas de una población con vector de parámetros $\boldsymbol{\theta} \in \mathrm{R}^{\,p}$. Sean $S_1 = s_1(X_1, X_2, \cdots, X_n)$, $S_2 = s_2(X_1, X_2, \cdots, X_n)$, ..., $S_r = s_r(X_1, X_2, \cdots, X_n)$ r estadísticos. Se dice que los r estadísticos son *suficientes* si y solo si la distribución condicional de $X_1, X_2, \cdots, X_n$ dados $S_1, S_2, \cdots S_r$ no depende de los parámetros $\theta_1, \theta_2, \cdots, \theta_p$.

La idea detrás de la suficiencia de un conjunto de estadísticos es la de reducir el número de observaciones originales *n*, por *r* estadísticos que contienen toda la información de los parámetros de la población. Esta propiedad es útil teóricamente hablando, pues permite simplificar ciertas expresiones. Un ejemplo de esto, es el teorema de Rao - Blackwell que se enuncia a continuación.

Teorema 2.4-1: Teorema de Rao - Blackwell

Sea X un vector aleatorio en $\mathrm{R}^{\,n}$, con función de densidad conjunta $f_X(X,\theta)$, donde θ es un vector de parámetros en $\mathrm{R}^{\,p}$. Sean $S_1, S_2, \cdots S_r$ estadísticos

suficientes, tal que $S_i = s_i(Y_1, Y_2, \cdots, Y_n), i = 1, \cdots, r$. Sea el estadístico $T = t\ (\ Y\ )$ un estimador insesgado de c (θ); y sea $T^* = E[T|S_1, S_2, \cdots, S_r]$ la esperanza condicional de T, dados $S_1, S_2, \cdots S_r$. Entonces:

- T^* es un estadístico;
- T^* es una función de los estadísticos suficientes $S_1, S_2, \cdots S_r$;
- T^* es un estimador insesgado de c (θ);
- $\text{var}(T^*) \leq \text{var}(T)$ para todo θ en R ^{p}. Además $\text{var}(T^*) < \text{var}(T)$ para al menos un valor de θ en R ^{p}, a menos que $T \equiv T^*$ (con probabilidad uno)

Este teorema establece la existencia de r estadísticos suficientes y un estimador insesgado, como condición suficiente para la existencia de un estimador UMVU. El simple enunciado de este teorema sería muy complicado sin el uso del concepto de estimadores suficientes.

La idea que generan los estadísticos suficientes es reducir al mínimo el número de estadísticos que se requieren para obtener toda la información de los parámetros a estimar.

2.5 ROBUSTEZ

Un estimador que cumple con los criterios vistos en este capítulo, hace supuestos sobre la población de la que se está tomando la muestra. Sin

embargo, si se modifica o altera estos supuestos, los estimadores pueden perder estas propiedades.

Definición 2.5-1: Estimador Robusto

Se dice que un estimador $\hat{\theta}$ es robusto si y solo si sus propiedades se conservan, cuando se modifican los supuestos iniciales bajo los que fue construido el estimador.

Los estimadores robustos son muy útiles, sobre todo cuando existe *contaminación* ó *valores aberrantes* en las observaciones. Los valores aberrantes son valores que se encuentran fuera del imagen esperado de las observaciones. Los datos que producen estos valores son conocidos como *contaminación*, y suelen presentarse en muchas situaciones reales.

La media aritmética $\bar{x}$ es un estimador eficiente, consistente y suficiente de la media poblacional μ. Pero con la presencia de valores aberrantes en las observaciones, una estimación obtenida por este estadístico para la media poblacional puede ser poco precisa. En estos casos, se puede probar que resulta mejor estimador de la media poblacional, la mediana de la muestra, puesto que la mediana es un estimador más robusto que la media, y no se ve afectado por la presencia de contaminación en los datos.

La "robustez perfecta" no es posible, puesto que si se modifican notoriamente los supuestos iniciales, los estimadores no van a conservar sus propiedades. Por

eso no se habla de que la mediana es robusta y la media no, sino que se dice que la mediana es más robusta que la media.

Ejemplo 2.5-1: Contaminación de los datos

Supongamos que deseamos estimar una media poblacional. Se toma una muestra de una población con distribución normal univariada con media μ y varianza σ^2. Supongamos que el $100(1-\alpha)\%$ de la veces se infiltra información de una población con distribución normal con media μ y varianza $k\sigma^2$. En este caso, si el valor de α es cercano a uno, la infiltración va a ser muy poca, y la media aritmética $\bar{x}$ sería un buen estimador de μ. Pero si $(1-\alpha)$ es significativamente mayor que cero, entonces la presencia de valores aberrantes va a ser mucho mayor, y por tanto $\bar{x}$ no va a ser una buena estimación de μ. Sin embargo, la estimación realizada a través de la mediana, será mucho más precisa, puesto que la población resultante sigue siendo simétrica.

Para estimar la media poblacional a través de la mediana muestral, un supuesto que se hace de la población de la cual se toma la muestra es el de simetría. En cambio que para estimar el mismo parámetro a través de la media aritmética, el supuesto es que la población de la cual se toma la muestra sea normal (a menos que número de datos sea grande). Si quitamos el supuesto de simetría, ni la media ni la mediana estimarían bien a μ. De aquí que la mediana es más robusta que la media, pero no es perfectamente robusta.

3 EL MODELO LINEAL GENERAL EN REGRESIÓN

3.1 INTRODUCCIÓN

En el capítulo 1, hay una breve presentación de lo que son los modelos lineales, y de los que es un modelo de regresión. De todos los modelos lineales, el modelo de regresión es el más usado. Los métodos de estimación utilizados en regresión, son aplicables a cualquier modelo lineal. Por esto, daremos especial énfasis a los modelos de regresión lineal.

La regresión trata de explicar una variable cuantitativa en términos de una o más variables cuantitativas. Cuando se emplea un modelo lineal para realizar dicha explicación, entonces la regresión es lineal.

Definición 3.1-1: Regresión

Sean $X_1, X_2, \cdots X_k$, k variables aleatorias, tal que $k \geq 1$. Sea Y una variable aleatoria medida sobre el mismo espacio probabilidad que las X_i. Se define la ecuación de regresión de Y sobre $X_1, X_2, \cdots X_k$ como la esperanza condicional:

$$y = E\left[Y | X_1, X_2, \cdots, X_k\right] \qquad \textbf{3.1-1}$$

En esta definición, la variable aleatoria Y se conoce como la variable a ser explicada, mientras que las variables aleatorias X_i son conocidas como las variables de explicación.

Ejemplo 3.1-1: Regresión

Supongamos que la función de densidad conjunta de las variables aleatorias x y y esta dada por:

$$f(x,y) = \begin{cases} kxy & 0 < x < y < 1 \\ 0 & \text{resto de } x \text{ y } y \end{cases} \qquad \textbf{3.1-2}$$

Puede demostrarse que el valor de la constante k tiene que ser igual a 8 para que la función sea de densidad.

El gráfico de esta función de densidad sería:

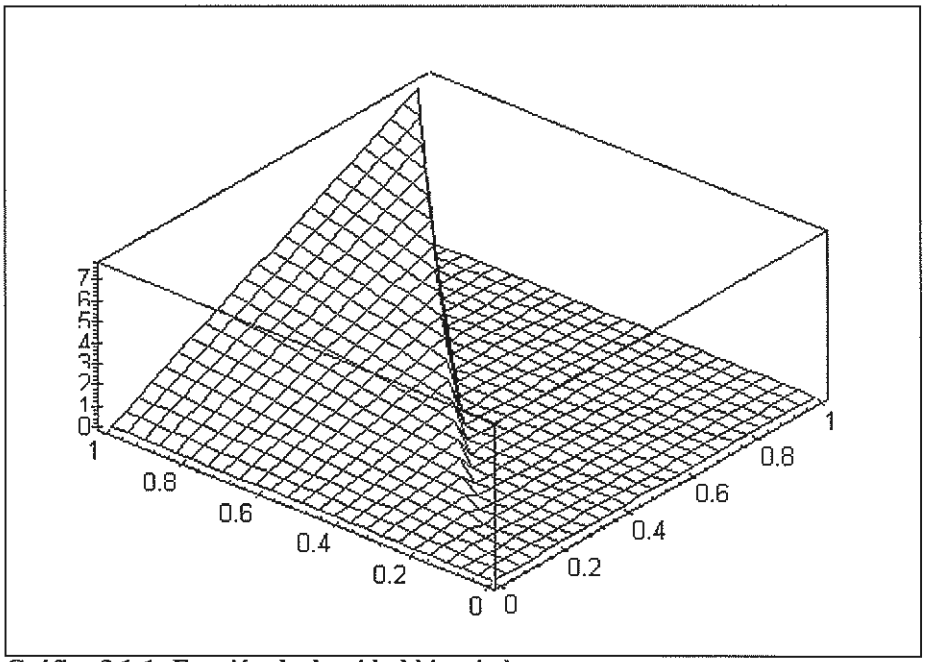


Gráfico 3.1-1: Función de densidad bivariada

A partir de esta regla de correspondencia, podemos obtener las funciones marginales y condicionales respectivas. Las reglas de correspondencia de estas funciones se muestran a continuación:

$$f_x(x) = 4x(1-x^2), \quad 0 < x < 1 \qquad \textbf{3.1-3}$$

$$f_{x|y}(x,y) = \frac{2x}{y^2}, \quad 0 < x < y, 0 < y < 1 \qquad \textbf{3.1-4}$$

$$f_y(y) = 4y^3, \quad 0 < y < 1 \qquad \textbf{3.1-5}$$

$$f_{y|x}(y,x) = \frac{2y}{1-x^2}, \quad x < y < 1, 0 < x < 1 \qquad \textbf{3.1-6}$$

Todas las funciones de densidad valen 0 en los intervalos no especificados.

Ahora, hallaremos la ecuación de regresión de X sobre Y como sigue:

$$E[X|Y=y]=\int_0^y \frac{2x^2}{y^2}dx=\frac{2}{3}y \qquad \textbf{3.1-7}$$

Luego, la ecuación de regresión de X sobre Y sería la siguiente:

$$x=\frac{2}{3}y, \quad 0<y<1 \qquad \textbf{3.1-8}$$

Esto no quiere decir que la relación de las variables X y Y sea la dada por esta ecuación, sino que el valor de esperado de x, para un determinado y, siempre es igual a dos tercios el valor de y.

De manera similar, se puede encontrar la ecuación de regresión de y sobre x. Esta ecuación está dada por:

$$y=\frac{2\left(1-x^3\right)}{3\left(1-x^2\right)}, \quad 0<x<1 \qquad \textbf{3.1-9}$$

Esta ecuación no se puede obtener despejando de la ecuación, ya que estas ecuaciones no establecen la relación de las variables aleatorias, sino las esperanzas condicionales.

En este ejemplo, el valor 2/3 que multiplica a y en la ecuación **3.1-8**, es conocido como un *parámetro* de la ecuación de regresión. De la misma manera, el 2 y el 3 de la segunda ecuación también tiene la característica de parámetros.

En general, el lado derecho de las ecuaciones de regresión se expresan en términos de las variables de explicación y de parámetros.

La forma general de una ecuación de regresión está dada por:

$$y = f(x_1, x_2, \cdots, x_n, \beta_0, \beta_1 \cdots, \beta_{p-1}) \qquad \textbf{3.1-10}$$

donde los β_i son parámetros desconocidos que debemos estimar, p es el número de parámetros del modelo, y n es el número de variables de explicación.

Cuando la función f definida anteriormente es lineal con respecto a los parámetros, entonces se dice que el modelo es un *modelo de regresión lineal*. Esto es

$$E[Y|X_1, X_2, \cdots, X_n] = \beta_0 f_0(x_1, x_2, \cdots, x_n) + \cdots + \beta_{p-1} f_{p-1}(x_1, x_2, \cdots, x_n) \qquad \textbf{3.1-11}$$

donde $f_0, f_1, \cdots, f_{p-1}$ son p funciones de R^n en R.

