UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

Notas de apoyo para el Curso de Análisis de Regresión

TESIS

QUE PARA OBTENER EL TÍTULO DE

Actuaria

PRESENTA

Karla Trujillo Fuentes

DIRECTOR DE TESIS: Mat. Margarita E. Chávez Cano

MÉXICO, D. F. 2000

UNIVERSIDAD NACIONAL AVENTMA DE MEXICO

MAT. MARGARITA ELVIRA CHÁVEZ CANO
Jefa de la División de Estudios Profesionales
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis:
Notas de apoyo para el Curso de Análisis de Regresión

realizado por Karla Trujillo Fuentes

Con número de cuenta 8724925-7 , pasante de la carrera de Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

A t e n t a m e n t e

Director de tesis Propietario	Mat. Margarita Elvira Chávez Cano
Propietario	M. en C. José Antonio Flores Díaz
Propietario	M. en C. Inocencio Rafael Madrid Ríos
Suplente	Act. María Luisa Nolasco Ochoa
Suplente	Act. Nancy Beatriz Zúñiga Hidalgo

Consejo Departamental de Matemáticas

M. en C. José Antonio Flores Díaz

Con toda mi gratitud a Margarita, por su apoyo, comprensión, paciencia y amistad

Índice General

1 El modelo lineal más simple 10

2 El modelo de regresión lineal simple 16

2.1 Estimación por mínimos cuadrados 18

2.1.1 Propiedades de los estimadores por mínimos cuadrados 22

2.2 Intervalos de confianza 32

2.2.1 Intervalo de confianza para β_1 32

2.2.2 Intervalo de confianza para β_0 33

2.2.3 Intervalo de confianza para σ^2 33

2.3 Coeficiente de correlación 34

2.4 Pruebas de hipótesis 41

2.5 Predicción 51

2.5.1 Intervalo de confianza de la respuesta media 51

2.5.2 Predicción de observaciones nuevas 54

2.5.3 Intervalo de predicción para la futura observación Y. 54

2.6 Inferencia simultánea en regresión lineal simple 56

2.6.1 Inferencia simultánea sobre los parámetros del modelo. 56

2.7 Gráficas de Y vs. X 61

2.8 Forma matricial del modelo lineal simple 63

3 Diagnóstico y medidas de qué tan adecuado es el modelo 69

3.1 Diagnóstico por residuales 74
3.1.1 Gráfica de residuales en papel probabilístico normal 74
3.1.2 Gráficas de residuales contra los valores ajustados $\hat{Y}$ 78
3.1.3 Gráfica de residuales contra una variable independiente omitida . 79
3.1.4 Gráficas contra las variables regresoras 80
3.1.5 Gráficas de residuales con respecto al tiempo 80
3.1.6 Efecto de las observaciones discordantes en el modelo 81
3.1.7 Transformación de variables 90
3.1.8 La prueba de carencia de ajuste 101

4 El modelo lineal general **110**

4.1 Las ecuaciones normales y su solución 115
4.2 El vector $\underline{\hat{Y}}$ y el vector de residuales $\underline{e}$ 118
4.3 Estimación de σ^2 123
4.4 Interpretación geométrica de los estimadores por mínimos cuadrados . . . 125
4.5 Precisión de los estimadores 126
4.6 Distribución de funciones lineales de variables aleatorias normales 133
4.7 Análisis de varianza y formas cuadráticas 135
4.8 Esperanza de formas cuadráticas 143
4.8.1 Varianzas estimadas o estimación de varianzas 145
4.9 Distribución de formas cuadráticas 146
4.10 Pruebas de hipótesis: La hipótesis lineal general 150
4.10.1 Casos especiales de la prueba lineal general 161
4.11 Intervalos de confianza 169
4.11.1 Intervalo de confianza para σ^2 170
4.11.2 Intervalo de confianza para $\underline{\lambda}'\underline{\beta}$ 171
4.11.3 Intervalo de confianza para $E(Y_0 \mid \underline{X}_0)$ 173

4.11.4 Intervalo de predicción para la respuesta media de futuras observaciones 174
4.11.5 Inferencia simultánea en regresión lineal múltiple 176

5 Selección de variables: El problema de definición del modelo 178

5.1 Consecuencias de la definición incorrecta del modelo 180
5.2 Criterios para evaluar modelos de regresión incompletos 185
5.2.1 Coeficiente de determinación múltiple 185
5.2.2 R^2 ajustado 187
5.2.3 Cuadrado medio de los residuales 187
5.2.4 Estadística C_p de Mallows 189
5.3 Usos del modelo de regresión y criterios para evaluarlo 192
5.4 Técnicas de cómputo para la selección de variables 193
5.4.1 Todas las regresiones posibles 193
5.4.2 Métodos de regresión stepwise 202
5.4.3 Comentarios generales acerca de los métodos de regresión stepwise 209
5.4.4 Criterios de decisión para detener un procedimiento de selección de variables 210
5.5 Algunas consideraciones finales 211

6 Elementos de álgebra lineal 214

6.1 Matrices 214
6.1.1 Definiciones básicas 214
6.1.2 Tipos especiales de matrices. 215
6.1.3 Operaciones con matrices 216
6.1.4 Teoremas acerca de la transposición de una matriz 220

6.1.5 Determinantes 223
6.1.6 Inversa (o recíproca) de una matriz 227
6.1.7 Matrices particionadas 231
6.1.8 Inversa de una matriz particionada 233
6.1.9 Determinantes de una matriz particionada 234
6.2 Dependencia lineal, rango y solución de ecuaciones homogéneas 235
6.3 Raíces y vectores característicos 240
6.4 Formas cuadráticas y matrices definidas positivas 243
6.5 Inversa generalizada e inversa condicional de una matriz 250
6.5.1 Inversa generalizada de una matriz 251
6.5.2 Inversa condicional de una matriz 252
6.6 Cálculo diferencial en notación matricial 253

7 Algunos resultados básicos de probabilidad y estadística 257
7.1 Operador suma y operador producto 257
7.1.1 Operador suma 257
7.1.2 Operador producto 258
7.2 Probabilidad 258
7.2.1 Teorema de la adición 258
7.2.2 Teorema de la multiplicación (o Teorema de Bayes) 258
7.2.3 Evento complemento 258
7.3 Variables aleatorias 259
7.3.1 Valor esperado 259
7.3.2 Varianza 260
7.3.3 Distribuciones de probabilidad conjunta, marginal y condicional 260
7.3.4 Covarianza 261
7.3.5 Variables aleatorias independientes 261

7.3.6 Funciones de variables aleatorias 262
7.3.7 Teorema del límite central 262
7.4 Algunas distribuciones de probabilidad 262
7.4.1 Distribución normal 262
7.4.2 Distribución χ^2 263
7.4.3 Distribución t 264
7.4.4 Distribución F 264
7.5 Estimación estadística 265
7.5.1 Propiedades de los estimadores 265
7.5.2 Estimadores máximo verosímiles 265
7.5.3 Estimadores por mínimos cuadrados 266
7.6 Inferencias acerca de la media de una población normal 266
7.6.1 Estimación por intervalos 267
7.6.2 Pruebas de hipótesis 267
7.6.3 Relación entre pruebas de hipótesis e intervalos de confianza 269
7.7 Comparación entre las medias de dos poblaciones normales 269
7.7.1 Muestras independientes 269
7.7.2 Estimación por intervalos 270
7.7.3 Pruebas de hipótesis 271
7.7.4 Observaciones apareadas 272
7.8 Inferencias acerca de la varianza de una población normal 272
7.8.1 Estimación por intervalos 272
7.8.2 Pruebas de hipótesis 273
7.9 Comparaciones entre las varianzas de dos poblaciones normales 274
7.9.1 Estimación por intervalos 274
7.9.2 Pruebas de hipótesis 275

Presentación

El trabajo que presentamos constituye el material para un curso introductorio al Análisis de Regresión Lineal. Supone por parte del lector, conocimientos básicos de Cálculo Diferencial e Integral, Probabilidad, Estadística y Álgebra Lineal.

En su preparación nos hemos esforzado principalmente en satisfacer las necesidades de los alumnos. Una lectura del Índice General mostrará que los temas que hemos incluido son los considerados en el nuevo plan de estudios de la asignatura denominada Análisis de Regresión, que se imparte como optativa en las carreras de Actuaría y Matemáticas de la Facultad de Ciencias.

Nuestro objetivo básico fue proporcionar toda la teoría requerida del Análisis de Regresión para el curso mencionado en un solo texto, ya que consideramos que no es conveniente, por lo general, el tener que complementar un libro de texto con material de otros libros.

El lector podrá también observar que cada uno de los temas es desarrollado con el rigor matemático requerido, pero sin perder de vista su correspondiente aplicación e interpretación.

El texto consta de siete capítulos. El primero de ellos, *El modelo lineal más simple*, pretende introducir al lector en los conceptos básicos necesarios para el manejo del Análisis de Regresión.

El capítulo 2, *El modelo de regresión lineal simple*, construye y desarrolla todos los elementos necesarios del modelo citado, explicando y aplicando sus principios fundamentales y, finalmente, señalando la necesidad de la notación matricial para el desarrollo de

modelos más generales.

El tercer capítulo, *Diagnóstico y medidas de qué tan adecuado es el modelo*, presenta una sinopsis de los problemas que surgen en la regresión por mínimos cuadrados y del impacto que éstos tienen en los resultados de la regresión. Asimismo, presenta métodos para encontrar tendencias en los datos que indiquen violaciones a las suposiciones fundamentales del modelo y técnicas para corregirlas. Este capítulo puede ser estudiado en el orden en el cual se presenta o bien al terminar el capítulo 4.

Este último, *El modelo lineal general*, desarrolla en notación matricial el modelo lineal múltiple, construyendo, a partir de las suposiciones fundamentales del modelo, la teoría necesaria para su aplicación.

El capítulo 5, *Selección de variables: El problema de definición del modelo*, proporciona principios relativos a la selección de modelos de regresión, incluyendo algunas técnicas y criterios de selección.

Los capítulos 6 y 7 son anexos de Algebra Lineal y Probabilidad y Estadística, respectivamente, que pueden servir para su consulta.

Se sugiere que, al terminar de estudiar el capítulo 4, se asigne a los estudiantes un proyecto, con datos preferentemente provenientes de un problema real, que puedan ser analizados exhaustivamente, no sólo con el fin de que apliquen las técnicas aprendidas durante el curso, sino de que desarrollen su intuición estadística y aprendan a presentar los resultados e interpretaciones por escrito.

Es evidente que la ayuda de algún software que aligere los cálculos será de gran ayuda durante el curso. Por ello, cabe señalar que la presente Tesis pertenece al primer paso de una línea de trabajo que pronto incluirá un paquete de cómputo, dividido por módulos, que pueda ser usado a la par del texto. Este paquete facilitará las operaciones inherentes a la obtención de los datos, permitiendo al mismo tiempo al estudiante ir observando, paso a paso, el proceso de construcción, análisis y definición del modelo.

Introducción

La modelación se refiere al desarrollo de expresiones matemáticas que describen en algún sentido el comportamiento de una variable de interés. Se considera que esta variable, llamada variable dependiente o respuesta, puede ser aproximada a partir de una relación funcional, en la cual aparecen todas aquellas variables que proveen información sobre el comportamiento de la misma; estas variables se incorporan al modelo como variables predictoras o explicativas y serán llamadas *variables independientes.* La relación funcional puede ser expresada como:

$$Y = f\ (\ X_1, X_2, ..., X_p\ ),$$

que de manera ideal, proporciona los valores de la respuesta Y.

A partir de este planteamiento surgen dos problemas:

a) La forma analítica de f puede ser desconocida o conocida pero muy complicada.

b) El número p de variables que intervienen en el estudio puede ser tan grande que sea imposible manipular adecuadamente a f

Las alternativas que se tienen son:

a) Aproximar a f mediante $f\ '$ (posiblemente un polinomio).

b) Ignorar todas aquellas variables cuya influencia sea considerada despreciable, reduciendo así el número de variables consideradas.

La alternativa b tiene como consecuencia que las variables ignoradas causen fluctuaciones en la respuesta. Estas fluctuaciones se consideran aleatorias, aun manteniendo fijos los valores de las variables consideradas. Así pues, a partir de a y b es posible establecer la siguiente relación:

$$Y = f\,'(X_1, X_2, ..., X_p) + \varepsilon$$

donde ε está determinada por los factores cuya influencia es considerada despreciable.

Además de las X's, los modelos involucran constantes desconocidas llamadas parámetros, que controlan el comportamiento del modelo. Estos parámetros serán denotados por letras griegas y estimados a partir de los datos.

La complejidad matemática del modelo y el grado hasta el cual sea un modelo realista, dependerá de cuánto se sepa acerca del proceso que está siendo estudiado.

En estudios preliminares de un proceso o en los casos donde la predicción es el objetivo primario, los modelos casi siempre caerán en la clase de modelos que son *lineales en los parámetros* . Esto es, los parámetros entran al modelo como coeficientes simples de las variables independientes. Tales modelos serán referidos como *modelos lineales.* Por otro lado, los modelos más reales son frecuentemente *no lineales en los parámetros* ; la mayoría de los modelos de crecimiento, por ejemplo, son modelos no lineales. Esta clase de modelos cae en dos categorías:

a) Modelos que pueden ser linealizados mediante una transformación apropiada sobre la variable dependiente, es decir, *modelos intrínsecamente lineales.*

b) Modelos que no pueden ser transformados.

La mayor parte del material que veremos, está dedicado a los modelos lineales y aquellos modelos no lineales que son intrínsecamente lineales.

Capítulo 1

El modelo lineal más simple

Considérese el siguiente modelo:

$$Y_i = \mu + \varepsilon_i \; ; \qquad i = 1, 2, ..., n$$
$$E(\varepsilon_i^2) = \sigma^2 \; ; \qquad E(\varepsilon_i) = 0; \qquad E(\varepsilon_i \varepsilon_j) = 0 \qquad i \neq j$$

Es decir, se considera que la variable bajo estudio fluctúa alrededor de un cierto valor. Las fluctuaciones aleatorias son no correlacionadas y tienen todas la misma dispersión (σ^2) alrededor de su media ($E(\varepsilon_i)$). Se pretende estimar μ y de ser posible realizar pruebas de hipótesis acerca de este valor, así como establecer límites de confianza para el mismo. Para llevar a cabo los dos últimos puntos será preciso hacer alguna suposición acerca de la distribución de ε_i .

Si en el problema de estimación deseamos restringirnos a la consideración de estimadores lineales e insesgados, el siguiente teorema da la pauta a seguir en la solución de este problema:

TEOREMA DE GAUSS MARKOV

Dentro de la clase de los estimadores lineales e insesgados para μ , el estimador obtenido por mínimos cuadrados es el mejor estimador lineal insesgado (Al mejor estimador lineal insesgado lo denotaremos por *M.E.L.I.*).

Cuando nos referimos a mejores estimadores, lo hacemos a estimadores de varianza mínima . Los estimadores lineales son de la forma: $\sum_{i=1}^{n} a_i Y_i$. La condición de insesgamiento da como resultado lo siguiente:

$$\begin{aligned} E\left[\sum_{i=1}^{n} a_i Y_i\right] &= \sum_{i=1}^{n} a_i E(Y_i) \\ &= \sum_{i=1}^{n} a_i (\mu + E(\varepsilon_i)) \\ &= \sum_{i=1}^{n} a_i \mu \\ &\Leftrightarrow \sum_{i=1}^{n} a_i = 1 \end{aligned}$$

La varianza del estimador está dada por:

$$\begin{aligned} Var\left(\sum_{i=1}^{n} a_i Y_i\right) &= \sum_{i=1}^{n} a_i^2 Var(Y_i) + \sum_{i \neq j} a_i a_j Cov(Y_i, Y_j) \\ &= \sum_{i=1}^{n} a_i^2 Var(\mu + \varepsilon_i) + \sum_{i \neq j} a_i a_j Cov(\mu + \varepsilon_i, \mu + \varepsilon_j) \\ &= \sum_{i=1}^{n} a_i^2 \sigma^2 + \sum_{i \neq j} a_i a_j E(\varepsilon_i \varepsilon_j) \\ &= \sum_{i=1}^{n} a_i^2 \sigma^2 \end{aligned}$$

Esta será mínima cuando lo sea $\sum a_i^2$, sujeta a la restricción $\sum a_i = 1$. Entonces debemos minimizar la suma de cuadrados, i. e.

$$\min\left\{\sum_{i=1}^{n} a_i^2 + \lambda\left(\sum_{i=1}^{n}(a_i - 1)\right)\right\} = Q$$

Para minimizar derívese con respecto a cada a_i

$$\begin{aligned} \frac{\partial Q}{\partial a_i} &= 2a_i + \lambda = 0 \qquad i = 1, .., n \\ \frac{\partial Q}{\partial \lambda} &= \sum_{i=1}^{n} a_i - 1 = 0 \end{aligned}$$

sumando las n derivadas de Q con respecto a las a_i se tiene

$$2\sum_{i=1}^{n} a_i + n\lambda = 0 \;\Rightarrow\; \lambda = -\frac{2}{n} \quad , \text{pues } \sum_{i=1}^{n} a_i = 1$$

sustituyendo este valor en $\frac{\partial Q}{\partial a_i}$ se tiene:

$$\frac{\partial Q}{\partial a_i} = 2a_i + \lambda = 2a_i - \frac{2}{n} = 0$$

de donde

$$a_i = \frac{1}{n} \qquad i = 1, 2, ..., n;$$

además $\frac{\partial^2 Q}{\partial a_i^2} = 2 > 0$. De donde los valores de a_i obtenidos hacen mínima la expresión de la varianza del estimador, el cual está dado por

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} Y_i = \bar{Y}$$

Para este caso, el método de mínimos cuadrados nos lleva al mismo estimador. En efecto, minimizar

$$\begin{aligned} Z &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \sum_{i=1}^{n} (Y_i - \mu)^2 \\ \frac{\partial Z}{\partial \mu} &= -2\sum_{i=1}^{n} Y_I - \mu = 0 \\ &\Rightarrow \sum_{i=1}^{n} Y_i - n\mu = 0 \\ &\Rightarrow \mu = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y} \end{aligned}$$

con lo que el teorema queda demostrado.

Si además se supone que la distribución de los errores es normal, i. e., $\varepsilon_i \sim N(o, \sigma^2)$, se sabe de la teoría estadística elemental lo siguiente:

$$\bar{Y} \sim N(\mu, \frac{\sigma^2}{n})$$

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

con $\bar{Y}$ y S^2 independientes. Entonces la prueba de hipótesis

$$Ho: \mu = 0 \qquad vs \qquad Ha: \mu \neq 0$$

Haciendo uso del cociente de verosimilitudes se tiene lo siguiente:

$$\lambda = \frac{\sup_{Ho} L(\mu, \hat{\sigma}^2)}{\sup_{Ho \cup Ha} L(\mu, \hat{\sigma}^2)} = \frac{\left(\frac{1}{2\pi\hat{\sigma}^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(Y_i - \hat{\mu}_0)^2}{\hat{\sigma}_0^2}\right\}}{\left(\frac{1}{2\pi\hat{\sigma}^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(Y_i - \hat{\mu})^2}{\hat{\sigma}^2}\right\}}$$

donde $\hat{\mu}_0 = 0$ por H_0 , $\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_0)^2 = \frac{1}{n}\sum_{i=1}^{n}Y_i^2$

$$\hat{\mu} = \bar{Y} \quad \text{y} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Y)^2$$

Sustituyendo estos valores en la expresión para λ se tiene:

$$\lambda = \frac{\left(2\pi\sum_{i=1}^{n}\frac{Y_i^2}{n}\right)^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\frac{\sum_{i=1}^{n}Y_i^2}{\sum_{i=1}^{n}Y_i^2}\right\}}{\left(2\pi\sum_{i=1}^{n}\frac{(Y_i - \bar{Y})^2}{n}\right)^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\right\}} = \left(\frac{\sum_{i=1}^{n}Y_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\right)^{-\frac{n}{2}}$$

de donde una función monótona no decreciente de λ es:

$$\lambda^{\frac{2}{n}} = \frac{\sum(Y_i - \bar{Y})^2}{\sum Y_i - n\bar{Y}^2 + n\bar{Y}^2} = \frac{\sum(Y_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2 + n\bar{Y}^2} = \frac{1}{1 + \frac{n\bar{Y}^2}{\sum(Y_i - \bar{Y})^2}}$$

que a su vez es una función monótona no decreciente de:

$$Z = \frac{n\bar{Y}^2}{(n-1)\frac{\sum(Y_i - \bar{Y})^2}{n-1}}$$

La hipótesis nula se rechaza si:

$$\lambda < C_1 \Leftrightarrow \lambda^{\frac{2}{n}} < C_2 \Leftrightarrow Z > C_3$$

Ahora bien, $\frac{n\bar{Y}^2}{\sigma^2} \sim \chi^2_{(1)}$ si Ho es cierta y $\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi^2_{(n-1)}$ si Ho no es cierta; la expresión del numerador tiene una distribución $\chi^2_{(1,\lambda)}$, que se conoce como una χ^2 no central; el parámetro de no centralidad está dado por

$$\lambda = \frac{\left(\sum_{i=1}^{n} \mu_i\right)^2}{2n} = \frac{1}{2n}\mu'\mu$$

Entonces si Ho es cierta, el cociente definido por Z multiplicado por $(n-1)$ tiene una distribución F central con 1 y $(n-1)$ grados de libertad.

$$Z' = \frac{(n\bar{Y}^2)\ /\ \sigma^2/1}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2\ /\sigma^2/(n-1)} \sim F_{(1,n-1)}$$

Con toda esta información ya esposible encontrar el valor de C_3 en las tablas de la distribución F.

Los pasos necesarios para la realización de la prueba pueden ser resumidos en la llamada *Tabla de Análisis de Varianza*, esta tabla consta de cinco columnas, a las cuales nos referiremos como sigue:

FV, fuente de variación;

GL, grados de libertad;

SC, suma de cuadrados;

CM, cuadrados medios;

F, valor de la estadística de prueba.

Para la prueba:

$$Ho: \mu = 0 \qquad vs \qquad Ha: \mu \neq 0$$

en el modelo lineal más simple, la Tabla de Análisis de Varianza está dada como sigue:

Tabla de Análisis de Varianza ($H_0 : \mu = 0$)

FV	GL	SC	CM	F
μ	1	$n\bar{Y}^2$	$n\bar{Y}^2$	$\dfrac{n\bar{Y}^2}{\dfrac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}}$
error	$n-1$	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$	$\dfrac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n-1}$	
total	n	$\sum_{i=1}^{n}Y_i^2$		

Capítulo 2

El modelo de regresión lineal simple

El modelo más simple involucra solamente una variable independiente y establece que la verdadera media de la variable dependiente cambia en razón constante cuando el valor de la variable independiente crece o decrece. De esta forma, la relación funcional entre la verdadera media de Y, $E(Y)$ y X es la ecuación de la línea recta

$$E(Y) = \beta_0 + \beta_1 X$$

donde β_0 es la intercepción de esta recta con el eje Y, el valor de $E(Y)$ cuando $X = 0$; β_1 es la pendiente de ella, la razón de cambio en $E(Y)$ por unidad de cambio en X.

En las situaciones prácticas o reales, la información con que se cuenta consta de n parejas de observaciones muestrales sobre X, Y, que pueden ser graficadas como se muestra en la figura 1.

La diferencia esencial que se observa a partir de esta figura es que en la práctica, la línea $\beta_0 + \beta_1 X$ es desconocida.

Las observaciones sobre la variable dependiente, Y_i, se supone que son observaciones aleatorias de poblaciones de variables aleatorias con la media dada por $E(Y_i)$. La desviación de una observación Y_i de su media poblacional $E(Y_i)$ (la línea desconocida), se toma en cuenta sumando un error aleatorio para dar el modelo estadístico

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

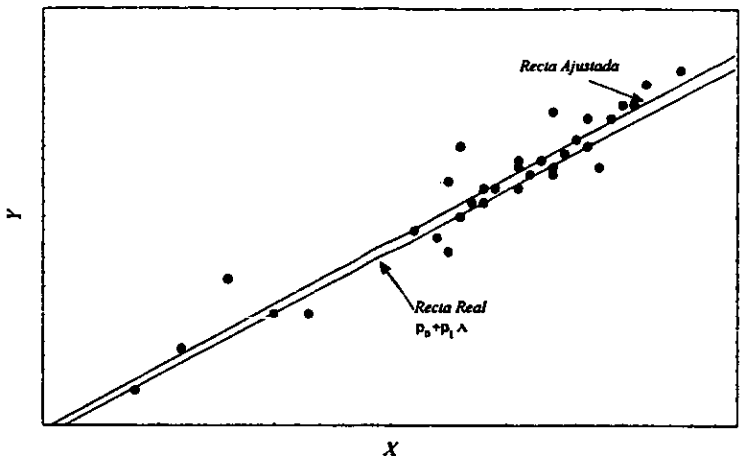


Figura 1: Recta real y recta ajustada sobre un diagrama de dispersión

Las X_i son las n observaciones sobre la variable independiente y se supone que son medidas sin error, esto es, se supone que los valores observados de X forman un conjunto de constantes conocidas. Las Y_i y las X_i son observaciones apareadas, medidas sobre cada unidad observacional.

Esencialmente, se tienen dos tipos de hipótesis que se hacen sobre el modelo, la *hipótesis estructural* y la *hipótesis distribucional*. La hipótesis estructural consiste en suponer que el modelo es lineal en los parámetros, esto es, los parámetros entran al modelo como coeficientes simples sobre las variables independientes o funciones de ellas. La hipótesis distribucional se refiere a las suposiciones que se hacen en relación a los errores aleatorios que aparecen en el modelo como ε_i ; como anteriormente se vio de manera implícita, se supone que la media de los ε_i es igual a cero, $E(\varepsilon_i) = 0$, ya que de manera natural se espera que en promedio no haya errores; se supone también que la varianza de los errores es constante, común y desconocida $Var\ (\varepsilon_i) = \sigma^2$; esto significa que se espera que las observaciones no se distribuyan de manera irregular alrededor de la línea media y de esta forma facilitar el desarrollo de la teoría. Obsérvese que $\sigma^2 = cte.$

refleja que los factores no controlados influyen de la misma manera sobre cada respuesta Y_i . Como ε_i es el único elemento aleatorio en el modelo, estas suposiciones implican que las Y_i son variables aleatorias, por lo tanto también tienen varianza común y son mutuamente independientes. Con el fin de construir intervalos de confianza y hacer pruebas de significancia, se introduce la hipótesis de que los errores aleatorios tienen distribución normal, lo cual implica que las Y_i también tienen distribución normal.

Las suposiciones acerca de los errores aleatorios son denotadas por:

$$\varepsilon_i \sim N(0, \sigma^2), \text{ independientes, } i = 1, 2, ..., n \text{ (notación de Wilks).}$$

2.1 Estimación por mínimos cuadrados

El modelo lineal simple

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

tiene dos parámetros, β_0 y β_1 , que serán estimados a partir de los datos. Con la hipótesis de varianza constante sobre los errores, aparece otro parámetro que no está incluido en el modelo, σ^2, pero que es necesario estimar también; el tratamiento para este parámetro se hará más adelante.

Si no hubiera error aleatorio en Y_i , podrían utilizarse cualesquiera dos parejas de observaciones para obtener explícitamente los valores de los parámetros. Sin embargo, la variación aleatoria de Y causa que cada pareja de datos dé diferentes resultados (todos los estimadores serían idénticos sólo si los datos observados cayeran exactamente sobre la línea recta). Se necesita un método que combine toda la información para dar una solución óptima de acuerdo a algún criterio.

El procedimiento o *método de mínimos cuadrados* tiene el siguiente criterio, conocido como el *principio de mínimos cuadrados* : La solución debe dar la suma de cuadrados de las desviaciones verticales de las Y_i observadas de los valores estimados más pequeña posible. Estas desviaciones son conocidas como los residuales, e_i, es decir

$$e_i = Y_i - \hat{Y}_i ; i = 1, 2, ..., n$$

Sean $\hat{\beta}_0$ y $\hat{\beta}_1$ los estimadores de los parámetros β_0 y β_1, respectivamente; sea

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \ ; i = 1, 2, ..., n$$

el valor estimado de Y para cada X_i, $i = 1, 2, ..., n$. Esta ecuación es conocida como la recta estimada o ajustada.

El principio de los mínimos cuadrados elige $\hat{\beta}_0$ y $\hat{\beta}_1$ que minimizan la suma de cuadrados de los residuales denotada como **SCE**

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{SCE}$$

Los estimadores para β_o y β_1 se obtienen utilizando las técnicas del cálculo diferencial para encontrar los valores que minimizan la **SCE**.

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_o - \hat{\beta}_1 X_i)^2$$

Derivando esta expresión con respecto a $\hat{\beta}_o$ y a $\hat{\beta}_1$ e igualando a cero, se tienen las ecuaciones

$$\begin{aligned} n\hat{\beta}_o + \left(\sum_{i=1}^{n} X_i\right)\hat{\beta}_1 &= \sum_{i=1}^{n} Y_i \\ \left(\sum_{i=1}^{n} X_i\right)\hat{\beta}_o + \left(\sum_{i=1}^{n} X_i^2\right)\hat{\beta}_1 &= \sum_{i=1}^{n} X_i Y_i \end{aligned}$$

Estas ecuaciones son conocidas como *ecuaciones normales*. Resolviéndolas simultáneamente para $\hat{\beta}_0$ y $\hat{\beta}_1$ se obtienen los estimadores de β_0 y β_1.

Multiplicando la primera ecuación por $\dfrac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$ y restando al resultado la segunda ecuación se tiene:

$$\begin{aligned} \hat{\beta}_1 \left(\bar{X} \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i^2\right) &= \bar{X} \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i Y_i \\ \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i (X_i - \bar{X})}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum X_i)^2}{n}} &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \end{aligned}$$

donde $\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - \frac{\sum_{i=1}^{n} X_i^2}{n}$ y $\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} Y_i(X_i - \bar{X})$

Ahora bien, dividiendo la primera ecuación por n y despejando $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{Y} - \bar{X}\hat{\beta}_1$$

Nota.- En algunos textos se utiliza la notación:

$$x_i = X_i - \bar{X}\,;\, y_i = Y_i - \bar{Y}$$

Estos estimadores de los parámetros dan la ecuación de regresión:

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

Ejemplo 1.- Considérense los datos obtenidos de un estudio dirigido por el Dr. A. S. Heagle en North Carolina State University. Dicho estudio analiza los efectos de la contaminación por ozono en granos de soya (tabla 1). Cuatro distintos niveles de ozono y la producción media de soya correspondiente fueron medidos. La dosis de ozono es la concentración promedio durante la época de crecimiento en partes por millón (ppm); la producción se reporta en gramos por planta.

Tabla 1: Resultados principales de la producción de soya (gm por planta), obtenidos como respuesta a los niveles indicados de exposición a ozono

i	ozono (ppm) X	producción (gm./plt)
1	.02	242
2	.07	237
3	.11	231
4	.15	201

Si suponemos que la producción de soya está relacionada linealmente con la cantidad de ozono, podemos aplicar el modelo antes desarrollado. De la tabla obtenemos

$$\sum_{i=1}^{4} X_i = .35 \qquad \sum_{i=1}^{4} Y_i = 911$$

$$\bar{X} = .0875 \qquad \bar{Y} = 227.75$$

$$\sum_{i=1}^{4} X_i^2 = .0399 \qquad \sum_{i=1}^{4} Y_i^2 = 208495$$

$$\sum_{i=1}^{4} X_i Y_i = 76.99$$

por lo cual los estimadores por mínimos cuadrados son

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{4} X_i Y_i - \frac{\left(\sum_{i=1}^{4} X_i\right)\left(\sum_{i=1}^{4} Y_i\right)}{n}}{\sum_{i=1}^{4} X_i^2 - \frac{\left(\sum_{i=1}^{4} X_i\right)^2}{n}} \\
&= \frac{76.99 - \frac{(.35)(911)}{4}}{.0399 - \frac{(3.5)^2}{4}} \\
&= -293.531
\end{aligned}$$

y

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Y} - \bar{X}\hat{\beta}_1 \\
&= 227.75 - (-293.531)(.0875) \\
&= 253.434
\end{aligned}$$

De esta manera, el modelo ajustado es:

$$\hat{Y} = 253.434 - 293.531X$$

La interpretación de $\hat{\beta}_1 = -293.531$ es que se espera que la producción media disminuya, puesto que la pendiente es negativa; esto es, la producción media disminuirá en aproximadamente 294 gramos por planta con cada unidad (ppm) de ozono que se agregue. Obsérvese que el rango de ozono va de .02 a .15, por lo cual no es razonable esperar que la misma tasa de decaimiento en la producción ocurra en, digamos, 1 ppm La intersección $\hat{\beta}_0 = 253.434$ es el valor de X en el cual la línea ajustada cruza el eje Y.