Si solo se tiene una sola variable de explicación y la ecuación es lineal con respecto a esa variable y a los parámetros, entonces se dice que el modelo es un *modelo de regresión lineal simple*. Esto es

$$E[Y|X] = \beta_0 + \beta_1 x \qquad \textbf{3.1-12}$$

Si las variables de explicación están elevadas a alguna potencia diferente de la unidad y/o se multiplican entre ellas, entonces el modelo es un *modelo de regresión polinómica.*

Si se tiene más de una variable de explicación, y la ecuación de regresión es lineal con respecto a los parámetros, entonces el modelo es un *modelo de regresión lineal múltiple.*

Existen otros modelos de regresión, como el modelo lineal splines, o el modelo no lineal. Como el modelo de regresión lineal simple es un caso particular del modelo de regresión lineal múltiple, en adelante utilizaremos el modelo de regresión lineal múltiple. La ecuación de regresión asociada con este modelo tendrá la forma:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} \qquad \textbf{3.1-13}$$

3.2 EL TEOREMA DE MARKOV

En esta sección, trataremos el marco teórico sobre el cual se basan algunos resultados para los modelos de regresión. Esto nos conducirá a la enunciación y demostración de un teorema, denominado Teorema de Markov, así como a otras consideraciones importantes sobre el modelo.

En primer lugar, supondremos que las variables de explicación se conocen de antemano.

Segundo, introduciremos en la ecuación un término de error, conocido como el *ruido* del modelo. Este término se lo incluye, por la imprecisión de las mediciones. Luego el modelo sería como sigue:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon \qquad \textbf{3.2-1}$$

donde ε es una variable aleatoria que representa el ruido del modelo.

Esta ecuación muestra la relación entre las variables x_i y la variable y. Esta no es la ecuación de regresión de y sobre las x_i, puesto que el lado derecho de esta ecuación no es la esperanza condicional de y dadas las x_i.

Tercero, se harán n repeticiones del modelo, de tal manera que ya no contamos con una sola ecuación, sino que con n ecuaciones. El modelo quedaría ahora de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i\,p-1} + \varepsilon_i, \quad i = 1, \cdots, n \qquad \textbf{3.2-2}$$

Estas ecuaciones tienen una representación matricial, de tal manera que todas las ecuaciones se pueden expresar por medio de una sola ecuación de forma matricial. Esto es:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \textbf{3.2-3}$$

donde

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p-1} \\ 1 & x_{21} & & x_{2p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & & x_{np-1} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

La matriz **X** es conocida como *matriz de diseño* del modelo lineal, mientras que la matriz β es el *vector de parámetros* del modelo. ε es el *vector de residuos* o *vector de error* del modelo.

Esta representación del modelo de regresión lineal múltiple, simplifica ciertas tareas. En lugar de usar grandes sumatorias, se utilizan operaciones de matriz.

Para efectos del modelo de regresión lineal, supondremos que el vector aleatorio ε, conocido también como vector de error, es un vector aleatorio con media cero y matriz de covarianza σ^2 **I**. **I** representa la matriz identidad de dimensión *n* x *n* y σ^2 es la *varianza del error*.

Teorema 3.2-1: Teorema de Markov

Sea $\beta \in M_{p \times 1}$ una matriz columna de parámetros desconocidos. Sea $\mathbf{X} \in M_{n \times p}$ una matriz de constantes conocidas. Sea $\varepsilon \in M_{n \times 1}$ un vector aleatorio tal que $E[\varepsilon] = \mathbf{0}$, $V[\varepsilon] = \sigma^2 \mathbf{I}$. Sea $\mathbf{Y} \in M_{n \times 1}$ una matriz columna tal que $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$.

Entonces $\mathbf{Y}$ es un vector aleatorio con media $\mathbf{0}$ y matriz de varianzas y covarianzas $\sigma^2\,\mathbf{I}$.

La demostración de este teorema se presenta a continuación. El producto matricial $\mathbf{X}\beta \in M_{n\ x\ 1}$ es un vector de constantes, que sumadas con vector aleatorio, resulta otro vector aleatorio. Luego, si $\mathbf{Y}$ es un vector aleatorio, entonces

$$E[\mathbf{Y}] = E[\mathbf{X}\beta + \varepsilon] = E[\mathbf{X}\beta] + E[\varepsilon] = \mathbf{X}\beta + \mathbf{0} = \mathbf{X}\beta$$

$$V[\mathbf{Y}] = E\left[(\mathbf{Y} - \mathbf{X}\beta)(\mathbf{Y} - \mathbf{X}\beta)^T\right] = E\left[(\mathbf{X}\beta + \varepsilon - \mathbf{X}\beta)(\mathbf{X}\beta + \varepsilon - \mathbf{X}\beta)^T\right]$$
$$= E\left[\varepsilon\varepsilon^T\right] = V[\varepsilon] = \sigma^2\mathbf{I}$$

Es decir, $\mathbf{Y}$ tiene la misma varianza que ε, pero con valor esperado $\mathbf{X}\beta$. Puede demostrarse que si el vector ε tiene distribución normal multivariada, entonces $\mathbf{Y}$ también tiene distribución normal multivariada, con la misma matriz de varianzas y covarianzas, pero con media en $\mathbf{X}\beta$.

3.3 EL MÉTODO DE MÍNIMOS CUADRADOS

Uno de los métodos más utilizados en la estimación de parámetros de un modelo lineal, es el método de mínimos cuadrados.

Supongamos el modelo

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i\,p-1} + \varepsilon_i, \quad i = 1, \cdots, n \qquad \textbf{3.3-1}$$

donde y_i es el valor observado, ε_i tiene distribución normal con media 0 y varianza σ^2. Además $\text{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$.

Entonces

$$E[y_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i\,p-1} \qquad \textbf{3.3-2}$$

Ahora, el valor con el que estimaremos las observaciones y_i es su valor esperado. La estimación de y_i se denota por $\hat{y}_i$. Como no se conocen los parámetros, entonces se emplearán las estimaciones de los parámetros para obtener el valor de $\hat{y}_i$. Esto es

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_{p-1} x_{i\,p-1} \qquad \textbf{3.3-3}$$

La diferencia entre el valor observado y el valor estimado, es igual al error correspondiente a esa observación, es decir

$$\varepsilon_i = y_i - \hat{y}_i \qquad \textbf{3.3-4}$$

Ahora, se requiere de una función diferenciable, de tal manera que el error total de la estimación, sea mínimo. Si minimizamos la suma del error, obtendríamos estimaciones con un error pequeño en valor relativo, pero de magnitud grande. Si minimizamos la suma de los valores absolutos del error, entonces la función resultante no va ser diferenciable.

Se define la suma cuadrática del error como la suma de los errores elevados al cuadrado. Se denota por Q, y es una medida del error total de la estimación.

El método de mínimos cuadrados consiste en establecer el valor de los β_i que minimicen la suma cuadrática del error. Esto se hace derivando parcialmente la suma cuadrática del error con respecto a cada parámetro, e igualando cada derivada a cero. Luego

$$Q = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{i\,p-1}\right)^2 \qquad \textbf{3.3-5}$$

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &= -2\sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{i\,p-1}\right) = 0 \\ \frac{\partial Q}{\partial \beta_j} &= -2\sum_{i=1}^{n} x_{ij} \left(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{i\,p-1}\right) = 0, \quad j = 1, \cdots, p-1 \end{aligned} \qquad \textbf{3.3-6}$$

Este es un conjunto de p ecuaciones con p incógnitas. La resolución de este sistema de ecuaciones, daría como resultado la estimación de los parámetros β_i por el método de mínimos cuadrados.

A este conjunto de ecuaciones, se le conoce como *ecuaciones normales del modelo.*

De la ecuación correspondiente a β_0, podemos obtener dos propiedades del método de mínimos cuadrados.

$$-2\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1 x_{i1}-\cdots-\beta_{p-1}x_{i\,p-1}\right)=0\Rightarrow\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)=0\Rightarrow\begin{cases}\sum_{i=1}^{n}\varepsilon_i=0\\ \sum_{i=1}^{n}y_i=\sum_{i=1}^{n}\hat{y}_i\end{cases}$$

3.3-7

Es decir, la suma total del error de estimación de las observaciones es cero, y la suma de los valores observados es igual a la suma de los valores estimados. Esto no quiere decir que los valores observados sean iguales a los valores estimados. También implica que la media aritmética del error es cero, y que la media aritmética de los valores observados es igual a la media aritmética de los valores estimados.

El vector de error ε se puede conocer solamente cuando se conoce el valor del vector β. Como el vector obtenido por el método anterior es una estimación de los parámetros, no podemos conocer el error, pero si podemos estimarlo. Al estimador de β se lo denota por ***b***, mientras que al estimador del vector de error ε, lo conoceremos como ***e***.

La media aritmética de los e_i, representada por $\bar{e}$ no es una variable aleatoria, a pesar que es resultado de una combinación lineal de variables aleatorias, puesto que $\bar{e}$ solo puede tomar un solo valor, cero, lo cual le da el carácter de *variable determinística*, o "*variable aleatoria con varianza cero*".

Las ecuaciones normales contiene algunos términos con sumatorias, y su resolución podría ser un poco tediosa. A continuación, se muestra un método matricial equivalente al de mínimos cuadrados.

La suma cuadrática del error, no es más que el producto interno del vector del error consigo mismo, y este se puede obtener multiplicando su transpuesta por el mismo. Luego se tiene que

$$Q=\sum_{i=1}^{n}\varepsilon_i{}^2=\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n\end{bmatrix}\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{bmatrix}=\varepsilon_1{}^2+\varepsilon_2{}^2+\cdots+\varepsilon_n{}^2=(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$$

3.3-8

Donde, de la representación matricial del modelo $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, se tiene que el $\varepsilon=\mathbf{Y} - \mathbf{X}\beta$. Es decir el vector $\hat{\mathbf{Y}}=\mathbf{X}\boldsymbol{\beta}$ representa el vector de los valores estimados.

La aplicación del método de mínimos cuadrados en su forma matricial sería

$$\frac{\partial Q}{\partial \boldsymbol{\beta}}=\frac{\partial}{\partial \boldsymbol{\beta}}\left(\mathbf{Y}^T\mathbf{Y}-\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y}+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right)=\mathbf{0} \qquad \textbf{3.3-9}$$

Aquí, el símbolo de derivada indica derivación parcial de Q con respecto a cada uno de los elementos de β. El tercer término es la transpuesta del segundo, y además ambas matrices son de dimensión 1 x 1. Esto quiere decir que ambos

términos son iguales. Con estas consideraciones procederemos a resolver la derivada y a igualar a cero

$$\frac{\partial}{\partial \boldsymbol{\beta}}\left(\mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right) = -2\mathbf{X}^T\mathbf{Y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \qquad \textbf{3.3-10}$$

Luego, las "ecuaciones normales" del modelo serían

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y} \qquad \textbf{3.3-11}$$

Luego, si **b** es la solución de este sistema de ecuaciones, entonces **b** sería la estimación de mínimos cuadrados del vector de parámetro β.

Estas ecuaciones son las mismas que se habían obtenido anteriormente, pero en notación simplificada.

Ejemplo 3.3-1: Aplicación del método de mínimos cuadrados

Supongamos un modelo de regresión lineal con dos variables de explicación, tal que el error tiene distribución normal con media cero y varianza 9, además los errores no están correlacionados entre sí, es decir cov(ε_i, ε_j) = 0, para i ≠ j. Simularemos los datos para este modelo con parámetros que conoceremos de antemano. Luego, supondremos que no conocemos estos parámetros y los estimaremos. De la misma manera se establecerá la diferencia entre el error real, y el error de estimación. Posteriormente, se verificará la propiedad enunciada anteriormente.

El modelo propuesto para este ejemplo es

$$y_i = 3 - 2x_{i1} + 4x_{i2} + \varepsilon_i \qquad \textbf{3.3-12}$$

Los valores para x_{i1} y x_{i2} son constantes, y sus valores serán los puntos de la maya formada por el cuadrado $0 < x_{i1} < 3$, $0 < x_{i2} < 3$ y subdivisiones del mismo en cuadrados de área 1. Los ε_i son generados como un error con distribución normal, con media 0 y varianza 9. Los y_i son obtenidos aplicando el modelo **3.3-12**. Los y_i teóricos se obtienen eliminando el término de error en dicho modelo. Luego la simulación sería como sigue:

i	x_{i1}	x_{i2}	y_i teórico	ε_i	y_i
1	0	0	3	0.072	3.07
2	0	1	7	0.675	7.68
3	0	2	11	-1.725	9.28
4	0	3	15	-1.479	13.52
5	1	0	1	5.504	6.50
6	1	1	5	-1.701	3.30
7	1	2	9	-5.256	3.74
8	1	3	13	6.581	19.58
9	2	0	-1	-0.890	-1.89
10	2	1	3	-4.221	-1.22
11	2	2	7	-3.993	3.01
12	2	3	11	4.553	15.55
13	3	0	-3	-3.790	-6.79
14	3	1	1	4.164	5.16
15	3	2	5	-2.424	2.58
16	3	3	9	-9.949	-0.95

Tabla 3.3-1

Ahora, estimaremos los parámetros de este modelo con el método de mínimos cuadrados.

Primero se construye la matriz de diseño. La transpuesta de esta sería

$$\mathbf{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 \\ 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 & 0 & 1 & 2 & 3 \end{bmatrix} \qquad \textbf{3.3-13}$$

Luego, la transpuesta de esta matriz se multiplica por esta matriz para obtener la matriz del sistema de ecuaciones normales.

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 16 & 24 & 24 \\ 24 & 56 & 36 \\ 24 & 36 & 56 \end{bmatrix} \qquad \textbf{3.3-14}$$

La matriz de observaciones sería

$$\mathbf{Y} = \begin{bmatrix} 3.07 \\ 7.68 \\ 9.28 \\ 13.52 \\ 6.50 \\ 3.30 \\ 3.74 \\ 19.58 \\ -1.89 \\ -1.22 \\ 3.01 \\ 15.55 \\ -6.79 \\ 5.16 \\ 2.58 \\ -0.95 \end{bmatrix} \qquad \textbf{3.3-15}$$

Para obtener el lado derecho del sistema de ecuaciones normales, se multiplica la transpuesta de la matriz de diseño por la matriz de observaciones.

$$\mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 82.1 \\ 64 \\ 195 \end{bmatrix} \qquad \textbf{3.3-16}$$

Ahora debemos resolver el sistema de ecuaciones de tres ecuaciones con tres incógnitas, donde las incógnitas son los elementos del vector β. La solución de este sistema, es el vector **b**, y es la estimación por el método de mínimos cuadrados para este modelo.

$$\begin{bmatrix} 16 & 24 & 24 \\ 24 & 56 & 36 \\ 24 & 36 & 56 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 82.1 \\ 64 \\ 195 \end{bmatrix} \qquad \textbf{3.3-17}$$

Este sistema tiene solución única dada por

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 4.16 \\ -3 \\ 3.6 \end{bmatrix} \qquad \textbf{3.3-18}$$

Luego, la ecuación del plano estimado sería

$$\hat{y}_i = 4.16 - 3x_1 + 3.6x_2 \qquad \textbf{3.3-19}$$

Lo cual es diferente al vector de parámetros originales. La diferencia entre los parámetros y su estimación sería

$$\boldsymbol{\beta} - \mathbf{b} = \begin{bmatrix} 3 \\ -2 \\ 4 \end{bmatrix} - \begin{bmatrix} 4.16 \\ -3 \\ 3.6 \end{bmatrix} = \begin{bmatrix} -1.16 \\ 1 \\ .4 \end{bmatrix} \qquad \textbf{3.3-20}$$

la norma de este vector es $\sqrt{(-1.16)^2 + 1^2 + 0.4^2} = 1.56$, mientras que la norma del vector original es 5.39, esto quiere decir que nos desviamos de los parámetros originales, aproximadamente un 29%.

En la columna de los errores del modelo anterior, podemos encontrar un valor aberrante (-9.949). Este valor hace que la estimación no sea del todo buena. Si se realiza la estimación de los parámetros por el método de mínimos cuadrados, sin considerar este valor, entonces la estimación sería

$$\mathbf{b} = \begin{bmatrix} 2.56 \\ -2.2 \\ 4.34 \end{bmatrix} \qquad \textbf{3.3-21}$$

A simple vista, estos valores se aproximan mejor a los parámetros que los de la estimación anterior. Si calculamos el vector diferencia entre los parámetros y los estimadores de los parámetros, se obtiene

$$\boldsymbol{\beta} - \mathbf{b} = \begin{bmatrix} 3 \\ -2 \\ 4 \end{bmatrix} - \begin{bmatrix} 2.56 \\ -2.2 \\ 4.34 \end{bmatrix} = \begin{bmatrix} .44 \\ .2 \\ -.34 \end{bmatrix} \qquad \textbf{3.3-22}$$

La norma de este vector es 0.6, la estimación se aleja de los parámetros originales, cuya norma es 5.39, aproximadamente un 11%.

Ahora verificaremos, sobre la primera estimación, la dos propiedades mencionadas sobre la media aritmética del error estimado y las medias aritméticas de las observaciones y los valores estimados.

i	y_i	$\hat{y}_i$	e_i
1	3.07	4.16487496	-1.09
2	7.68	7.76774069	-0.09
3	9.28	11.3706064	-2.10
4	13.52	14.9734721	-1.45
5	6.50	1.20711974	5.30
6	3.30	4.80998547	-1.51
7	3.74	8.4128512	-4.67
8	19.58	12.0157169	7.56
9	-1.89	-1.75063548	-0.14
10	-1.22	1.85223025	-3.07
11	3.01	5.45509598	-2.45
12	15.55	9.0579617	6.49
13	-6.79	-4.7083907	-2.08
14	5.16	-1.10552497	6.27
15	2.58	2.49734076	0.08
16	-0.95	6.10020649	-7.05
Suma	**82.12**	**82.12**	**0**
Media	**5.13**	**5.13**	**0**

Tabla 3.3-2

Como vemos, la media de los valores observados y los valores estimados son la misma. Además la media del error estimado es igual a cero

Nótese la sensibilidad de este método ante la presencia de valores aberrantes. En este caso, eliminando un dato, la estimación resultó ser mucho mejor. Sin embargo, la práctica de eliminar datos, puede constituirse en nociva, puesto que este dato bien podría ser producto del azar, o podría contener información relevante acerca de la población de la cual se muestrea.

3.4 PRUEBAS PARA LOS MODELOS DE REGRESIÓN

No todos los modelos lineales se ajustan perfectamente a los datos. Además, bien podría ser que ciertas variables de explicación no influyan sobre la variable explicada. También podría darse el caso que ninguna variable de explicación influya sobre la variable explicada. Otro caso sería que el modelo no se ajuste a los datos, en cuyo caso el error no solo será por el error de medición, sino también por *falta de ajuste* del modelo hacia los datos. También se requieren tener ciertas medidas para comparar diferentes modelos.

Una de las herramientas utilizadas para realizar inferencias acerca de los parámetros es la tabla de *análisis de varianza (ANOVA)*. La tabla ANOVA está conformada por 5 columnas: Fuente de variación, grados de libertad, sumas cuadráticas, medias cuadráticas y Valor F.

En los modelos de regresión, tenemos usualmente tres fuentes de variación. A continuación se muestra un bosquejo de una tabla ANOVA para los modelos de regresión.

Fuentes de Variación	Grados de libertad	Sumas Cuadráticas	Medias Cuadráticas	Valores F
Regresión	p - 1	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	$MCR = \frac{SSR}{p-1}$	$\frac{MCR}{MCE}$
Error	n - p	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$MCE = \frac{SSE}{n-p}$	
Total	n - 1	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$		

Tabla 3.4-1

La versión matricial de esta tabla se muestra a continuación:

Fuentes de Variación	Grados de libertad	Sumas Cuadráticas	Medias Cuadráticas	Valores F
Regresión	p - 1	$SSR = \mathbf{b}^T\mathbf{X}^T\mathbf{Y} - n\bar{y}^2$	$MCR = \frac{SSR}{p-1}$	$\frac{MCR}{MCE}$
Error	n - p	$SSE = \mathbf{Y}^T\mathbf{Y} - \mathbf{b}^T\mathbf{X}^T\mathbf{Y}$	$MCE = \frac{SSE}{n-p}$	
Total	n - 1	$SST = \mathbf{Y}^T\mathbf{Y} - n\bar{y}^2$		

Tabla 3.4-2

Algunos programas de computadora, tienen por salida una versión más desglosada de esta tabla. La idea es realizar una medida del grado de contribución que tiene cada parámetro sobre la variación total de la regresión.

En primer lugar, la suma cuadrática total no es como la de la tabla 3.4-2, sino que se considera la suma cuadrática total como la suma de todas las observaciones de la variable *Y*.

Después, se define la suma cuadrática $SC(b_0, b_1, \cdots, b_k)$, $0 \leq k \leq p$, como la suma cuadrática de regresión que hubiera tenido el modelo, si en lugar de tomar p parámetros, se hubieran tomado solamente k.

Puede demostrarse que

$$SC(b_0) = n\bar{Y}^2 \qquad \textbf{3.4-1}$$

Se puede probar que SC(b_0) tiene distribución *Ji cuadrada* con 1 grado de libertad

Defínase ahora,

$$SC(b_{k+1}, b_{k+2}, \cdots, b_q | b_0, b_1, \cdots, b_k) = SC(b_0, b_1, \cdots, b_q) - SC(b_0, b_1, \cdots, b_k) \qquad \textbf{3.4-2}$$

Puede probarse que esta suma cuadrática tiene distribución *Ji cuadrada* con *q-k* grados de libertad.

Puede demostrarse que la suma de regresión definida en la tabla 3.4-2 es igual a $SC(b_1, b_2, \cdots, b_p | b_0)$. Además, empleando inducción matemática, se puede demostrar de la ecuación **3.4-2** que

$$SC(b_1, b_2, \cdots, b_p | b_0) = SC(b_1 | b_0) + SC(b_2 | b_0, b_1) + \cdots + SC(b_p | b_0, b_1, \cdots, b_{p-1}) \qquad \textbf{3.4-3}$$

Luego, el desglose de la tabla 3.4-2 sería

Fuentes de Variación	Grados de libertad	Sumas Cuadráticas	Medias Cuadráticas	Valores F
b_1	1	$SC(b_1 \mid b_0)$	$SC(b_1 \mid b_0)$	$\frac{SC(b_1 \mid b_0)}{MCE}$
b_2	1	$SC(b_2 \mid b_0, b_1)$	$SC(b_2 \mid b_0, b_1)$	$\frac{SC(b_2 \mid b_0, b_1)}{MCE}$
$\vdots$				
b_p	1	$SC(b_p \mid b_0, \cdots, b_{p-1})$	$SC(b_p \mid b_0, \cdots, b_{p-1})$	$\frac{SC(b_p \mid b_0, \cdots, b_{p-1})}{MCE}$
Regresión	p - 1	$SC(b_1, b_2, \cdots, b_p \mid b_0)$	$MCR = \frac{SSR}{p-1}$	$\frac{MCR}{MCE}$
Error	n - p	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$MCE = \frac{SSE}{n-p}$	
Total Ajustado	n - 1	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$		
Ajuste(b_0)	1	$SC(b_0) = n\bar{Y}^2$	$SC(b_0)$	$\frac{SC(b_0)}{MCE}$
Total	n	$\sum_{i=1}^{n} y_i^2$		

Tabla 3.4-3

De la tabla anova, se obtienen algunas medidas del modelo.

La media cuadrática del error (MCE) es un estimador insesgado de la varianza del error, si se cumplen los supuestos iniciales. Esto se demuestra en el capítulo 4.

Se puede demostrar que la suma cuadrática del error sigue una distribución *Ji cuadrada* con *n-p* grados de libertad. Además, la suma cuadrática de regresión,

tiene una distribución *Ji cuadrada* con p - 1 grados de libertad. También, cada una de las sumas cuadráticas correspondientes a cada estimador, tiene una distribución *Ji cuadrada* con 1 grado de libertad.

Como todos los valores F de la tabla 3.4-3, son obtenidos del cociente de una Ji cuadrado dividida entre sus grados de libertad, y otra Ji cuadrado dividida para sus grados de libertad, entonces todos los valores F tienen una distribución F.

Los valores F son estadísticos utilizado para hacer una inferencia en la que se prueba si las variables de explicación tienen influencia sobre la variable explicada.

Para el caso del valor F de regresión, la hipótesis nula esta dada por $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$, y la hipótesis alterna H_1 establece que por lo menos uno de los β no es cero. Si el valor p de esta prueba es bajo, el modelo es válido, caso contrario las variables de explicación no tienen influencia sobre la variable explicada. Este estadístico F tiene una distribución F con p - 1 grados de libertad en el numerador y n - p grados en el denominador.

El valor p para las pruebas de hipótesis con poblaciones con distribución F, son el integral desde el valor F hasta infinito, de la densidad da la variable aleatoria F.

Para el caso de cada valor F de $b_1, b_2, \cdots, b_p$, sirven para hacer inferencias respecto a cada parámetro. Cada uno de estos estadísticos F tiene 1 grado de libertad en el numerador y $n - p$ grado de libertad en el denominador.

Un indicador empleado para comparar varios modelos es la *potencia de explicación del modelo*. Este indicador se denota por R^2 y es el cociente entre la suma cuadrática de regresión y la suma cuadrática total.