En este caso, como el valor más bajo del nivel de ozono es .02 se puede considerar como una extrapolación interpretar a $\hat{\beta}_0$ como el valor estimado de la producción cuando no existe contaminación por ozono.

2.1.1 Propiedades de los estimadores por mínimos cuadrados

Los estimadores por mínimos cuadrados, $\hat{\beta}_0$ y $\hat{\beta}_1$, tienen varias propiedades estadísticas importantes. Examinemos primero la propiedad de insesgamiento.

$$\begin{aligned} E(\hat{\beta}_1) &= E\left(\frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \\ &= \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sum_{i=1}^{n}(X_i - \bar{X})E(Y_i) \\ &= \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \sum_{i=1}^{n}(X_i - \bar{X})(\beta_o + \beta_1 X_i) \\ &= \beta_1 \end{aligned}$$

ya que $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$ y $\sum_{i=1}^{n} X_i(X_i - \bar{X}) = \sum_{i=1}^{n}(X_i - \bar{X})^2$

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{Y} - \hat{\beta}_1\bar{X}) \\ &= \frac{\sum_{i=1}^{n} E(Y_i)}{n} - E(\hat{\beta}_1)\bar{X} \\ &= \frac{\sum_{i=1}^{n}(\beta_0 + \beta_1 X_i)}{n} - \beta_1\bar{X} \\ &= \beta_0 + \beta_1\bar{X} - \beta_1\bar{X} \\ &= \beta_0 \end{aligned}$$

Por lo tanto, $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores insesgados.

Note que $\hat{\beta}_0$ y $\hat{\beta}_1$ son lineales en el sentido de que son combinaciones lineales de las $Y_i's$, esto es,

$$\hat{\beta}_1 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) Y_i$$

y

$$\hat{\beta}_0 = \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{X}(Xi - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) Y_i$$

Con estas expresiones y bajo la hipótesis de independencia se tiene:

$$\begin{aligned} Var(\hat{\beta}_1) &= \sum_{i=1}^{n} \left(\frac{(Xi - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)^2 Var(Y_i) \\ &= \sigma^2 \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \end{aligned}$$

$$\begin{aligned} Var(\hat{\beta}_0) &= \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{X}(Xi - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)^2 Var(Y_i) \\ &= \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - \frac{\bar{X}(Xi - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)^2 \\ &= \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \sigma^2 \end{aligned}$$

El problema que surge ahora es determinar cómo se comportan conjuntamente $\hat{\beta}_0$ y $\hat{\beta}_1$, por lo que es de interés el cálculo de la covarianza.

Como $\hat{\beta}_0$ y $\hat{\beta}_1$ son combinaciones lineales de las $Y_i's$, podemos proponer una solución general para la covarianza de las funciones lineales:

Sea U una combinación lineal de las $Y_i's$ y W otra combinación lineal de las $Y_i's$, esto es,

$$U = \sum_{i=1}^{n} a_i Y_i \qquad \text{y} \qquad W = \sum d_i Y_i$$

La covarianza entre U y W está dada por

$$\begin{aligned}
Cov\,(U,W) &= E\,((U - E(U))\,(W - E(W))) \\
&= E\left((\sum_{i=1}^{n} a_i Y_i - \sum_{i=1}^{n} a_i E(Y_i))(\sum_{i=1}^{n} d_i Y_i - \sum_{i=1}^{n} d_i E(Y_i))\right) \\
&= E\left((\sum_{i=1}^{n} a_i\,(Y_i - E(Y_i))\,)\,(\sum_{i=1}^{n} d_i\,(Y_i - E(Y_i)))\right) \\
&= \sum_{i=1}^{n} a_i d_i\,E\,((Y_i - E(Y_i))^2) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i d_j\,E\,((Y_i - E(Y_i))(Y_j - E(Y_j))) \\
&= \sum_{i} a_i d_i\,Var(Y_i) + \sum_{i}\sum_{j} a_i d_j\,Cov(Y_i, Y_j)
\end{aligned}$$

donde $Cov\,(Y_i, Y_j) = 0$ puesto que las $Y_i's$ son independientes. Por lo tanto

$$Cov\,(U,W) = \sum a_i d_i\,Var\,(Y_i)$$

Usando este resultado tenemos

$$\begin{aligned}
Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov\left(\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) Y_i\ ,\ \sum_{i=1}^{n}\left(\frac{1}{n} - \frac{\bar{X}\,(Xi - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) Y_i\right) \\
&= \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\left(\frac{1}{n} - \frac{\bar{X}\,(Xi - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\right) Var\,(Y_i) \\
&= \sigma^2\left(\frac{1}{n}\sum_{i=1}^{n}\frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} - \bar{X}\,\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}\right) \\
&= \sigma^2\left(-\frac{\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \neq 0
\end{aligned}$$

Esto significa que $\hat{\beta}_0$ y $\hat{\beta}_1$ no son independientes. Las propiedades de los estimadores pueden resumirse en el siguiente teorema.

Teorema de Gauss-Markov: En el modelo de regresión lineal simple

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i;\ \ \varepsilon_i \sim N\,(0,\,\sigma^2),\ \text{independientes}$$

los estimadores por mínimos cuadrados para β_0 y β_1 son los mejores estimadores lineales insesgados (mejores en el sentido de varianza mínima) .

Demostración: Probaremos primero que el teorema se cumple en el caso de $\hat{\beta}_1$. Para ello proponemos un estimador $\hat{\beta}_1' \neq \hat{\beta}_1$ que satisfaga las condiciones requeridas; es decir, $\hat{\beta}_1'$ debe ser una combinación lineal de las $Y_i's$, ser insesgado y de varianza mínima. Sea $\hat{\beta}_1' = \sum_{i=1}^{n} c_i Y_i$. Aplicando a éste la esperanza se tiene

$$\begin{aligned} E\left(\hat{\beta}_1'\right) &= E\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i E\left(Y_i\right) \\ &= \sum_{i=1}^{n} c_i \left(\beta_0 + \beta_1 X_i\right) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i X_i \end{aligned}$$

donde $E\left(\hat{\beta}_1'\right) = \beta_1$ si y sólo si $\sum_{i=1}^{n} c_i = 0$ y $\sum_{i=1}^{n} c_i X_i = 1$
Se desea, además, minimizar la varianza de este estimador:

$$Var\left(\hat{\beta}_1'\right) = Var\left(\sum_{i=1}^{n} c_i Y_i\right) = \sum_{i=1}^{n} c_i^2 Var\left(Y_i\right) = \sum_{i=1}^{n} c_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^{n} c_i^2$$

donde σ^2 es una constante desconocida. Así, el problema puede plantearse como minimizar $\sum_{i=1}^{n} c_i^2$ sujeto a las condiciones de insesgamiento, es decir,

$$\text{minimizar } \sum_{i=1}^{n} c_i^2 \text{ sujeto a } \sum_{i=1}^{n} c_i = 0 \text{ y } \sum_{i=1}^{n} c_i X_i - 1 = 0$$

Apliquemos el método de multiplicadores de Lagrange.
Sea $\phi = \sum_{i=1}^{n} c_i^2 - 2\lambda\left(\sum_{i=1}^{n} c_i\right) - 2\gamma\left(\sum_{i=1}^{n} c_i X_i - 1\right)$

$$\begin{aligned} \frac{\partial \phi}{\partial \lambda} &= -2\sum_{i=1}^{n} c_i &&= 0 \Rightarrow \sum_{i=1}^{n} c_i = 0 \\ \frac{\partial \phi}{\partial \gamma} &= -2\left(\sum_{i=1}^{n} c_i X_i - 1\right) &&= 0 \Rightarrow \sum_{i=1}^{n} c_i X_i = 1 \\ \frac{\partial \phi}{\partial c_i} &= 2c_i - 2\lambda - 2\gamma X_i &&= 0 \Rightarrow c_i - \lambda - \gamma X_i = 0 \\ & & & \qquad i = 1, 2, \ldots, n \end{aligned}$$

Haciendo la suma sobre i, se tiene

$$\sum_{i=1}^{n} c_i - n\lambda - \gamma \sum_{i=1}^{n} X_i = 0$$

Usando la condición $\sum_{i=1}^{n} c_i = 0$ en la ecuación anterior se tiene

$$\lambda = -\gamma \bar{X}$$

Retomando la tercera ecuación y multiplicándola por X_i se tiene:

$$\begin{aligned} c_i X_i - \lambda X_i - \gamma X_i^2 &= 0 \\ \sum_{i=1}^{n} c_i X_i - \lambda \sum_{i=1}^{n} X_i - \gamma \sum_{i=1}^{n} X_i^2 &= 0 \end{aligned}$$

donde $\sum_{i=1}^{n} c_i X_i = 1$, por lo cual, sustituyendo λ :

$$1 + \gamma \bar{X} \sum_{i=1}^{n} X_i - \gamma \sum_{i=1}^{n} X_i^2 = 0$$

lo cual ocurre si y sólo si

$$\gamma \left(\bar{X} \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} X_i^2 \right) = -1$$

de donde

$$\gamma = \frac{1}{\sum_{i=1}^{n} X_i^2 - \bar{X} \sum_{i=1}^{n} X_i} = \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

por lo tanto, de la tercera ecuación :

$$\begin{aligned} c_i &= \lambda + \gamma X_i \\ &= -\gamma \bar{X} + \gamma X_i = \gamma (X_i - \bar{X}) \\ &= \frac{X_i - \bar{X}}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \end{aligned}$$

y puesto que $\hat{\beta}_1' = \sum_{i=1}^{n} c_i Y_i$ se tiene que $\hat{\beta}_1' = \dfrac{\sum_{i=1}^{n} (X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} Y_i = \hat{\beta}_1$, que es lo que queríamos demostrar. De manera análoga se construye el estimador para β_0.

Con el fin de formular intervalos de confianza y pruebas de hipótesis, obtendremos a continuación la distribución de $\hat{\beta}_0$ y $\hat{\beta}_1$.

Bajo la hipótesis $\varepsilon_i \sim N(0, \sigma^2)$, se tiene que $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ también tiene distribución normal; i. e.

$$Y_i \sim N\left(\beta_0 + \beta_1 X_i,\ \sigma^2\right)$$

por ser una combinación lineal de ε_i, variable aleatoria normal para cada i.

De la misma forma, por ser $\hat{\beta}_0$ y $\hat{\beta}_1$ combinaciones lineales de variables aleatorias normales, su distribución también es normal con los parámetros obtenidos anteriormente; entonces:

$$\hat{\beta}_0 \sim N\left(\beta_0, \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\sigma^2\right)$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

La distribución conjunta de estos estimadores es una normal bivariada, es decir:

$$\begin{pmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{pmatrix} \sim N\left(\begin{pmatrix}\beta_0 \\ \beta_1\end{pmatrix}, \sum\right)$$

donde $\sum$ es la matriz de varianzas y covarianzas

$$\sum = \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} & \frac{-\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \\ \frac{-\bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2} & \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \end{pmatrix}$$

La teoría y los detalles correspondientes se verán más adelante.

Bajo la hipótesis de normalidad puede verificarse que los estimadores obtenidos por el método de máxima verosimilitud para β_0 y β_1, coinciden con los estimadores correspondientes obtenidos por mínimos cuadrados. En efecto, calculemos la densidad conjunta de las $Y_i's$ (la función de verosimilitud):

$$\begin{aligned} L(Y_1, Y_2, ..., Y_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2\right\} \end{aligned}$$

Si aplicamos a la ecuación anterior la función ln se tiene

$$\ln\left(L(Y_1, Y_2, ..., Y_n)\right) = \frac{-n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 X_i\right)^2$$

que es la función que deseamos maximizar; de este modo

$$\frac{\partial \ln L(Y_1, Y_2, ..., Y_n)}{\partial \beta_0} = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2\left(Y_i - \beta_0 - \beta_1 X_i\right) = 0 \quad \ldots 1$$

$$\frac{\partial \ln L(Y_1, Y_2, ..., Y_n)}{\partial \beta_1} = \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2\left(Y_i - \beta_0 - \beta_1 X_i\right) Xi = 0 \quad \ldots 2$$

$$\frac{\partial \ln L(Y_1, Y_2, ..., Y_n)}{\partial \sigma^2} = -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 X_i\right)^2 = 0 \quad \ldots 3$$

Obsérvese que de la ecuación (1)

$$\sum_{i=1}^{n} Y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} X_i$$

y de la segunda ecuación

$$\sum_{i=1}^{n} X_i Y_i = \beta_0 \sum_{i=1}^{n} X_i + \beta_1 \sum_{i=1}^{n} X_i^2$$

que son las ecuaciones normales.

Por otro lado, ya que σ^2 , la varianza del error, es un parámetro adicional desconocido, el estimador máximo verosímil correspondiente es

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} e_i^2
\end{aligned}$$

Este estimador no es insesgado, como veremos a continuación

$$E\left(\hat{\sigma}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right)$$

donde

$$\begin{aligned}
e_i &= Y_i - \hat{Y}_i \\
&= \beta_0 + \beta_1 X_i + \varepsilon_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \\
&= (\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) X_i + \varepsilon_i
\end{aligned}$$

por lo que

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n}\left(-(\hat{\beta}_0-\beta_0)-(\hat{\beta}_1-\beta_1)X_i+\varepsilon_i\right)^2 \\ &= \sum_{i=1}^{n}\left((\hat{\beta}_0-\beta_0)+(\hat{\beta}_1-\beta_1)X_i-\varepsilon_i\right)^2\end{aligned}$$

de donde, desarrollando el cuadrado

$$\begin{aligned}E\left(\sum_{i=1}^{n} e_i^2\right) &= E(\sum_{i=1}^{n}(\hat{\beta}_0-\beta_0)^2+(\hat{\beta}_1-\beta_1)^2\sum_{i=1}^{n}X_i^2+\sum_{i=1}^{n}\varepsilon_i^2+2(\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}X_i \\ &\quad -2(\hat{\beta}_0-\beta_0)\sum_{i=1}^{n}\varepsilon_i-2(\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}X_i\varepsilon_i)\end{aligned}$$

$$\begin{aligned}&= nE(\hat{\beta}_0-\beta_0)^2+\sum_{i=1}^{n}X_i^2E(\hat{\beta}_1-\beta_1)^2+\sum_{i=1}^{n}E(\varepsilon_i^2)+2\sum_{i=1}^{n}X_iE\left((\hat{\beta}_0-\beta_0)(\hat{\beta}_1-\beta_1)\right) \\ &\quad -2E\left((\hat{\beta}_0-\beta_0)\sum_{i=1}^{n}\varepsilon_i\right)-2E\left((\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}X_i\varepsilon_i\right)\end{aligned}$$

$$\begin{aligned}&= nVar\,(\hat{\beta}_0)+\sum_{i=1}^{n}X_i^2Var\,(\hat{\beta}_1)+n\sigma^2+2\sum_{i=1}^{n}X_iCov(\hat{\beta}_0,\hat{\beta}_1) \\ &\quad -2E\left((\hat{\beta}_0-\beta_0)\sum_{i=1}^{n}\varepsilon_i\right)-2E\left((\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}X_i\varepsilon_i\right)\end{aligned}$$

$$\begin{aligned}&= n\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)\sigma^2+\sum_{i=1}^{n}X_i^2\frac{\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}+n\sigma^2 \\ &\quad -2\sum_{i=1}^{n}X_i\frac{\bar{X}\sigma^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}-2E\left((\hat{\beta}_0-\beta_0)\sum_{i=1}^{n}\varepsilon_i\right)-2E\left((\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}X_i\varepsilon_i\right)\end{aligned}$$

Dado que $Y_i=\beta_0+\beta_1X_i+\varepsilon_i$

$$\begin{aligned}\hat{\beta}_0 &= \sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{X}(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)Y_i \\ &= \beta_0+\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{X}(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)\varepsilon_i\end{aligned}$$

y

$$\begin{aligned}\hat{\beta}_1 &= \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)Y_i \\ &= \beta_1+\frac{\sum_{i=1}^{n}(X_i-\bar{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\end{aligned}$$

entonces,

$$E\left((\hat{\beta}_0-\beta_0)\sum_{i=1}^{n}\varepsilon_i\right)=E\left(\left(\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{X}(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)\varepsilon_i\right)\left(\sum_{i=1}^{n}\varepsilon_i\right)\right)$$

Desarrollando esta expresión y aplicando la esperanza se tiene

$$E\left((\hat{\beta}_0-\beta_0)\sum_{i=1}^{n}\varepsilon_i\right)=\sigma^2\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{X}(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)=\sigma^2$$

De manera similar tenemos

$$\begin{aligned}E\left((\hat{\beta}_1-\beta_1)\sum_{i=1}^{n}X_i\varepsilon_i\right) &= E\left(\left(\frac{\sum_{i=1}^{n}(X_i-\bar{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)\left(\sum_{i=1}^{n}X_i\varepsilon_i\right)\right) \\ &= \frac{1}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\left(\sigma^2\sum_{i=1}^{n}X_i(X_i-\bar{X})\right) \\ &= \sigma^2\end{aligned}$$

por lo tanto

$$
\begin{aligned}
E\left(\sum_{i=1}^{n} e_i^2\right) &= \sigma^2 + \frac{n\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sigma^2 + \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sigma^2 + n\sigma^2 \\
&\quad -2n\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sigma^2 - 2\sigma^2 - 2\sigma^2 \\
&= (n-3)\sigma^2 + \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sigma^2 - \frac{n\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sigma^2 \\
&= \frac{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sigma^2 + (n-3)\sigma^2 \\
E\left(\sum_{i=1}^{n} e_i^2\right) &= (n-2)\sigma^2
\end{aligned}
$$

de donde, $E\,(\,\hat{\sigma}^2\,) \neq \sigma^2$, esto es, $\hat{\sigma}^2$ obtenido por máxima verosimilitud no es insesgado. Sin embargo, es posible construir un estimador insesgado a partir de la última expresión obtenida; en efecto, si

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

entonces

$$E\,(\hat{\sigma}^2) = \sigma^2$$

además se tiene que

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

donde los grados de libertad de la distribución, así como el denominador del estimador insesgado es $(n-2)$. Esta cantidad corresponde a n, el número de observaciones o tamaño de la muestra y a 2, que es el número de parámetros en el modelo. La demostración de este resultado se hará más adelante.

2.2 Intervalos de confianza

Las características generales de una línea recta están dadas por la intersección con el eje de las Y's y la pendiente, que corresponden a β_0 y β_1 respectivamente. Es importante estudiar estas características a través de los intervalos de confianza, ya que así sabremos si la recta considerada pasa por el origen o no y si tiene pendiente distinta de cero.

2.2.1 Intervalo de confianza para β_1

Como

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) \Rightarrow \frac{\hat{\beta}_1 - \beta_1}{\sigma\sqrt{\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}} \sim N(0,1)$$

donde σ es desconocido, por lo que, para construir una cantidad pivotal consideramos a la variable

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

entonces

$$\begin{aligned}
\frac{\hat{\beta}_1 - \beta_1}{\sigma\sqrt{\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}} \Big/ \sqrt{\frac{\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2}}{(n-2)}} &= \frac{(\hat{\beta}_1 - \beta_1)\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}{\sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}} \\
&= \frac{(\hat{\beta}_1 - \beta_1)\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}{\hat{\sigma}} \\
&= T \sim t_{(n-2)}
\end{aligned}$$

por lo tanto

$$P\left[t_{(n-2)}^{\frac{\alpha}{2}} < T < t_{(n-2)}^{1-\frac{\alpha}{2}}\right] = 1-\alpha$$

donde $t_{(n-2)}^{1-\frac{\alpha}{2}}$ y $t_{(n-2)}^{\frac{\alpha}{2}}$ son los cuantiles $1-\frac{\alpha}{2}$ y $\frac{\alpha}{2}$ de una distribución t con $(n-2)$ grados de libertad y $t_{(n-2)}^{1-\frac{\alpha}{2}} = -t_{(n-2)}^{\frac{\alpha}{2}}$.

Por lo tanto, el intervalo del $(1-\alpha) \times 100\%$ de confianza para β_1 está dado por

$$\hat{\beta}_1 \pm t_{(n-2)}^{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

2.2.2 Intervalo de confianza para β_0 .

De manera similar al caso anterior tenemos

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sigma^2\right)$$

y

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

Si $t_{(n-2)}^{\frac{\alpha}{2}}$ y $t_{(n-2)}^{1-\frac{\alpha}{2}}$ denotan los cuantiles $\frac{\alpha}{2}$ y $1-\frac{\alpha}{2}$ de una distribución t con $(n-2)$ grados de libertad y considerando que $t_{(n-2)}^{1-\frac{\alpha}{2}} = -t_{(n-2)}^{\frac{\alpha}{2}}$, el intervalo del $(1-\alpha) \times 100\%$ de confianza para β_0 está dado por

$$\hat{\beta}_0 \pm t_{(n-2)}^{1-\frac{\alpha}{2}} \hat{\sigma} \frac{\sqrt{\sum_{i=1}^{n} X_i^2}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

2.2.3 Intervalo de confianza para σ^2

Dado que

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} = (n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

se tiene

$$P\left[\chi^{2}_{(n-2)\frac{\alpha}{2}} < (n-2)\frac{\hat{\sigma}^2}{\sigma^2} < \chi^{2}_{(n-2)1-\frac{\alpha}{2}}\right] = 1-\alpha$$

Entonces, un intervalo del $(1-\alpha)\times 100\%$ de confianza para σ^2 está dado por

$$\left(\frac{(n-2)\hat{\sigma}^2}{\chi^{2}_{(n-2)1-\frac{\alpha}{2}}}, \frac{(n-2)\hat{\sigma}^2}{\chi^{2}_{(n-2)\frac{\alpha}{2}}}\right)$$

2.3 Coeficiente de correlación

Un tema muy importante en los modelos lineales es el que discute la variación conjunta de dos o más variables y responde a la pregunta: ¿Qué tan estrechamente están asociadas las variables?, o de otra manera, ¿cuál es el grado de asociación entre las variables?.

Las técnicas que se han desarrollado para medir el grado de asociación entre variables, son conocidas como *métodos de correlación*. Este nombre refleja la práctica generalizada de hablar acerca de *medidas de correlación* . Consecuentemente, cuando se hace un análisis para determinar la cantidad de correlación, se dice que se ha efectuado un análisis de correlación. La medida de correlación es usualmente conocida como *coeficiente de correlación*.

Debido a la naturaleza del concepto, es claro que está estrechamente relacionado con el concepto de regresión. Así, para una ecuación de regresión dada, se verá que es razonable esperar que un coeficiente de correlación medirá qué tan bien se ajusta a los datos la ecuación de regresión.

Para dar una expresión apropiada del coeficiente de correlación, consideremos primero que dicho coeficiente se define entre dos variables aleatorias X y Y como

$$\rho_{xy} = \frac{Cov\,(X\,,Y\,)}{\sigma_x\sigma_y} = \frac{E\,((X-E(X))(Y-E(Y))}{\sigma_x\sigma_y}\,,$$

donde σ_x y σ_y denotan las desviaciones estándar de X y Y respectivamente. De esta definición se demuestra que $-1 < \rho_{xy} < 1$.

En análisis de regresión lineal es de interés calcular el coeficiente de correlación muestral, es decir, el coeficiente de correlación para la muestra

$$(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$$

Dicha muestra puede ser considerada como la pareja de vectores en $\Re^n$

$$\underline{X}' = (X_1, X_2, ..., X_n) \ , \ \underline{Y}' = (Y_1, Y_2, ..., Y_n)$$

Decimos que dos vectores, $\underline{X}$ y $\underline{Y}$ son ortogonales si el producto $\underline{X}'\underline{Y}$ o $\underline{Y}'\underline{X}$ es cero. Geométricamente, dos vectores ortogonales son perpendiculares entre sí o forman ángulo recto en el origen. Dos vectores linealmente dependientes forman ángulos de 0 o de 180 grados en el origen. Todos los otros ángulos reflejan vectores que no son ortogonales ni linealmente dependientes. En general, si θ representa el ángulo entre dos vectores, el coseno de éste es

$$\cos\theta = \frac{\underline{X}'\underline{Y}}{\sqrt{\underline{X}'\underline{X}}\sqrt{\underline{Y}'\underline{Y}}} = \frac{\underline{X}'\underline{Y}}{\|\underline{X}\| \ \|\underline{Y}\|}$$

Si los elementos de cada vector tienen media cero, el coseno del ángulo formado por los dos vectores es la correlación entre las dos columnas de datos en los vectores. Por lo tanto, la ortogonalidad de dos vectores corresponde a una correlación igual a cero entre los elementos en los dos vectores. Si dos vectores son linealmente dependientes, el coeficiente de correlación entre los elementos de los dos vectores será 1 ó -1, dependiendo de si los vectores tienen la misma dirección o direcciones contrarias.

Así, la expresión del coeficiente de correlación muestral, conocida como coeficiente de correlación de Pearson es

$$r_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

el cual debe satisfacer

$$-1 \leq r_{xy} \leq 1$$

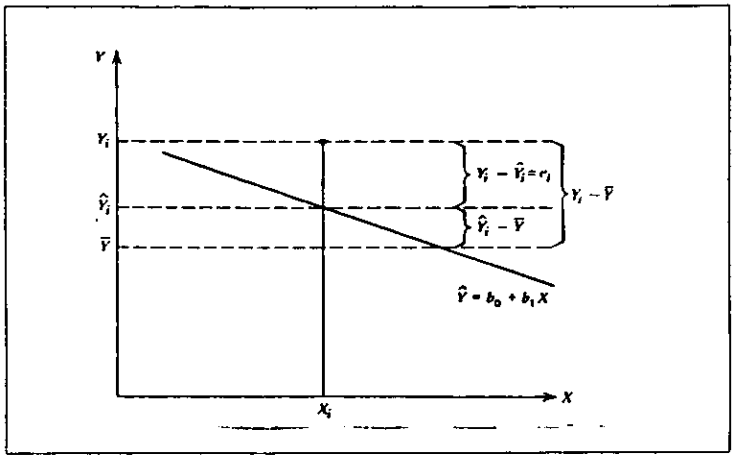


Figura 2: Interpretación geométrica de la igualdad fundamental del análisis de varianza

Para demostrar que ésta se cumple, veremos antes lo que se llama la *igualdad fundamental del análisis de varianza* (o la partición de $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$), para lo cual consideraremos la igualdad $Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$

En esta expresión $Y_i - \bar{Y}$ nos indica la distancia vertical entre Y_i y $\bar{Y}$. Elevando al cuadrado ambos lados de la igualdad y sumando sobre i se tiene

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})\right)^2$$

desarrollando el segundo miembro se tiene

$$\begin{aligned}\sum_{i=1}^{n}\left((Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})\right)^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 \\ &\quad +2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})\end{aligned}$$

donde

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$$

ya que

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y}) = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)\hat{\beta}_1(X_i-\bar{X})$$

En efecto

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_0+\hat{\beta}_1 X_i \\ \bar{Y} &= \hat{\beta}_0+\hat{\beta}_1\bar{X} \\ \Rightarrow\ \hat{Y}_i-\bar{Y} &= \hat{\beta}_1(X_i-\bar{X}) \end{aligned}$$

entonces

$$\begin{aligned} \sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y}) &= \hat{\beta}_1\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(X_i-\bar{X}) \\ &= \hat{\beta}_1\sum_{i=1}^{n}(X_i-\bar{X})\Big((Y_i-\bar{Y})-\hat{\beta}_1(X_i-\bar{X})\Big) \end{aligned}$$

puesto que

$$\begin{aligned} Y_i-\hat{Y}_i &= (Y_i-\bar{Y})+(\bar{Y}-\hat{Y}_i) \\ &= (Y_i-\bar{Y})-(\hat{Y}_i-\bar{Y}) = (Y_i-\bar{Y})-\hat{\beta}_1(X_i-\bar{X}) \end{aligned}$$

por lo tanto

$$\begin{aligned} \sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y}) &= \hat{\beta}_1\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})-\hat{\beta}_1^2\sum_{i=1}^{n}(X_i-\bar{X})^2 \\ &= \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) \\ &\quad -\frac{\left(\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})\right)^2}{\left(\sum_{i=1}^{n}(X_i-\bar{X})^2\right)^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \end{aligned}$$

$$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y}) = 0$$

de donde

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$$

La notación que se usa para esta igualdad es:

$$\mathbf{SCT = SCE + SCR}$$

donde **SCE** es la suma de cuadrados de residuales o suma de cuadrados del error y **SCR** es la suma de cuadrados debida a la regresión. **SCT** es conocida como la suma de cuadrados total o suma de cuadrados corregida por la media.

Esto significa que la variabilidad total de los valores de Y alrededor de su media muestral puede ser descompuesta en dos partes, la primera es la variación de los valores de $\hat{Y}$ alrededor de su media; note que

$$\begin{aligned} \hat{Y}_i &= \hat{\beta}_0 + \hat{\beta}_1 X_i \\ \Rightarrow \quad \overline{\hat{Y}} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \end{aligned}$$

y

$$\begin{aligned} \overline{\hat{Y}} &= \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \\ &= (\bar{Y} - \hat{\beta}_1 \bar{X}) + \hat{\beta}_1 \bar{X} \\ &= \bar{Y} \end{aligned}$$

esto es

$$\overline{\hat{Y}} = \bar{Y}$$

Generalmente se dice que $\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$ es la suma de cuadrados *debida a* o *explicada por* la influencia lineal de X. La segunda componente es la variación residual o *no explicada* de los valores de Y alrededor de la línea de ajuste.

Considere a continuación que

$$
\begin{aligned}
(\hat{Y}_i - \bar{Y}) &= \hat{\beta}_1 (X_i - \bar{X}) \\
\Rightarrow \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 &= \hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 \\
\Rightarrow \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} &= \hat{\beta}_1^2 \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \\
&= \frac{\left(\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \right)^2}{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \right)^2} \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}
\end{aligned}
$$

es decir

$$
\frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \frac{\left(\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \right)^2}{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \right) \left(\sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right)} = r_{xy}^2
$$

Esto significa que en el caso del modelo lineal simple, el coeficiente de correlación al cuadrado, denotado por R^2, es igual a la proporción de la variación total en torno a la media $\bar{Y}$ explicada por la regresión; idealmente se espera que la suma de cuadrados debida a la regresión sea mucho mayor que la suma de cuadrados del error, es decir, que R^2 tenga un valor muy cercano a uno. Este es llamado también coeficiente de determinación.

Ahora bien,

$$
\begin{aligned}
\sum_{i=1}^{n} (Y_i - \bar{Y})^2 &= \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \\
\Rightarrow \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} &= \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} + \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}
\end{aligned}
$$

por lo tanto

$$
r_{xy}^2 = 1 - \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}
$$

de donde se deduce que el valor máximo de r_{xy}^2 es 1, lo cual ocurre sólo cuando

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 0$$

y esto último ocurre cuando y sólo cuando cada una de las $Y_i - \hat{Y}_i = e_i$ es igual a cero y, por tanto, los puntos están sobre una línea recta. Por lo anterior, los límites de r son ± 1, donde el signo queda determinado por el signo del término

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

El valor mínimo de r^2 es cero y ocurre cuando

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

esto es, cuando la línea de regresión es $\hat{Y} = \bar{Y}$ y la variación explicada es cero, ya que

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

La suma de cuadrados total no corregida de las Y_i , $SC_{Total} = \sum_{i=1}^{n} Y_i^2$, puede ser particionada de manera similar

$$\begin{aligned} Y_i &= \hat{Y}_i + e_i \\ \Rightarrow \sum_{i=1}^{n} Y_i^2 &= \sum_{i=1}^{n} \hat{Y}_i^2 + \sum_{i=1}^{n} e_i^2 \\ SC_{Total} &= SC_{(Modelo)} + SC_{(Res.)} \end{aligned}$$

dado que el término $\sum_{i=1}^{n} \hat{Y}_i e_i$ se hace cero. El término $\sum_{i=1}^{n} \hat{Y}_i^2$ es la suma de cuadrados aportada por el modelo y $\sum_{i=1}^{n} e_i^2$ es la suma de cuadrados de la parte no explicada por el modelo.

Las formas de cálculo más convenientes son

$$\begin{aligned} SC_{(Modelo)} &= \sum_{i=1}^{n} Y_i^2 - \sum_{i=1}^{n} e_i^2 \\ SC_{(Res)} &= \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} Y_i^2 - SC_{(Modelo)} \end{aligned}$$

La partición de la suma de cuadrados total no corregida, puede ser expresada en términos de la suma de cuadrados corregida, restando $n\bar{Y}^2$ de cada lado de la igualdad. Entonces

$$\begin{aligned} \sum_{i=1}^{n} Y_i^2 - n\bar{Y} &= \left(\sum_{i=1}^{n} \hat{Y}_i^2 - n\bar{Y}\right) + \sum_{i=1}^{n} e_i^2 \\ \Rightarrow \sum_{i=1}^{n} (Y_i - \bar{Y})^2 &= \hat{\beta}_1 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} e_i^2 \end{aligned}$$

que corresponde a la partición vista anteriormente.

2.4 Pruebas de hipótesis

En el aspecto que concierne a las pruebas de hipótesis, es de gran importancia la que se refiere a la pendiente, planteando

$$Ho : \beta_1 = 0 \ \ vs. \ \ Ha : \beta_1 \neq 0$$

Esta dice que la variable dependiente Y no muestra ni incremento ni decremento lineal cuando cambia la variable independiente. En algunos casos, la naturaleza del problema sugerirá otros valores para la hipótesis nula. Un desarrollo preliminar de dicha prueba es el que sigue.

Dado que

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right)$$

se tiene

$$\frac{\hat{\beta}_1 - \beta_1}{\frac{\sigma}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}}} \sim N(0,1)$$

por lo tanto

$$\frac{(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(1)}$$

Por otro lado también se sabe que

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

Estas dos variables son independientes (como se verá más adelante), por lo tanto

$$\frac{\frac{(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} (X_i - \bar{X})^2}{\sigma^2}}{\frac{\sum_{i=1}^{n} e_i^2}{\frac{\sigma^2}{(n-2)}}} = \frac{(\hat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n} (X_i - \bar{X})^2}{\frac{\sum_{i=1}^{n} e_i^2}{(n-2)}} = F \sim F_{(1,n-2)}$$

Bajo Ho la F se reduce a

$$F = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2}{\frac{\sum_{i=1}^{n} e_i^2}{(n-2)}} = \frac{Q_1}{\frac{Q_2}{(n-2)}}$$

donde

$$\hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

De esta forma, la estadística F puede utilizarse para llevar a cabo la prueba; note que las sumas de cuadrados que aparecen en la expresión, corresponden a las sumas de cuadrados en las cuales queda particionada $\sum_{i=1}^{n} (Y_i - \bar{Y})^2$, la suma de cuadrados total.

Tabla de Análisis de Varianza ($Ho: \beta_1 = 0$)

Fuente de variación	Grados de libertad	Sumas de Cuadrados	Cuadrados Medios	F
Regresión	1	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = Q_1$	$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	$\dfrac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\dfrac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$
Residuales	$n-2$	$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = Q_2$	$\dfrac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}$	
Total	$n-1$	$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = Q_1 + Q_2$		

Todos los cálculos para obtener la estadística F se resumen en una tabla, llamada Tabla de Análisis de Varianza (tabla anterior).

Ejemplo 2: La industria del curtido se enfrenta a un problema complejo, el control de la contaminación del agua. Los desperdicios de esta industria se caracterizan por altos valores de demanda de oxígeno bioquímico, sólidos volátiles y otros contaminantes. Considere los datos experimentales que aparecen en la tabla 2.

Tabla 2: Porcentaje de reducción en los sólidos totales X y porcentaje de reducción en la demanda de oxígeno químico$Y$ en desperdicio químicamente tratado

X	Y	X	Y	X	Y	X	Y
3	5	30	35	36	38	41	41
7	11	31	30	36	34	42	40
11	21	31	40	37	36	42	44
15	16	32	32	38	38	43	37
18	16	33	34	39	37	44	44
27	28	33	32	39	36	45	46
29	27	34	34	39	45	46	46
30	25	36	37	40	39	47	49
						50	51

Estos se obtuvieron de 33 muestras de desperdicio químicamente tratado durante el estudio *Chemical Treatment of Spent Vegetable Tan Liquor*, realizado en Virginia Polytecnic Institute y State University, en 1970. Se registraron las lecturas de X, porcentaje de reducción en los sólidos totales y de Y, porcentaje de reducción en la demanda de oxígeno químico para las 33 muestras. Los datos de la tabla 2 se graficaron en la figura 3 para obtener un diagrama de dispersión.

En una primera inspección de este diagrama, se observa que los datos siguen muy claramente una cierta línea recta, lo cual indica que la suposición de linealidad parece razonable. En este mismo diagrama se han dibujado dos rectas; una corresponde a la línea ajustada o estimada y la otra a la recta teórica o desconocida, que ha sido esbozada basándose en experiencia anterior.

De los datos de la tabla 2 obtenemos:

$$\sum_{i=1}^{33} X_i = 1104 \ , \ \sum_{i=1}^{33} Y_i = 1124 \ , \ \sum_{i=1}^{33} X_i Y_i = 41355 \ , \ \sum_{i=1}^{33} X_i^2 = 41086$$

de donde $\hat{\beta}_0 = 3.829640$, $\hat{\beta}_1 = 0.903643$

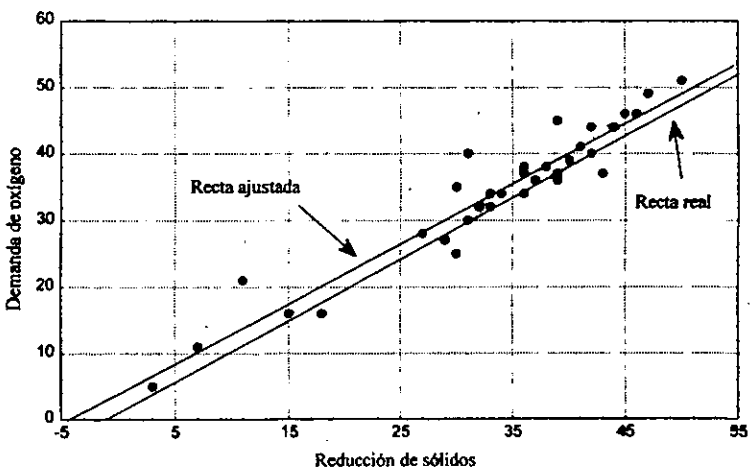


Figura 3: Recta ajustada y recta real en el diagrama de dispersión del ejemplo 2

De esta forma, la recta de regresión está dada por

$$\hat{Y} = 3.8296 + 0.9036X$$

en la cual, se redondean los coeficientes a cuatro decimales. Tenemos también

$$\sum_{i=1}^{33}(X_i - \bar{X})^2 = 4152.18 \ , \ \sum_{i=1}^{33}(Y_i - \bar{Y})^2 = 3713.88 \ , \ \sum_{i=1}^{33}(X_i - \bar{X})(Y_i - \bar{Y}) = 3752.09$$

Donde:

X: Reducción de Sólidos (%)

Y: Demanda de oxígeno químico

Observe que

$$\begin{aligned}
\sum_{i=1}^{n} e_i{}^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \\
&= \sum_{i=1}^{n}\left((Y_i - \bar{Y}) - \hat{\beta}_1(X_i - \bar{X})\right)^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - 2\hat{\beta}_1\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) + \hat{\beta}_1^2\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - 2\hat{\beta}_1\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \\
&\quad + \frac{\left(\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})\right)^2}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}\sum_{i=1}^{n}\left(\acute{X}_i - \bar{X}\right)^2 \\
&= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})
\end{aligned}$$

Generalmente se toma la siguiente notación para simplificar las expresiones

$$S_{xx} = \sum_{i=1}^{n}(X_i - \bar{X})^2 \ , \ S_{yy} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \ , \ S_{xy} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

De este modo, podemos escribir: $\sum_{i=1}^{n} e_i{}^2 = S_{yy} - \hat{\beta}_1 S_{xy}$, lo cual simplifica los cálculos, puesto que nos evita calcular cada $\hat{Y}_i$. El intervalo del 95% de confianza para β_0 es

$$0.2131 < \beta_0 < 7.4462$$

de donde se infiere que la línea no pasa por el origen y puede seguir considerándose a β_0 en el modelo. El intervalo del 95% de confianza para β_1 es

$$0.8011 < \beta_1 < 1.0061$$

es decir, $\beta_1 \neq 0$ y por lo tanto tiene sentido la regresión; además, este resultado sugiere que $\beta_1 = 1$.

Tabla de análisis de varianza $(Ho:\ \beta_1 = 0)$

Fuente de variación	Grados de libertad	Sumas de cuadrados	Cuadrados medios	F
Regresión	1	3,390.38	3,390.38	
Error	31	323.49	10.4356	324.899
Total	32	3,713.88		

$$\sum_{i=1}^{n} e_i^2 = S_{yy} - \hat{\beta}_1 S_{xy} = 3,713.88 - (0.9036)(3,752.09)$$

$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \sum_{i=1}^{n} e_i^2 = 3,390.38$$

El cuantil .95 de una $F_{(1,31)}$ es 4.17; el cuantil .99 de una $F_{(1,31)}$ es 7.56. Por lo tanto, la prueba es altamente significativa, esto es, rechazamos Ho con un alto grado de confianza.

Para calcular r ó r^2, notemos primero que:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} = \frac{S_{xy}}{S_{xx}} \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} = \hat{\beta}_1 \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}}$$

por lo tanto

$$r^2 = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

es la proporción de la variación total en Y que queda explicada por el ajuste de la regresión. Sustituyendo los valores en nuestro problema

$$r^2 = (0.9036)^2 \frac{4,152.18}{3,713.88} = 0.9128$$

lo cual indica que la mayor parte de la variabilidad corresponde al ajuste de la regresión.

Así, se rechaza la hipótesis $Ho : \beta_1 = 0$ al nivel de significancia α si $F > F^{1-\alpha}_{(1,n-2)}$ donde $F^{1-\alpha}_{(1,n-2)}$ denota el cuantil $(1-\alpha)$ de una distribución F con 1 y $n-2$ grados de libertad.

El desarrollo formal para llevar a cabo la prueba de hipótesis $Ho : \beta_1 = 0$ es por razón de verosimilitudes:

Se definen primero los espacios paramétricos:

$$\Theta_{Ho} = \{(\beta_0, \beta_1, \sigma^2) \mid \beta_0 \in \Re,\ \beta_1 = 0,\ 0 < \sigma^2 < \infty\}$$

y

$$\Theta = \{(\beta_0, \beta_1, \sigma^2) \mid \beta_0, \beta_1 \in \Re,\ 0 < \sigma^2 < \infty\}$$

$L(\Theta_{Ho})$ y $L(\Theta)$ denotan a las funciones de verosimilitud en los espacios correspondientes y $L(\widehat{\Theta_{Ho}})$ y $L(\widehat{\Theta})$ denotan a las funciones evaluadas en el punto donde alcanzan su valor máximo. Entonces

$$\begin{aligned} L(\Theta_{Ho}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \beta_0)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0)^2\right\} \\ \ln L(\Theta_{Ho}) &= -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0)^2 \end{aligned}$$

derivando e igualando a cero esta expresión se tiene

$$\begin{aligned} \frac{\partial \ln L(\Theta_{Ho})}{\partial \beta_0} &= \frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(Y_i - \beta_0) = 0 \\ \frac{\partial \ln L(\Theta_{Ho})}{\partial \sigma^2} &= -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(Y_i - \beta_0)^2 = 0 \end{aligned}$$

De este sistema se obtiene

$$\hat{\beta}_0 = \bar{Y}$$

y

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

por lo tanto

$$\begin{aligned} L(\widehat{\Theta_{Ho}}) &= \left(\frac{n}{2\pi\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{n}{2\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\sum_{i=1}^{n}(Y_i-\bar{Y})^2\right\} \\ &= \left(\frac{n}{2\pi\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{n}{2}\right\} \end{aligned}$$

Por otro lado

$$\begin{aligned} L(\Theta) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_i)^2\right\} \\ \ln L(\Theta) &= -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_i)^2 \\ \frac{\partial \ln L(\Theta)}{\partial\beta_0} &= \frac{1}{2\sigma^2}\sum_{i=1}^{n}2(Y_i-\beta_0-\beta_1X_i) = 0 \\ \frac{\partial \ln L(\Theta)}{\partial\beta_1} &= \frac{1}{2\sigma^2}\sum_{i=1}^{n}2(Y_i-\beta_0-\beta_1X_i)(X_i) = 0 \\ \frac{\partial \ln L(\Theta)}{\partial\sigma^2} &= -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_i)^2 = 0 \end{aligned}$$

Resolviendo el sistema tenemos

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} \; ; \; \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$$

y

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}e_i{}^2$$

Sustituyendo $\hat{\beta}_0$, $\hat{\beta}_1$ y $\hat{\sigma}^2$ en la función de verosimilitud y considerando que $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, se obtiene

$$L(\hat{\Theta}) = \left(\frac{n}{2\pi \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} \right)^{\frac{n}{2}} \exp\left\{ -\frac{n}{2\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \right\}$$

$$= \left(\frac{n}{2\pi \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} \right)^{\frac{n}{2}} \exp\left\{ -\frac{n}{2} \right\}$$

De esta forma, la razón de verosimilitudes es la siguiente

$$\lambda = \frac{L(\widehat{\Theta_{Ho}})}{L(\hat{\Theta})} = \frac{\left(\frac{n}{2\pi \sum_{i=1}^{n} (Y_i - \bar{Y})^2} \right)^{\frac{n}{2}}}{\left(\frac{n}{2\pi \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} \right)^{\frac{n}{2}}} = \left(\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \right)^{\frac{n}{2}} \leq \lambda_0$$

lo cual ocurre si y sólo si

$$\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} \leq \lambda_0^{\frac{2}{n}}$$

donde, por la partición de la suma de cuadrados total tenemos

$$\frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2} = \frac{1}{1 + \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}}$$

por lo tanto, la desigualdad anterior puede escribirse como

$$\frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \geq c$$

donde

$$\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

y

$$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$$

donde, bajo Ho

$$\hat{\beta}_1 \sim N\left(0, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)$$

entonces

$$\frac{\frac{\hat{\beta}_1}{\sigma}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}}} \sim N(0,1)$$

$$\Rightarrow \frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{(1)}$$

De lo anterior se obtiene

$$F = \frac{\frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{1}}{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}} = \frac{\frac{\hat{\beta}_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2}}{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sigma^2}} \sim F_{(1,n-2)}$$

Por lo tanto la regla decisión es: rechazar $Ho : \beta_1 = 0$ al nivel de significancia α si el valor de la estadística F excede al cuantil $(1-\alpha)$ de una distribución $F_{(1,n-2)}$, i. e., si

$$F > F^{1-\alpha}_{(1,n-2)}$$

Como puede verse, el resultado es el mismo que se obtuvo anteriormente y de la misma forma, el resumen del desarrollo efectuado para esta prueba queda expresado en la tabla de análisis de varianza correspondiente.

2.5 Predicción

Una vez obtenida la ecuación de regresión (o línea ajustada), la pregunta natural que surge es: ¿Cómo va a ser utilizada la ecuación de regresión ?. El objetivo primario es proveer una buena descripción del comportamiento de la variable dependiente (o de respuesta); esto se logra haciendo una interpretación adecuada de la ecuación de regresión. En segundo lugar, es de gran interés la predicción de respuestas futuras y la estimación de respuestas medias , así como la extrapolación o predicción de respuestas fuera del rango de los datos.

2.5.1 Intervalo de confianza de la respuesta media

Un uso importante de un modelo de regresión es estimar la respuesta media $E(Y)$ para un valor particular de la variable regresora o independiente X. Sea X_0 el nivel de esta variable para el cual deseamos estimar la respuesta media, es decir $E(Y \mid X_0)$. Suponemos que X_0 es cualquier valor de X sobre el rango de los datos originales sobre X usados para ajustar el modelo.

Un estimador puntual de $E(Y \mid X_0)$ que sea insesgado es

$$\widehat{E(Y \mid X_0)} = \hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 \ .$$

donde $\hat{\beta}_0$ y $\hat{\beta}_1$ son los estimadores por mínimos cuadrados. Para obtener un intervalo del $(1-\alpha)\times 100\%$ de confianza para $E(Y \mid X_0)$, notamos primero que $\hat{Y}_0$ es una combinación lineal de las $Y_i's$, por lo tanto, es normalmente distribuido. La varianza de $\hat{Y}_0$ es

$$
\begin{aligned}
\mathrm{Var}(\hat{Y}_0) &= \mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 X_0) = \mathrm{Var}(\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_0) = \mathrm{Var}(\bar{Y} + \hat{\beta}_1 (X_0 - \bar{X})) \\
&= \mathrm{Var}\left(\sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} (X_0 - \bar{X}) \right) Y_i \right) \\
&= \sum_{i=1}^{n} \left(\frac{1}{n} + \frac{(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} (X_0 - \bar{X}) \right)^2 \mathrm{Var}(Y_i)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \sigma^2 \left(\frac{1}{n^2} + \frac{(X_i - \bar{X})^2}{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2 \right)^2} (X_0 - \bar{X})^2 + 2 \frac{1}{n} \frac{(X_i - \bar{X})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} (X_0 - \bar{X}) \right) \\
&= n \frac{\sigma^2}{n^2} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \sigma^2 \\
&= \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \sigma^2
\end{aligned}
$$

Tenemos entonces

$$
\frac{\hat{Y}_0 - E(Y \mid X_0)}{\sqrt{\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \sigma^2}} \sim N(0,1) \quad \text{y} \quad \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi^2_{(n-2)}
$$

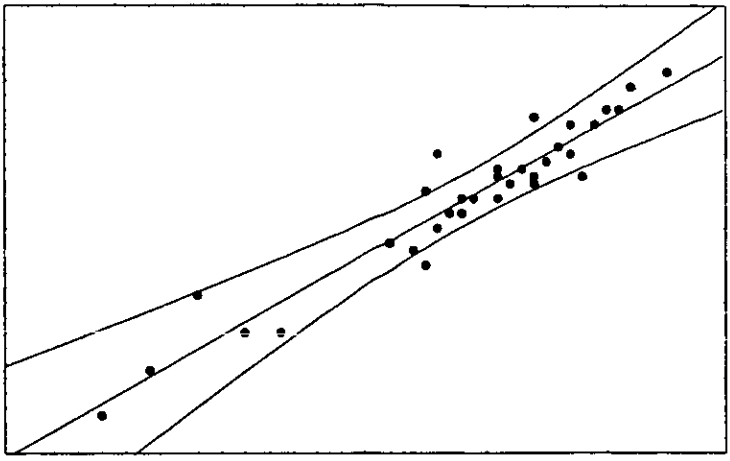


Figura 4: Representación gráfica del intervalo de confianza para $E(Y \mid X_0)$

lo cual implica

$$t = \frac{\hat{Y}_0 - E(Y \mid X_0)}{\sqrt{\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)\sigma^2}} \Big/ \sqrt{\frac{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sigma^2}}{n-2}} = \frac{\hat{Y}_0 - E(Y \mid X_0)}{\hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}} \sim t_{(n-2)}$$

donde

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \mathbf{CME}$$

Por lo tanto, un intervalo del $(1-\alpha) \times 100\%$ de confianza sobre la respuesta media en el punto $X = X_0$ es

$$\hat{Y}_0 - t_{(n-2)}^{1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\left(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)} < E(Y \mid X_0) < \hat{Y}_0 + t_{(n-2)}^{1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\left(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)}$$

Note que la amplitud del intervalo es función de X_0. La amplitud mínima del intervalo ocurre en el punto $X_0 = \bar{X}$.

2.5.2 Predicción de observaciones nuevas

Si X_0 es el valor de X, entonces

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

es el estimador puntual del nuevo valor de respuesta Y_0.

2.5.3 Intervalo de predicción para la futura observación Y.

Note que la variable aleatoria

$$\begin{aligned} \psi &= (Y_0 - \hat{Y}_0) \sim N(0, Var(\psi)) \\ Var(\psi) &= Var(Y_0 - \hat{Y}_0) \\ &= Var(Y_0) + Var(\hat{Y}_0) \end{aligned}$$

ya que la observación futura Y_0 es independiente de $\hat{Y}_0$. Por lo tanto

$$Var(\psi) = \sigma^2 + \sigma^2\left(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right) = \sigma^2\left(1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)$$

El intervalo de predicción del $(1-\alpha)\times 100\%$ de confianza sobre una observación futura en X_0 es

$$\hat{Y}_0 - t_{(n-2)}^{1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}\left(1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)} < Y_0 < \hat{Y}_0 + t_{(n-2)}^{1-\frac{\alpha}{2}}\sqrt{\hat{\sigma}\left(1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)}$$

donde $t_{(n-2)}^{1-\frac{\alpha}{2}}$ es el cuantil $1-\frac{\alpha}{2}$ de una distribución t con $n-2$ grados de libertad.

Este intervalo tiene amplitud mínima en $X_0 = \bar{X}$ y se hace más ancho a medida que $|X_0 - \bar{X}|$ se incrementa.

Comparando los intervalos que acabamos de calcular, observamos que el intervalo de predicción en X_0 es siempre más ancho que el intervalo de confianza en X_0, debido a que el intervalo de predicción depende de dos cosas: El error del modelo ajustado y el error asociado con observaciones futuras.

Podemos generalizar el intervalo de predicción para encontrar un intervalo de predicción del $(1-\alpha) \times 100\%$ de confianza sobre la media de m futuras observaciones sobre la respuesta en $X = X_0$. Sea $\bar{Y}_0$ la media de m futuras observaciones en $X = X_0$. Un estimador puntual de $\bar{Y}_0$ es:

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$$

El intervalo de predicción del $(1-\alpha) \times 100\%$ sobre $\bar{Y}_0$ es

$$\hat{Y}_0 - t_{(n-2)}^{1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}\left(\frac{1}{m} + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)} < \bar{Y}_0 <$$

$$\hat{Y}_0 + t_{(n-2)}^{1-\frac{\alpha}{2}} \sqrt{\hat{\sigma}\left(\frac{1}{m} + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$$

Ejemplo 3.- Utilizando los datos de la tabla 2, construiremos un intervalo del 95% de confianza para $E(Y \mid X_0)$ con $X_0 = 20$. A partir de la ecuación de regresión se encuentra que:

$\hat{Y}_0 = 3.8296 + (0.9036)(20) = 21.9025$

$\bar{X} = 33.4545,\ S_{xx} = 4152.18,\ \hat{\sigma} = 3.2296$

$$21.095 - (2.045)(3.2296)\sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}} < Y_0 <$$

$$21.095 + (2.045)(3.2296)\sqrt{1 + \frac{1}{33} + \frac{(20 - 33.4545)^2}{4152.18}}$$

lo cual ocurre sólo cuando

$$14.2508 < Y_0 < 27.9392$$

2.6 Inferencia simultánea en regresión lineal simple

Además de los intervalos individuales para $\hat{\beta}_0$ y $\hat{\beta}_1$, así como predicción, muchos problemas requieren que se construyan varios intervalos de confianza o de predicción utilizando los mismos datos muestrales. En estos casos, el investigador está interesado en especificar el coeficiente de confianza que sirva simultáneamente para el conjunto entero de intervalos de confianza. Los intervalos de confianza o predicción que son todos ciertos simultáneamente con probabilidad $1-\alpha$, son llamados intervalos de confianza o predicción simultáneos.

Usualmente, si tenemos un intervalo del 95% de confianza para β_0 y otro intervalo de confianza para β_1, pero se quiere hacer inferencia sobre ambos conjuntamente, una posibilidad sería construir intervalos del 95% de confianza para ambos parámetros. Sin embargo, si estas estimaciones por intervalo son independientes, la probabilidad asociada a ambos conjuntamente es $(0.95)^2 = 0.9025$. Además, dado que los intervalos están construidos usando el mismo conjunto de datos, no son independientes. Esto introduce una nueva complicación en la determinación del nivel de confianza para el conjunto de proposiciones.

2.6.1 Inferencia simultánea sobre los parámetros del modelo.

Considere la estimación de β_0 y β_1 con una región del $(1-\alpha) \times 100\%$ de confianza. El modelo es:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \varepsilon \\ &= \beta_0' + \beta_1(X - \bar{X}) + \varepsilon \end{aligned}$$

donde $\beta_0' = \beta_0 + \beta_1 \bar{X}$

Los estimadores por mínimos cuadrados para β_0' y β_1 son

$$\hat{\beta}_0' = \bar{Y} \quad \text{y} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Estos estimadores son insesgados y sus varianzas son

$$Var(\hat{\beta}_0') = \frac{\sigma^2}{n} \qquad \text{y} \qquad Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Bajo las hipótesis usuales de normalidad, las variables aleatorias

$$\frac{\hat{\beta}_0' - \beta_0'}{\sqrt{\frac{\sigma^2}{n}}} \quad \text{y} \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}$$

son independientes, ya que $Cov(\hat{\beta}_0', \hat{\beta}_1) = 0$, entonces

$$\left(\frac{\hat{\beta}_0' - \beta_0'}{\sqrt{\frac{\sigma^2}{n}}}\right)^2 = \frac{n(\hat{\beta}_0' - \beta_0')^2}{\sigma^2} \sim \chi^2_{(1)}$$

$$\left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}\right)^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2(\hat{\beta}_1 - \beta_1)^2}{\sigma^2} \sim \chi^2_{(1)}$$

Como $\hat{\beta}_0'$ y $\hat{\beta}_1$ son independientes, estas dos variables aleatorias, cuya distribución es ji-cuadrada, son independientes. Así, usando la propiedad aditiva de esta distribución, encontramos que

$$\frac{n(\hat{\beta}_0' - \beta_0')^2}{\sigma^2} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2(\hat{\beta}_1 - \beta_1)^2}{\sigma^2} \sim \chi^2_{(2)}$$

además, $\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} \sim \chi^2_{(n-2)}$ y puede demostrarse que esta variable es independiente de $\hat{\beta}_0'$ y $\hat{\beta}_1$. Por lo tanto

$$\frac{\left(\frac{n(\hat{\beta}_0' - \beta_0')^2}{\sigma^2} + \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2(\hat{\beta}_1 - \beta_1)^2}{\sigma^2}\right)\frac{1}{2}}{\frac{\sum_{i=1}^{n} e_i^2}{\frac{\sigma^2}{n-2}}}$$

$$= \frac{n(\hat{\beta}_0' - \beta_0')^2 + \sum_{i=1}^{n}(X_i - \bar{X})^2(\hat{\beta}_1 - \beta_1)^2}{\frac{2\sum_{i=1}^{n} e_i^2}{(n-2)}} \sim F_{(2,n-2)}$$

Si substituimos ahora $\hat{\beta}_0' = \hat{\beta}_0 + \hat{\beta}_1\bar{X}$ y $\beta_0' = \beta_0 + \beta_1\bar{X}$ tenemos

$$\frac{n(\hat{\beta}_0 - \beta_0)^2 + 2\sum_{i=1}^{n} X_i(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + \sum_{i=1}^{n} X_i^2(\hat{\beta}_1 - \beta_1)^2}{\frac{2\sum_{i=1}^{n} e_i^2}{(n-2)}}$$

y como

$$P\left\{\frac{n(\hat{\beta}_0 - \beta_0)^2 + 2\sum_{i=1}^{n} X_i(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + \sum_{i=1}^{n} X_i^2(\hat{\beta}_1 - \beta_1)^2}{\frac{2\sum_{i=1}^{n} e_i^2}{(n-2)}} \leq F_{(2,n-2)}^{1-\alpha}\right\}$$
$$= 1 - \alpha$$

es válida para todos los valores de β_0 y β_1, la región del $(1-\alpha) \times 100\%$ de confianza para ambos parámetros es

$$\frac{n(\hat{\beta}_0 - \beta_0)^2 + 2\sum_{i=1}^{n} X_i(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) + \sum_{i=1}^{n} X_i^2(\hat{\beta}_1 - \beta_1)^2}{\frac{2\sum_{i=1}^{n} e_i^2}{(n-2)}} \leq F_{(2,n-2)}^{1-\alpha}$$

La desigualdad anterior define una elipse que con muestreo repetido contendrá a β_0 y β_1 simultáneamente el $(1-\alpha) \times 100\%$ de las veces.

La expresión del numerador en forma matricial es

$$\left(\hat{\beta}_0 - \beta_0,\ \hat{\beta}_1 - \beta_1 \right) \begin{pmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \end{pmatrix}$$

ya que es una forma cuadrática. Esta notación facilita el cálculo y el manejo de las expresiones, como se verá en regresión múltiple.

Ejemplo 4.- El motor de un cohete es construido mediante la unión de un propulsor de ignición y uno de soporte dentro de una cámara metálica. El esfuerzo cortante presente en la unión entre ambos es una característica importante al medir la calidad del motor. Se sospecha que dicho esfuerzo está relacionado con la edad en semanas del lote de propulsores de soporte. Veinte observaciones de edad del propulsor y el correspondiente esfuerzo en la unión se muestran en la tabla 3.

Tabla 3: Esfuerzo cortante vs. edad en semanas del propulsor de un cohete

Obs.	Esfuerzo (psi.)	Edad del propulsor (semanas)	Obs.	Esfuerzo (psi.)	Edad del propulsor (semanas)
i	Y_i	X_i	i	Y_i	X_i
1	2158.70	15.50	11	2165.20	13.00
2	1678.15	23.75	12	2399.55	3.75
3	2316.00	8.00	13	1779.80	25.00
4	2061.30	17.00	14	2336.75	9.75
5	2207.50	5.50	15	1765.30	22.00
6	1708.30	19.00	16	2053.50	18.00
7	1784.70	24.00	17	2414.40	6.00
8	2575.00	2.50	18	2200.50	12.50
9	2357.90	7.50	19	2654.20	2.00
10	2256.70	11.00	20	1753.70	21.50

Para estimar los parámetros del modelo se calculan

$$\begin{aligned} S_{xx} &= \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \\ &= 4667.69 - \frac{71422.56}{20} \\ &= 1106.56 \end{aligned}$$

y

$$\begin{aligned} S_{xy} &= \sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n} \\ &= 528492.64 - \frac{(267.25)\,(42627.15)}{20} \\ &= -41112.65 \end{aligned}$$

de lo anterior se obtiene

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-41112.65}{1106.56} = -37.15$$

y

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 2131.3575 - (-37.15)\,13.3625 = 2627.82$$

El modelo ajustado es

$$\hat{Y} = 2627.82 - 37.15X$$

A continuación hallaremos una región del 95% de confianza para β_0 y β_1. Los valores para CME, $\sum_{i=1}^{n} x_i$ y $\sum_{i=1}^{n} x_i^2$ se dan a continuación.

$$\begin{aligned} \text{CME} &= \frac{\text{SCE}}{n-2} \\ &= \frac{S_{yy} - \hat{\beta}_1 S_{xy}}{n-2} \\ &= \frac{166402.62}{18} \\ &= 9244.59 \end{aligned} \qquad \begin{aligned} \sum_{i=1}^{n} x_i &= 267.25 \\ \sum_{i=1}^{n} x_i^2 &= 4677.69 \end{aligned}$$

Con lo anterior se obtiene

$$\frac{\left[\begin{array}{c} 20\,(2627.82-\beta_0)^2+2\,(267.25)\,(2627.82-\beta_0)\,(-37.15-\beta_1) \\ +\,(4677.69)\,(-37.15-\beta_1)^2 \end{array}\right]}{[2\,(9244.59)]} \leq 3.55$$

como la región de confianza, puesto que $F_{.05}^{18} = 3.55$. La figura 5 muestra a la elipse definida por la ecuación anterior.