Ejemplo 3.4-1: Tabla ANOVA

En este ejemplo, construiremos la tabla ANOVA del ejemplo de la sección anterior. Para este ejemplo, p es igual a 3 y n = 16. La tabla que se muestra a continuación, está basada en las tablas expuestas anteriormente.

Fuentes de Variación	**Grados de libertad**	**Sumas Cuadráticas**	**Medias Cuadráticas**	**Valores F**
Regresión	2	434.58	217.29	10.5389
Error	13	268.03	20.62	
Total	15	702.61		

La media cuadrática del error es 20.62, lo cual quiere decir que la estimación de la varianza para este modelo es 20.62, lo cual es bastante lejano a 9 que es la varianza verdadera. En la práctica, no sabemos que tan cercana está la estimación de la varianza de la varianza poblacional.

El valor p correspondiente al valor F es .0019, con lo cual podemos afirmar, que con alta confianza, se rechaza la hipótesis nula H_0.

El valor R^2 para este modelo es $\frac{434.58}{702.61} \cdot 100\% = 61.85\%$.

3.5 *ANÁLISIS DEL VECTOR ALEATORIO ERROR*

Hemos dicho sobre el error que su media debe ser **0** y su matriz de varianzas y covarianzas $\sigma^2\mathbf{I}$. Esto quiere decir, que la varianza de cada elemento del vector de error, es la misma. Además, la correlación entre cada componente es cero.

Puede demostrarse que si la distribución del vector de error es normal, entonces sus elementos son independientes.

No siempre tenemos esta condición. En muchas circunstancias, la varianza del error no está distribuida de esa manera, sino más bien existe correlación entre los elementos del error. No todos estos casos se pueden tratar por medio de modelos lineales, pero algunos de ellos sí.

Supongamos que $V[\varepsilon]=\sigma^2\mathbf{G}$ donde **G** es una matriz simétrica definida positiva. Es posible demostrar que existe una matriz $\mathbf{G}^{-1/2}$ tal que $\mathbf{G}^{-1/2}\,\mathbf{G}^{-1/2} = \mathbf{G}^{-1}$. La matriz $\mathbf{G}^{-1/2}$ así definida, se denomina "raíz cuadrada" de $\mathbf{G}^{-1}$. Podemos hacer un cambio de variables tal que $\mathbf{Z} = \mathbf{G}^{-1/2}\,\mathbf{Y}$, de tal manera que

$$\mathbf{Z} = \mathbf{G}^{-1/2}\mathbf{Y} = \mathbf{G}^{-1/2}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = (\mathbf{G}^{-1/2}\mathbf{X})\boldsymbol{\beta} + \mathbf{G}^{-1/2}\boldsymbol{\varepsilon} = \mathbf{U}\boldsymbol{\beta} + \boldsymbol{\omega}$$

Este cambio de variable resulta en nuevo modelo, con matriz diseño **U** y matriz de error ω. El valor esperado y la varianza de este nuevo termino de error sería

$$E[\boldsymbol{\omega}] = E\left[\mathbf{G}^{-1/2}\boldsymbol{\varepsilon}\right] = \mathbf{G}^{-1/2}E[\boldsymbol{\varepsilon}] = \mathbf{G}^{-1/2}\mathbf{0} = \mathbf{0}$$

$$V[\boldsymbol{\omega}] = V\left[\mathbf{G}^{-1/2}\boldsymbol{\varepsilon}\right] = E\left[\left(\mathbf{G}^{-1/2}\boldsymbol{\varepsilon}\right)\left(\mathbf{G}^{-1/2}\boldsymbol{\varepsilon}\right)'\right] = \mathbf{G}^{-1/2}E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right]\mathbf{G}^{-1/2}$$
$$= \sigma^2\mathbf{G}^{-1/2}\mathbf{G}\mathbf{G}^{-1/2} = \sigma^2\mathbf{G}^{-1/2}\mathbf{G}^{1/2}\mathbf{G}^{1/2}\mathbf{G}^{-1/2} = \sigma^2\mathbf{I}$$

Luego, este nuevo modelo cumple con las condiciones del modelo original.

Para el modelo que se trata aquí, la varianza es constante, y el error no está correlacionado. A la característica de varianza constante se la conoce como *Homocedasticidad.* Cuando para cada elemento no es la misma, se llama *Heterocedasticidad.* En algunos casos, el problema de la heterocedasticidad se puede resolver mediante algún cambio de variable, como en denominado mínimos cuadrados ponderados. Pero, no siempre se puede hacer esto, y a veces hay que recurrir a otros artificios.

Existen ciertas pruebas estadísticas para medir la homocedasticidad del modelo, así como para medir la *autocorrelación* del ruido. El término autocorrelación se refiere a la correlación de los elementos del vector de ruido. Los elementos de este vector son observaciones de la misma variable, de ahí el prefijo auto.

Existen pruebas como la prueba de Durbin - Watson para medir la *correlación en serie* del ruido. La correlación en serie se refiere a la correlación entre dos elementos consecutivos del ruido. El más usado es el test de Durbin - Watson,

el cual se encuentra en la mayoría de los paquetes informáticos que realizan modelos lineales.

Este test, consiste en el cálculo del estadístico

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^{\,2}} \qquad \textbf{3.5-1}$$

Los paquetes de computación, suelen calcular el valor *p* de este estadístico, de tal manera que se puedan hacer inferencias respecto a la autocorrelación del error.

Algunas de las propiedades mencionadas, se puede observar gráficamente, aunque esto no constituye una prueba analítica. A continuación se muestran y describen algunos gráficos de ε_i versus i. En ellos se explican las distintas situaciones que podrían presentarse.

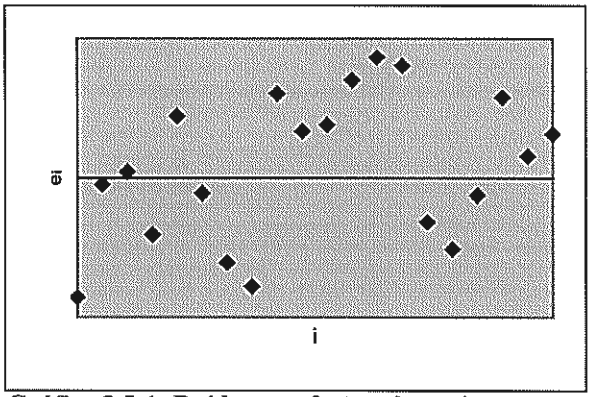


Gráfico 3.5-1: Ruido con efectos aleatorios.

En este gráfico, el error no sigue ningún patrón especial, es decir es aleatorio y las observaciones son independientes entre ellas.

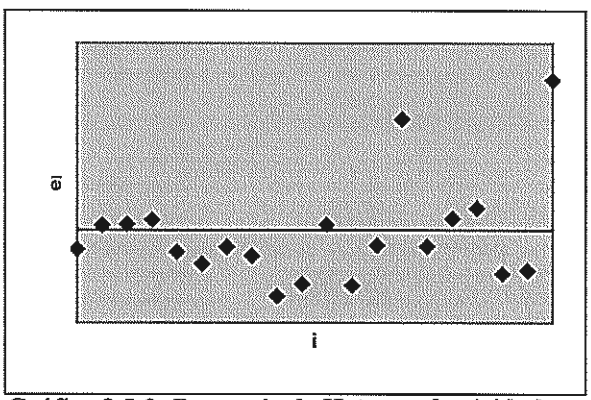


Gráfico 3.5-2: Presencia de Heterocedasticidad.

En este gráfico, las observaciones tienden a tener una mayor variación para i más grande. Esto quiere decir que la varianza no es constante, sino que aumenta progresivamente. Esto es un caso claro de heterocedasticidad. Podría resolverse con un cierto cambio de las variables del error.

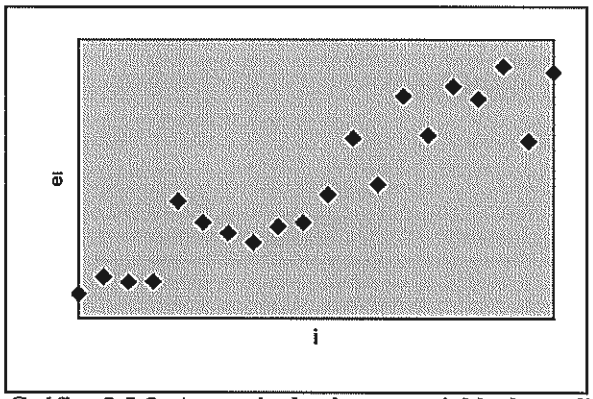


Gráfico 3.5-3: Ausencia de alguna variable de explicación

Aquí, la variación del error es uniforme, pero la media creciente. Esto puede deberse a la falta de variables de explicación.

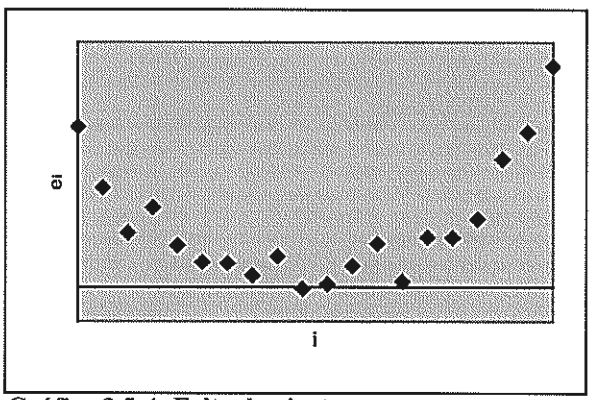


Gráfico 3.5-4: Falta de ajuste

En este gráfico, la varianza es constante en todo el modelo, pero la media tiende a formar un parábola. Este es un caso clásico de falta de ajuste del modelo, y puede deberse a la falta de potencias de las variables de explicación o de interacciones entre ellas (modelo polinómico).

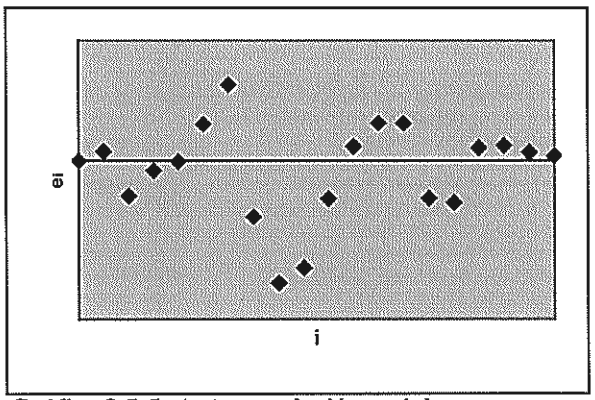


Gráfico 3.5-5: Autocorrelación serial

En este gráfico, los puntos se "siguen" entre sí. Esto indica una correlación en serie del error. En este caso, ya no podemos aplicar el modelo lineal, sino más bien se emplean modelos de series temporales.

La importancia de analizar el error radica en que los diferentes métodos de estimación están basados en el error, se trata siempre de minimizarlo.

4 MÉTODOS DE ESTIMACIÓN DE PARÁMETROS PARA MODELOS LINEALES: MÁXIMA VEROSIMILITUD, MÍNIMOS CUADRADOS, PROCEDIMIENTOS NO PARAMÉTRICOS

Hemos visto ya algunas propiedades útiles de los modelos lineales. Hemos visto también el método de mínimos cuadrados para estimación de parámetros. En este capítulo se introducirán otros dos métodos, así como una breve descripción de las propiedades de cada uno. Además, se mostrarán otras propiedades del método de mínimos cuadrados.

4.1 EL MÉTODO DE MÁXIMA VEROSIMILITUD

El método de máxima verosimilitud es un procedimiento general empleado para estimar parámetros. Se puede probar que por este método de estimación se llega a las mismas ecuaciones que por el método de mínimos cuadrados. Para

demostrar esto, daremos una breve introducción del método, como funciona y sus propiedades.

Definición 4.1-1: Función de verosimilitud

Sea $X_1, X_2, \cdots, X_n$ una sucesión de variables aleatorias independientes e idénticamente distribuidas, tomadas de una población con vector de parámetros $\theta \in \mathrm{R}^p$. Sea **X** un vector aleatorio cuyos elementos corresponden a la sucesión de variables aleatorias antes definida. Sea $f(X,\theta)$ la función de distribución multivariada correspondiente al vector aleatorio **X** y al vector de parámetros θ. Entonces, se dice que $f(X,\theta)$ es la función de verosimilitud correspondiente a **X** y θ.

El método de estimación de máxima verosimilitud, consiste hallar el valor del vector θ en términos del vector **X** tal que el valor de $f(\mathbf{X},\theta)$ sea máximo.