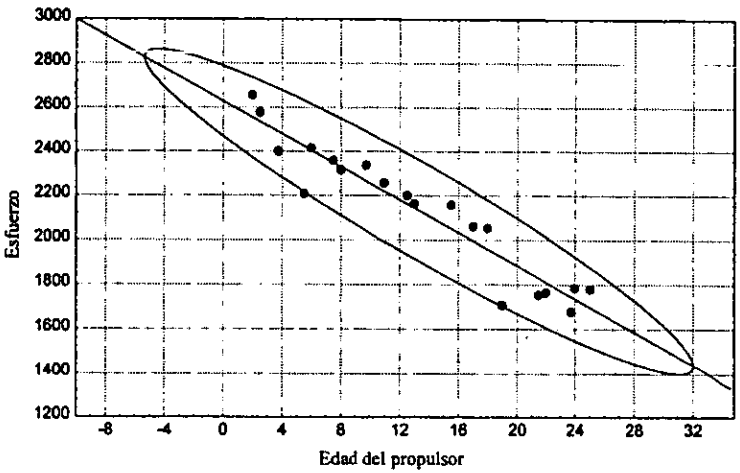


Figura 5: Representación gráfica del intervalo simultáneo de confianza para β_0 y β_1 en el ejemplo 4

2.7 Gráficas de Y vs. X

Una manera útil de comenzar el análisis del modelo es graficando la variable independiente X contra la variable dependiente Y. Esta gráfica puede servir tanto para sugerir una relación entre las variables como para manifestar posibles violaciones a los supuestos del modelo.

Usualmente la variable regresora se grafica en el eje horizontal y la variable independiente en el eje vertical.

Ejemplo 5, datos de Forbes.- Entre los años 1840 y 1850, un físico escocés, James D. Forbes, deseaba estimar la altura sobre el nivel del mar a partir de mediciones del punto de ebullición del agua. Él sabía que la altura podía ser determinada considerando la presión atmosférica, medida con un barómetro, donde presiones pequeñas correspondían a alturas grandes. En los experimentos descritos estudió la relación entre la presión atmosférica y el punto de ebullición del agua. Su interés en éste problema estaba motivado por la dificultad de transportar los frágiles barómetros de aquella época. Medir el punto de ebullición del agua daría a los viajeros una manera rápida de estimar la altura.

Forbes recolectó datos en Escocia y Los Alpes. Después de elegir un lugar, ensamblaba sus aparatos y medía el punto de ebullición del agua y la presión. Las presiones eran registradas en pulgadas de mercurio y el punto de ebullición del agua en grados Farenheit. Los datos para 17 lugares se reproducen en la siguiente tabla.

Datos de Forbes

Punto de ebullición	Presión atmosférica	Punto de ebullición	Presión atmosférica
194.5	20.79	201.3	24.01
194.3	20.79	203.6	25.14
197.9	22.4	204.6	26.57
198.4	22.67	209.5	28.49
199.4	23.15	208.6	27.76
199.9	23.35	210.7	29.04
200.9	23.89	211.9	29.88
201.1	23.99	212.2	30.06
201.4	24.02		

Examínese la gráfica de dispersión correspondiente a los datos de Forbes que se encuentra en la figura 6.

La impresión general que se obtiene de ella es que parece razonable que una línea recta es la función que mejor describe la relación entre las variables; sin embargo, más

adelante discutiremos algunos métodos que nos muestran que es posible hallar un modelo de regresión que describa más correctamente la relación entre el punto de ebullición y la presión atmosférica en el experimento de Forbes.

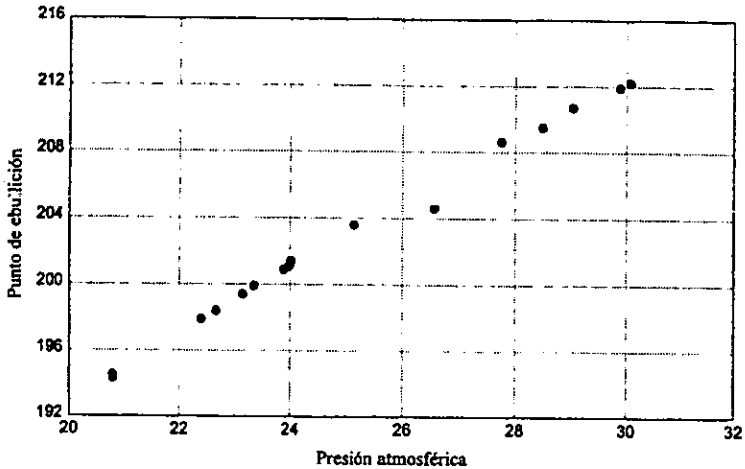


Figura 6: Diagrama de dispersión de los Datos de Forbes

2.8 Forma matricial del modelo lineal simple

En el presente capítulo hemos construido el modelo de regresión lineal simple en notación algebraica; a continuación introduciremos la notación matricial para éste. Esta notación nos permitirá más adelante continuar con la construcción de modelos más generales.

Sabemos que

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

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

$i = 1, 2, ..., n$

Si escribimos cada uno de los n elementos se tiene

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_1 + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_2 + \varepsilon_2 \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 X_n + \varepsilon_n
\end{aligned}
$$

que es equivalente a la siguiente expresión matricial

$$
\begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_0 + \beta_1 X_n + \varepsilon_n \end{pmatrix}
$$

Sea

$$
\underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{pmatrix}
$$

se tiene

$$
\underline{Y} = \begin{pmatrix} \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_0 + \beta_1 X_n + \varepsilon_n \end{pmatrix}
$$

descomponiendo el miembro derecho

$$\underline{Y} = \begin{pmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_0 + \beta_1 X_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & X_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix}$$

Si denotamos por

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & X_n \end{pmatrix}$$

$$\underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix}$$

se tiene la siguiente expresión de la recta ajustada

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$$

Note que $\underline{Y}$ es un vector de dimensión $n \times 1$, $\mathbf{X}$ es una matriz de $n \times 2$, $\underline{\beta}$ es un vector de 2×1 y $\underline{\varepsilon}$ es un vector de $n \times 1$.

Para denotar la hipótesis distribucional, definamos la esperanza de un vector $\underline{Z}$, de tamaño n, cuyas entradas z_i son variables aleatorias como:

$$E(\underline{Z}) = \begin{pmatrix} E(z_1) \\ E(z_2) \\ . \\ . \\ . \\ E(z_n) \end{pmatrix}$$

La esperanza de una matriz se define como la matriz de las esperanzas, esto es:

$$E(A) = \begin{pmatrix} E(a_{11}) & E(a_{12}) & \cdots & E(a_{1n}) \\ E(a_{21}) & E(a_{22}) & \cdots & E(a_{2n}) \\ . & & & . \\ . & & & . \\ . & & & . \\ E(a_{m1}) & E(a_{m2}) & \cdots & E(a_{mn}) \end{pmatrix}$$

Necesitamos también definir la varianza de un vector; sea $\underline{Z}$ un vector de tamaño n

$$Var(\underline{Z}) = E\left((\underline{Z} - E(\underline{Z}))(\underline{Z} - E(\underline{Z}))'\right)$$

que es la matriz

$$E\begin{pmatrix} (z_1 - E(z_1))^2 & (z_1 - E(z_1))(z_2 - E(z_2)) & \cdots & (z_1 - E(z_1))(z_n - E(z_n)) \\ (z_2 - E(z_2))(z_1 - E(z_1)) & (z_2 - E(z_2))^2 & \cdots & (z_2 - E(z_2))(z_n - E(z_n)) \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ (z_n - E(z_n))(z_1 - E(z_1)) & (z_n - E(z_n))(z_2 - E(z_2)) & \cdots & (z_n - E(z_n))^2 \end{pmatrix}$$

esto es:

$$Var(\underline{Z}) = \begin{pmatrix} Var(z_1) & Cov(z_1,z_2) & \cdots & Cov(z_1,z_n) \\ Cov(z_1,z_2) & Var(z_2) & \cdots & Cov(z_2,z_n) \\ \vdots & & & \\ Cov(z_1,z_n) & Cov(z_2,z_n) & \cdots & Var(z_n) \end{pmatrix}$$

Que es llamada la matriz de varianzas y covarianzas. Obsérvese que ésta es una matriz simétrica y que si los elementos del vector fueran independientes, se tendría la matriz diagonal

$$Var(\underline{Z}) = \begin{pmatrix} Var(z_1) & 0 & \cdots & 0 \\ 0 & Var(z_2) & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & Var(z_n) \end{pmatrix}$$

Así, tenemos la notación

$$\underline{\varepsilon} \sim N(0, \sigma^2 I_n)$$

$$E(\underline{\varepsilon}) = \underline{0}$$

$$Var\,(\underline{\varepsilon}) = \begin{pmatrix} Var(\varepsilon_1) & 0 & \cdots & 0 \\ 0 & Var(\varepsilon_2) & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & Var(\varepsilon_n) \end{pmatrix}$$

Por otro lado, la densidad conjunta de los errores

$$f\,(\varepsilon_1, \varepsilon_2, ..., \varepsilon_n) = f\,(\varepsilon_1) \cdot f\,(\varepsilon_2) \cdot ... f\,(\varepsilon_n)$$

puede ser escrita de la siguiente forma:

$$\begin{aligned} f(\underline{\varepsilon}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\varepsilon_i)^2\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\varepsilon_i)^2\right\} \end{aligned}$$

donde $f(\underline{\varepsilon}) = f(\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)$

La suma de cuadrados de los errores puede ser denotada por

$$\underline{\varepsilon}'\underline{\varepsilon} = \sum_{i=1}^{n}(\varepsilon_i)^2 = \underline{\varepsilon}' I_n \underline{\varepsilon}$$

con lo cual la densidad conjunta de los errores queda expresada de la siguiente manera:

$$\begin{aligned} f(\underline{\varepsilon}) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\underline{\varepsilon}'\left(\frac{1}{\sigma^2}I_n\right)\underline{\varepsilon}\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2}\underline{\varepsilon}'(\sigma^2 I_n)^{-1}\underline{\varepsilon}\right\} \\ &= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} (|\sigma^2 I_n|)^{-1} \exp\left\{-\frac{1}{2}\underline{\varepsilon}'(\sigma^2 I_n)^{-1}\underline{\varepsilon}\right\} \end{aligned}$$

Capítulo 3

Diagnóstico y medidas de qué tan adecuado es el modelo

Cuando se ha seleccionado un modelo para una aplicación, como el modelo de regresión lineal simple, usualmente no se tiene la certeza de que el modelo es adecuado para esa aplicación, ya que puede estarse violando una o más hipótesis, como la linealidad del modelo o la normalidad de los errores. Por lo tanto, es importante saber si el modelo es adecuado para los datos antes de que se tengan otros análisis de éste basados en ellos. Ordinariamente el análisis que se lleva a cabo es sobre los residuales; como se ha definido antes, el residual e_i es la diferencia entre el valor observado y el valor estimado o ajustado:

$$e_i = Y_i - \hat{Y}_i;\ i = 1,\ 2,\ ...n$$

El análisis de residuales se desarrolla para verificar si se viola una o más de las hipótesis principales que se postulan para el desarrollo de la teoría del análisis de regresión.

Las violaciones de las hipótesis que son estudiadas por medio de los residuales son:

1. La función de regresión no es lineal
2. Los errores no tienen varianza constante
3. Los errores no son independientes

4. El modelo se ajusta bien a la mayoría de las observaciones excepto por una o unas pocas observaciones discordantes (*outliers*).

5. Los errores no se distribuyen normales

6. Una o más variables independientes importantes han sido omitidas del modelo.

Cuando se hace un ajuste, hay que considerar siempre que la validez de las suposiciones es dudosa, por lo que se tendría que examinar entonces si el modelo es adecuado. Las violaciones graves de las suposiciones pueden producir un modelo inestable en el sentido de que una muestra diferente podría llevar a un modelo totalmente diferente con conclusiones opuestas. Usualmente no se pueden detectar separaciones de las hipótesis subyacentes examinando las estadísticas estándares, tales como F ó R^2 . Las hipótesis son propiedades globales del modelo y como tales no garantizan que éste sea adecuado.

Vistos los residuales como errores observados, es natural que cualquier alejamiento de las hipótesis subyacentes sobre los errores debe reflejarse en ellos. Vistos como errores de ajuste, éstos proporcionan una medida de la variabilidad no explicada por el modelo de regresión.

Considerando lo anterior, podemos afirmar que si nuestro modelo ajustado es correcto, los residuales deberían mostrar un comportamiento que confirmase las suposiciones hechas o, al menos, no mostrar una tendencia que invalidase alguna de ellas. Así, los residuales, después de ser examinados nos deberían permitir concluir:

a) Alguna de las suposiciones del modelo parece ser violada

b) Ninguna de las suposiciones del modelo parece ser violada

Nótese que b) no significa que las suposiciones sean correctas, simplemente que, a la luz del análisis de los datos, no se tiene ninguna razón para suponer que alguna no lo sea.

Los residuales tienen varias propiedades importantes: Tienen media cero; esto puede deducirse de la propiedad de que

1. $\sum_{i=1}^{n} e_i = 0$
2. $\sum_{i=1}^{n} x_i e_i = 0$
3. $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$

Estas propiedades se siguen directamente de las ecuaciones normales cuando se estima por mínimos cuadrados.

Otra propiedad importante de los residuales es que su varianza estimada es:

$$\frac{\sum_{i=1}^{n} (e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \mathbf{CME} = \hat{\sigma}^2$$

Los residuales no son independientes pero, si se dispone de una muestra suficientemente grande de observaciones, esta carencia de independencia puede ser ignorada. Algunas veces es útil trabajar con los residuales estandarizados, que son definidos como sigue

$$d_i = \frac{e_i}{\sqrt{\mathbf{CME}}}$$

Los residuales estandarizados tienen media cero y varianza muy cercana a uno. Esta última ecuación pondera a los residuales dividiéndolos por su desviación estándar promedio; sin embargo, en algunos conjuntos de datos los residuales pueden tener desviaciones estándar que difieran significativamente. En el modelo lineal simple:

$$\begin{aligned} Var(e_i) &= Var\left(Y_i - \hat{Y}_i\right) \\ &= Var\left(Y_i\right) + Var\left(\hat{Y}_i\right) - 2Cov\left(Y_i, \hat{Y}_i\right) \\ &= \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \right) - 2Cov\left(Y_i, \hat{Y}_i\right) \end{aligned}$$

donde

$$Cov\left(Y_i, \hat{Y}_i\right) = Cov\left(Y_i, \bar{Y} + \frac{S_{xy}}{S_{xx}}\left(X_i - \bar{X}\right)\right) = \sigma^2 \left(\frac{1}{n} + \frac{\left(X_i - \bar{X}\right)^2}{S_{xx}} \right)$$

de lo cual se obtiene

$$Var(e_i) = \sigma^2 \left(1 - \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}\right)\right)$$

A partir de lo anterior se define a los residuales studentizados como sigue:

$$r_i = \frac{e_i}{\sqrt{\text{CME}\left[1 - \left(\frac{1}{n} + \frac{(X_i - \bar{X})}{S_{xx}}\right)\right]}}$$

En muestras pequeñas los residuales studentizados son a menudo más apropiados que los residuales estandarizados, puesto que las diferencias en la varianza de ellos pueden ser mayores; sin embargo, cuando se dispone de una muestra suficientemente grande, la diferencia entre ambos suele ser muy pequeña.

Para ilustrar la conveniencia de examinar cuidadosamente al modelo con el fin de detectar violaciones a las hipótesis fundamentales de éste, se presenta a continuación una serie de cuatro conjuntos diferentes de datos, construidos por Anscomb (1973); cada uno de ellos presenta un comportamiento distinto, pero todos tienen la misma recta ajustada y valores de R^2.

Tabla 1: Datos de Anscomb

Obs	X_1	Y_1	X_2	Y_2	X_3	Y_3	X_4	X_5
1	10	8.04	10	9.14	10	7.46	8	6.58
2	8	6.95	8	8.14	8	6.77	8	5.76
3	13	7.58	13	8.74	13	12.74	8	7.71
4	9	8.81	9	8.77	9	7.11	8	8.84
5	11	8.33	11	9.26	11	7.81	8	8.47
6	14	9.96	14	8.10	14	8.84	8	7.04
7	6	7.24	6	6.13	6	6.08	8	5.25
8	4	4.26	4	3.10	4	5.39	19	12.50
9	12	10.84	12	9.13	12	8.15	8	5.56
10	7	4.82	7	7.26	7	6.42	8	7.91
11	5	5.68	5	4.74	5	5.73	8	6.89

En el caso de cada uno de los conjuntos antes descritos $\hat{\beta}_0 = 3$, $\hat{\beta}_1 = 0.5$, $R^2 = 0.66$. Obsérvense las gráficas de dispersión de ellos que se encuentran en la figura 1.

Un análisis basado exclusivamente en las estadísticas estándares no hubiera sido suficiente para detectar estas diferencias en la tendencia y hubiera conducido a conclusiones incorrectas.

A lo largo de este capítulo se presentan varias formas de examinar los residuales con el fin de detectar violaciones al modelo. Algunas de ellas son gráficas, las cuales además de ser sencillas de llevar a cabo, son sumamente reveladoras en los casos en los que existan violaciones a las hipótesis del modelo.

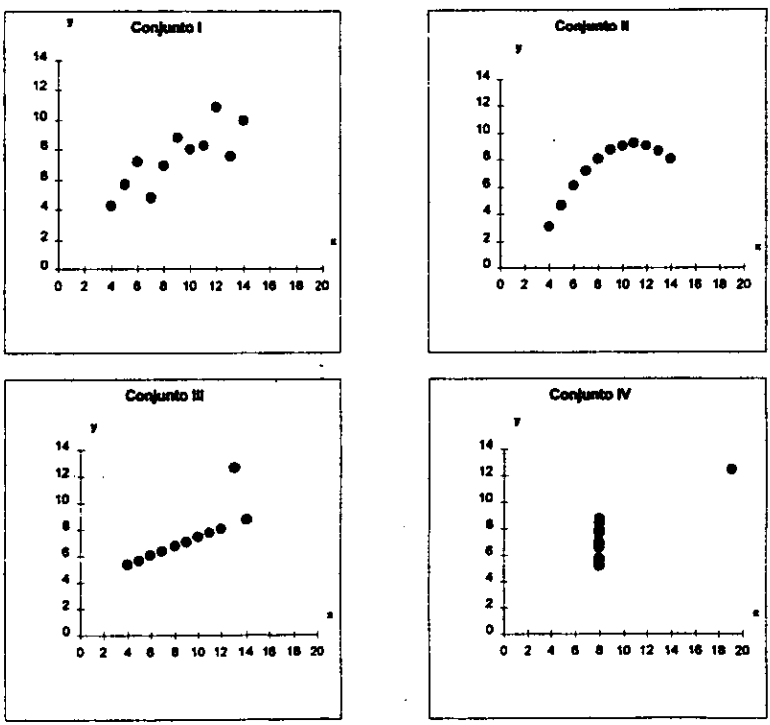


Figura 1: Diagramas de dispersión de los Datos de Anscomb

3.1 Diagnóstico por residuales

Las gráficas de residuales (o residuales estandarizados) que se utilizan para el diagnóstico son:

1. Gráfica de residuales en papel probabilístico normal
2. Gráfica de residuales contra valores estimados o ajustados
3. Gráfica de residuales contra una variable independiente omitida
4. Gráficas de residuales contra las variables regresoras
5. Gráficas de residuales con respecto al tiempo

3.1.1 Gráfica de residuales en papel probabilístico normal

Los residuales pueden producir información concerniente a la validez de la hipótesis de normalidad de los errores. Cualquier lector que desee estudiar los detalles del proceso de diagnóstico, debe revisar primero porqué es importante detectar desviaciones de la normalidad.

Aunque pequeñas desviaciones de la normalidad no suelen afectar demasiado al modelo, un comportamiento muy distinto del normal es más serio, puesto que las pruebas de hipótesis y los intervalos de confianza dependen de esta suposición. Además, si los errores provienen de una distribución cuya cola sea más o menos pesada que la de la normal, el ajuste puede mostrarse muy sensible a un pequeño grupo de observaciones. Los errores con distribuciones de cola pesada frecuentemente generan observaciones discordantes que arrastran a la regresión.

La eficiencia de los criterios de ajuste del modelo, tales como las estadísticas $\hat{\sigma}^2$ y R^2 se mantiene ya sea que los errores se distribuyan normalmente o no; asimismo, tampoco se requiere de esta hipótesis para estimar a los parámetros por mínimos cuadrados. Sin

embargo, los estimadores por mínimos cuadrados funcionan de un modo más eficiente bajo condiciones de independencia, normalidad y varianza constante.

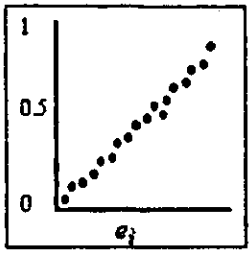
De lo anterior, un analista debe concluir que los estimadores por mínimos cuadrados usados en condiciones distintas a la de normalidad no son los mejores y pueden ser mejorados. La metodología que aquí se presenta para detectar desviaciones de la normalidad se basa en información directa relativa a las estadísticas de orden.

Sean $z_1, ..., z_n$ observaciones independientes de una distribución normal, con media μ y varianza σ^2. Considere al conjunto ordenado:

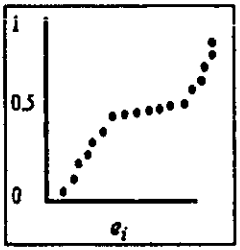
$$z_{(1)} \leq z_{(2)} \leq ... \leq z_{(n)}$$

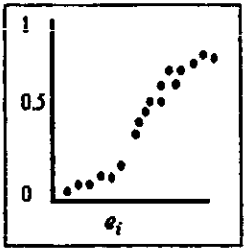
Cada una de estas $z_{(i)}$ tiene su propia distribución, con diferente μ y σ^2. Si $\mu_{(i)}$ es la media de la i-ésima variable aleatoria ordenada con media 0 y varianza 1, entonces $E\left(z_{(i)}\right) = \mu + \sigma\mu_{(i)}$. De esta manera, si las $z's$ son verdaderamente normales, la regresión de $z_{(i)}$ en $\mu_{(i)}$ debe ser una línea recta. Las $\mu_{(i)}$ son llamadas *rankits*. Idealmente, se desearía usar la ecuación anterior para graficar los residuales ordenados contra μ_i. Esta clase de gráficas puede elaborarse manualmente en papel probabilístico normal u obtenerse muy fácilmente con ayuda de casi cualquier paquete estadístico.

Un analista debe considerar que el tratamiento dado a los residuales en las gráficas descritas anteriormente no es estrictamente válido. Como ya hemos mencionado, los residuales no son independientes y, más aun, en general no tienen varianza común. Una metodología más razonable es usar los residuales studentizados o bien, estandarizados. En ambos casos, una gráfica ideal es una línea recta con pendiente igual a uno e intercepción igual a cero (figura 2a). Además, si los errores se distribuyen normalmente, entonces aproximadamente un 68% de los residuales estandarizados deben caer entre -1 y 1 y aproximadamente el 95% de ellos deberían caer entre -2 y 2. Una desviación importante de este promedio debe ser considerada una violación a la suposición de normalidad. Si n es pequeña, debemos reemplazar los límites ± 1 y ± 2 por los valores correspondientes en una distribución t_{n-2}. Esta clase de examen a los residuales también nos ayuda a detectar observaciones discordantes.

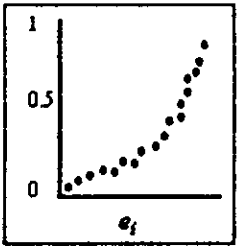


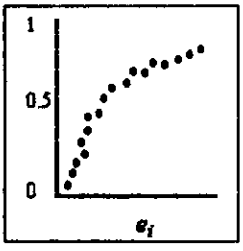
(a) ideal





(b) distribución de "cola pesada" *(c) distribución de "cola ligera"*





(d) inclinación positiva *(e) inclinación negativa*

Figura 2: Gráficas en papel normal

Podemos describir algunas formas que suelen tomar estas gráficas y que revelan ciertas características de la distribución de los errores (fig. 2b a 2e).

La figura 2b muestra una curva pronunciada en ambos extremos, lo cual indica que las colas de esta distribución son muy pesadas para ser considerada normal. En cambio, la figura 2c muestra un aplanamiento en los extremos, lo cual es un signo típico de muestras cuya distribución posee colas más ligeras que las de una normal. La figuras 2d y 2e

muéstran comportamientos asociados a *inclinación* positiva y negativa, respectivamente.

Como las muestras tomadas de una distribución normal no forman líneas rectas exactamente, es necesario adquirir cierta experiencia para interpretar adecuadamente nuestra gráfica. Algunos autores proporcionan esta clase de gráficas. Otra alternativa es tomar n observaciones aleatorias normalmente distribuidas (donde n es el número de observaciones en nuestro análisis), procedentes de alguna tabla previamente construida y graficarlas. Esto puede ayudarnos a adquirir una idea bastante clara de qué desviación de la línea recta es aceptable. Con el fin de evitar llegar a conclusiones incorrectas el analista debe mostrarse muy cauto al examinar los residuales. Probablemente la mayor dificultad estriba precisamente en la lectura e interpretación de las gráficas en papel normal puesto que éstas involucran muchas dificultades.

Una desviación de la línea recta normal puede ser resultado de una mala especificación del modelo más que de errores no distribuidos normales. La validez de las gráficas en papel normal depende de la correcta especificación del modelo. Además, si la muestra es pequeña puede ser extremadamente difícil detectar cualquier desviación de la normalidad, aun cuando ésta exista. Esto es debido a que la estructura de los residuales es tal que un e_i específico no contiene la información requerida acerca de su ε_i correspondiente. Es muy sencillo ilustrar que e_i, el i-ésimo residual, es una función de todos los $\varepsilon_i's$ y no sólo de ε_i. Sea $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ ($\mathbf{H}$ es conocida como la matriz sombrero, *hat*. Algunas propiedades sumamente útiles de ella son descritas en el capítulo referente al modelo múltiple). Considere el vector de residuales

$$\begin{aligned} \underline{Y} - \mathbf{X}\hat{\underline{\beta}} &= (\mathbf{I} - \mathbf{H})\underline{Y} \\ &= (\mathbf{I} - \mathbf{H})\left(\mathbf{X}\underline{\beta} + \underline{\varepsilon}\right) \\ &= \underline{\varepsilon} + \mathbf{X}\underline{\beta} - \mathbf{H}\mathbf{X}\underline{\beta} - \mathbf{H}\underline{\varepsilon} \\ &= \underline{\varepsilon} - \mathbf{H}\underline{\varepsilon} \end{aligned}$$

Sea h_{ij} el (i,j)-ésimo elemento de la matriz $\mathbf{H}$, e_i puede escribirse del siguiente modo:

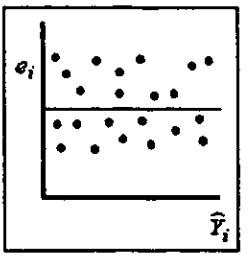
$$e_i = \varepsilon_i - \sum_{j=1}^{n} h_{ij}\varepsilon_j$$

De esta ecuación se concluye que un residual específico, e_i, no es sólo función de ε_i sino una combinación lineal de todos los errores. En muestras pequeñas h_{ij}, $i \neq j$, no es insignificante. Cuando n crece e_i es arrastrado por el error del modelo ε_i. En el caso de tamaños moderados y pequeños de muestras, existe una tendencia del lado derecho de la ecuación anterior a tener una distribución cercana a la normal, aun cuando ε_i no sea normal. Usualmente las muestras pequeñas ($n \leq 16$) producen gráficas que se desvían mucho de la línea recta. Para muestras grandes ($n \geq 20$), las gráficas tienen un comportamiento más aceptable. Por lo anterior, no se debe ser demasiado optimista en relación con el éxito que se tenga al detectar desviaciones de la normalidad en el caso de muestras pequeñas; en general se necesitan cerca de 20 observaciones para obtener gráficas que se puedan interpretar adecuadamente.

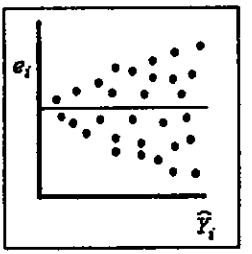
Cuando el número de residuales es demasiado grande, se deben tomar para la gráfica en papel normal sólo las observaciones más pequeñas. Por ejemplo, si se tienen 200 observaciones, podríamos graficar la 10^a menor, la 20^a, etc., hasta llegar a la 180^a y luego graficar las observaciones cuyo comportamiento las revista de interés.

3.1.2 Gráficas de residuales contra los valores ajustados $\hat{Y}$

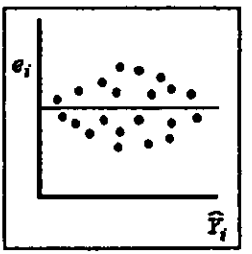
Una gráfica de los residuales contra los valores ajustados $\hat{Y}$ es útil para detectar varios tipos de anomalías en el modelo. Si esta clase de gráficas se parecen a la figura 3a, entonces no existen defectos obvios en el modelo. Si las gráficas son parecidas a las figuras 3b y 3c, entonces concluimos que la varianza de los errores no es constante. El patrón de embudo de la figura 3b indica que la varianza es una función creciente de Y. El patrón de doble arco en la figura 3c ocurre cuando Y es una proporción entre cero y 1. Una gráfica curvada como la de la figura 3d, indica no linealidad. Esto puede significar que se necesitan variables adicionales en el modelo, tales como un término cuadrático o uno mixto. Es posible también que algunas transformaciones en las variables del modelo sean requeridas.



(a) satisfactorio



(b) patrón de "embudo" varianza no constante



(c) patrón de doble arco varianza no constante

(d) gráfica curvada no - linealidad

Figura 3: Gráficas de residuales

Las gráficas de residuales contra el valor ajustado $\hat{Y}$ pueden revelar también uno o más residuales demasiado grandes. Estos puntos son, por supuesto, posibles observaciones discordantes. Residuales grandes que ocurran en los valores extremos de $\hat{Y}$ pueden también indicar varianza no constante o que la verdadera relación entre Y y X no es lineal. Estas posibilidades deben estudiarse cuidadosamente antes de considerar tales puntos como observaciones discordantes.

3.1.3 Gráfica de residuales contra una variable independiente omitida

Graficar los residuales contra una variable independiente omitida, también puede revelar qué tan adecuado es el modelo. Por supuesto, una gráfica de esta clase sólo puede

construirse si los niveles de la variable regresora omitida son conocidos. Cualquier patrón sistemático que muestre esta gráfica indica que el modelo puede ser mejorado añadiendo la nueva variable.

3.1.4 Gráficas contra las variables regresoras

Graficar los residuales contra las variables regresoras es también útil. Estas gráficas suelen exhibir comportamientos similares al de las gráficas de residuales contra los valores ajustados. En este caso, las anomalías ilustradas en la figura pueden indicar

1. Varianza no constante (fig. 3b)
2. Efecto lineal de alguna de las variables regresoras (fig. 3c)
3. Se requieren términos extra en el modelo, por ejemplo, un término cuadrático en X_j o una transformación sobre las Y_i (fig. 3d)

3.1.5 Gráficas de residuales con respecto al tiempo

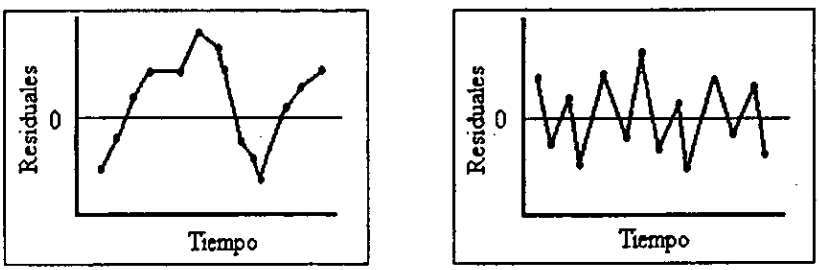
Supóngase que los residuales de nuestro fenómeno ocurren en un orden determinado e igualmente distribuidas en el tiempo. Los residuales así graficados deberían dar la apariencia de una banda horizontal en torno a t, lo cual indicaría que cuando el tiempo pasa, no hay ningún efecto sobre las observaciones. Si los residuales muestran alguna de las tendencias descritas en la figura 3, podríamos concluir que:

1. La varianza no es constante, sino que crece con respecto al tiempo (fig. 3b)
2. Un término lineal debería haber sido incluido en el modelo (fig. 3c)
3. Un término lineal y uno cuadrático deberían haber sido incluidos en el modelo (fig. 3d)

Por supuesto que pueden ocurrir combinaciones y variaciones de estas tendencias.

Si las observaciones no ocurrieran a intervalos iguales de tiempo y se conocieran los verdaderos, se deben graficar de acuerdo a ellos.

Cuando se efectúa este tipo de análisis, la secuencia en la cual se grafican los residuales podría indicar que los errores en un momento dado están correlacionados con otros de períodos distintos. La correlación entre los errores del modelo en diferentes periodos de tiempo recibe el nombre de autocorrelación; estas tendencias se ilustran en la figura 4. Un aspecto como el de 4a indica autocorrelación positiva, en tanto que el de la figura 4b es un comportamiento típico de autocorrelación negativa. La presencia de autocorrelación es una seria violación a las suposiciones básicas el modelo.



(a) autocorrelación positiva *(b) autocorrelación negativa*

Figura 4: Gráficas de residuales vs. tiempo mostrando autocorrelación

3.1.6 Efecto de las observaciones discordantes en el modelo

Las observaciones discordantes o aberrantes, también conocidas por su nombre en inglés, *outliers*, son observaciones con residuales muy grandes, es decir, residuales muy grandes comparados con los del resto de las observaciones. Residuales que sean considerablemente mayores en valor absoluto que los otros, digamos 3 ó 4 veces la desviación estándar a partir de la media, deben ser considerados provenientes de posibles outliers. Estos puntos parecen no pertenecer al resto de los datos y dependiendo de sus características pueden tener influencias moderadas o severas en la regresión; las gráficas en papel normal y las

de residuales contra X y $\hat{Y}$ son útiles para detectar este tipo de puntos.

Estos residuales deben ser cuidadosamente examinados con el fin de conocer la razón de su comportamiento extremo. Algunos de ellos pueden ser resultado de errores en la transcripción o la obtención de los datos, incluyendo posibles fallas durante el análisis, registro incorrecto de los datos o uso de un instrumento de medición impreciso o defectuoso. Si éste es el caso, entonces tales datos deberían ser corregidos -siempre que sea posible- o bien ser removidos del modelo. Sin embargo, podrían ser también observaciones genuinas, altamente significativas y sugestivas, por lo cual merecen una atención especial. Esto nos lleva a remarcar el hecho de que se deben tener evidencias importantes y firmes, obviamente no estadísticas, que nos aseguren que los outliers son procedentes de errores para proceder a descartarlos.

Cuando los outliers son observaciones plausibles, descartarlos del modelo puede ser peligroso, puesto que aun cuando esto mejore el ajuste de él, puede restarle precisión en la estimación y la predicción. En otros casos encontramos que los outliers son incluso más importantes para la ecuación de regresión que el resto de los datos, influyendo seriamente en la estimación de los parámetros.

Las observaciones discordantes pueden ayudarnos también a detectar qué tan adecuado es el modelo, por ejemplo, en su ajuste a la hipótesis lineal.

El efecto de los outliers en la regresión puede examinarse fácilmente reestimando los parámetros del modelo, habiendo removido previamente las observaciones discordantes y observando los efectos que ello tenga en él. Es también de sumo valor intentar entender las circunstancias que pudieron haber generado tan grandes residuales.

Los parámetros y las estadísticas principales son sumamente sensibles ante las observaciones discordantes; por lo tanto, cuando están presentes valores extremos en las observaciones, es posible que sólo una parte muy pequeña de los datos tenga un efecto significativo sobre la ecuación de regresión. Un investigador debe entonces evaluar si este desequilibrio es o no aceptable. La importancia de algún subconjunto de datos en particular puede ser evaluada empíricamente calculando el modelo ajustado incluyendo

y sin incluir a éste. Se considera que el modelo es adecuado si, después de remover un pequeño grupo de observaciones (cualquier grupo), la relación parece aun apropiada; esto es, si la regresión se permea de todas las observaciones y no depende sustancialmente sólo de algunas de ellas.

Esta clase de análisis, basada en los residuales y en la posible supresión de algunas observaciones, suele llevarnos a conclusiones muy distintas de las que se obtienen sólo con las estadísticas básicas.

Ejemplo 1.- El éxito de un programa en la televisión comercial está determinado en parte por un sistema de nivel de audiencia (o, por su nombre en inglés, *rating*), que es un intento de medir la habilidad del programa para atraer y mantener televidentes. En términos reales, el rating genera interés en los patrocinadores y por consiguiente, ganancias a la estación. El director de una cierta estación televisiva, preocupado acerca del rating de un noticiario, desea realizar un estudio para identificar los factores que lo afectan. Adicionalmente a los factores obvios tales como formato, secciones especiales y atractivo personal de los conductores, se ha sugerido un efecto remanente del programa que lo precede. Con el fin de cuantificar dicho efecto, se tomó una muestra aleatoria de niveles previos de audiencia en varias regiones y pertenecientes a diferentes fechas de los últimos dos años. Los datos son parejas ordenadas (X, Y), donde Y denota el rating del noticiario y X el del programa que lo precede. Dichos datos se reportan en la tabla 2.

El primer paso que se sigue en el análisis es graficar los datos, Y *vs.* X. Dicha gráfica se muestra en la figura 5. Al examinarla observamos una tendencia de los datos que sugiere que es razonable proponer un modelo lineal para representarlos; sin embargo, muestra también cuatro observaciones que se alejan considerablemente de los datos y despierta la sospecha de que pudiera tratarse de posibles outliers.

Al realizar los cálculos correspondientes se llega a la siguiente recta ajustada:

$$\hat{Y} = 1.706 + 0.665X$$

Tabla 2: Datos de audiencia televisiva

Obs.	*X*	*Y*	*Obs.*	*X*	*Y*	*Obs.*	*X*	*Y*
1	2.5	3.80	11	4.50	5.80	21	6.50	7.00
2	2.7	4.10	12	4.70	3.80	22	6.70	5.40
3	2.9	5.80	13	4.90	4.75	23	6.90	6.10
4	3.10	4.80	14	5.10	3.90	24	7.10	6.50
5	3.30	5.70	15	5.30	6.20	25	7.30	6.10
6	3.50	4.40	16	5.50	4.35	26	7.50	4.75
7	3.70	4.80	17	5.70	4.15	27	2.50	1.00
8	3.90	3.60	18	5.90	4.85	28	2.70	1.20
9	4.10	5.50	19	6.10	6.20	29	7.30	9.50
10	4.30	4.15	20	6.30	3.80	30	7.50	9.00

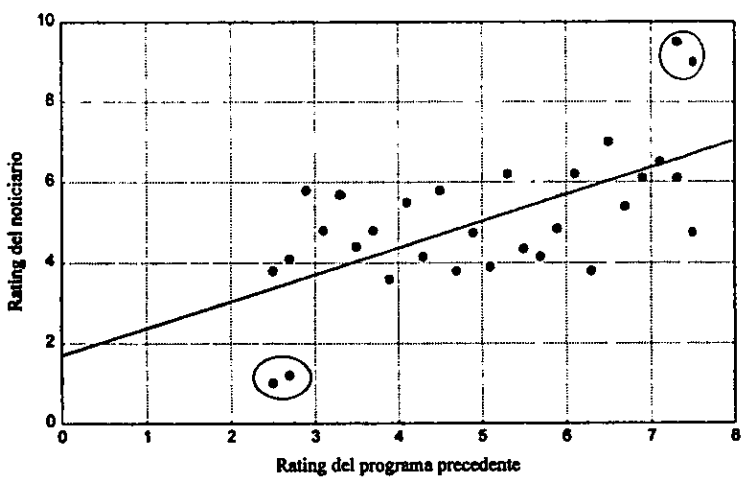


Figura 5: Gráfica de dispersión con la recta ajustada (datos de audiencia televisiva)

donde $R^2 = 0.396$ y $\hat{\sigma} = 1.402$. Obsérvese que el valor de la pendiente es muy cercano a cero; esta situación nos sorprende puesto que el director de la estación aseguró que aproximadamente el 40% de la variación en el rating del noticiario puede ser explicado

por el rating del programa que lo precede, mientras que nosotros hallamos que por cada punto en el rating que aumente el programa precedente al noticiario, el rating de éste aumentará sólo 0.66 de punto; asimismo, notamos que el coeficiente de determinación es muy pequeño. Al graficar los residuales estandarizados contra X y contra $\hat{Y}$ (Figura 6 y Figura 7, respectivamente) volvemos a encontrar estas cuatro observaciones; en el resto de los datos no parece haber evidencia de que alguna de las hipótesis del modelo esté siendo seriamente violada, es decir, muestran un comportamiento bastante aceptable; sin embargo, en la gráfica de residuales contra los valores ajustados notamos una cierta tendencia de los datos a tener residuales grandes para valores bajos de $\hat{Y}$ y viceversa.

Las 4 observaciones arriba mencionadas muestran una tendencia exactamente opuesta. Esto nos hace sospechar que estas cuatro observaciones no pertenecen al resto de los datos. Podemos incluso pensar que estos datos inflan a la pendiente de la recta ajustada.

Por lo anterior, estos puntos deben ser considerados outliers y ser objeto de un análisis muy cuidadoso. Para conocer el efecto que tienen éstos sobre el modelo, procedemos a calcular los estimadores de los parámetros y las estadísticas principales excluyéndolos del conjunto de los datos; la recta ajustada a la que se llega es:

$$\hat{Y} = 3.713 + 0.260X$$

con $R^2 = 0.161$ y $\hat{\sigma} = 0.925$. Los resultados obtenidos confirman nuestra sospecha de que los outliers arrastran significativamente a la regresión y que el ajuste de los datos al modelo no mejora, pues el coeficiente de determinación disminuye a casi un tercio; es decir, los resultados obtenidos usando solamente los datos que se comportan de manera aceptable nos llevan a pensar que el rating del programa precedente al noticiario tiene un efecto insignificante sobre el rating de éste.

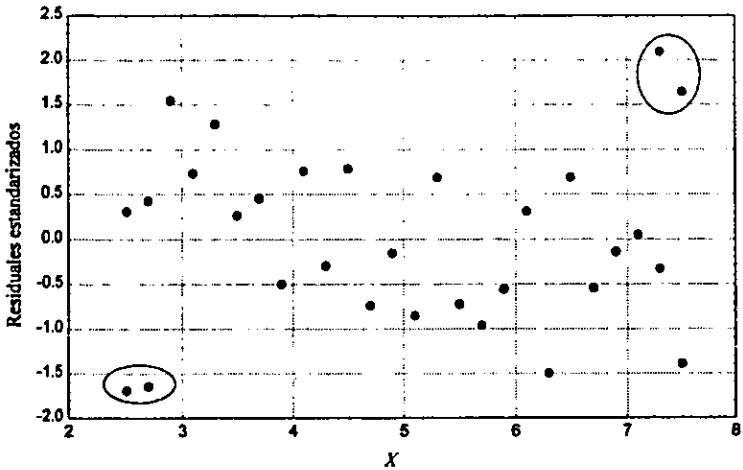


Figura 6: Gráfica de residuales estandarizados vs. X (datos de audiencia televisiva)

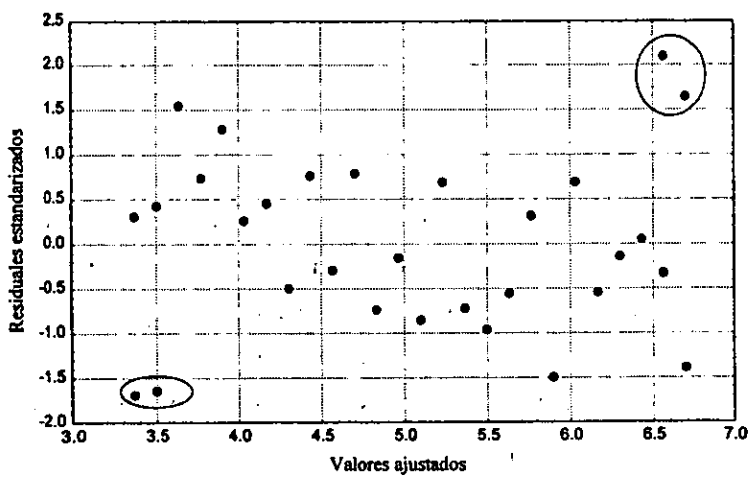


Figura 7: Gráfica de residuales estandarizados vs. los valores ajustados (datos de audiencia televisiva)

Ejemplo 2.- La tabla 3 muestra los residuales estandarizados y studentizados para el ejemplo 4 del Capítulo 1 (propulsores de un cohete). Fueron obtenidos usando las siguientes fórmulas

$$d_i = \frac{e_i}{\sqrt{\hat{\sigma}^2}} = \frac{e_i}{\sqrt{9244.59}} = \frac{e_i}{96.15}$$

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2\left(1 - \left(\frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{xx}}\right)\right)}} = \frac{e_i}{\sqrt{9244.59\left(1 - \left(\frac{1}{20} + \frac{(X_i - 13.3625)^2}{1106.56}\right)\right)}}$$

$i = 1, 2, ..., 20$

Mediante una inspección a la tabla notamos que las observaciones 5 y 6 tienen residuales muy grandes. Obsérvese la gráfica de los residuales en papel normal (Figura 8); ésta refuerza la sospecha de que estos datos son posibles outliers.

Tabla 3: Residuales estandarizados y studentizados

(datos de los propulsores de un cohete)

Obs.	e_i	d_i	r_i	*Obs.*	e_i	d_i	r_i
1	106.76	1.11	1.14	11	20.37	0.21	0.22
2	−67.27	−0.70	−0.76	12	−88.95	−0.93	−0.99
3	−14.59	−0.15	−0.16	13	80.82	0.84	0.92
4	65.09	0.68	0.70	14	71.17	0.74	0.76
5	−215.98	−2.25	−2.38	15	−45.15	−0.47	−0.50
6	−213.60	−2.22	−2.32	16	94.44	0.98	1.02
7	48.56	0.51	0.55	17	9.50	0.10	0.10
8	40.06	0.42	0.45	18	37.10	0.39	0.40
9	8.73	0.09	0.09	19	100.68	1.05	1.15
10	37.57	0.39	0.40	20	−75.32	−0.78	−0.83

Al examinar las gráficas de residuales contra los valores ajustados $\hat{Y}$ y contra la variable regresora X (Figuras 9 y 10, respectivamente) volvemos a notar que estos datos se diferencian de los demás. Aparte de ellos, no existe evidencia alguna en las gráficas que nos haga pensar que alguna de las hipótesis del modelo está siendo seriamente violada.

Note también que la observación número 5 ocurre en un valor relativamente bajo de X, mientras que la observación 6 pertenece a un valor alto de ella; es decir, ocurren cada una en un extremo de X y parecen estar separadas del resto de los datos.

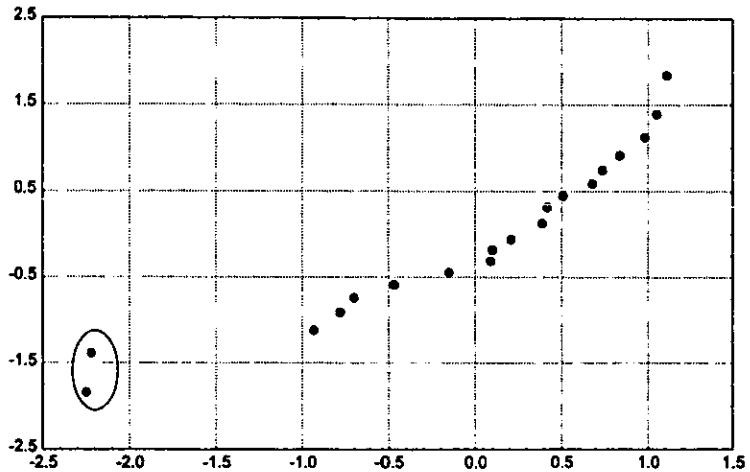


Figura 8: Gráfica en papel normal (datos de los propulsores de un cohete)

Esto nos lleva a sospechar que la información proveniente de estos dos datos influye significativamente en las propiedades del modelo. Para investigar qué tanta influencia tienen estos datos sobre el modelo se calculan la recta ajustada, el coeficiente de determinación y el cuadrado medio del error, sin incluirlos en el análisis. La recta ajustada obtenida usando el conjunto completo es $\hat{Y} = 2627.82 - 37.15X$ con $R^2 = 0.9018$ y $\mathbf{CME} = 9244.59$; por otro lado, los resultados obtenidos al descartar las observaciones 5 y 6 son $\hat{Y} = 2658.97 - 37.69X$ con $R^2 = 0.9578$ y $\mathbf{CME} = 3964.63$.

Dado que las estimaciones de los parámetros no cambian sustancialmente al descartar las observaciones 5 y 6, podemos concluir que éstas no tienen demasiada influencia en el ajuste del modelo.

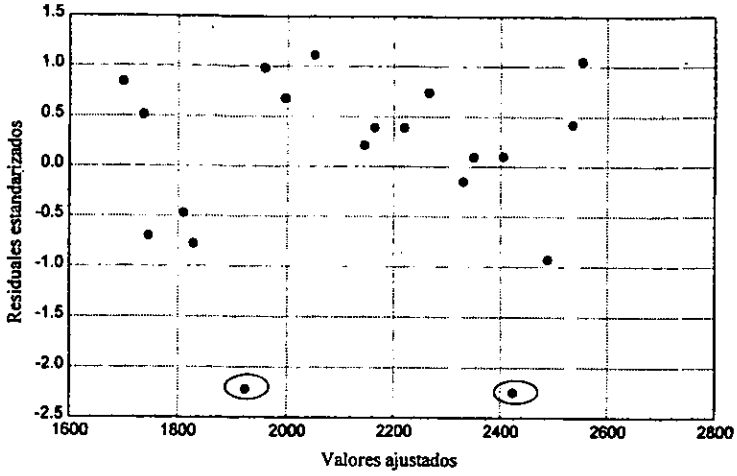


Figura 9: Gráfica de residuales estandarizados vs. los valores ajustados (datos de los propulsores de un cohete)

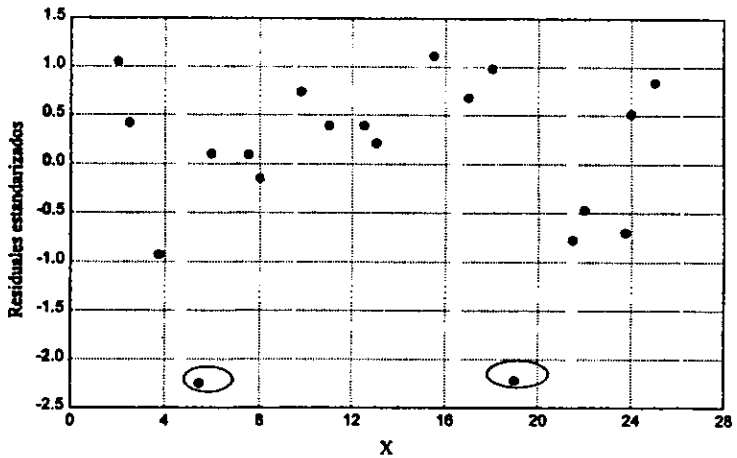


Figura 10: Gráfica de residuales estandarizados vs. X (datos de los propulsores de un cohete)

Sin embargo, podemos observar que el cuadrado medio del error disminuyó a casi un tercio de su valor en el modelo original y que R^2 creció ligeramente. Por lo anterior, se decide no descartar ninguna de las observaciones, en la confianza relativa de que incluirlas no afecta seriamente las propiedades del modelo.

3.1.7 Transformación de variables

Para satisfacer las suposiciones del modelo de regresión, algunas veces es necesario trabajar con variables transformadas, en lugar de hacerlo con las variables originales. Las transformaciones pueden deberse a muchas razones, las cuales pueden ser resumidas como sigue:

a) Consideraciones teóricas pueden especificar previamente que la relación entre las variables no es lineal, por ejemplo, en el caso de modelos de crecimiento poblacional

b) La variable dependiente Y puede tener una distribución de probabilidad cuya varianza dependa de su esperanza. De este modo la varianza de Y cambiará cuando X cambie, esto es, la varianza no será constante. La distribución de Y usualmente no es normal bajo estas condiciones, lo cual invalida las pruebas de significancia (aunque esto no ocurre en una magnitud muy importante cuando se trata de muestras grandes), puesto que éstas están basadas en la hipótesis de normalidad. La carencia de constancia en el error puede producir estimadores insesgados, pero no precisos. En la práctica, las transformaciones son escogidas para asegurar la constancia de la varianza, éstas son llamadas transformaciones estabilizadoras de la varianza; por coincidencia, estas transformaciones son también útiles en la normalización de las variables.

c) Al examinar los residuales se sospecha que se requiere una transformación.

Transformaciones para corregir no-linealidad detectada en las gráficas de residuales

Una de las suposiciones básicas del análisis de regresión es que el modelo que describe a los datos es lineal. Por consideraciones teóricas previas, o bien, después de examinar un diagrama de dispersión de Y contra X, podemos llegar a la conclusión de que el modelo no es lineal. Sin embargo, es posible que el modelo apropiado sea linealizable y que mediante una transformación adecuada se llegue a un modelo lineal. En la tabla 4 se listan algunas de ellas, con la gráfica correspondiente en la figura 11.

Tabla 4: Funciones Linealizables con las transformaciones correspondientes

Función	*Transformación*	*Forma lineal*	*Figura*
$Y = \beta_0 X^{\beta_1}$	$Y' = \log Y,\ X' = \log X$	$Y' = \log \beta_0 + \beta_1 X$	11*a*, *b*
$Y = \beta_0 e^{\beta_1 X}$	$Y' = \ln Y$	$Y' = \ln \beta_0 + \beta_1 X$	11*c*, *d*
$Y = \beta_0 + \beta_1 \log X$	$X' = \log X$	$Y = \beta_0 + \beta_1 X'$	11*e*, *f*
$Y = \dfrac{X}{\beta_0 X - \beta_1}$	$Y' = \dfrac{1}{y},\ X' = \dfrac{1}{X}$	$Y' = \beta_0 - \beta_1 X'$	11*g*, *h*
$Y = \dfrac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$	$Y' = \ln\left(\dfrac{Y}{1-Y}\right)$	$Y' = \beta_0 + \beta_1 X$	11*i*

Ejemplo 3.-Los datos que aparecen en la tabla 5 representan el número de bacterias sobrevivientes (en unidades de 100) en conteos por recipiente (cajas de Petri) de un experimento con bacterias marinas expuestas a radiación por rayos X de 200 kilovolts, durante un número de periodos que fluctúa entre 1 y 15, cada uno formado por intervalos de 6 minutos. Las bacterias bajo estudio no forman cadenas o grupos, por lo cual el número de ellas puede ser estimado directamente a partir de la superficie que ocupan en cada recipiente. Este experimento fue llevado a cabo para examinar la acción de los rayos X bajo un campo constante de radiación.

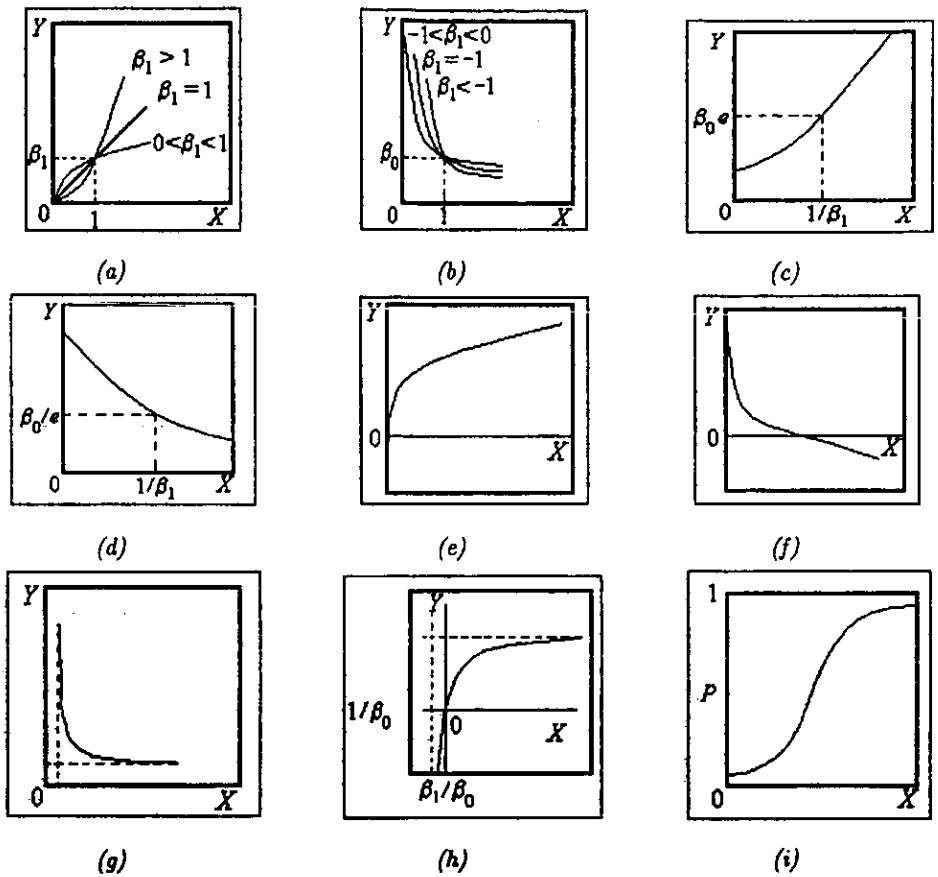


Figura 11: Gráfica de funciones linealizables

Si la teoría es aplicable, el logaritmo del número de sobrevivientes debe aparecer como una línea recta si se grafica contra t. Si n_t representa el número de bacterias sobrevivientes después de un tiempo t de exposición a la radiación se tiene

$$n_t = n_0 e^{\beta t}$$

donde n_0 y β son parámetros que deben ser estimados. Los parámetros n_0 y β tienen interpretaciones físicas muy simples: n_0 es el número de bacterias al inicio del experimento

y β es la tasa de muerte de bacterias.

Tabla 5: Número de bacterias sobrevivientes
después del tiempo indicado de exposición a rayos X

Número de bacterias	*Tiempo*	*Número de bacterias*	*Tiempo*
355	1	56	9
211	2	38	10
197	3	36	11
166	4	32	12
142	5	21	13
106	6	19	14
104	7	15	15
60	8		

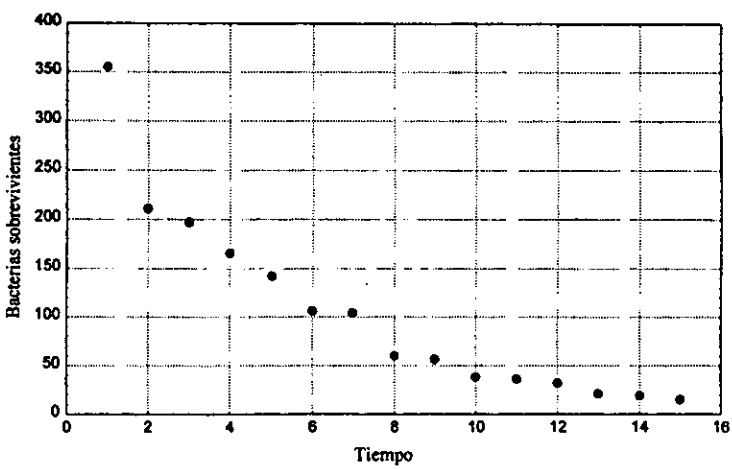


Figura 12: Gráfica de Y con respecto al tiempo (ejemplo 3)

Si introducimos el error aleatorio ε_t se tiene

$$\ln n_t = \beta_0 + \beta_1 t + \varepsilon_t$$

y podemos ahora aplicar el método de mínimos cuadrados. Con el fin de que ε_t, el término de error, sea aditivo en el modelo transformado, éste debe aparecer como un factor en el modelo original. La representación correcta del modelo debe ser:

$$n_t = n_0 e^{\beta t} \varepsilon_t'$$

donde ε_t' es el error aleatorio multiplicativo. Obsérvese que $\varepsilon_t = \ln \varepsilon_t'$. En el modelo de regresión lineal simple que hemos analizado se requiere que ε_t esté normalmente distribuido, lo cual implica que ε_t' se distribuya lognormal. En la práctica, después de haber ajustado el modelo transformado se examinan los residuales para saber si la suposición hecha acerca del modelo es razonable.

El primer paso en el análisis consiste en graficar los datos, n_t vs. t. La gráfica (figura 12), sugiere que la verdadera relación entre las variables no es lineal. Sin embargo, ajustemos el modelo lineal simple y analicemos las consecuencias que ello implica. El modelo es:

$$n_t = \beta_0 + \beta_1 t + \varepsilon_t$$

con las hipótesis usuales. La recta ajustada a la cual se llega es:

$$\hat{n}_t = 259.58 - 19.46t$$

donde $R^2 = 0.823$. Nótese que,a pesar de tenerse un valor de R^2 alto, el modelo no es adecuado. Esto resulta evidente al examinar la gráfica de residuales contra la variable independiente (figura 13); en efecto, los residuales para $t = 2, ..., 11$ son negativos, los residuales para $t = 12, .., 15$ son negativos y en $t = 1$ parece existir un outlier.

Dado que la relación entre n_t y t no es lineal, se propone trabajar con la variable transformada $\ln n_t$ (que ya había sido sugerida previamente por consideraciones teóricas).

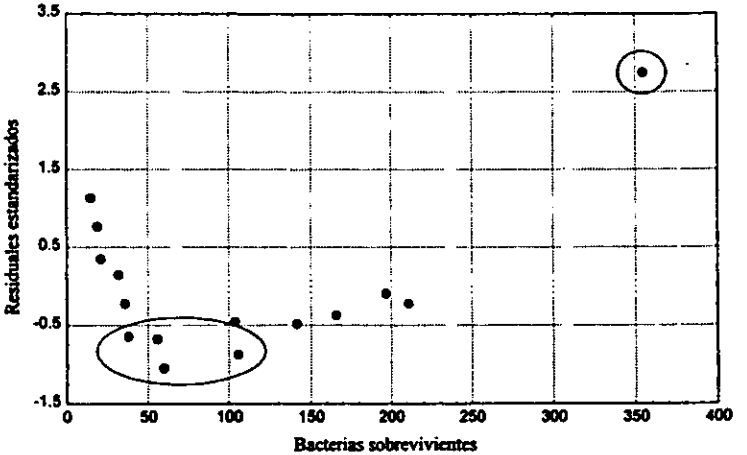


Figura 13: Residuales estandarizados vs. Y (ejemplo 3)

Al examinar la gráfica de $\ln n_t$ vs. t (figura 14), observamos que la transformación aplicada a n_t mejora el ajuste a la hipótesis de linealidad.

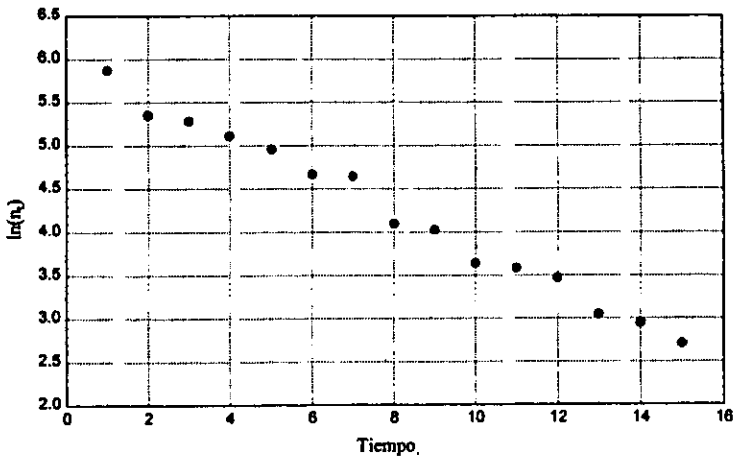


Figura 14: $\ln(n_t)$ *vs. t (ejemplo 3)*

El modelo ajustado transformado es:

$$\hat{n}_t = 5.973 - 0.218t$$

con $R^2 = 0.9884$. La inspección a la gráfica de los residuales (después de aplicar la transformación), contra la variable regresora, en la figura 15 , no revela ningún patrón que nos haga pensar que se viola la hipótesis de linealidad. De este modo, concluimos que el logaritmo del número de bacterias sobrevivientes a la acción de radiación está relacionado linealmente con éste.