En algunos casos, la función de verosimilitud es diferenciable en θ, por lo que para hallar su valor máximo, de deriva parcialmente la función de verosimilitud con respecto a cada uno de los elementos del vector de parámetros, y se iguala cada ecuación a cero.

Usualmente, las funciones de verosimilitud vienen dadas en términos de multiplicaciones, y su diferenciación puede ser complicada. En estos casos, se busca una función continua y monótona, la cual se "compone" con la función de verosimilitud, para simplificar la expresión original. La nueva función así

construida tendrá los mismo puntos críticos que la función de verosimilitud original. Lo usual es escoger el logaritmo natural, puesto que esta función convierte los productos en suma, y es monótona y diferenciable en todos los reales positivos.

Para simplificación de notación utilizaremos la notación matricial del modelo lineal.

Supongamos que se tiene el modelo lineal en su forma matricial

$$\mathbf{Y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \qquad \textbf{4.1-1}$$

correspondiente a n variables aleatorias independientes e idénticamente distribuidas de un modelo con p parámetros. Supóngase además que el vector de errores tiene una distribución normal multivariada con media $\mathbf{0}$ y matriz de varianzas y covarianzas $\sigma^2\mathbf{I}$. Entonces, el vector Y tiene también distribución normal multivariada, con la misma matriz de varianzas y covarianzas que ε, pero con media en $X\beta$.

La función de verosimilitud correspondiente es

$$f(\mathbf{Y},\boldsymbol{\beta}) = (2\pi)^{-n/2}\det(\sigma^2\mathbf{I})^{-1/2}\exp\left(-\frac{1}{2}(\mathbf{Y}-\mathbf{X\beta})^T(\sigma^2\mathbf{I})^{-1}(\mathbf{Y}-\mathbf{X\beta})\right) \qquad \textbf{4.1-2}$$

donde la expresión $\det(\sigma^2\mathbf{I})$ se refiere al determinante de la matriz $\sigma^2\mathbf{I}$.

Aplicando logaritmo natural la expresión **4.1-2** y simplificándola, se obtiene

$$\ln(f(\mathbf{Y},\boldsymbol{\beta})) = -\frac{n}{2}\ln(2\pi) - \ln(\sigma) - \frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \qquad \textbf{4.1-3}$$

luego, obtenemos las derivadas parciales de esta expresión y las igualamos a cero

$$\frac{\partial}{\partial\boldsymbol{\beta}}\ln(f(\mathbf{Y},\boldsymbol{\beta})) = -\frac{1}{2\sigma^2}\frac{\partial}{\partial\boldsymbol{\beta}}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = -\frac{1}{2\sigma^2}\frac{\partial Q}{d\boldsymbol{\beta}} = 0 \qquad \textbf{4.1-4}$$

donde Q es la suma cuadrática del error. Como el primer factor es diferente de cero, el segundo factor tiene que ser igual a cero. Luego

$$\frac{\partial Q}{\partial\boldsymbol{\beta}} = 0 \qquad \textbf{4.1-5}$$

que corresponden a las ecuaciones normales del modelo (**3.3-9**).

4.2 *PROPIEDADES DE LOS ESTIMADORES DE MÍNIMOS CUADRADOS*

El capítulo anterior hace una introducción del método de los mínimos cuadrados. En esta sección veremos algunas propiedades de este método, así como la distribución de los estimadores obtenidos por este método.

Teorema 4.2-1: Propiedades de los estimadores de mínimos cuadrados

1. La media de los valores observados es igual a la media de los valores estimados.

2. La media aritmética de los errores estimados es igual a cero.

3. Los estimadores obtenidos por el método de mínimos cuadrados son los mismos que los de máxima verosimilitud.

4. Los estimadores obtenidos por el método de mínimo cuadrados tienen la característica de UMVU.

De estas propiedades, hemos demostrado ya las tres primeras.

Para demostrar la cuarta, demostraremos en primer lugar que el estimador obtenido por el método de mínimos cuadrados es insesgado. En efecto

$$E[\mathbf{b}] = E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right] = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E[\mathbf{Y}] = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta} \qquad \textbf{4.2-1}$$

Nos interesa demostrar ahora que el elemento b_j del vector **b** es el estimador insesgado de mínima varianza del elemento βj del vector β. El elemento b_j de puede obtener multiplicando por la izquierda de **b**, una matriz $\mathbf{a} \in M_{px1}$, tal que el elemento en la posición $j + 1$ del vector **a** sea uno, y los elementos en las demás posiciones sean cero. De hecho, podemos generalizar la demostración, suponiendo que **a** pueda ser cualquier vector. Entonces, la expresión que deseamos estimar tendría la forma

$$\begin{bmatrix} a_0 & a_1 & \cdots & a_{p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} = a_0\beta_0 + \cdots + a_{p-1}\beta_{p-1} \qquad \textbf{4.2-2}$$

A continuación demostraremos que el estimador insesgado de mínima varianza de $\mathbf{a}^T\beta$ es $\mathbf{a}^T\mathbf{b}$, es decir el estimador de mínimos cuadrados.

Recordemos que $\mathbf{b}$ es la solución del sistema $\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$. Supongamos que este sistema tienen solución única; por tanto, la matriz $\mathbf{X}^T\mathbf{X}$ es invertible. Entonces tenemos

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \qquad \textbf{4.2-3}$$

Supongamos que deseamos estimar el vector $\theta = \mathbf{X}\boldsymbol{\beta}$, y que θ solo puede tomar valores en Θ. Luego, el estimador de mínimos cuadrados de este vector sería $\hat{\theta} = \mathbf{X}\mathbf{b} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$. Este estimador, es una función lineal del vector aleatorio $\mathbf{Y}$. Para simplificar la demostración, supondremos que

$$\mathbf{P} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \qquad \textbf{4.2-4}$$

de esta manera, el estimador de mínimos cuadrados de θ quedaría como $\hat{\theta} = \mathbf{P}\mathbf{Y}$.

Ahora, $E\left[\hat{\theta}\right] = E[\mathbf{X}\mathbf{b}] = \mathbf{X}E[\mathbf{b}] = \mathbf{X}\boldsymbol{\beta} = \theta$, es decir el vector $\hat{\theta}$ es un estimador insesgado del vector θ. Supongamos ahora que $\mathbf{c}$ es cualquier vector de dimensión *n* x *1*. Demostraremos que el estimador insesgado con menor varianza de $\mathbf{c}^T\theta$ es $\mathbf{c}^T\hat{\theta}$, es decir el de mínimos cuadrados.

Tenemos que

$$E\left[\mathbf{c}^T\hat{\theta}\right] = \mathbf{c}^T E\left[\hat{\theta}\right] = \mathbf{c}^T\theta \qquad \textbf{4.2-5}$$

luego, es un estimador insesgado.

Supongamos ahora que $\mathbf{d}^T\mathbf{Y}$ es cualquier estimador insesgado de $\mathbf{c}^T\theta$ diferente de $\mathbf{c}^T\hat{\theta}$. Entonces, el valor esperado de este estimador sería $\mathbf{c}^T\theta$, es decir

$$\mathbf{c}^T\theta = E\left[\mathbf{d}^T\mathbf{Y}\right] = \mathbf{d}^T E[\mathbf{Y}] = \mathbf{d}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{d}^T\theta \qquad \textbf{4.2-6}$$

Luego

$$\mathbf{c}^T\theta = \mathbf{d}^T\theta \Rightarrow (\mathbf{c}-\mathbf{d})^T\theta = \mathbf{0} \qquad \textbf{4.2-7}$$

lo cual quiere decir que el vector **c** - **d** es ortogonal θ, para cualquier θ pertenciente al espacio Θ.

Ahora, puede demostrarse que el recorrido de **P** (espacio columna de **P**), es igual al recorrido de **X**. Además como $\mathbf{X}\beta = \theta$, para todo β en R^p, luego el recorrido de **X** sería igual a Θ. De aquí que cualquier vector que sea ortogonal a todos los vectores en Θ, es también ortogonal a **X** y a **P**. Luego

$$\mathbf{P}(\mathbf{c}-\mathbf{d}) = \mathbf{0} \Rightarrow \mathbf{Pc} = \mathbf{Pd} \qquad \textbf{4.2-8}$$

Ahora,

$$\mathbf{P}^T = \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T = \mathbf{X}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)^T\mathbf{X}^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{P} \qquad \textbf{4.2-9}$$

y

$$\mathbf{PP} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T = \mathbf{P} \qquad \textbf{4.2-10}$$

es decir, la matriz **P** es simétrica e idempotente. Puede demostrarse que **I** - **P** también es simétrica e idempotente

Luego,

$$\begin{aligned}&\text{var}\left[\mathbf{c}^T\hat{\theta}\right] = \text{var}\left[\mathbf{c}^T\mathbf{PY}\right] = \text{var}\left[\left(\mathbf{P}^T\mathbf{c}\right)^T\mathbf{Y}\right] = \text{var}\left[\left(\mathbf{Pc}\right)^T\mathbf{Y}\right] = \text{var}\left[\left(\mathbf{Pd}\right)^T\mathbf{Y}\right]\\ &= \left(\mathbf{Pd}\right)^T\text{var}[Y]\mathbf{Pd} = \mathbf{d}^T\mathbf{P}\left(\sigma^2\mathbf{I}\right)\mathbf{Pd} = \sigma^2\mathbf{d}^T\mathbf{PPd} = \sigma^2\mathbf{d}^T\mathbf{Pd}\end{aligned}$$

Entonces

$$\begin{aligned}&\text{var}\left[\mathbf{d}^T\mathbf{Y}\right] - \text{var}\left[\mathbf{c}^T\hat{\theta}\right] = \mathbf{d}^T\,\text{var}[\mathbf{Y}]\mathbf{d} - \text{var}\left[\mathbf{c}^T\hat{\theta}\right] = \mathbf{d}^T\sigma^2\mathbf{Id} - \sigma^2\mathbf{d}^T\mathbf{Pd}\\ &= \sigma^2\left(\mathbf{d}^T\mathbf{d} - \mathbf{d}^T\mathbf{Pd}\right) = \sigma^2\mathbf{d}^T\left(\mathbf{I} - \mathbf{P}\right)\mathbf{d} = \sigma^2\mathbf{d}^T\left(\mathbf{I} - \mathbf{P}\right)^T\left(\mathbf{I} - \mathbf{P}\right)\mathbf{d} = \sigma^2\mathbf{d}_1{}^T\mathbf{d}_1 \geq 0\end{aligned}$$

de aquí que

$$\text{var}\left[\mathbf{d}^T\mathbf{Y}\right] \geq \text{var}\left[\mathbf{c}^T\hat{\theta}\right] \qquad \textbf{4.2-11}$$

Es decir, que la varianza de cualquier otro estimador insesgado de $\mathbf{c}^T\theta$ diferente de $\mathbf{c}^T\hat{\theta}$, es mayor o igual que la varianza de $\mathbf{c}^T\hat{\theta}$. Esto implica que $\mathbf{c}^T\hat{\theta}$ es el estimador insesgado de mínima varianza de $\mathbf{c}^T\theta$, para cualquier vector constante **c**.

Ahora, supongamos que $\mathbf{c}^T = \mathbf{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$, entonces $\mathbf{c}^T\theta = \mathbf{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{a}^T\boldsymbol{\beta}$, y el estimador insesgado de mínima varianza de $\mathbf{c}^T\theta$, sería $\mathbf{c}^T\hat{\theta} = \mathbf{c}^T\mathbf{X}\mathbf{b} = \mathbf{a}^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{a}^T\mathbf{b}$. Luego, el estimador UMVU de $\mathbf{a}^T\beta$ sería $\mathbf{a}^T\mathbf{b}$, que es el estimador obtenido por el método de mínimos cuadrados.

Teorema 4.2-2: Valor esperado de la media cuadrática del error

Sea $\mathbf{Y}=\mathbf{X}\beta+\varepsilon$ un modelo lineal tal que ε tiene distribución normal *n*-multivariada, con media $\mathbf{0}$ y varianza $\sigma^2\mathbf{I}$, $\beta \in \mathrm{R}^p$. Sea $\mathbf{b}$ el estimador de mínimos cuadrados del vector β. Entonces

$$E\left[\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{n-p}\right] = \sigma^2 \qquad \textbf{4.2-12}$$

Para probar este teorema, utilizaremos la matriz $\mathbf{P}$ definida en **4.2-4**. Esta matriz es simétrica e idempotente (**4.2-9** y **4.2-10**). Puede demostrarse que la matriz $\mathbf{I}$ - $\mathbf{P}$ también es simétrica e idempotente.

Además utilizaremos los siguientes resultados

- Sea $\mathbf{X}$ un vector aleatorio y $\mathbf{A}$ una matriz simétrica, entonces $\mathrm{var}\left(X^T AX\right) = tr\left(A\,\mathrm{var}(X)\right) + E\left[X^T\right]AE[X]$. **4.2-13**

- $tr(\mathbf{I}-\mathbf{P}) = n-p$. **4.2-14**

- $(\mathbf{I}-\mathbf{P})\mathbf{X}=\mathbf{0}$ **4.2-15**

Entonces

$$E\left[\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{n-p}\right]=\frac{1}{n-p}E\left[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right] \quad \textbf{4.2-16}$$

luego

$$\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}=\mathbf{Y}-\mathbf{P}\mathbf{Y}=(\mathbf{I}-\mathbf{P})\mathbf{Y} \quad \textbf{4.2-17}$$

lo cual implica

$$(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})=\mathbf{Y}^T(\mathbf{I}-\mathbf{P})(\mathbf{I}-\mathbf{P})\mathbf{Y}=\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y} \quad \textbf{4.2-18}$$

Aplicando esta expresión en **4.2-16** se obtiene

$$\frac{1}{n-p}E\left[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right]=\frac{1}{n-p}E\left[\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y}\right] \quad \textbf{4.2-19}$$

Usando el resultado **4.2-13** se tiene

$$\begin{aligned}\frac{1}{n-p}E\left[\mathbf{Y}^T(\mathbf{I}-\mathbf{P})\mathbf{Y}\right]&=tr\left((\mathbf{I}-\mathbf{P})\mathrm{var}(\mathbf{Y})\right)+E\left[\mathbf{Y}^T\right](\mathbf{I}-\mathbf{P})E[\mathbf{Y}]\\&=\frac{1}{n-p}\left(\sigma^2 tr(\mathbf{I}-\mathbf{P})+\boldsymbol{\beta}^T\mathbf{X}^T(\mathbf{I}-\mathbf{P})\mathbf{X}\boldsymbol{\beta}\right)\end{aligned} \quad \textbf{4.2-20}$$

Empleando los resultados **4.2-14** y **4.2-15** se tiene

$$E\left[\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{n-p}\right]=\frac{1}{n-p}\left(\sigma^2(n-p)+\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{0}\boldsymbol{\beta}\right)=\sigma^2\frac{n-p}{n-p}=\sigma^2$$

Teorema 4.2-3: Distribución del estimador de mínimos cuadrados

Sea $\mathbf{Y}=\mathbf{X}\beta+\varepsilon$ un modelo lineal tal que ε tiene distribución normal *n*-multivariada, con media $\mathbf{0}$ y varianza $\sigma^2\mathbf{I}$, $\beta \in \mathbb{R}^p$. Sea $\mathbf{b}$ el estimador de mínimos cuadrados del vector β. Entonces $\mathbf{b}$ tiene una distribución normal multivariada, con media β y matriz de varianzas y covarianzas $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

Para demostrar este teorema, nos basamos en la distribución de $\mathbf{Y}$, que es normal multivariada con media $\mathbf{X}\beta$ y varianza $\sigma^2\mathbf{I}$. Si $\mathbf{b}$ es el estimador de mínimos cuadrados de β, entonces

$$\mathbf{b}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \qquad \textbf{4.2-21}$$

luego, $\mathbf{b}$ tiene distribución normal *p*-multivariada, tal que su media es β y su varianza

$$\begin{aligned}\boldsymbol{\Sigma}&=\mathrm{var}(\mathbf{b})=\mathrm{var}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}\right)=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\,\mathrm{var}(\mathbf{Y})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\\&=\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{X}\right)^{-1}=\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\end{aligned} \qquad \textbf{4.2-22}$$

En la realidad, no conocemos σ^2, sin embargo lo podemos estimar a través de la media cuadrática del error, es decir

$$\hat{\sigma}^2=MCE=\frac{1}{n-p}SCE=\frac{1}{n-p}\left(\mathbf{Y}^T\mathbf{Y}-\mathbf{b}^T\mathbf{X}^T\mathbf{Y}\right) \qquad \textbf{4.2-23}$$

Luego, si **S** es el estimador de la matriz de varianzas y covarianzas del vector **b**, se puede probar que

$$\hat{\Sigma} = \mathbf{S} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} MCE \qquad \textbf{4.2-24}$$

Teorema 4.2-4: Distribución muestral del estimador de mínimos cuadrados

Sea $\mathbf{Y}=\mathbf{X}\beta+\varepsilon$ un modelo lineal tal que ε tiene distribución normal *n*-multivariada, con media **0** y varianza $\sigma^2\mathbf{I}$, $\beta \in \mathrm{R}^p$. Sea **b** el estimador de mínimos cuadrados del vector β. Si $\mathbf{S} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} MCE$ es el estimador de la matriz de varianzas y covarianzas del vector **b**. Entonces, la variable aleatoria $U = n\left(\mathbf{b}-\boldsymbol{\beta}\right)^T \mathbf{S}^{-1}\left(\mathbf{b}-\boldsymbol{\beta}\right)$ tiene una distribución T^2, con p y $(n\text{-}p)$ grados de libertad.

La demostración de este teorema, escapa del alcance de este trabajo.

Puede probarse que la matriz $\mathbf{S} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} MCE$ es definida positiva. Luego, la expresión $n\left(\mathbf{b}-\boldsymbol{\beta}\right)^T \mathbf{S}^{-1}\left(\mathbf{b}-\boldsymbol{\beta}\right) = k^2$ es una *forma cuadrática* determinada por una matriz definida positiva, e igualada a una constante positiva. Por tanto representa una elipsoide en R^p. Si $p = 2$, la ecuación representaría una elipse.

Para definir una región de confianza, se tiene que esta variable aleatoria, con (1-α)100% de confianza, siempre es menor que $T^2_{\alpha,p,n-p}$. Así la región determinada por

$$n(\mathbf{b}-\boldsymbol{\beta})^T \mathbf{S}^{-1}(\mathbf{b}-\boldsymbol{\beta}) \leq T^2{}_{\alpha,p,n-p} \qquad \textbf{4.2-25}$$

define una elipsoide de confianza para β, o intervalo simultáneo de confianza para los elementos de β.

Se podría construir intervalos de confianza para cada parámetro, empleando la distribución *t* de student, pero los parámetros, en general, están correlacionados, y los intervalos de confianza simultáneos, no consisten en los rectángulos definidos por los intervalos separados, sino que hay que considerar la correlación entre la variables. Si la correlación de los elementos de **b** es nula, entonces es mejor utilizar los rectángulos definidos por los intervalos de confianza de cada elemento de **b**, para realizar los intervalos de confianza simultáneos de para β.

4.3 PROCEDIMIENTOS NO PARAMÉTRICOS

Una clase de procedimientos más robustos, pero menos potentes, que el método de mínimos cuadrados para estimar parámetros de un modelo lineal, son los procedimiento no paramétricos.

Estos, no hacen supuesto alguno sobre la distribución del vector de error. Más bien se pueden aplicar para cualquier distribución del error que sea simétrica. Esto hace, que este estimador sea más robusto, pero en cambio es menos eficiente que el estimador de mínimos cuadrados. En estos procedimientos, no se trata de minimizar el error, ni la varianza.

A continuación, describiremos un método no paramétrico para estimar los parámetros de un modelo de regresión lineal simple.

Supongamos que tenemos el modelo

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i; \quad i = 1, \cdots, n \qquad \textbf{4.3-1}$$

Para estimar los parámetros de este modelo, procederemos de la siguiente manera:

Sea $N = \binom{n}{2}$ el número de parejas que se pueden tomar de los n datos.

Por cada pareja, calcúlese el valor

$$S_{ij} = \frac{y_j - y_i}{x_j - x_i}; \quad 1 \le i < j \le n \qquad \textbf{4.3-2}$$

Luego

$$b_1 = \hat{\beta}_1 = \text{mediana}\{S_{ij}\} \qquad \textbf{4.3-3}$$

Es decir, el estimador del parámetro de *pendiente* de la recta que representa el modelo, es la mediana de todas las pendientes que se pueden formar de las distintas rectas determinadas por las diferentes parejas de puntos.

De manera similar, la estimación del parámetro de *elevación natural* sería

$$b_0 = \hat{\beta}_0 = \text{mediana}\{y_i - S_{ij}x_i\} \qquad \textbf{4.3-4}$$

Es decir, la estimación del parámetro de elevación natural de la recta que representa el modelo es la mediana de todas las intersecciones del eje Y de las diferentes rectas que determinan las diferentes parejas de puntos.

Nótese que es indiferente si se reemplaza y_i y x_i por y_j y x_j, puesto que los dos puntos pertenecen a la misma recta, y la pendiente es S_{ij}.

Ejemplo 4.3-1: Simulación de un modelo de regresión simple.

Supongamos el modelo

$$y = 4 + 7x + \varepsilon$$

Supongamos que hacemos una corrida de este modelo, con $n = 20$, y que ε tiene una distribución uniforme con parámetros –3 y 3. Luego, la media del error sería cero, y su varianza sería 3.

Tenemos entonces

i	x_i	ε_i	y_i
1	4	-0.309	31.691
2	4.2	-1.606	31.794
3	4.4	1.436	36.236
4	4.6	-2.603	33.597
5	4.8	-0.966	36.634
6	5	1.904	40.904
7	5.2	-2.810	37.59
8	5.4	1.515	43.315
9	5.6	-1.320	41.88
10	5.8	1.685	46.285
11	6	-1.331	44.669
12	6.2	-2.844	44.556
13	6.4	1.753	50.553
14	6.6	0.588	50.788
15	6.8	0.859	52.459
16	7	-2.906	50.094
17	7.2	2.031	56.431
18	7.4	-0.942	54.858
19	7.6	-2.360	54.84
20	7.8	1.214	59.814

Tabla 4.3-1

Ahora, estimemos los parámetros $\beta_0 = 4$ y $\beta_1 = 7$, por el método descrito en esta sección.

Se pueden formar 190 parejas diferentes de los puntos observados. Si formamos las 190 parejas, y obtenemos la pendiente en cada una, y luego sacamos la mediana de todas ellas, se obtiene

$$\hat{\beta}_1 = 7.081$$

De igual manera, hacemos la misma tarea, pero para la elevación natural, obteniéndose

$$\hat{\beta}_0 = 3.222$$

Ahora, los estimadores obtenidos por el método de mínimos cuadrados serían:

$$b_0 = 3.043; \quad b_1 = 7.103$$

En este caso, la estimación por el método no paramétrico se acerca más a los parámetros originales que la estimación de mínimos cuadrados, pero la diferencia no es mayor. Si obtenemos la diferencia porcentual de ambas estimaciones, tenemos que la del método no paramétrico es 9.7%, mientras que la del método de mínimos cuadrados es 11.94%.

Cuando existen valores aberrantes en las observaciones, las estimaciones obtenidas mediante procedimientos no paramétricos, son en general más cercanas a los parámetros verdaderos, que las estimaciones obtenidas por el método de mínimos cuadrados.

Para el ejemplo anterior, no existían valores aberrantes, sin embargo la distribución del error no fue normal, y por tanto no se cumplían los supuestos del método de mínimos cuadrados. Sin embargo, la estimación fue precisa, pero la del método no paramétrico fue mejor.

5 EL MÉTODO DE LA "MÍNIMA MEDIANA DE LOS CUADRADOS" (MMC) COMO PROCEDIMIENTO PARA ESTIMAR PARÁMETROS EN UN MODELO LINEAL

5.1 INTRODUCCIÓN

El método de mínimos cuadrados, si bien es eficiente, no es robusto. Inclusive, cuando se cumplen todos los supuestos que se hacen para poder aplicarlo, éste puede fallar, omite la presencia de valores aberrantes.

Hoy en día, se busca métodos de estimación de parámetros, que tengan mayor *ruptura*. El término *ruptura* se refiere a la capacidad que tiene un método, para hacer estimaciones que no sean afectadas por la presencia de valores aberrantes.

No debemos confundir los términos robustez u ruptura. El uno se refiere a cambios en las condiciones iniciales, y el otro se refiere a la presencia de

valores aberrantes. Lo que sí se puede afirmar, es que una alta ruptura, implica un mayor grado de robustez.

En 1984, Rousseeuw propuso un método, que es mucho más robusto que el de mínimos cuadrados, en cuanto a su ruptura. Este es el método de la *Mínima mediana de los Cuadrados*, o MMC.