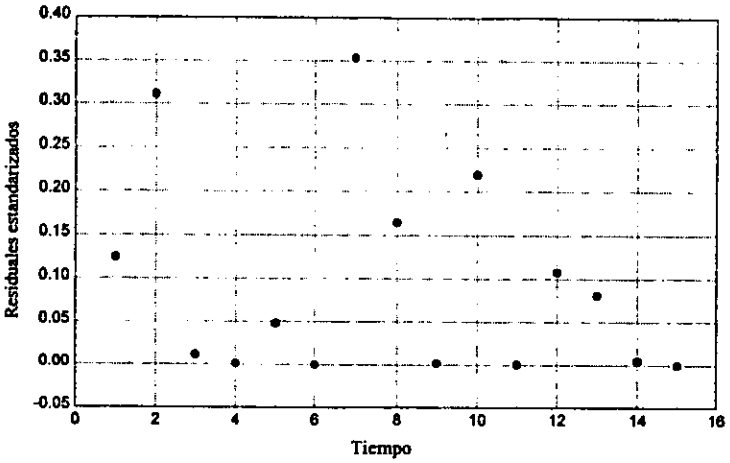


Figura 15: Gráfica de residuales estandarizados vs. t (ejemplo 3)

Transformaciones para estabilizar la varianza

Cuando la varianza del error no es constante, se dice que el error es heterocedástico. La heterocedasticidad puede ser removida a través de una transformación adecuada.

Una causa común de la violación a la hipótesis de varianza constante (que es, como se recordará, una de las principales suposiciones del modelo de regresión) es que la variable dependiente Y siga una distribución de probabilidad cuya varianza dependa de la media. La distribución Binomial y la distribución Poisson son dos funciones de densidad con tales características.

Cuando la relación entre la media y la varianza de una variable aleatoria es conocida, es posible encontrar una transformación simple de la variable, la cual haga a la varianza aproximadamente constante, esto es, que estabilice a la varianza.

Las transformaciones que se describen en la siguiente tabla no sólo estabilizan a la varianza, sino que también tienen la propiedad de acercar considerablemente la distribución de la variable transformada a la distribución normal.

Tabla 6: Transformaciones para estabilizar varianza

Distribución de probabilidad de y	*Relación entre* σ^2 *y* μ	*Transformación*	*Varianza después de la transformación*
Poisson	μ	$\sqrt{Y}$ *ó* $\left(\sqrt{Y}+\sqrt{Y+1}\right)$	0.25
Binomial	$\frac{\mu(1-\mu)}{n}$	$\text{sen}^{-1}\sqrt{Y}$ *(grados)*	$\frac{821}{n}$
		$\text{sen}^{-1}\sqrt{Y}$ *(radianes)*	$\frac{0.25}{n}$
Binomial negativa	$\mu+\lambda^2\mu^2$	$\lambda^{-1}\,\text{sen}\,h^{-1}\left(\lambda\sqrt{Y}\right)$ *ó*	
		$\lambda^{-1}\,\text{sen}\,h^{-1}\left(\lambda\sqrt{Y}+0.5\right)$	0.25

En algunas ocasiones, nos podemos valer de consideraciones teóricas previas o de experiencia previa para elegir una transformación adecuada para estabilizar la varianza. Sin embargo, en muchos casos no se tienen datos previos que nos lleven a suponer que la varianza no es constante y nuestro primer indicio lo proporcionan las gráficas de los residuales, en cuyo caso, una transformación adecuada debe ser elegida empíricamente.

Es sumamente importante detectar y corregir la heterocedasticidad del modelo. Si este problema no es eliminado, los estimadores por mínimos cuadrados serán insesgados pero no serán de varianza mínima, es decir, los coeficientes de la regresión tendrán errores estándar demasiado grandes. El efecto de las transformaciones arriba descritas es dar mayor precisión a los parámetros del modelo e incrementar la sensibilidad de las pruebas estadísticas.

Ejemplo 4.- En un estudio de 27 establecimientos industriales de diferentes tamaños se registró el número de trabajadores supervisados X y el número de supervisores Y.

Tabla 7: Número de trabajadores supervisados X
y de supervisores Y para 27 establecimientos industriales

Obs.	X	Y	*Obs.*	X	Y	*Obs*	X	Y
1	294	30	10	697	78	19	700	106
2	247	32	11	688	80	20	850	128
3	267	37	12	630	84	21	980	130
4	358	44	13	709	88	22	1025	160
5	423	47	14	627	97	23	1021	97
6	311	49	15	615	100	24	1200	180
7	450	56	16	999	109	25	1250	112
8	534	62	17	1022	114	26	1500	210
9	438	68	18	1015	117	27	1650	135

Se decidió estudiar la relación entre estas dos variables y se propuso el modelo $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ como un acercamiento inicial.

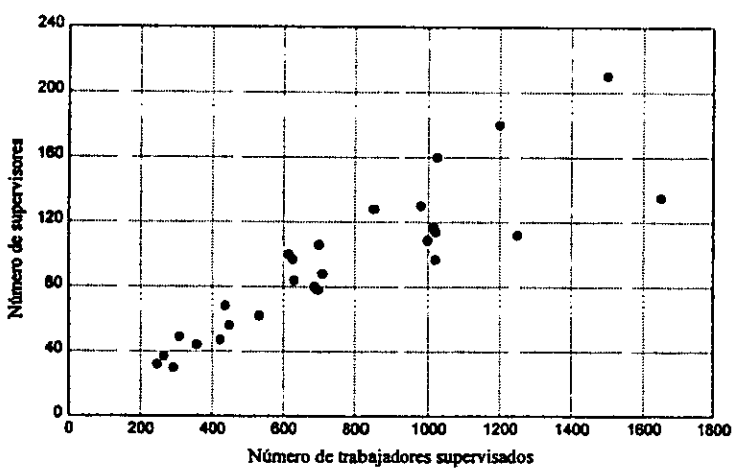


Figura 15: Gráfica de Y vs. X (ejemplo 4)

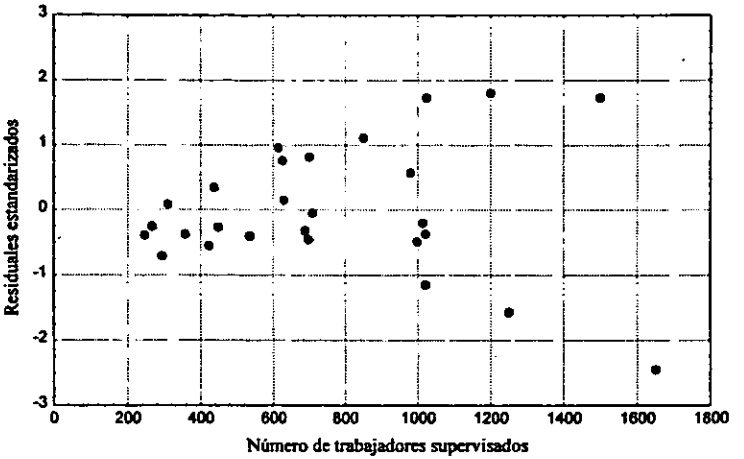


Figura 16: Gráfica de residuales estandarizados vs. X (ejemplo 4)

La recta ajustada a la que se llegó fue la siguiente:

$$\hat{Y} = 14.448 + 0.115X$$

Tanto la gráfica de Y vs. X (figura 15), como la de residuales estandarizados contra la variable regresora (figura 16), muestran que la varianza del error tiende a crecer cuando la variable independiente crece. Los residuales muestran una tendencia a caer en una banda que diverge al moverse en el sentido positivo del eje X. En general, si la banda en la cual caen los residuales diverge (se vuelve más ancha) cuando X crece, entonces la varianza del error crece al crecer X; si la banda converge (tiende a estrecharse), la varianza del error decrece al crecer X.

En muchas aplicaciones industriales, biológicas y económicas, cuando se detecta heterocedasticidad, a menudo se encuentra que la desviación estándar de los residuales tiende a crecer cuando X crece; basándonos en esta observación empírica y en la información que nos proporcionan las gráficas, podemos suponer que en el presente ejemplo la desviación estándar de los residuales es proporcional a X:

$$Var\,(e_i) = k^2 X_i^2,\ k > 0$$

Considérese el modelo

$$\frac{Y_i}{X_i} = \frac{\beta_0}{X_i} + \beta_1 + \frac{\varepsilon_i}{X_i}$$

el cual se obtiene dividiendo ambos lados de la ecuación por X_i. Se tiene entonces un nuevo conjunto de variables y coeficientes:

$$Y' = \frac{Y}{X},\ X' = \frac{1}{X},\ \beta_0' = \beta_1,\ \beta_1' = \beta_0,\ \varepsilon_i' = \frac{\varepsilon_i}{X_i}$$

y el modelo

$$Y_i' = \beta_0' + \beta_1' X_i' + \varepsilon_i'$$

Note que en el modelo transformado la varianza de ε_i' es constante e igual a k^2. Si nuestra suposición $Var\,(\varepsilon_i) = k^2 X_i^2,\ k > 0$ es válida, entonces, para ajustar apropiadamente el modelo, debemos trabajar con las variables transformadas $Y' = \frac{Y}{X}$ y $X' = \frac{1}{X}$. Como las variables dependiente e independiente, respectivamente.

Si el modelo ajustado para los datos transformados es

$$\frac{\hat{Y}}{X} = \hat{\beta}_0' + \frac{\hat{\beta}_1'}{X}$$

el modelo ajustado en términos de las variables originales es

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_0 X$$

La ecuación de regresión calculada para las variables transformadas es:

$$\frac{\hat{Y}}{X} = 0.121 + \frac{3.803}{X}$$

En términos de las variables originales se tiene

$$\hat{Y} = 3.803 + 0.121X$$

Los residuales obtenidos después de ajustar el modelo transformado se hallan graficados en la figura 17. En ella los residuales parecen estar distribuidos aleatoriamente y caer entre dos bandas paralelas al eje $\frac{Y}{X}$, que es la variable dependiente del modelo transformado. La distribución de los residuales no muestra ningún patrón distintivo y podemos concluir que el modelo transformado es adecuado. Nuestra suposición acerca de los errores parece ser correcta; el modelo transformado tiene errores homocedásticos.

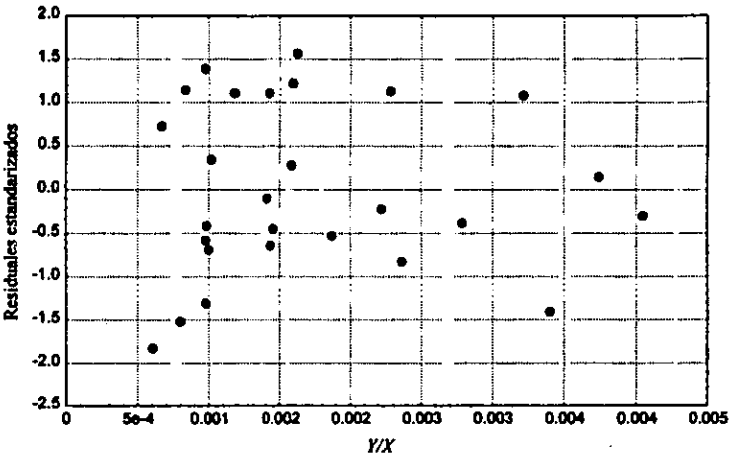


Figura 17: Gráfica de residuales estandarizados vs. Y/X (ejemplo 4)

3.1.8 La prueba de carencia de ajuste

Hemos enfatizado que una línea ajustada mediante una regresión es una línea que se calcula basándose en un cierto modelo o suposición, misma que no debe ser aceptada ciegamente, sino ser considerada tentativamente. En ciertas circunstancias podemos verificar si el modelo es o no correcto. En primer lugar, podemos examinar las consecuencias de un modelo incorrecto. Recuérdese que $e_i = Y_i - \hat{Y}_i$ es el i-ésimo residual, es decir, el residual en X_i. Esta es la cantidad en la cual la verdadera observación Y_i difiere del valor ajustado $\hat{Y}_i$. Como ya hemos visto, los residuales contienen toda la información

disponible en el sentido en el cual el modelo falla para explicar la variación observada en la variable dependiente.

Sea $\eta_i = E(Y_i)$ el valor dado por el verdadero modelo, cualquiera que éste sea, en $X = X_i$. Podemos entonces escribir:

$$\begin{aligned} Y_i - \hat{Y}_i &= \left(Y_i - \hat{Y}_i\right) - E\left(Y_i - \hat{Y}_i\right) + E\left(Y_i - \hat{Y}_i\right) \\ &= \left[\left(Y_i - \hat{Y}_i\right) - \left[\eta_i - E\left(\hat{Y}_i\right)\right]\right] + \left[\eta_i - E\left(\hat{Y}_i\right)\right] \\ &= q_i + B_i \end{aligned}$$

donde

$$q_i = \left(Y_i - \hat{Y}_i\right) - \left[\eta_i - E\left(\hat{Y}_i\right)\right] \text{ y } B_i = \eta_i - E\left(\hat{Y}_i\right)$$

B_i es el sesgo del error en $X = X_i$. Si el modelo es correcto, entonces $\eta_i = E(\hat{Y}_i)$ y B_i es igual a cero. Si el modelo no es correcto, $\eta_i \neq E(\hat{Y}_i)$ y B_i es distinto de cero, pero tiene un valor que depende del modelo verdadero y de X_i. La cantidad q_i es una variable aleatoria con media cero, puesto que:

$$\begin{aligned} E(q_i) &= E\left[\left(Y_i - \hat{Y}_i\right) - \left[\eta_i - E\left(\hat{Y}_i\right)\right]\right] \\ &= E(Y_i) - E\left(\hat{Y}_i\right) - \left[\eta_i - E\left(\hat{Y}_i\right)\right] \\ &= \eta_i - E\left(\hat{Y}_i\right) - \left[\eta_i - E\left(\hat{Y}_i\right)\right] \\ &= 0 \end{aligned}$$

lo cual sucede sea o no correcto el modelo, esto es, si $\eta_i = E(\hat{Y}_i)$ o si $\eta_i \neq E(\hat{Y}_i)$.

Se puede probar que $\sum_{i=1}^{n} q_i$ tiene un valor esperado igual a $(n-2)\,\sigma^2$. A partir de ello se demuestra que el cuadrado medio del error

$$\text{CME} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-2}$$

tiene un valor esperado igual a σ^2 si el modelo es correcto o igual a $\sigma^2 + \dfrac{\sum_{i=1}^{n}\left(\eta_i - E\left(\hat{Y}_i\right)\right)^2}{n-2}$ si no lo es.

Si el modelo es correcto, es decir, si $B_i = 0$, entonces los residuales son desviaciones aleatorias q_i; sin embargo, si el modelo no es correcto, esto es, si $B_i \neq 0$, entonces los residuales contienen componentes aleatorias y sistemáticas. Podemos referirnos a ellas como los componentes de sesgo y varianza del error de los residuales, respectivamente. Asimismo, el **CME** tenderá a estar *inflado* y no proporcionará una medida satisfactoria de la variación aleatoria presente en las observaciones.

En el caso del modelo de regresión simple, el sesgo del error puede ser detectado examinando una gráfica de los datos. Sin embargo, cuando el modelo es más complicado o involucra más variables esto no es posible. Si se dispone de una estimación previa de σ^2, es decir, una estimación obtenida a partir de experiencia previa acerca de la variación del fenómeno estudiado, podemos verificar (o probar mediante una prueba F) si el **CME** es significativamente mayor que esta estimación previa. Si lo es, diremos que existe carencia de ajuste y podremos pensar que la forma actual del modelo no es adecuada. Si no se dispone de ninguna estimación previa de σ^2 pero se han presentado observaciones repetidas de Y (2 ó más) para el mismo valor de X, podemos usar estas repeticiones para obtener un estimador de σ^2. De dicho estimador se dice que representa el error puro, puesto que si el valor de X es igual para dos observaciones, sólo la variación aleatoria puede influir los resultados y generar diferencias entre ellos. Tales diferencias proveerán un estimador de σ^2 que sea mucho más confiable que el que pudiéramos obtener de cualquier otra fuente. Es por esta razón que es sensato, al diseñar experimentos, partir de observaciones repetidas.

Es importante comprender que las observaciones repetidas deben ser auténticas y no sólo repeticiones de la misma lectura. Por ejemplo, supongamos que estamos tratando de relacionar, por métodos de regresión, las variables Y, coeficiente de inteligencia y X, estatura de la persona. Una repetición auténtica debería ser obtenida si medimos los respectivos I. Q. de dos personas que tengan exactamente la misma estatura. Si, por el contrario, medimos dos veces el I. Q. de una misma persona, ésta no sería una repetición auténtica, sino sólo un punto confirmado. Este último podría proveer información acerca

de la variación del método de prueba, el cual es parte de la variación en σ^2, pero no podría proveerla acerca de la variación del I. Q. entre personas de la misma estatura, que es la σ^2 de nuestro problema. En experimentos químicos, una sucesión de lecturas hechas durante la permanencia de las condiciones de un cierto experimento no proporciona repeticiones genuinas. Supongamos, por ejemplo, que Y representa la viscosidad de una cierta sustancia y X la temperatura de ésta. Se obtendrían repeticiones auténticas realizando n_i veces el experimento que llevara a la sustancia a la temperatura X_i y midiendo cada vez la viscosidad, y no ejecutando una sola vez el experimento y midiendo n_i veces la viscosidad.

Cuando hay observaciones repetidas en los datos necesitamos notación adicional para tomar en cuenta las múltiples observaciones en Y para el mismo nivel de X. Suponga que se tienen n_i observaciones en la respuesta Y al nivel X_i, $i = 1, 2, ..., m$. Sea Y_{ij} la j-ésima observación en la respuesta al nivel X_i ,$j = 1, 2, ..., n_i$. Se tienen $n = \sum_{i=1}^{m} n_i$ observaciones en total. La prueba se desarrolla particionando la **SCE** en dos componentes

$$\mathbf{SCE} = SC_{EP} + SC_{CA}$$

donde SC_{EP} es la suma de cuadrados debida al error puro y SC_{CA} es la suma de cuadrados debida a la carencia de ajuste. Para desarrollar esta partición note que el (ij)-ésimo residual es

$$Y_{ij} - \hat{Y}_i = (Y_{ij} - \bar{Y}_i) + \left(\bar{Y}_i - \hat{Y}_i\right)$$

donde $\bar{Y}_i$ es el promedio de las respuestas de las n_i observaciones en X_i. Elevando al cuadrado ambos lados de la ecuación anterior y sumando sobre i y j se obtiene:

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \hat{Y}_i\right)^2 = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_i\right)^2 + \sum_{i=1}^{m} n_i\left(\bar{Y}_i - \hat{Y}_i\right)^2$$

El lado izquierdo de la ecuación anterior es la suma de cuadrados de los residuales, **SCE**. Las dos componentes del lado derecho miden el error puro y la carencia de ajuste. Obsérvese que la suma de cuadrados del error puro,

$$SC_{EP} = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2,$$

se obtiene calculando la suma de cuadrados corregida de las observaciones repetidas en cada nivel i de X y sumándolas sobre los m niveles de ésta. Si la suposición de varianza constante se satisface, ésta es una medida del error puro independiente del modelo, puesto que para calcularla sólo se usa la variación de las $Y's$ en cada nivel de X . Dado que se tienen $n_i - 1$ grados de libertad asociados al error puro en X_i, la suma de cuadrados del error puro tiene asociados

$$\sum_{i=1}^{m}(n_i - 1) = n - m$$

grados de libertad. La suma de cuadrados debida a la carencia de ajuste,

$$SC_{CA} = \sum_{i=1}^{m} n_i \left(\bar{Y}_i - \hat{Y}_i\right)^2$$

es una suma ponderada de los cuadrados de las desviaciones entre la respuesta media $\bar{Y}_i$ al nivel X_i y su valor ajustado correspondiente. Si los valores ajustados $\hat{Y}_i$ están cerca de la respuesta media $\bar{Y}_i$, entonces tenemos evidencia fuerte que indica que la relación funcional entre las variables debe ser lineal. Si por el contrario los valores ajustados $\hat{Y}_i$ difieren considerablemente de la respuesta media $\bar{Y}_i$, entonces podemos pensar que la verdadera función de regresión no es lineal. La suma de cuadrados debida a la carencia de ajuste SC_{CA} tiene $m - 2$ grados de libertad asociados, puesto que hay m diferentes niveles de X y se pierden 2 grados de libertad debido a los parámetros que se deben estimar.

La estadística para esta prueba es:

$$F = \frac{SC_{CA}/(m-2)}{SC_{EP}/(n-m)} = \frac{CM_{CA}}{CM_{EP}}$$

El valor esperado de CM_{EP} es σ^2 y el valor esperado de CM_{CA} es:

$$E\left(CM_{CA}\right) = \sigma^2 + \frac{\sum_{i=1}^{m} n_i \left[E(Y_i) - \beta_0 - \beta_1 X_i\right]}{m-2}$$

Si la verdadera relación funcional entre las variables es lineal,entonces $E(Y_i) = \beta_0 + \beta_1 X_i$ y $E(CM_{CA}) = \sigma^2$; si no es así, $E(Y_i) \neq \beta_0 + \beta_1 X_i$ y $E(CM_{CA}) > \sigma^2$. Si la verdadera función de regresión es lineal, la estadística F_0 tiene una distribución F con $m-2$ y $n-m$ grados de libertad.

Así, si $F_0 > F^{\alpha}_{(m-2,n-m)}$, se concluye que el modelo presenta carencia de ajuste. Esta prueba puede ser anexada fácilmente al análisis de varianza elaborado para probar la significancia global de la regresión.

Si la prueba es significativa, ello indica que el modelo es inadecuado. Deben entonces hacerse estudios que revelen cómo y dónde ocurre esta falla. Si, idealmente, la prueba no es significativa, concluimos que no parece haber razón para dudar del ajuste del modelo. Ejemplo 5.- La línea ajustada $\hat{Y} = 1.436 + 0.338X$ fue estimada a partir de los datos de la tabla 8.

Tabla 8

Obs.	Y	X	*Obs.*	Y	X	*Obs.*	Y	X
1	2.3	1.3	9	1.7	3.7	17	3.5	5.3
2	1.8	1.3	10	2.8	4.0	18	2.8	5.3
3	2.8	2.0	11	2.8	4.0	19	2.1	5.3
4	1.5	2.0	12	2.2	4.0	20	3.4	5.7
5	2.2	2.7	13	5.4	4.7	21	3.2	6.0
6	3.8	3.3	14	3.2	4.7	22	3.0	6.0
7	1.8	3.3	15	1.9	4.7	23	3.0	6.3
8	3.7	3.7	16	1.8	5.0	24	5.9	6.7

La tabla de análisis de varianza se muestra en la tabla 9. Note que el valor de F para la prueba de significancia global aun no ha sido calculado, puesto que aun no sabemos si el modelo presenta carencia de ajuste.

Tabla 9: Tabla de Análisis de Varianza

Fuente de Variación	*Grados de libertad*	*Sumas de Cuadrados*	*Cuadrados Medios*	*F*
Regresión	1	6.326	6.326	6.569
Residuales	22	21.192	0.963	
Total	23	27.518		

La prueba es significativa si no muestra carencia de ajuste. En este paso debemos hallar el error puro y por consiguiente la carencia de ajuste.

1. La suma de cuadrados del error puro debida a las repeticiones en $X = 1.3$ es $\frac{(2.3-1.8)^2}{2} = 0.125$, con un grado de libertad.

2. La suma de cuadrados del error puro debida a las repeticiones en $X = 4.7$ es:

$$\begin{aligned}(5.4)^2 + (3.2)^2 + (1.9)^2 - 3\left[(5.4+3.2+1.9)/3\right]^2 &= 43.01 - (10.5)^2/3 \\ &= 43.01 - 36.75 \\ &= 6.26\end{aligned}$$

con 2 grados de libertad. Cálculos similares nos llevan a las siguientes cantidades:

Tabla 10

Nivel de X	$\sum_{i=1}^{n_i} \left(Y_{ij} - \bar{Y}_i\right)^2$	*Grados de libertad*
1.3	0.125	1
2.0	0.845	1
3.3	2.000	1
3.7	2.000	1
4.0	0.240	2
4.7	6.260	2
5.3	0.980	2
6.0	0.020	1
Total	12.470	11

Podemos entonces reescribir la tabla de análisis de varianza. El valor obtenido para F en la prueba de carencia de ajuste, $F = 0.699$ no es significativo, puesto que es menor que uno. En efecto, al examinar las tablas correspondientes nos damos cuenta que todos los cuantiles en esta distribución son iguales o mayores que la unidad.

Tabla 11: Tabla de análisis de varianza que incluye la prueba de carencia de ajuste

Fuente de Variación	*Grados de Libertad*	*Sumas de Cuadrados*	*Cuadrados Medios*	F
Regresión	1	6.326	6.326	6.569
Residuales	22	21.192	0.963	
Carencia de Ajuste	11	8.722	0.793	0.699
Error Puro	11	12.470	1.134	
Total	23	27.518		

A la luz de este análisis no tenemos motivos para dudar que el modelo sea adecuado y podemos usar $S^2 = 0.963$ como un estimador de σ^2 para llevar a cabo la prueba de significancia global. Esta última sólo es válida si el modelo no exhibe carencia de ajuste. Para enfatizar este punto, resumiremos los pasos a seguir cuando nuestros datos contienen observaciones repetidas.

1. Ajuste el modelo, obtenga la tabla de análisis de varianza usual. No lleve a cabo la prueba de significancia global aun.

2. Calcule la suma de cuadrados del error puro.

3. Desarrolle la prueba de carencia de ajuste. Si el modelo exhibe una carencia de ajuste significativa vaya al paso siguiente; de no ser así, no se tiene razón para dudar de la validez del modelo. En este caso, vaya al último paso.

4. Detenga el análisis del modelo ajustado y considere alternativas para mejorarlo examinando los residuales. No lleve a cabo la prueba de significancia global y no

intente obtener intervalos de confianza. Las suposiciones en las que están basados estos cálculos no son ciertas si se encuentra carencia de ajuste en el modelo.

5. Lleve a cabo la prueba de significancia global, obtenga intervalos de confianza y examine los residuales.

Note que el hecho de que el modelo haya pasado estas dos pruebas no significa que sea correcto. Simplemente significa que es un modelo plausible en el cual no se ha encontrado carencia de ajuste.

Capítulo 4

El modelo lineal general

Un modelo de regresión que involucra más de una variable independiente es llamado un modelo de regresión múltiple. La mayor parte de los modelos usan más de una variable independiente para explicar el comportamiento de la variable dependiente; por ejemplo, suponga que la vida útil de una cortadora depende de la velocidad y de la profundidad del corte. Un modelo de regresión múltiple que podría describir esta relación es:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Donde Y denota la vida útil de la cortadora y X_1 y X_2 la velocidad y la profundidad del corte, respectivamente. Este es un modelo lineal de regresión, con dos variables independientes. El término *lineal* es usado porque la ecuación anterior describe una función lineal en β_1 y β_2.

El modelo lineal puede ser extendido para incluir cualquier número de variables independientes:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i$$

La notación incluye ahora un subíndice en X y β para identificar cada variable independiente y su coeficiente de regresión correspondiente. Hay k variables independientes, e incluyendo a β_0, el término de intercepción, $(k+1)$ parámetros que necesitan ser estimados.

Las suposiciones usuales del método de mínimos cuadrados se aplican en este caso: Se supone que las ε_i son independientes y tienen varianza común y constante σ^2. Para construir pruebas de significancia e intervalos de confianza, se supone también que los errores aleatorios se distribuyen normalmente. Se supone también que las variables independientes son medidas sin error.

El método de mínimos cuadrados aplicado a este modelo requiere que los estimadores encontrados para los $(k+1)$ parámetros sean tales que

$$\begin{aligned} \mathbf{SCE} &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_k X_{ik}\right)\right)^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2} - ... - \hat{\beta}_k X_{ik}\right)^2 \end{aligned}$$

la suma de cuadrados del error, sea mínima.

Los $\hat{\beta}_i$, $i = 1, 2, ..., k$ son los estimadores de los parámetros. Los valores de $\hat{\beta}_i$ que minimizan a **SCE** se obtienen al igualar a cero la derivada de **SCE** con respecto a cada $\hat{\beta}_i$. Esto da por resultado $(k+1)$ ecuaciones normales que deben ser resueltas simultáneamente para obtener los estimadores por mínimos cuadrados de los $(k+1)$ parámetros.

Como se puede observar, el problema se vuelve cada vez más complicado cuando el número de variables independientes aumenta. La notación algebraica se torna entonces particularmente fastidiosa. Es por ello que será usada la notación matricial para desarrollar los resultados de regresión del modelo lineal general.

El modelo lineal general para relacionar una variable dependiente a k variables independientes es:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} + \varepsilon_i$$

El subíndice i denota la unidad observacional a partir de la cual las observaciones en Y y las k variables independientes fueron tomadas. El segundo subíndice designa la variable independiente. El tamaño de la muestra será denotado por n, $i = 1, 2, ..., n$ y k denota

el número de variables independientes. Existen $(k+1)$ parámetros β_j, $j = 0, 1, 2, ...k$ para ser estimados si se incluye un término intercepción β_0. Por conveniencia usaremos $k' = (k+1)$ y supondremos que $n \geq k'$, es decir, que se tienen más observaciones que parámetros.

Así, para expresar el modelo lineal general en notación matricial se necesitan cuatro matrices:

$\underline{Y}$: El vector columna de tamaño $(n \times 1)$ que consta de las observaciones en la variable independiente Y_i;

$\mathbf{X}$: La matriz de orden $(n \times k')$, formada por una columna cuyas entradas son todas uno, seguida de los k vectores columna de las observaciones independientes;

$\underline{\beta}$: Un vector columna de tamaño $(k' \times 1)$ cuyas entradas son los parámetros que serán estimados y

$\underline{\varepsilon}$: El vector de tamaño $(n \times 1)$ de los errores aleatorios.

Con estas definiciones, el modelo lineal puede ser escrito como

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$$

o bien

$$\underbrace{\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}}_{n\times 1} = \underbrace{\begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & X_{23} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & X_{n3} & \cdots & X_{nk} \end{pmatrix}}_{n\times k'} \underbrace{\begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}}_{k'\times 1} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{n\times 1}$$

Cada columna de $\mathbf{X}$ contiene los valores para una variable independiente en particular. Los elementos de un renglón particular de $\mathbf{X}$, por ejemplo, el renglón r, son los coeficientes de los parámetros correspondientes en $\underline{\beta}$, los cuales dan $E(Y_r)$. Note que el coeficiente de regresión de β_0 es constante e igual a uno para todas las observaciones; de esta forma, el vector columna $\underline{1}$ es la primera columna de $\mathbf{X}$. Al multiplicar el primer renglón de $\mathbf{X}$ por $\underline{\beta}$ y añadiendo el primer elemento de $\underline{\varepsilon}$, se confirma que el modelo para la primera observación es :

$$Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + ... + \beta_k X_{1k} + \varepsilon_1$$

Los vectores $\underline{Y}$ y $\underline{\varepsilon}$ son vectores aleatorios, es decir, los elementos de estos vectores son constantes desconocidas que serán estimadas a partir de los datos. Cada elemento β_i de $\underline{\beta}$ es un coeficiente parcial de regresión que refleja el cambio en la variable dependiente Y por unidad de cambio en la i-ésima variable independiente, X_i, suponiendo que todas las demás variables independientes se mantienen constantes. La definición de cada coeficiente parcial de regresión depende del conjunto de variables independientes incluidas en el modelo.

Las suposiciones usuales acerca de ε_i se expresan ahora en términos del vector aleatorio $\underline{\varepsilon}$. Se dice que $\underline{\varepsilon}$ tiene una distribución normal multivariada, cuya media es el vector $\underline{0}$ (de orden $(n \times 1)$). La varianza de un elemento individual ε_i, es generalizada a la matriz de varianzas y covarianzas del vector $\underline{\varepsilon}$. Recuérdese que la matriz de varianzas y covarianzas de cualquier vector aleatorio de n elementos, se define como una matriz simétrica, de orden $(n \times n)$, cuyos elementos en la diagonal principal son iguales a las varianzas de las variables aleatorias; y los (i, j)-ésimos elementos que se hallan fuera de la diagonal principal, son las covarianzas entre ε_i y ε_j. Por ejemplo, si $\underline{Z}$ es un vector aleatorio de tamaño n, cuyos elementos son las variables aleatorias z_1, z_2, ...,z_n, la matriz de varianzas y covarianzas de $\underline{Z}$ es:

$$Var(\underline{Z}) = \begin{pmatrix} Var(z_1) & Cov(z_1, z_2) & \cdots & Cov(z_1, z_n) \\ Cov(z_1, z_2) & Var(z_2) & \cdots & Cov(z_2, z_n) \\ \vdots & \vdots & & \vdots \\ Cov(z_1, z_n) & Cov(z_2, z_n) & \cdots & Var(z_n) \end{pmatrix}$$

Obsérvese que si los elementos del vector fueran independientes, se tendría la matriz

$$\text{diagonal}$$
$$Var(\underline{Z}) = \begin{pmatrix} Var(z_1) & 0 & \cdots & 0 \\ 0 & Var(z_2) & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & Var(z_n) \end{pmatrix}$$

Así, las dos suposiciones usuales sobre ε_i (que las ε_i son independientes con varianza común y constante, σ^2) implican que la varianza de $\underline{\varepsilon}$, $Var(\underline{\varepsilon})$ sea

$$Var(\underline{\varepsilon}) = \begin{pmatrix} Var(\varepsilon_1) & 0 & \cdots & 0 \\ 0 & Var(\varepsilon_2) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & Var(\varepsilon_n) \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 \mathbf{I}_n$$

De esta forma, se tiene la notación

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

Dado que los elementos de $\mathbf{X}$ y $\underline{\beta}$ son constantes, el producto $\mathbf{X}\underline{\beta}$ en el modelo es un conjunto de constantes que se sumarán al vector de errores aleatorios $\underline{\varepsilon}$. Así, $\underline{Y}$ es un vector aleatorio cuya media es el vector $\mathbf{X}\underline{\beta}$ y cuya matriz de varianzas y covarianzas es $\sigma^2 \mathbf{I}_n$; en efecto:

$$E(\underline{Y}) = E(\mathbf{X}\underline{\beta} + \underline{\varepsilon}) = E(\mathbf{X}\underline{\beta}) + E(\underline{\varepsilon}) = \mathbf{X}\underline{\beta}$$

y

$$Var(\underline{Y}) = Var(\mathbf{X}\underline{\beta} + \underline{\varepsilon}) = Var(\mathbf{X}\underline{\beta}) + Var(\underline{\varepsilon}) = \sigma^2 \mathbf{I}_n$$

$Var(\underline{Y}) = Var(\underline{\varepsilon})$ puesto que añadir una constante a una variable aleatoria no altera su varianza. Cuando la distribución de $\underline{\varepsilon}$ es una normal multivariada, la distribución de $\underline{Y}$ también lo es. De esta manera:

$$\underline{Y} \sim N(\mathbf{X}\underline{\beta}, \sigma^2 \mathbf{I}_n)$$

Este resultado se basa en la suposición de que el modelo lineal que se está usando es el correcto. Si variables independientes importantes han sido omitidas o si la forma funcional del modelo no es correcta, $\mathbf{X}\underline{\beta}$ no es el valor esperado de $\underline{Y}$.

Ejemplo 1.- En los datos acerca de contaminación por ozono en granos de soya (ejemplo 1 del capítulo 1) se tiene

$$\mathbf{X} = \begin{pmatrix} 1 & 0.02 \\ 1 & 0.07 \\ 1 & 0.11 \\ 1 & 0.15 \end{pmatrix} \qquad \underline{Y} = \begin{pmatrix} 242 \\ 237 \\ 231 \\ 201 \end{pmatrix} \qquad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

y $\underline{\varepsilon}$ es el vector de los cuatro errores aleatorios (no observables).

4.1 Las ecuaciones normales y su solución

Para estimar los coeficientes de regresión del modelo lineal general se usará también el método de mínimos cuadrados; aplicado a este modelo requiere que los estimadores encontrados para los $(k+1)$ parámetros sean tales que el cuadrado medio del error sea mínimo. Así, el objetivo es encontrar un vector

$$\underline{\hat{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

tal que **SCE** sea mínimo. Para resolver este problema, definamos al vector

$$\underline{e} = \underline{Y} - \underline{\hat{Y}} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} \left(Y_1 - \hat{Y}_1\right) \\ \left(Y_2 - \hat{Y}_2\right) \\ \vdots \\ \left(Y_n - \hat{Y}_n\right) \end{pmatrix}$$

y note que

$$\underline{e}'\underline{e} = \mathbf{SCE} = \sum_{i=1}^{n} e_i^2$$

Se tiene también

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

De esta forma, se desea minimizar a $\sum_{i=1}^{n} e_i^2 = \underline{Y}'\underline{Y}-\underline{Y}'\mathbf{X}\underline{\hat{\beta}}-\underline{\hat{\beta}}'\mathbf{X}'\underline{Y}+\underline{\hat{\beta}}'\mathbf{X}'\mathbf{X}\underline{\hat{\beta}}$ y lo haremos utilizando las técnicas del cálculo diferencial. Derivando la expresión anterior con respecto a $\underline{\hat{\beta}}$ e igualando a cero, se obtiene:

$$
\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \underline{\hat{\beta}}} = -2\mathbf{X}'\underline{Y}+2\mathbf{X}'\mathbf{X}\underline{\hat{\beta}} = 0
$$

es decir

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

lo cual implica

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

donde $\hat{\beta}$ es un vector de tamaño k', mismo que se obtiene del sistema de ecuaciones normales. El producto $\mathbf{X}'\mathbf{X}$ genera una matriz cuadrada de tamaño k', en cuya diagonal principal se encuentran las sumas de cuadrados de cada una de las variables independientes y, fuera de ella, las sumas de los productos cruzados entre ellas. La forma general es:

$$
\mathbf{X}'\mathbf{X} = \begin{pmatrix}
n & \sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i2} & \cdots & \sum_{i=1}^{n} X_{ik} \\
\sum_{i=1}^{n} X_{i1} & \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1}X_{i2} & \cdots & \sum_{i=1}^{n} X_{i1}X_{ik} \\
\sum_{i=1}^{n} X_{i2} & \sum_{i=1}^{n} X_{i1}X_{i2} & \sum_{i=1}^{n} X_{i2}^2 & \cdots & \sum_{i=1}^{n} X_{i2}X_{ik} \\
\vdots & \vdots & \vdots & & \vdots \\
\sum_{i=1}^{n} X_{ik} & \sum_{i=1}^{n} X_{i1}X_{ik} & \sum_{i=1}^{n} X_{i2}X_{ik} & \cdots & \sum_{i=1}^{n} X_{ik}^2
\end{pmatrix}
$$

En el caso del modelo lineal simple, la matriz $\mathbf{X}'\mathbf{X}$ consiste solamente de la matriz de (2×2) superior izquierda.

Los elementos del producto $\mathbf{X}'\underline{Y}$ son las sumas de los productos entre cada variable independiente y la variable dependiente:

$$\mathbf{X}'\underline{Y} = \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_{i1}Y_i \\ \sum_{i=1}^{n} X_{i2}Y_i \\ \vdots \\ \sum_{i=1}^{n} X_{ik}Y_i \end{pmatrix}$$

El primer elemento, $\sum_{i=1}^{n} Y_i$, es la suma de los productos entre el vector de unos (la primera columna de $\mathbf{X}$) y $\underline{Y}$. De nuevo, si solamente está involucrada una variable independiente en el modelo, el vector $\mathbf{X}'\underline{Y}$ está formado únicamente por los dos primeros elementos.

Las ecuaciones normales tienen solución única; ésta existe si y solamente si la inversa de $\mathbf{X}'\mathbf{X}$ existe, es decir, si y sólo si $\mathbf{X}'\mathbf{X}$ es de rango completo. Esto es, no deben existir dependencias lineales entre las variables independientes. La implicación práctica de esta condición es que no deben existir redundancias en la información contenida en $\mathbf{X}$. Por ejemplo, la cantidad de nitrógeno en la dieta algunas veces puede expresarse como la cantidad de proteínas multiplicada por una constante. Dado que la misma información se reporta de dos distintas formas, se crea una dependencia lineal si ambas se incluyen en $\mathbf{X}$.

Note que

$$\underline{\hat{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\underline{Y}) = \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\underline{Y}$$

muestra que los estimadores de los coeficientes de regresión son funciones lineales de la variable dependiente $\underline{Y}$, donde los coeficientes de estas funciones están dados por los

elementos de los renglones de $\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)$. Puesto que las X's son valores constantes, los valores esperados de los coeficientes de regresión involucran solamente la esperanza de $\underline{Y}$. Si el modelo $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$ es correcto, la esperanza de $\underline{Y}$ es $\mathbf{X}\underline{\beta}$ y la esperanza de $\hat{\underline{\beta}}$ es

$$\begin{aligned} E(\hat{\underline{\beta}}) &= E\left(\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\underline{Y}\right) \\ &= \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)E(\underline{Y}) \\ &= \underbrace{\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{X}}_{\mathbf{I}_n}\underline{\beta} \\ &= \underline{\beta} \end{aligned}$$

Es decir, $\hat{\underline{\beta}}$ es un estimador insesgado de $\underline{\beta}$.

Ejemplo 2.- Se ilustrarán las operaciones matriciales usando $\mathbf{X}$ y $\underline{Y}$ de los datos de Heagle (ejemplo anterior):

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 4 & 0.3500 \\ 0.3500 & 399 \end{pmatrix} \qquad \mathbf{X}'\underline{Y} = \begin{pmatrix} 911 \\ 76.99 \end{pmatrix}$$

y

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1.07547 & -9.43396 \\ -9.43396 & 107.81671 \end{pmatrix}$$

Los estimadores de los coeficientes de regresión son

$$\hat{\underline{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{Y} = \begin{pmatrix} 253.434 \\ -293.531 \end{pmatrix}$$

4.2 El vector $\hat{\underline{Y}}$ y el vector de residuales $\underline{e}$

En un conjunto de datos, el vector de los valores estimados de la variable dependiente $\underline{Y}$ es

$$\hat{\underline{Y}} = \mathbf{X}\hat{\underline{\beta}}$$

Esta es la manera más fácil de calcular $\hat{\underline{Y}}$. Sin embargo, con el fin de desarrollar algunos resultados, es útil expresar a $\hat{\underline{Y}}$ como una función lineal de $\underline{Y}$, sustituyendo $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{Y}$ por $\hat{\underline{\beta}}$. Así tenemos:

$$\begin{aligned}\hat{\underline{Y}} &= \left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\underline{Y} \\ &= \mathbf{H}\underline{Y}\end{aligned}$$

donde $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ es una matriz cuadrada de tamaño n determinada completamente por los valores de las variables independientes. Esta matriz, conocida también como la matriz *sombrero* (hat), juega un papel muy importante en el análisis de regresión.

A continuación examinaremos algunas de sus propiedades, que serán de utilidad más adelante:

1. $\mathbf{H}$ y $(\mathbf{I}_n - \mathbf{H})$ son matrices simétricas e idempotentes
2. $rango\,(\mathbf{I}_n - \mathbf{H}) = tr\,(\mathbf{I}_n - \mathbf{H}) = n - k'$
3. $(\mathbf{I}_n - \mathbf{H})\,\mathbf{X} = 0$ en el modelo de regresión

Demostración

1. $\mathbf{H}$ y $(\mathbf{I}_n - \mathbf{H})$ son trivialmente simétricas. Probemos la idempotencia de $\mathbf{H}$:

$$\begin{aligned}\mathbf{H}^2 &= \left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right) \\ &= \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underbrace{\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}}_{\mathbf{I}_n}\mathbf{X}' \\ &= \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' \\ &= \mathbf{H}\end{aligned}$$

A partir de la idempotencia de $\mathbf{H}$ se puede probar la idempotencia de $(\mathbf{I}_n - \mathbf{H})$:

$$\begin{aligned}\left(\mathbf{I}_n - \mathbf{H}\right)^2 &= \mathbf{I}_n^2 - 2\mathbf{I}_n\mathbf{H} + \mathbf{H}^2 \\ &= \mathbf{I}_n - 2\mathbf{H} + \mathbf{H} \\ &= \mathbf{I}_n - \mathbf{H}\end{aligned}$$

2. Para probar esta propiedad necesitamos antes probar el siguiente resultado:

a) Si P es una matriz simétrica entonces P es una matriz idempotente de rango r si y sólo si P tiene r valores propios iguales a uno y $(n-r)$ valores propios iguales a cero.

Demostración: Dado que $P^2 = P$, entonces, $P\underline{X} = \lambda\underline{X}$ $(\underline{X} \neq \underline{0})$ implica que $\lambda\underline{X}'\underline{X} = \underline{X}'P\underline{X} = \underline{X}P^2\underline{X} = (P\underline{X})'(P\underline{X}) = \lambda^2\underline{X}'\underline{X}$, y $\lambda(\lambda-1) = 0$. Es decir, los vectores propios de P son cero y uno. Dado que en una matriz simétrica el rango de ésta es igual al número de valores propios distintos de cero, P tiene r valores propios iguales a uno y $(n-r)$ iguales a cero. Recíprocamente, si los valores propios de P son iguales a cero y uno, entonces podemos suponer sin pérdida de generalidad que los primeros r valores propios son iguales a uno. Así, existe una matriz ortogonal T tal que

$$T'PT = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix} = \Lambda \quad \text{o} \quad P = T\Lambda T'$$

Por lo tanto $P^2 = T\Lambda T'T\Lambda T' = T\Lambda^2 T' = T\Lambda T' = P$, y $rango(P) = r$.

El resultado anterior nos permite afirmar que $rango(P) = tr(P) = r$. Por lo anterior, dado que $(\mathbf{I}_n - \mathbf{H})$ es simétrica e idempotente, se tiene

$$rango\,(\mathbf{I}_n - \mathbf{H}) = tr\,(\mathbf{I}_n - \mathbf{H}) = tr(\mathbf{I}_n) - tr(\mathbf{H}) = n - tr\left(\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{\mathbf{I}_{k'}}\right) = n - k'$$

3. A partir de la definición de $\mathbf{H}$, se tiene:

$$(\mathbf{I}_n - \mathbf{H})\,\mathbf{X} = \mathbf{X} - \mathbf{X}\underbrace{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}}_{\mathbf{I}_{k'}} = \mathbf{X} - \mathbf{X} = 0$$

Note que la esperanza de $\hat{\underline{Y}}$ es

$$\begin{aligned} E(\hat{\underline{Y}}) &= E\left[\left(\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\right)\underline{Y}\right] \\ &= \mathbf{H}E\,(\underline{Y}) = \mathbf{H}\mathbf{X}\underline{\beta} = \mathbf{X}\underline{\beta} \end{aligned}$$

Es decir, $\underline{\hat{Y}}$ es un estimador insesgado de las medias de $\underline{Y}$ para los valores de X en el conjunto de los datos. El hecho de que $\mathbf{HX} = \mathbf{X}$ se puede verificar a partir de la definición de $\mathbf{H}$:

$$\mathbf{HX} = \left(\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\mathbf{X} = \mathbf{X}\left(\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\right) = \mathbf{X}$$

Ejemplo 3.- En los datos del ejemplo anterior:

$$\begin{aligned}\mathbf{H} &= \begin{pmatrix} 1 & 0.02 \\ 1 & 0.07 \\ 1 & 0.11 \\ 1 & 0.15 \end{pmatrix} \begin{pmatrix} 1.07547 & -9.43396 \\ -9.43396 & 107.81671 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0.02 & 0.07 & 0.11 & 0.15 \end{pmatrix} \\ &= \begin{pmatrix} .741240 & .377358 & .086253 & -.204852 \\ .377358 & .283019 & .207547 & .132075 \\ .086253 & .207547 & .304582 & .401617 \\ -.204852 & .132075 & .401617 & .671159 \end{pmatrix}\end{aligned}$$

Así, por ejemplo,

$$\hat{Y}_1 = .741\,Y_1 + .377\,Y_2 + .086\,Y_3 - .205\,Y_4$$

El vector de residuales $\underline{e}$ refleja la carencia de consistencia entre los valores observados $\underline{Y}$ y los estimados, $\underline{\hat{Y}}$:

$$\underline{e} = \underline{Y} - \underline{\hat{Y}}$$

Existen otras formas de expresar a $\underline{e}$ que nos serán útiles más adelante, como:

$$\begin{aligned}\underline{e} &= \underline{Y} - \mathbf{X}\underline{\hat{\beta}} \\ &= \underline{Y} - \mathbf{H}\underline{Y} \\ &= (\mathbf{I} - \mathbf{H})\underline{Y}\end{aligned}$$

Recuérdese que los estimadores por mínimos cuadrados minimizan la suma de cuadrados de los residuales, es decir, $\underline{\hat{\beta}}$ ha sido elegido de tal forma que $\underline{e}'\underline{e}$ sea mínimo.

Dado que $(\mathbf{I} - \mathbf{H})$ es simétrica e idempotente, la esperanza del vector de residuales es

$$
\begin{aligned}
E(\underline{e}) &= E\left((\mathbf{I}-\mathbf{H})\underline{Y}\right) \\
&= (\mathbf{I}-\mathbf{H})\mathbf{X}\underline{\beta} \\
&= (\mathbf{X}-\mathbf{HX})\mathbf{X}\underline{\beta} \\
&= (\mathbf{X}-\mathbf{X})\underline{\beta} \\
&= \underline{0}
\end{aligned}
$$

Es decir, los residuales observados son variables aleatorias con media cero.

Así, la variable respuesta $\underline{Y}$ está dividida en dos partes: Aquella que aporta el modelo, $\underline{\hat{Y}}$, y la que aportan los residuales, $\underline{e}$. El hecho de que estas dos partes sumen $\underline{Y}$ es evidente, puesto que $\underline{e}$ se obtiene por diferencia; o bien, puede ser demostrado como sigue:

$$\underline{\hat{Y}} + \underline{e} = \mathbf{H}\underline{Y} + (\mathbf{I}-\mathbf{H})\underline{Y} = (\mathbf{H}+\mathbf{I}-\mathbf{H})\underline{Y} = \underline{Y}$$

Ejemplo 4.- Continuando con el ejemplo 3, obtenemos:

$$\underline{\hat{Y}} = \begin{pmatrix} 1 & 0.02 \\ 1 & 0.07 \\ 1 & 0.11 \\ 1 & 0.15 \end{pmatrix} \begin{pmatrix} 253.434 \\ -293.531 \end{pmatrix} = \begin{pmatrix} 247.563 \\ 232.887 \\ 221.146 \\ 209.404 \end{pmatrix}$$

Los residuales son:

$$\underline{e} = \underline{Y} - \underline{\hat{Y}} = \begin{pmatrix} -5.563 \\ 4.113 \\ 9.854 \\ -8.404 \end{pmatrix}$$

Los resultados del ejemplo del ozono están resumidos en la tabla que sigue

Tabla 1

X_i	Y_i	$\hat{Y}_i$	e_i
.02	242	247.563	−5.563
.07	237	232.887	4.113
.11	231	221.146	9.854
.15	201	209.404	−8.404

4.3 Estimación de σ^2

Como en el caso del modelo lineal simple, debemos desarrollar un estimador de σ^2 a partir de la suma de cuadrados de los residuales.

$$\begin{aligned} \text{SCE} &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\ &= \sum_{i=1}^{n} e_i^2 \\ &= \underline{e}'\underline{e} \end{aligned}$$

Sustituyendo $\underline{e} = \underline{Y} - \mathbf{X}\underline{\hat{\beta}}$ tenemos

$$\begin{aligned} \text{SCE} &= \left(\underline{Y} - \mathbf{X}\underline{\hat{\beta}}\right)' \left(\underline{Y} - \mathbf{X}\underline{\hat{\beta}}\right) \\ &= \underline{Y}'\underline{Y} - \underline{\hat{\beta}}'\mathbf{X}'\underline{Y} - \underline{Y}'\mathbf{X}\underline{\hat{\beta}} + \underline{\hat{\beta}}'\mathbf{X}'\mathbf{X}\underline{\hat{\beta}} \\ &= \underline{Y}'\underline{Y} - 2\underline{\hat{\beta}}'\mathbf{X}'\underline{Y} + \underline{\hat{\beta}}'\mathbf{X}'\mathbf{X}\underline{\hat{\beta}} \end{aligned}$$

Dado que $\mathbf{X}'\mathbf{X}\underline{\hat{\beta}} = \mathbf{X}'\underline{Y}$ la ecuación anterior puede escribirse de la siguiente manera

$$\text{SCE} = \underline{Y}'\underline{Y} - \underline{\hat{\beta}}'\mathbf{X}'\underline{Y}$$

La suma de cuadrados de los residuales tiene $(n - k')$ grados de libertad asociados, puesto que existen k' parámetros en el modelo que se van a estimar. Así, el cuadrado medio del error es

$$\text{CME} = \frac{\text{SCE}}{n - k'}$$

Se propone a CME como un estimador de σ^2, es decir, $\hat{\sigma}^2 = \text{CME}$. A continuación probaremos que $\hat{\sigma}^2$ es un estimador insesgado de σ^2.

En efecto, recuérdese que

$$\underline{e} = \mathbf{I} - \mathbf{H}$$

Por otro lado

$$\begin{aligned}(n-k')\,\mathbf{CME} &= \underline{Y}'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\underline{Y} \\ &= \underline{Y}'(\mathbf{I}-\mathbf{H})^2\underline{Y} \\ &= \underline{Y}'(\mathbf{I}-\mathbf{H})\underline{Y}\end{aligned}$$

Para calcular la esperanza del cuadrado medio de los residuales, usaremos el siguiente resultado acerca del valor esperado de una forma cuadrática:

Teorema 1.- Sea $\underline{V}$ un vector aleatorio de tamaño n, sea A una matriz simétrica de tamaño $(n \times n)$. Si $E(\underline{V}) = \underline{\theta}$ y $Var(\underline{V}) = \Sigma$, entonces

$$E(\underline{V}'A\underline{V}) = tr(A\Sigma) + \underline{\theta}'A\underline{\theta}$$

Demostración.- $E(\underline{V}'A\underline{V}) = E\left[(\underline{V}-\underline{\theta})'A(\underline{V}-\underline{\theta}) + \underline{\theta}'A\underline{Y} + \underline{Y}'A\underline{\theta} - \underline{\theta}'A\underline{\theta}\right]$. Ahora, $\underline{Y}'A\underline{\theta} = (\underline{Y}'A\underline{\theta})' = \underline{\theta}'A\underline{Y}$ y

$$E(\underline{\theta}'A\underline{Y}) = \underline{\theta}'AE(\underline{Y}) = \underline{\theta}'A\underline{\theta}$$

Por lo anterior

$$\begin{aligned}E(\underline{V}'A\underline{V}) &= E\left[(\underline{V}-\underline{\theta})'A(\underline{V}-\underline{\theta})\right] + \underline{\theta}'A\underline{\theta} \\ &= \sum_i\sum_j a_{ij}E\left[(Y_i-\theta_i)(Y_j-\theta_j)\right] + \underline{\theta}'A\underline{\theta} \\ &= \sum_i\sum_j a_{ij}\sigma_{ij} + \underline{\theta}'A\underline{\theta} \\ &= tr(A\Sigma) + \underline{\theta}'A\underline{\theta}\end{aligned}$$

Usando este teorema y el hecho de que $(\mathbf{I}-\mathbf{H})\mathbf{X} = 0$ se tiene

$$\begin{aligned}E\left[(n-k')\,\mathbf{CME}\right] &= E\left[\underline{Y}'(\mathbf{I}-\mathbf{H})\underline{Y}\right] \\ &= tr\left(\sigma^2(\mathbf{I}-\mathbf{H})\right) \\ &= \sigma^2(n-k')\end{aligned}$$

Es decir, $\hat{\sigma}^2 = \mathbf{CME} = \dfrac{\underline{e}'\underline{e}}{n-k'}$ es un estimador insesgado de σ^2.

4.4 Interpretación geométrica de los estimadores por mínimos cuadrados

A veces resulta de utilidad una interpretación geométrica intuitiva de los estimadores por mínimos cuadrados. Puede pensarse en el vector de observaciones $\underline{Y}' = (Y_1, Y_2, ..., Y_n)$, como en un vector del origen al punto A en la siguiente figura. Note que $Y_1, Y_2, ..., Y_n$ forman las coordenadas de un espacio muestral n-dimensional. El espacio muestral de la figura 1 es tridimensional ($n = 3$).

La matriz $\mathbf{X}$ está formada por k' vectores columna de tamaño n $(\underline{1}, \underline{X}_1, \underline{X}_2, ..., \underline{X}_k)$. Cada uno de estos vectores definen un vector a partir del origen en el espacio muestral. Estos k' vectores forman un espacio k'-dimensional llamado *el espacio de estimación.*

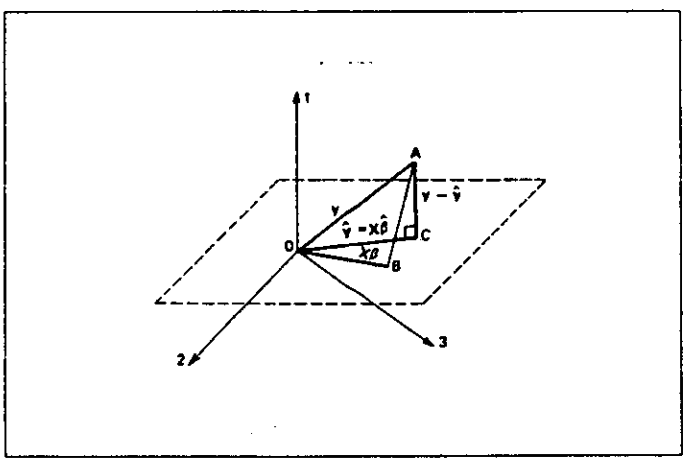


Figura 1: Interpretación geométrica de los estimadores por mínimos cuadrados

El espacio de estimación cuando $k' = 2$ se muestra en la figura 1. Cualquier punto en este espacio puede representarse como una combinación lineal de los vectores $\underline{1}, \underline{X}_1, \underline{X}_2, ..., \underline{X}_k$. Así, cualquier punto en el espacio de estimación es de la forma $\mathbf{X}\underline{\beta}$. Sea B el punto determinado por el vector $\mathbf{X}\underline{\beta}$ en la figura 1. La distancia al cuadrado de A a B es precisamente

$$\begin{aligned} \underline{e}'\underline{e} &= \left(\underline{Y} - \hat{\underline{Y}}\right)' \left(\underline{Y} - \hat{\underline{Y}}\right) \\ &= \left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right)' \left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right) \end{aligned}$$

Por lo tanto, minimizar la distancia cuadrada del punto A al espacio de estimación, significa hallar el punto del espacio de estimación que se encuentre más cerca de A. La distancia al cuadrado es mínima cuando el punto en el espacio de estimación es la intersección entre el espacio de estimación y la línea ortogonal a éste que pasa por A. Este es el punto C de la figura 1 y está definido por el vector $\hat{\underline{Y}} = \mathbf{X}\hat{\underline{\beta}}$. De esta manera, dado que $\underline{Y} - \hat{\underline{Y}} = \underline{Y} - \mathbf{X}\hat{\underline{\beta}}$ es perpendicular al espacio de estimación, podemos escribir

$$\mathbf{X}'\left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right) = \underline{0}$$

o bien

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

que es el sistema de ecuaciones normales.

4.5 Precisión de los estimadores

$\hat{\underline{\beta}}$, $\hat{\underline{Y}}$ y $\underline{e}$ son vectores aleatorios, puesto que son funciones lineales de $\underline{Y}$, que es un vector aleatorio. Esta condición nos ha permitido calcular sus valores esperados, a saber,

$$E\left(\hat{\underline{\beta}}\right) = \underline{\beta}, \quad E\left(\hat{\underline{Y}}\right) = \mathbf{X}\underline{\beta} \quad \text{y} \quad E(\underline{e}) = \underline{0}$$

En esta sección hallaremos las matrices de varianzas y covarianzas de estos vectores, usando las reglas del cálculo de varianzas de funciones lineales de vectores aleatorios.

El resultado general de la varianza de funciones lineales de variables aleatorias puede extenderse a la notación matricial. Suponga que $\underline{Y}$ es un vector aleatorio cuya matriz de varianzas y covarianzas es $Var(\underline{Y})$. Sea $U = \underline{a}'\underline{Y}$, cualquier función lineal del vector $\underline{Y}$, donde $\underline{a}'$ es un vector columna cuyos elementos son los coeficientes que definen la función lineal. La varianza de U puede expresarse en términos de la varianza de $\underline{Y}$ como

$$Var\,(U) = \sigma^2 U = \underline{a}'\,(Var\,(\underline{Y}))\,\underline{a}$$

Si $Var\,(\underline{Y}) = \mathbf{I}\sigma^2$ (que es una de las suposiciones básicas del modelo), se tiene

$$\sigma^2\,(U) = \underline{a}'\,(\mathbf{I}\sigma^2)\,\underline{a} = \underline{a}'\underline{a}\sigma^2$$

Note que $\underline{a}'\underline{a}$ es la suma de cuadrados de los coeficientes de la función lineal, $\sum_{i=1}^{n} a_i^2$, el cual es el mismo resultado que el obtenido en el caso del modelo lineal simple.

Muchas funciones lineales de Y pueden ser consideradas simultáneamente si se extiende $\underline{a}$ a una matriz A de coeficientes de tamaño $k \times n$, donde cada renglón de A contiene los coeficientes de una función lineal. Así, $U = A\underline{Y}$, es un vector columna que contiene k nuevas variables aleatorias, cada una de las cuales es una función lineal del vector aleatorio $\underline{Y}$.

La matriz de varianzas y covarianzas para U es cuadrada de orden k y está dada por

$$Var(U) = A\,(Var\,(\underline{Y}))$$

o bien, cuando $Var\,(\underline{Y}) = \mathbf{I}\sigma^2$,

$$Var\,(U) = AA'\sigma^2$$

El i-ésimo elemento de la diagonal principal de AA' es la suma de cuadrados de los coeficientes de la i-ésima función lineal. Este coeficiente, multiplicado por σ^2, da la varianza de la i-ésima función lineal. El elemento (i,j)-ésimo (fuera de la diagonal) es la suma de los productos de los coeficientes de la i-ésima y la j-ésima función lineal y, cuando se multiplica por σ^2, da la covarianza entre las dos funciones lineales.

La demostración de la forma general de la varianza de funciones lineales de vectores aleatorios se muestra aquí como un ejercicio del álgebra de matrices. Por definición, la matriz de varianzas y covarianzas de un vector aleatorio $\underline{Y}$ es:

$$Var\,(\underline{Y}) = E\left[(\underline{Y} - E\,(\underline{Y}))\,(\underline{Y} - E\,(\underline{Y}))'\right]$$

donde la función esperanza es aplicada a todos los elementos de la matriz que está en el argumento. El producto en el argumento del operador esperanza en la ecuación anterior es una matriz cuadrada de orden n cuyos elementos en la diagonal principal son de la forma $(Y_i - E(Y_i))^2$ y los elementos fuera de ella son de la forma $(Y_i - E(Y_i))(Y_j - E(Y_j))$. Por definición, las esperanzas de estos elementos son varianzas y covarianzas, respectivamente. Esta definición de matriz de varianzas y covarianzas será usada para hallar la varianza de $U = A\underline{Y}$. Por definición, la matriz de varianzas y covarianzas de U es

$$Var(U) = E\left\{[U - E(U)][U - E(U)]'\right\}$$

Sustituyendo $A\underline{Y}$ por U y factorizando, se obtiene

$$\begin{aligned} Var(U) &= E\left\{[A\underline{Y} - E(A\underline{Y})][A\underline{Y} - E(A\underline{Y})]'\right\} \\ &= E\left\{A[\underline{Y} - E(\underline{Y})][\underline{Y} - E(\underline{Y})]'A'\right\} \\ &= AE\left\{[\underline{Y} - E(\underline{Y})][\underline{Y} - E(\underline{Y})]'\right\}A' \\ &= A(Var(\underline{Y}))A' \end{aligned}$$

La factorización de productos matriciales debe efectuarse cuidadosamente; recuérdese que el producto de matrices no es conmutativo. Por ello, A se factoriza por el lado izquierdo (tomándola del primer factor entre corchetes) y por el lado derecho (a partir de la transpuesta del segundo factor entre corchetes). El lector debe también recordar que, al transponer un producto, se invierte el orden de los factores. Dado que A es una matriz de constantes, ésta puede sacarse del operador esperanza. Esto deja una matriz interior que es, por definición, $Var(\underline{Y})$.