Éste método, a diferencia de los procedimientos no paramétricos, sí considera al error. De hecho, en este método se trata de minimizar una función que mida el error del modelo. La idea es la misma que en mínimos cuadrados, pero la función objetivo es otra (véase **5.2-2**).

Puede demostrarse que esta función objetivo, si bien es continua, no es diferenciable en todos los puntos, y tiene muchos mínimos locales.

Se han propuesto diferentes algoritmos para estimar parámetros por el método MMC, pero estos algoritmos requieren de procesos largos para realizar la estimación, y usualmente su funcionamiento ha sido engorroso.

Sin embargo, con la velocidad de las computadoras modernas, y las grandes capacidades de memoria, a bajo costo, cada vez importa menos el número de instrucciones que realice un algoritmo o la cantidad de memoria que ocupe.

Uno de los algoritmos más usados para estimar parámetros por el método MMC, es el algoritmo denominado PROGRESS, propuesto por Rousseeuw y

Leroy en 1984. Con este algoritmo, no se puede encontrar el estimador LMS, pero si se puede encontrar una aproximación.

En 1992, Hettmansperger y Sheather probaron que el algoritmo PROGRESS es muy insensible cuando se realizan pequeños cambios en los datos.

En 1993, Hawkins y Simonoff, diseñaron un algoritmo, denominado MVELMS, que provee también una aproximación del estimador MMC, pero para el caso de 2 parámetros, el algoritmo provee el estimador "exacto" MMC.

En 1993, Stromberg propuso un algoritmo que encuentra el estimador MMC exacto, para cualquier número de parámetros, y que en el ejemplo que Hettmansperger y Sheather demostraron que PROGRESS era inestable, el algoritmo de Stomberg fue más estable. Sin embargo, en el mismo año, Stromberg concluye que su algoritmo es inestable con ciertos conjunto de datos.

En este trabajo emplearemos el algoritmo de Stromberg, para realizar las estimaciones por el método MMC.

5.2 EL MÉTODO MMC

En el método de mínimos cuadrados, se intenta hallar el estimador que minimice la suma cuadrática total. Esto es

$$\min_{\beta}\left\{\sum_{i=1}^{n}\varepsilon_i^{\,2}(\beta)\right\} \qquad \textbf{5.2-1}$$

Es decir, cada β posible, tiene su propio vector de error ($\varepsilon(\beta) = \mathbf{Y} - \mathbf{X}\beta$). Luego, el estimador de mínimos cuadrados **b** es tal que, la función $\varepsilon^T(\beta)\,\varepsilon(\beta)$ tiene un mínimo absoluto en $\beta = \mathbf{b}$.

En el método de la mínima mediana de los cuadrados (MMC), lo que cambia es la función objetivo. En lugar de ser la suma cuadrática del error, tenemos como función objetivo la mediana de los elementos del vector de error correspondiente a β, es decir

$$\min_{\beta}\left\{mediana\left(\varepsilon_i^{\,2}(\beta)\right)\right\} \qquad \textbf{5.2-2}$$

Esta función es continua, pero no es diferenciable, es menos sensible a los valores aberrantes, y por tanto tiene mayor ruptura, y es más robusta.

Estadísticamente, la mediana muestral es

$$\widetilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}; & n \text{ impar} \\ \dfrac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right); & n \text{ par} \end{cases}$$

Para lograr la más alta ruptura posible en la estimación, en lugar de minimizar la mediana, minimizaremos el estadístico de orden *h*, tal que *h* podría ser

$$h = \left\lfloor \frac{n}{2} \right\rfloor + \left\lfloor \frac{p+1}{2} \right\rfloor \qquad \textbf{5.2-3}$$

donde p es el número de parámetros del modelo lineal, y n es el número de observaciones.

En 1990, Cook y Hawkins recomendaron que se debe investigar la estabilidad de un estimador MMC sobre un imagen de valores de h.

Nótese, que por la definición de h, el estadístico de orden h, es muy cercano a la mediana. De hecho, si n es impar, y solo tengo dos parámetros, este estadístico coincidiría con la mediana.

Entonces, el estimador MMC se reduce a

$$\min_{\beta}\left\{\varepsilon_h{}^2(\beta)\right\} \qquad \textbf{5.2-4}$$

Este estimador, no es muy fácil de calcular. A continuación describiremos como se calcula el estimador MMC, según el algoritmo de Stromberg.

Primero, el estimador MMC, minimiza el cuadrado del *h-ésimo* residuo mayor, o el cuadrado del residuo de orden h, de un conjunto dado de datos. Entonces, el estimador MMC debe minimizar el máximo de los cuadrados de los residuos, para algún subconjunto de tamaño h. Esto quiere decir, que el estimador MMC vendría a ser una estimación por el método de Chebyshev" (ó minimax), pues intenta hallar un subconjunto de los datos originales que minimice el máximo de los cuadrados residuales.

Aparentemente, habría que examinar todos los conjuntos de tamaño h, de los n datos. El número de subconjuntos a examinar sería $\binom{n}{h}$, lo cual es bastante elevado, porque h es un valor cercano a la mitad de n.

Sin embargo, Cheney demostró en 1962, que el estimador de Chebyshev de todos los datos, es el estimador de Chebyshev para un subconjunto de tamaño p+1. Esto quiere decir, solamente hay que revisar $\binom{n}{p+1}$, lo cual es mucho menor que $\binom{n}{h}$.

Luego para cada subconjunto de tamaño p+1, tendremos una matriz de diseño y una matriz de observaciones. Por cada β, tendremos un residuo e_i (β).

Supongamos ahora que para un subconjunto E, la matriz de diseño es $\mathbf{X}_E$ y la matriz de observaciones es $\mathbf{Y}_E$. Sea $\hat{\boldsymbol{\beta}}_{EMC}$, el estimador de mínimos cuadrados correspondiente a los puntos del conjunto E, esto es

$$\hat{\boldsymbol{\beta}}_{EMC} = \left(\mathbf{X}_E{}^T \mathbf{X}_E\right)^{-1} \mathbf{X}_E{}^T \mathbf{Y}_E \qquad \textbf{5.2-5}$$

A continuación, defínase δ como

$$\delta = \frac{\sum_{i=1}^{p+1} e_i{}^2\left(\hat{\boldsymbol{\beta}}_{EMC}\right)}{\sum_{i=1}^{p+1} \left| e_i\left(\hat{\boldsymbol{\beta}}_{EMC}\right)\right|} \qquad \textbf{5.2-6}$$

Ahora, sea **s** un vector columna de dimensión p+1, tal que $s_i = \text{sgn}\left(e_i\left(\hat{\boldsymbol{\beta}}_{EMC}\right)\right)$. Luego, la estimación de Chebyshev para el conjunto de datos E sería la solución del siguiente sistema lineal:

$$\left(\mathbf{X}_E{}^T\mathbf{X}_E\right)\hat{\boldsymbol{\beta}}_{EC} = \mathbf{X}_E{}^T\left(\mathbf{Y}_E - \delta\mathbf{s}\right) \qquad \textbf{5.2-7}$$

El algoritmo e Stromberg consiste en examinar todos los subconjuntos posibles de tamaño p+1, y para cada subconjunto calcular $\hat{\boldsymbol{\beta}}_{EC}$. Luego el estimador LMS exacto, $\hat{\boldsymbol{\beta}}_{MMC}$ sería aquel $\hat{\boldsymbol{\beta}}_{EC}$, que aplicado con los datos originales, produzca el menor error cuadrático de orden h.

5.3 CONSIDERACIONES SOBRE EL MÉTODO MMC

El algoritmo descrito anteriormente, funciona bastante bien con un número pequeño de datos, y un número pequeño de parámetros. Sin embargo, El número de sistemas lineales que hay que resolver crece considerablemente, conforme el número de datos es mayor. Por ejemplo, con tres parámetros, el número de sistemas lineales cuatro por cuatro por resolver para 30 datos es $2\binom{30}{4} = 54{,}810$, y para 90 datos es $2\binom{90}{4} = 5{,}110{,}380$. Es decir, al triplicar el número de datos, el número de sistemas lineales por resolver se multiplica aproximadamente por 100.

Para una computadora moderna, esto no es mucho problema, porque el tiempo en que se lo haría es relativamente pequeño.

Sin embargo, en este trabajo, la intención es simular un modelo de regresión, y estimar los parámetros por el método MMC, además de otros métodos. Si se simula un número suficientemente grande de veces el modelo de regresión, esta simulación podría demorar bastante tiempo, aún en las computadoras más modernas.

A continuación, se presenta un ejemplo, para ilustrar el método mostrado en la sección anterior

Ejemplo 5.3-1: Método MMC

Supongamos el modelo lineal

$$y = 3 - 2x + \varepsilon \qquad \textbf{5.3-1}$$

donde ε tiene distribución normal con media cero y varianza uno. Si hacemos una corrida de este modelo para $n = 4$, entonces

i	x_i	y_i	ε_i
1	1	1.4518	0.4518
2	1.5	-0.087	-0.087
3	2	-1.837	-0.837
4	2.5	-3.298	-1.298

Tabla 5.3-1

Luego, para hallar el estimador MMC, examinaremos los $\binom{4}{3} = 4$ subconjuntos posibles de tamaño 3 de los datos, esto es { 1, 2, 3 }, { 1, 2, 4 }, { 1, 3, 4 }y { 2, 3, 4}.

Para cada subconjunto, encontraremos el estimador de Chebyshev. Por ejemplo, para el primer subconjunto tendremos

$$\mathbf{X}_E = \begin{bmatrix} 1 & 1 \\ 1 & 1.5 \\ 1 & 2 \end{bmatrix} \qquad \mathbf{Y}_E = \begin{bmatrix} 1.452 \\ -0.087 \\ -1.837 \end{bmatrix}$$

Luego,

$$\mathbf{X}_E{}^T\mathbf{X}_E = \begin{bmatrix} 3 & 4.5 \\ 4.5 & 7.25 \end{bmatrix} \qquad \mathbf{X}_E{}^T\mathbf{Y}_E = \begin{bmatrix} -0.472 \\ -2.352 \end{bmatrix}$$

Tenemos ahora el sistema

$$\begin{bmatrix} 3 & 4.5 \\ 4.5 & 7.25 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} -0.472 \\ -2.352 \end{bmatrix}$$

La solución de este sistema es

$$\hat{\boldsymbol{\beta}}_{ELS} = \begin{bmatrix} 4.776 \\ -3.289 \end{bmatrix}$$

Luego

$$\delta = 0.053 \qquad \mathbf{s} = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}$$

Entonces, el nuevo sistema es

$$\begin{bmatrix} 3 & 4.5 \\ 4.5 & 7.25 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} -0.419 \\ -2.273 \end{bmatrix}$$

La solución de este nuevo sistema es

$$\hat{\boldsymbol{\beta}}_{EC} = \begin{bmatrix} 4.793 \\ -3.289 \end{bmatrix}$$

Si aplicamos esta estimación de los parámetros a los datos originales, se tiene el siguiente vector de error:

$$e = \begin{bmatrix} -0.053 \\ 0.053 \\ -0.053 \\ 0.13 \end{bmatrix}$$

Para este caso $h = 3$. Si tomamos los cuadrados de los elementos de este vector, y obtenemos el tercer menor, este sería 0.0029.

Luego, repetimos la operación para cada subconjunto:

E	β_0	β_1	e_h^2
{ 1, 2, 3 }	4.793	-3.289	0.002804
{ 1, 2, 4 }	4.641	-3.167	0.000503
{ 1, 3, 4 }	4.557	-3.167	0.003728
{ 2, 3, 4 }	4.658	-3.211	0.000025

Tabla 5.3-2

Entonces, el que tiene menor error cuadrático de orden 3 es

$$\hat{\boldsymbol{\beta}}_{LMS} = \begin{bmatrix} 4.658 \\ -3.211 \end{bmatrix}$$

La estimación por mínimos cuadrados para β, es

$$\mathbf{b} = \begin{bmatrix} 4.658 \\ -3.2 \end{bmatrix}$$

Como vemos en este ejemplo, los estimadores obtenidos por ambos métodos son similares, ambos están alejados de los parámetros originales. Esto se debe a que hay muy pocos datos para realizar la estimación.

6 ANÁLISIS COMPARATIVO POR SIMULACIÓN DE LOS DIFERENTES MÉTODOS CONSIDERADOS

6.1 INTRODUCCIÓN

Actualmente, el método de estimación de parámetros de modelos lineales más ampliamente usado, es el método MC (mínimos cuadrados). El método MMC (mínima mediana de los cuadrados) es nuevo, y se supone que es más robusto que el método MC. El estudio comparativo que se muestra en esta sección, se enfoca principalmente en estos dos métodos.

Para la realización del estudio comparativo, se ha diseñado un programa de computadora. El lenguaje para el archivo fuente de este programa es Visual Basic. El programa está diseñado bajo Windows 95, y muestra un diseño modular. Esto hace que sea de fácil uso, y de fácil edición.

El programa tiene incorporado dos algoritmos de regresión: MC y MMC. También tiene una opción en la que se generan números aleatorios de alguna población.

La opción de simulación del programa, realiza un cierto número de "corridas", o repeticiones de un experimento teórico. En cada corrida, se generan números aleatorios que representan los errores del modelo. Con estos errores generados, procedemos a calcular los valores observados del modelo. Luego, se estiman los parámetros del modelo, por el método que el usuario haya escogido. Entonces, pone en una tabla los parámetros estimados correspondientes a esa corrida, para al final de todas la corridas, guardarlos.

El programa también tiene un algoritmo para estimar la matriz de varianzas y covarianzas de un grupo de datos. Esta opción se utiliza sobre una tabla en la que se encuentren las estimaciones de los parámetros de todas la corridas de alguna simulación. De esta manera, se estiman las varianzas de los estimadores de cada parámetro.

Para efectos de esta investigación, es necesaria la convergencia de las simulaciones. Esto tiene que ver con el número de corridas de cada simulación. Se realizaron pruebas con 100, 200 y 500 corridas para analizar la convergencia de las simulaciones. Para cada uno de los número de corridas, se probaron tamaños de muestra de 16, 30 y 49. La convergencia de las simulaciones, se dio para las 200 corridas, y tamaño de muestra 30. Por esto, todos los resultados

expuestos en esta investigación, están basados en simulaciones de 200 corridas con tamaño de muestra 30.

El modelo de regresión empleado en todas las simulaciones es

$$y_i = 3 - 4x_{1i} + 2x_{2i} \qquad \textbf{6.1-1}$$

es decir $\beta_0 = 3$, $\beta_1 = -4$ y $\beta_2 = 2$. La varianza del error para todas las simulaciones es uno, esto es, el error tiene distribución normal estándar.

Además, se analizan los casos en los que la población de la que se toma la muestra, presenta algún tipo de contaminación. En estos casos, se estudia también la "*cobertura*" de los intervalos de confianza de cada estimación. El término *cobertura* se refiere a una medida de la cantidad de intervalos de confianza, que contienen el parámetro poblacional.

6.2 COMPORTAMIENTO DE LOS ESTIMADORES EN CONDICIONES NORMALES

La tabla que se muestra a continuación, está clasificada en las filas, por el tamaño de la muestra, y en las columnas por el número de corridas. En cada elemento de la tabla, se encuentran el promedio de los estimadores de los parámetros correspondiente a un número de corridas y tamaño muestral. Los intervalos de confianza definidos son de 1 desviación estándar. Estas estimaciones son hechas con el método de mínimos cuadrados.

de corridas

Tamaño de la muestra	100	200	500
16	2.936±0.181	3.097±0.147	3.018±0.0157
	-3.992±0.037	-4.017±0.034	-4.000±0.036
	2.014±0.054	1.988±0.047	1.994±0.049
30	3.031±0.103	2.980±0.103	3.012±0.106
	-4.010±0.016	-3.990±0.016	-4.000±0.016
	2.000±0.031	1.994±0.030	1.996±0.029
49	3.011±0.077	3.031±0.075	3.011±0.084
	-3.998±0.011	-4.003±0.011	-4.002±0.011
	1.993±0.016	1.994±0.014	1.998±0.016

Tabla 6.2-1

Obsérvese que los valores de la primera columna de cada elemento de la tabla, son cercanos a 3, -4 y 2, es decir, a los parámetros. La segunda columna de cada elemento, tiene solo valores positivos. Esto es porque representa la estimación de la varianza correspondiente a cada estimador de parámetros.

A continuación mostramos una tabla similar a la anterior, pero las estimaciones son realizadas por el método de mínima mediana de los cuadrados.

de corridas

Tamaño de la muestra	100	200	500
16	2.641±0.387	3.150±0.385	2.934±0.354
	-3.926±0.084	-4.003±0.100	-3.989±0.090
	2.083±0.105	1.944±0.097	2.016±0.095
30	2.884±0.235	3.017±0.275	3.020±0.259
	-4.005±0.048	-3.998±0.047	-4.008±0.046
	2.037±0.064	1.993±0.068	2.010±0.066
49	3.073±0.202	2.974±0.208	
	-4.018±0.034	-3.989±0.036	
	2.001±0.033	1.994±0.034	

Tabla 6.2-2

Nótese que el promedio de las estimaciones es cercano al valor de los parámetros. Además, las desviaciones de esta tabla son significativamente mayores que las desviaciones de la tabla 6.2-1. Esto muestra que la varianza de los estimadores obtenidos por el método de mínimos cuadrados es menor que la varianza de los estimadores obtenidos por el método de la mínima mediana de los cuadrados.

Luego, bajo condiciones normales, la estimación obtenida por el método de mínimos cuadrados es más precisa que la estimación obtenida por el método de la mínima mediana de los cuadrados.

6.3 COMPORTAMIENTO DE ESTIMADORES CON POBLACIONES CONTAMINADAS

La generación de errores sin contaminación, se realizaba en la sección anterior, empleando una distribución normal estándar. La contaminación se realiza tomando en promedio el $100\alpha\%$ de las veces, muestras de una población normal, con media cero y varianza 16.

En la tabla que se muestra a continuación, se encuentran los valores de las estimaciones de los parámetros, tanto para el método de mínimos cuadrados (MC), como para el método de la mínima mediana de los cuadrados (MMC). Además, se muestran los porcentajes de cobertura de intervalos de confianza para cada caso. Estos intervalos son de 68% de confianza. Cada fila, corresponde a un nivel de contaminación diferente.

	Mínimos Cuadrados		Mínima Mediana de los Cuadrados	
0	2.980±0.103	68.00%	3.017±0.275	69.00%
	-3.990±0.016	65.50%	-3.998±0.047	68.50%
	1.994±0.030	66.00%	1.993±0.068	66.00%
0.05	3.014±0.139	68.50%	2.930±0.258	68.50%
	-4.007±0.021	73.00%	-4.002±0.047	70.00%
	2.004±0.040	71.50%	2.028±0.061	68.00%
0.10	2.957±0.178	75.00%	2.930±0.265	72.00%
	-4.000±0.032	73.50%	-4.002±0.045	70.00%
	2.011±0.049	71.00%	2.028±0.069	71.50%
0.15	2.937±0.202	70.50%	3.011±0.290	70.00%
	-3.994±0.036	69.00%	-3.998±0.048	70.50%
	2.020±0.054	68.50%	2.001±0.071	70.00%
0.20	3.116±0.254	71.00%	2.939±0.257	70.50%
	-3.999±0.038	68.50%	-3.997±0.049	71.50%
	1.975±0.060	69.50%	2.020±0.069	69.00%

Tabla 6.3-1

Como puede apreciarse, las coberturas son similares, tanto para mínimos cuadrados, como para mínima mediana de los cuadrados. Además, nótese que lo peor que le puede pasar a la varianza del estimador de mínimos cuadrados, es llegar a ser la misma que la varianza del estimador de mínima mediana de los cuadrados.

A continuación, se muestra un tabla donde, el nivel de contaminación es 0.10, pero los datos del cual se toma contaminar la muestra, tienen una distribución normal con medias diferentes de cero, pero varianza uno. Las filas representan las medias de la contaminación.

	Mínimos Cuadrados		Mínima Mediana de los Cuadrados	
0	2.980±0.103	68.00%	3.017±0.275	69.00%
	-3.990±0.016	65.50%	-3.998±0.047	68.50%
	1.994±0.030	66.00%	1.993±0.068	66.00%
1	3.081±0.113	63.00	3.037±0.306	66.00
	-3.998±0.019	68.00	-3.981±0.052	65.00
	2.004±0.030	64.00	2.000±0.069	68.00
2	3.169±0.131	73.50%	2.994±0.307	84.50%
	-3.996±0.025	83.00%	-3.995±0.054	85.00%
	2.006±0.035	81.50%	2.009±0.071	87.00%

Tabla 6.3-2

CONCLUSIONES

1. Al comparar las tablas 6.2-1 y 6.2-2, podemos apreciar que bajo condiciones en la que se cumplan los supuestos iniciales, el estimador de mínimos cuadrados resulta preferible que el estimador de la mínima mediana de los cuadrados. Para este caso, los resultados teóricos, resultaron concordantes con los resultados experimentales (simulaciones).

2. De la tabla 6.3-1, podemos concluir que el estimador de la mínima mediana de los cuadrados es mucho más robusto que el de mínimos cuadrados, puesto que la varianza del estimador MMC no varía significativamente, a pesar de la contaminación. En cambio, la varianza del estimador MC, aumenta conforme se aumenta el nivel de contaminación. Esto, también concuerda con la teoría.

3. En cuanto al sesgo, ambos estimadores presentan sesgos numéricamente iguales. Además, el sesgo en cada uno es nulo. Por tanto, podemos concluir que los dos estimadores son insesgados.

4. A pesar que la varianza del estimador MC aumenta, y la del estimador MMC se mantiene, la varianza del estimador MC siempre es estadísticamente menor que la del estimador MMC. El único caso en el que las varianzas de ambos estimadores son numéricamente iguales, es cuando el nivel de contaminación es alto (0.20). Esto quiere decir, que a pesar de la contaminación, el estimador MC es tan bueno como el estimador MMC.

5. En la tabla 6.3-2, es la que contenía contaminación con desplazamiento. La varianza del estimador MMC tuvo un aumento mayor que la varianza del estimador MC. Esto quiere decir, que para este tipo de contaminación, ambos métodos son similares en robustez. Otra vez, resulta mejor el estimador de mínimos cuadrados.

6. Según la teoría expuesta en este trabajo, el estimador MMC es más robusto que el de MC, y supuestamente el método MMC debería ser mejor en condiciones de contaminación que MC. La robustez se cumple como se predijo, pero la varianza del estimador MMC es tan grande, que a pesar de no aumentar con la contaminación, de todos modos siempre es mayor que la varianza del estimador MC, la cual aumenta, conforme aumenta la contaminación. Se Esperaba que el estimador MMC sea mucho mejor que el estimador MC, en circunstancias de contaminación de datos. Sin embargo, aquí se muestra que, no siempre se obtiene lo que uno espera hallar.

RECOMENDACIONES

1. Se recomienda utilizar en la mayoría de los casos el estimador de mínimos cuadrados, puesto que este método demuestra supremacía sobre el método de la mínima mediana de los cuadrados.

2. Solo cuando se conoce de antemano que la contaminación es significativamente grande (≥ 0.20), entonces se recomienda utilizar estimadores de mínima mediana de los cuadrados.

3. Cuando se sospecha de contaminación de datos, o cuando se detecta un nivel pequeño de valores aberrantes en las observaciones, es mejor emplear el método de mínimos cuadrados, porque la varianza de este esos casos, es menor que la del método de la mínima mediana de los cuadrados, aunque ambos aparentan ser insesgados

4. Cuando el tamaño de la muestra es grande, es mejor emplear el método de mínimos cuadrados, porque su varianza es menor que la del método de la mínima mediana de los cuadrados.

BIBLIOGRAFÍA

1. Rao C. Radhakrishna, *Linear Statistical Inference and Its Applications*, John Wiley & Sons, Estados Unidos, 1973.

2. Graybill Franklin A., *Theory And Application of the Linear Model*, Duxbury Press, California, 1976.

3. Draper N. R., Smith H., *Applied Regression Analysis*, John Wiley & Sons, Estados Unidos, 1980.

4. Seber G., *Linear Regression Analysis*, John Wiley & Sons, Estados Unidos, 1980.

5. Mitra Amitava, *Fundamentals of Quality Control and Improvement*, Macmillan, NewYork, ____.

6. Hollander Myles, Wolfe Douglas A., *Nonparametric Statistical Methods*, John Wiley & Sons, Estados Unidos, 1973.

7. Hawkins Douglas M., Simonoff Jeffrey S., Stromberg Arnold J., *Distributing a computationally intensive estimator: the case of exact LMS Regression*, Computational Statistics, Tennessee, 1994.

8. Fishman George S., *Principles of Discrete Event Simulation*, John Wiley & Sons, Estados Unidos, 1978.