Ejemplo 5.- La varianza de la media de una muestra aleatoria con n observaciones, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ se calculará usando el resultado general que se acaba de obtener. $\bar{Y}$ puede expresarse en forma matricial como sigue:

$$\bar{Y} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)\underline{Y}$$

De esta manera, $\bar{Y}$ es una función lineal de $\underline{Y}$, donde el vector de coeficientes $\underline{a}$ es

$$\underline{a} = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)$$

De esta forma, si $Var\,(\underline{Y}) = \mathbf{I}\sigma^2$, se tiene

$$\begin{aligned} Var\,(\underline{Y}) &= \underline{a}' Var\,(\underline{Y})\,\underline{a} \\ &= \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)\mathbf{I}\sigma^2 \begin{pmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix} = n\left(\frac{1}{n}\right)^2\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

A continuación, usaremos el resultado general de varianzas de funciones lineales de vectores aleatorios para hallar las matrices de varianzas y covarianzas de $\hat{\underline{\beta}}$, $\hat{\underline{Y}}$ y $\underline{e}$.

La matriz de coeficientes en $\underline{Y}$ que da por resultado $\hat{\underline{\beta}}$ es $A = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$; así,

$$\begin{aligned} Var\left(\hat{\underline{\beta}}\right) &= \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\left(Var\,(\underline{Y})\right)\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\ &= \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)'\sigma^2 \end{aligned}$$

Dado que la transpuesta de un producto es el producto de las transpuestas en el orden inverso y que $(\mathbf{X}'\mathbf{X})$ es simétrica, obtenemos

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

Es decir, las varianzas y covarianzas de los estimadores de los coeficientes de regresión están dadas por los elementos de $(\mathbf{X}'\mathbf{X})^{-1}$ multiplicados por σ^2. Los elementos en la diagonal principal de la matriz $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ son las varianzas de los coeficientes de regresión, en el orden en el que aparecen en $\underline{\beta}$. Los elementos fuera de ella son las covarianzas entre los coeficientes de regresión.

Ejemplo 6.- En nuestro ejemplo,

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1.0755 & -9.4340 \\ -9.4340 & 107.8167 \end{pmatrix}$$

Así, $Var\left(\hat{\beta}_0\right) = 1.0755\,\sigma^2$ y $Var\left(\hat{\beta}_1\right) = 107.8167\sigma^2$. La covarianza entre $\hat{\beta}_0$ y $\hat{\beta}_1$ es $Cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = -9.4340\,\sigma^2$.

La matriz de varianzas y covarianzas de $\hat{Y}$ puede hallarse usando la relación $\hat{Y} = \mathbf{X}\underline{\hat{\beta}}$ ó $\hat{Y} = \mathbf{H}\underline{Y}$. Recuérdese que $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. Aplicando el resultado general para las varianzas de funciones lineales a la primera relación, se tiene

$$\begin{aligned} Var\left(\hat{Y}\right) &= Var\left(\mathbf{X}\underline{\hat{\beta}}\right) = \mathbf{X}Var\left(\underline{\hat{\beta}}\right)\mathbf{X}' \\ &= \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2 \\ &= \mathbf{H}\sigma^2 \end{aligned}$$

Si se usa la segunda relación se tiene

$$\begin{aligned} Var\left(\hat{Y}\right) &= \mathbf{H}\left(Var\left(\underline{Y}\right)\right)\mathbf{H}' \\ &= \mathbf{H}\mathbf{H}'\sigma^2 \\ &= \mathbf{H}\sigma^2 \end{aligned}$$

dado que $\mathbf{H}$ es simétrica e idempotente. Por lo anterior, al multiplicar $\mathbf{H}$ por σ^2 se obtienen las varianzas y covarianzas de todos los valores ajustados. Dado que $\mathbf{H}$ es una matriz cuadrada de orden n, a menudo es poco funcional trabajar con ella. Las varianzas de cualquier subconjunto de los valores ajustados $\hat{Y}_i$ pueden ser halladas usando solamente los renglones de $\mathbf{X}$, digamos $\mathbf{X}_r$ que correspondan a los puntos de interés. Si usamos el resultado general se tiene

$$Var\left(\hat{Y}_r\right) = \mathbf{X}_r Var\left(\underline{\hat{\beta}}\right)\mathbf{X}_r' = \mathbf{X}_r\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}_r'\sigma^2$$

Las varianzas dadas por $\mathbf{H}\sigma^2$ son las varianzas apropiadas de las $\hat{Y}_i$ cuando éstas son usadas para estimar las medias de Y para los valores dados de las variables independientes. Para predecir observaciones futuras, cada una de las varianzas deberá ser aumentada por σ^2 para contar la varianza de la cantidad que se va a pronosticar. De esta forma, la matriz de varianzas y covarianzas para predicción es

$$Var(\hat{Y}_{pred}) = (\mathbf{I} + \mathbf{H})\,\sigma^2$$

La matriz de varianzas y covarianzas de $\underline{e} = (\mathbf{I} - \mathbf{H})\underline{Y}$, el vector de residuales, es

$$Var(\underline{e}) = (\mathbf{I} - \mathbf{H})\sigma^2$$

usando el hecho de que $(\mathbf{I} - \mathbf{H})$ es una matriz simétrica e idempotente.

Ejemplo 7.- La matriz H fue calculada para los datos de Heagle en el ejemplo 3. Se tiene entonces

$$\begin{aligned} Var\left(\hat{Y}\right) &= \mathbf{H}\sigma^2 \\ &= \begin{pmatrix} .741240 & .377358 & .086253 & -.204852 \\ .377358 & .283019 & .207547 & .132075 \\ .086253 & .207547 & .304582 & .401617 \\ -.204852 & .132075 & .401617 & .671159 \end{pmatrix} \sigma^2 \end{aligned}$$

La matriz de varianzas y covarianzas de los residuales se obtiene utilizando $Var(\underline{e}) = (\mathbf{I} - \mathbf{H})\sigma^2$. Así,

$$\begin{aligned} Var(e_1) &= (1 - 0.741)\sigma^2 &&= 0.259\sigma^2 \\ Var(e_2) &= (1 - 0.305)\sigma^2 &&= 0.695\sigma^2 \\ Cov(e_1, e_3) &= -Cov\left(\hat{Y}_1, \hat{Y}_3\right) &&= -0.086\sigma^2 \end{aligned}$$

Es importante notar que las varianzas de los residuales observados no son iguales a σ^2 y que las covarianzas entre ellos no son cero. La suposición de varianza común y constante y de covarianzas iguales es válida para las ε_i y no para las e_i.

La varianza de cualquier valor ajustado particular $\hat{Y}_i$ y la varianza del correspondiente e_i siempre suman σ^2, porque

$$\mathbf{H}\sigma^2 + (\mathbf{I} - \mathbf{H})\sigma^2 = \mathbf{I}\sigma^2$$

Dado que las varianzas no pueden ser negativas, cada elemento en la diagonal de H debe estar entre cero y uno: $0 < v_{ii} < 1$, donde v_{ii} es el i-ésimo elemento de la diagonal principal de H. Por lo anterior, la varianza de cualquier $\hat{Y}_i$ es siempre menor que σ^2, la varianza de las observaciones individuales. Esto muestra la ventaja de ajustar un

modelo de respuesta continuo, suponiendo que el modelo es correcto, simplemente usando los datos observados como estimadores de la media de Y para los valores dados de X. Esta precisión tan alta del modelo proviene del hecho de que cada $\hat{Y}_i$ usa información proveniente de los puntos que la rodean. La precisión del modelo puede ser realmente impresionante. En el ejemplo anterior, la precisión obtenida en los estimadores de las medias de los dos valores intermedios de ozono usando la ecuación de respuesta lineal fueron $0.283\sigma^2$ y $0.305\sigma^2$. Para lograr el mismo grado de precisión sin usar el modelo de regresión lineal, se hubieran requerido más de tres observaciones en cada nivel de ozono.

La ecuación anterior implica que los puntos con una alta precisión en $\hat{Y}_i$ (varianza baja) tendrán poca precisión en e_i (varianza alta) y vice-versa. Belsley, Kuh y Welsch (1980) muestran que los elementos en la diagonal principal de $\mathbf{H}$ pueden interpretarse como medidas de la distancia de los puntos correspondientes al centro del espacio de X (o $\bar{X}$, en el caso de una sola variable independiente). Puntos lejanos al centro del espacio X tienen valores v_{ii} relativamente grandes y, por lo tanto, relativamente poca precisión en $\hat{Y}_i$ y precisión más alta en e_i. Varianzas pequeñas de los residuales que se hallan lejos del centro de los datos indican que el modelo tiende a estar cerca de los valores observados para estos puntos. Este aspecto de $\mathbf{H}$ puede ser usado para detectar puntos con demasiada influencia en el modelo.

Hemos expresado a las varianzas (y covarianzas) como múltiplos de σ^2. Los coeficientes están determinados completamente por la matriz $\mathbf{X}$, una matriz de constantes que depende del modelo que se va a ajustar y de los niveles de las variables independientes en el estudio. En algunos experimentos los niveles de las variables independientes están sujetos al control del investigador y pueden ser conocidas antes de que el experimento se lleve a cabo. La eficiencia de diseños experimentales alternativos puede evaluarse calculando $(\mathbf{X}'\mathbf{X})^{-1}$ y $\mathbf{H}$ para cada diseño; se preferirá aquél que posea las varianzas menores para las cantidades de interés.

4.6 Distribución de funciones lineales de variables aleatorias normales

La suposición de que los errores aleatorios ε_i se distribuyen normalmente, implica que los valores observados de la variable dependiente Y_i también se distribuyen normalmente.

Es un resultado general que cualquier función de variables aleatorias normalmente distribuidas se distribuye normalmente (Véase Pfeiffer, Paul E.; Concepts of Probability Theory; 2a Edición; Edit. Dover; Nueva York, 1978; p. 263).

Los estimadores de los coeficientes de regresión, los valores ajustados y los residuales observados son funciones lineales de las observaciones originales en la variable dependiente. Consecuentemente, la suposición de normalidad de $\underline{\varepsilon}$ también implica que los vectores aleatorios $\underline{Y}$, $\hat{\underline{\beta}}$, $\hat{\underline{Y}}$ y $\underline{e}$ poseen cada uno una distribución normal multivariada. La media y la varianza de estos vectores ya fueron calculadas anteriormente. Así, con la suposición $\underline{\varepsilon} \sim N(\underline{0}, \mathbf{I}\sigma^2)$, las distribuciones de los vectores aleatorios más importantes en la regresión por mínimos cuadrados se puede resumir como sigue:

$$\begin{aligned}
\underline{Y} &\sim N\left(\mathbf{X}\underline{\beta}, \mathbf{I}\sigma^2\right) \\
\hat{\underline{\beta}} &\sim N\left(\underline{\beta}, \left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2\right) \\
\hat{\underline{Y}} &\sim N\left(\mathbf{X}\underline{\beta}, \mathbf{H}\sigma^2\right) \\
\underline{e} &\sim N\left(\underline{0}, \left(\mathbf{I}-\mathbf{H}\right)\sigma^2\right) \\
\hat{\underline{Y}}_{pred} &\sim N\left(\mathbf{X}\underline{\beta}, \left(\mathbf{I}+\mathbf{H}\right)\sigma^2\right)
\end{aligned}$$

En general, ninguno de los elementos en los vectores aleatorios calculados en la regresión por mínimos cuadrados será independiente de los otros y los elementos fuera de la diagonal principal de cada una de las matrices de varianzas y covarianzas son distintos de cero.

Las pruebas de hipótesis convencionales y los intervalos de confianza para los estimadores de los parámetros está basada en la suposición de que los estimadores se distribuyen normalmente. Es decir, la suposición de normalidad de los errores aleatorios es clave para estos propósitos. Sin embargo, esta suposición no es necesaria para la estimación por mínimos cuadrados. Aun en la ausencia de normalidad, los estimadores por mínimos

cuadrados son los mejores estimadores lineales insesgados (mejores en el sentido de tener la varianza mínima entre todos los estimadores lineales insesgados); en efecto, probemos el siguiente

Teorema 2.- Sea $\hat{\underline{\theta}}$ el estimador por mínimos cuadrados de $\underline{\theta} = \mathbf{X}\underline{\beta}$. Entonces, en la clase de los estimadores insesgados de $\underline{c}'\underline{\theta}$, $\underline{c}'\hat{\underline{\theta}}$ es el único estimador con varianza mínima (decimos que $\underline{c}'\hat{\underline{\theta}}$ es el mejor estimador lineal insesgado de $\underline{c}'\underline{\theta}$).

Demostración.- $\hat{\underline{\theta}} = \mathbf{X}\hat{\underline{\beta}} = \mathbf{H}\hat{\underline{\theta}} = \mathbf{H}\underline{Y}$, donde $\mathbf{HX} = \mathbf{X}$. Entonces $E\left(\underline{c}'\hat{\underline{\theta}}\right) = \underline{c}'\mathbf{H}\underline{\theta} = \underline{c}'\underline{\theta}$, es decir, $\underline{c}'\hat{\underline{\theta}}$ es un estimador lineal insesgado de $\underline{c}'\underline{\theta}$. Sea $\underline{d}'\underline{Y}$ otro estimador lineal insesgado de $\underline{c}'\underline{\theta}$. Entonces $\underline{c}'\underline{\theta} = E\left(\underline{d}'\underline{Y}\right) = \underline{d}'\underline{\theta}$; es decir, $(\underline{c} - \underline{d})'\,\underline{\theta} = 0$, por lo tanto, $\mathbf{H}\left(\underline{c} - \underline{d}\right) = \underline{0}$ y $\mathbf{H}\underline{c} = \mathbf{H}\underline{d}$.

Por otro lado, usando el hecho de que $\mathbf{H}$ e $(\mathbf{I} - \mathbf{H})$ son simétricas e idempotentes:

$$\begin{aligned} Var\left((\mathbf{H}\underline{d})'\,\underline{Y}\right) &= \sigma^2\underline{d}'\mathbf{H}'\mathbf{H}\underline{d} \\ &= \sigma^2\underline{d}'\mathbf{H}^2\underline{d} \\ &= \sigma^2\underline{d}'\mathbf{H}\underline{d} \end{aligned}$$

de esta forma

$$\begin{aligned} Var\left(\underline{d}'\underline{Y}\right) - Var\left(\underline{c}'\hat{\underline{\theta}}\right) &= Var\left(\underline{d}'\underline{Y}\right) - Var\left((\mathbf{H}\underline{d})'\,\underline{Y}\right) \\ &= \sigma^2\left(\underline{d}'\underline{d} - \underline{d}'\mathbf{H}\underline{d}\right) \\ &= \sigma^2\underline{d}'\left(\mathbf{I} - \mathbf{H}\right)\underline{d} \\ &= \sigma^2\underline{d}'\left(\mathbf{I} - \mathbf{H}\right)'\left(\mathbf{I} - \mathbf{H}\right)\underline{d} \geq 0 \end{aligned}$$

donde la igualdad con cero ocurre si y sólo si $\left(\mathbf{I} - \mathbf{H}\right)\underline{d} = \underline{0}$ ó $\underline{d} = \mathbf{H}\underline{d} = \mathbf{H}\underline{c}$. Por lo tanto, $\underline{c}'\hat{\underline{\theta}}$ tiene varianza mínima y es único.

Si la hipótesis de normalidad se mantiene, los estimadores por mínimos cuadrados coinciden con los estimadores por máxima verosimilitud. En efecto, la función de densidad conjunta de $\underline{Y}$ es

$$
\begin{aligned}
f(\underline{Y}) &= f(Y_1)\cdot f(Y_2)\cdot \ldots \cdot f(Y_n) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(Y_i-\beta_0-\beta_1 X_{i1}-\ldots-\beta_k X_{ik})^2\right\} \\
&= \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1 X_{i1}-\ldots-\beta_k X_{ik})^2\right\} \\
&= \frac{1}{2\pi^{n/2}}\frac{1}{(\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(\underline{Y}-\mathbf{X}\underline{\beta})'(\underline{Y}-\mathbf{X}\underline{\beta})\right\}
\end{aligned}
$$

Resolviendo $\frac{\partial \ln(f(\underline{Y}))}{\partial \underline{\beta}} = 0$ y $\frac{\partial \ln(f(\underline{Y}))}{\partial \sigma^2} = 0$ se hallarán los estimadores máximo verosímiles.

$$
\begin{aligned}
\ln(f(\underline{Y})) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\underline{Y}-\mathbf{X}\underline{\beta})'(\underline{Y}-\mathbf{X}\underline{\beta}) \\
\frac{\partial \ln(f(\underline{Y}))}{\partial \underline{\beta}} &= -\frac{1}{2\sigma^2}\frac{\partial}{\partial \underline{\beta}}(\underline{Y}'\underline{Y}-\underline{Y}'\mathbf{X}\underline{\beta}-\underline{\beta}'\mathbf{X}'\underline{Y}+\underline{\beta}'\mathbf{X}'\mathbf{X}\underline{\beta}) \\
&= -\frac{1}{2\sigma^2}\frac{\delta}{\partial \underline{\beta}}(\underline{Y}'\underline{Y}-2\underline{\beta}'\mathbf{X}'\underline{Y}+\underline{\beta}'\mathbf{X}'\mathbf{X}\underline{\beta}) \\
&= -\frac{1}{2\sigma^2}(-2\mathbf{X}'\underline{Y}+2\mathbf{X}'\mathbf{X}\underline{\beta}) = 0
\end{aligned}
$$

lo cual implica $\mathbf{X}'\underline{Y} = \mathbf{X}'\mathbf{X}\underline{\beta}$, que es el sistema de ecuaciones normales. Obsérvese que este método nos provee de un estimador máximo verosímil para σ^2, que se obtiene resolviendo $\frac{\partial \ln(f(\underline{Y}))}{\partial \sigma^2} = 0$.

$$
\begin{aligned}
\frac{\partial \ln(f(\underline{Y}))}{\partial \sigma^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}(\underline{Y}-\mathbf{X}\underline{\beta})'(\underline{Y}-\mathbf{X}\underline{\beta}) \\
\Rightarrow \hat{\sigma}^2 &= \frac{1}{n}\left(\underline{Y}-\mathbf{X}\hat{\underline{\beta}}\right)'\left(\underline{Y}-\mathbf{X}\hat{\underline{\beta}}\right) \\
&= \frac{1}{n}\left(\underline{Y}-\hat{\underline{Y}}\right)'\left(\underline{Y}-\hat{\underline{Y}}\right) \\
&= \frac{1}{n}\underline{e}'\underline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i^2
\end{aligned}
$$

Note que este estimador es distinto del estimador insesgado que se calculó anteriormente.

4.7 Análisis de varianza y formas cuadráticas

Los estimadores de los coeficientes de regresión, las medias estimadas y los residuales han sido presentados en notación matricial; se demostró que todos ellos son funciones

lineales de las observaciones originales en la variable de respuesta, $\underline{Y}$. En esta sección mostraremos que todas las sumas de cuadrados y productos son funciones cuadráticas de $\underline{Y}$. Esto significa que cada suma de cuadrados puede ser escrita como $\underline{Y}'A\underline{Y}$, donde A es una matriz de coeficientes que llamaremos *matriz de definición*. De $\underline{Y}'A\underline{Y}$ diremos que es una *forma cuadrática* en $\underline{Y}$ (El lector encontrará una breve introducción a las formas cuadráticas en el apéndice A).

El objetivo del modelo ajustado es explicar, en la medida de lo posible, la variación en la variable dependiente a partir de la información contenida en las variables independientes. Las contribuciones de las variables independientes al modelo son medidas por particiones de la suma de cuadrados total de $\underline{Y}$ que pueden ser atribuidas a ó explicadas por las variables independientes. Cada componente de la partición de las sumas de cuadrados es una forma cuadrática en $\underline{Y}$. Los grados de libertad asociados a una suma de cuadrados en particular y la ortogonalidad entre diferentes sumas de cuadrados se determinan por las matrices de definición en las formas cuadráticas. La forma matricial de una suma de cuadrados simplifica los cálculos si se tiene acceso a algún paquete estadístico o de álgebra matricial o a una hoja de cálculo. Las esperanzas y varianzas de las sumas de cuadrados también pueden determinarse fácilmente de esta forma.

Se puede verificar que el vector de observaciones en la variable dependiente $\underline{Y}$ puede particionarse como sigue:

$$\underline{Y} = \hat{\underline{Y}} + \underline{e}$$

Esta partición será usada para obtener una partición similar de la suma de cuadrados total de la variable dependiente:

$$\underline{Y}'\underline{Y} = \sum_{i=1}^{n} Y_i^2$$

Esta es una forma cuadrática cuya matriz de definición es la matriz identidad:

$$\underline{Y}'\underline{Y} = \underline{Y}'\mathrm{I}\underline{Y}$$

La matriz identidad es idempotente y su traza igual a su orden, indicando que la suma de cuadrados total (no corregida) tiene tantos grados de libertad como elementos en el vector $\underline{Y}$. La matriz identidad es la única matriz idempotente de rango completo. Dado que

$$\underline{Y} = \hat{\underline{Y}} + \underline{e}$$

se tiene

$$\underline{Y}'\underline{Y} = \left(\hat{\underline{Y}} + \underline{e}\right)' \left(\hat{\underline{Y}} + \underline{e}\right) = \hat{\underline{Y}}'\hat{\underline{Y}} + \hat{\underline{Y}}'\underline{e} + \underline{e}'\hat{\underline{Y}} + \underline{e}'\underline{e}$$

Sustituyendo $\hat{\underline{Y}} = \mathbf{H}\underline{Y}$ y $\underline{e} = (\mathbf{I} - \mathbf{H})\underline{Y}$ se obtiene

$$\begin{aligned} \underline{Y}'\underline{Y} &= (\mathbf{H}\underline{Y})'(\mathbf{H}\underline{Y}) + (\mathbf{H}\underline{Y})'[(\mathbf{I} - \mathbf{H})\underline{Y}] + [(\mathbf{I} - \mathbf{H})\underline{Y}]'(\mathbf{H}\underline{Y}) \\ &\quad + [(\mathbf{I} - \mathbf{H})\underline{Y}]'[(\mathbf{I} - \mathbf{H})\underline{Y}] \\ &= \underline{Y}'\mathbf{H}'\mathbf{H}\underline{Y} + \underline{Y}'\mathbf{H}'(\mathbf{I} - \mathbf{H})\underline{Y} + \underline{Y}'(\mathbf{I} - \mathbf{H})'\mathbf{H}\underline{Y} + \underline{Y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\underline{Y} \end{aligned}$$

Ambas, $\mathbf{H}$ y $(\mathbf{I} - \mathbf{H})$ son simétricas e idempotentes, de tal forma que $\mathbf{H}'\mathbf{H} = \mathbf{H}$ y $(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H}) = (\mathbf{I} - \mathbf{H})$. Los dos términos centrales de la ecuación anterior son cero puesto que las dos formas cuadráticas son ortogonales una con la otra:

$$\mathbf{H}'(\mathbf{I} - \mathbf{H}) = \mathbf{H} - \mathbf{H} = 0$$

De esta forma:

$$\underline{Y}'\underline{Y} = \underline{Y}'\mathbf{H}\underline{Y} + \underline{Y}'(\mathbf{I} - \mathbf{H})\underline{Y} = \hat{\underline{Y}}'\hat{\underline{Y}} + \underline{e}'\underline{e}$$

La suma de cuadrados total no corregida ha sido particionada en dos formas cuadráticas cuyas matrices de definición son $\mathbf{H}$ y $(\mathbf{I} - \mathbf{H})$, respectivamente. $\hat{\underline{Y}}'\hat{\underline{Y}}$ es la parte de $\underline{Y}'\underline{Y}$ que puede ser atribuida al modelo ajustado y la denotaremos por $\mathbf{SC}_{mod}$. El segundo término, $\underline{e}'\underline{e}$, es la parte de $\underline{Y}'\underline{Y}$ que no puede ser explicada por el modelo, es la suma de cuadrados de residuales después de ajustar el modelo y será denotada por $\mathbf{SC}_{res}$. La ortogonalidad de las formas cuadráticas asegura que $\mathbf{SC}_{mod}$ y $\mathbf{SC}_{res}$ sean particiones

aditivas. Los grados de libertad asociados con cada una de ellas depende del rango de sus matrices de definición. El rango de $\mathbf{H} = \left[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]$ está determinado por el rango de $\mathbf{X}$, que es también el número de parámetros en $\underline{\beta}$. Así, los grados de libertad de $\mathbf{SC}_{mod}$ es k' cuando el modelo es de rango completo.

Hemos probado que $rango(\mathbf{H}) = tr(\mathbf{H}) = k'$ y que $rango(\mathbf{I}_n - \mathbf{H}) = tr(\mathbf{I}_n) - tr(\mathbf{H}) = n - k'$. El subíndice en $\mathbf{I}$ indica el orden de la matriz identidad. Los grados de libertad de $\mathbf{SC}_{res}$ son $n-k'$ y son obtenidos utilizando la aditividad de las dos particiones u observando el hecho de que $tr(\mathbf{I}_n - \mathbf{H}) = tr(\mathbf{I}_n) - tr(\mathbf{H}) = n - k'$. El orden de esta matriz identidad es n.

Las expresiones para las formas cuadráticas de la expresión anterior son formas propias para definición; muestran la naturaleza de las sumas de cuadrados que se están calculando. Sin embargo, existen formas de estas expresiones que facilitan los cálculos. La forma más conveniente de $\mathbf{SC}_{mod} = \underline{\hat{Y}}'\underline{\hat{Y}}$ es

$$\mathbf{SC}_{mod} = \underline{\hat{\beta}}'\mathbf{X}'\underline{Y}$$

Esta igualdad se prueba sustituyendo $\mathbf{X}\underline{\hat{\beta}}$ por la primera $\underline{\hat{Y}}$ y $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{Y}$ por la segunda. Así, la suma de cuadrados debida al modelo puede obtenerse sin calcular el vector de valores ajustados o la matriz de $n \times n$, $\mathbf{H}$. El vector $\underline{\hat{\beta}}$ es de orden menor que $\underline{\hat{Y}}$ y $\mathbf{X}'\underline{Y}$ ya estará calculado para cuando $\underline{\hat{\beta}}$ lo esté. Dado que estas dos particiones son aditivas, la forma de cálculo más simple para $\mathbf{SC}_{res}$ es por diferencia:

$$\mathbf{SC}_{res} = \underline{Y}'\underline{Y} - \mathbf{SC}_{mod}$$

Las formas de cálculo y de definición para esta partición de la suma de cuadrados total está resumida en la tabla 2.

Tabla 2: Resumen del análisis de varianza para el análisis de regresión

(Prueba de significancia global)

Fuente de Variación	*Grados de Libertad*	*Suma de Fórmula de Definición*	*Cuadrados Fórmula de Cálculo*
Total (no corregida)	$rango\,(\mathbf{I}_n) = n$	$\underline{Y}'\underline{Y}$	$\underline{Y}'\underline{Y}$
Debida al modelo	$rango\,(\mathbf{H}) = k'$	$\underline{\hat{Y}}'\underline{\hat{Y}} = \underline{Y}'\mathbf{H}\underline{Y}$	$\underline{\hat{\beta}}'\mathbf{X}'\underline{Y}$
Residuales	$rango\,(\mathbf{I}_n - \mathbf{H}) = (n - k')$	$\underline{e}'\underline{e} = \underline{Y}'\,(\mathbf{I} - \mathbf{H})\,\underline{Y}$	$\underline{Y}'\underline{Y} - \underline{\hat{\beta}}'\mathbf{X}'\underline{Y}$

Ejemplo 8.- Continuando con el ejemplo 7; la partición de la suma de cuadrados total será ilustrada usando los datos de ozono de Heagle. La suma de cuadrados total no corregida, con cuatro grados de libertad es

$$\begin{aligned}\underline{Y}'\underline{Y} &= \begin{pmatrix} 242, & 237, & 231, & 201, \end{pmatrix} \begin{pmatrix} 242 \\ 237 \\ 231 \\ 201 \end{pmatrix} \\ &= 242^2 + 237^2 + 231^2 + 201^2 \\ &= 208,495\end{aligned}$$

La suma de cuadrados debida al modelo, $\mathbf{SC}_{mod}$ se obtiene a partir de la fórmula de definición:

$$\begin{aligned}\underline{\hat{Y}}'\underline{\hat{Y}} &= \begin{pmatrix} 247.563, & 232.887, & 221.146, & 209.404 \end{pmatrix} \begin{pmatrix} 247.563 \\ 232.887 \\ 221.146 \\ 209.404 \end{pmatrix} \\ &= 247.563^2 + 232.887^2 + 221.146^2 + 209.404^2 \\ &= 208,279.39\end{aligned}$$

Usando la fórmula de cálculo (que es más conveniente), se tiene

$$\underline{\hat{\beta}}'\mathbf{X}'\underline{Y} = (253.434,\ -293.531)\begin{pmatrix} 911 \\ 76.99 \end{pmatrix} = 208,279.39$$

La fórmula de definición para la suma de cuadrados de los residuales da

$$\underline{e}'\underline{e} = \begin{pmatrix} -5.563, & 4.113, & 9.854, & -8.404, \end{pmatrix}\begin{pmatrix} -5.563 \\ 4.113 \\ 9.854 \\ -8.404 \end{pmatrix} = 215.61$$

Si usamos la fórmula de cálculo se tiene

$$\mathbf{SC}_{res} = \underline{Y}'\underline{Y} - \mathbf{SC}_{mod} = 208,495 - 208,279.39 = 215.61$$

La suma total de cuadrados no corregida ha sido particionada en la suma de cuadrados debida al modelo completo y la suma de cuadrados debida a los residuales. Sin embargo, usualmente se está más interesado en explicar la variación de $\underline{Y}$ en torno a su media que en su variación en torno al cero; otro punto de gran interés es cuánta de la información proveniente de las variables independientes contribuye a esta explicación. Si no se tiene información disponible a partir de las variables independientes, el mejor valor pronosticado de $\underline{Y}$ es el mejor estimador disponible de la media poblacional.

Cuando se dispone de variables independientes, la cuestión de interés es cuánta información contenida en ellas contribuye a la predicción de $\underline{Y}$, además de aquella proveniente de la media total de $\underline{Y}$.

La medida de la información adicional que aportarán las variables independientes, es la diferencia entre $\mathbf{SC}_{mod}$ cuando las variables independientes están incluidas en el modelo y $\mathbf{SC}_{mod}$ cuando no lo están. El modelo sin variables independientes contiene sólo un parámetro, la media total de $\underline{Y}$, μ. Cuando μ es el único parámetro en el modelo, $\mathbf{SC}_{mod}$ será denotado por $\mathbf{SC}_{\mu}$.

$\mathbf{SC}_{\mu}$ es llamado frecuentemente *el factor de corrección*. La suma de cuadrados adicional, que es aportada por las variables independientes, será llamada *suma de cuadrados debida a la regresión*, y denotada por $\mathbf{SC}_{reg}$. Así,

$$\mathbf{SC}_{reg} = \mathbf{SC}_{mod} - \mathbf{SC}_{\mu}$$

donde $\mathbf{SC}_{mod}$ es la suma de cuadrados debida al modelo, incluyendo las variables independientes.

La suma de cuadrados debida solamente a μ, $\mathbf{SC}_{\mu}$, será determinada usando notación matricial, con el fin de mostrar el desarrollo de las matrices de definición para las formas cuadráticas. El modelo, cuando μ es el único parámetro, puede escribirse aun como $\underline{Y} = X\beta + \varepsilon$, pero ahora $\underline{X}$ es sólo un vector columna cuyas entradas son todas iguales a uno y $\beta = \mu$ es un escalar. Si denotamos al vector columna de unos como $\underline{1}$, se tiene

$$\hat{\beta} = (\underline{1}'\underline{1})^{-1}\,\underline{1}'\underline{Y} = \frac{1}{n}\underline{1}'\underline{Y} = \bar{Y}$$

y

$$\begin{aligned}
\mathbf{SC}_{\mu} &= \hat{\beta} = (\underline{1}'\underline{Y}) = \frac{1}{n}(\underline{1}'\underline{Y})'(\underline{1}'\underline{Y}) \\
&= \left(\frac{1}{n}\right)\underline{Y}'(\underline{1}\,\underline{1}')\underline{Y} \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i^2
\end{aligned}$$

la suma de cuadrados corregida por la media.

El producto $\underline{1}\,\underline{1}'$ es una matriz cuadrada de orden n cuyos elementos son todos iguales a uno; denotemos a esta matriz por J. De esta forma, la matriz de definición para la forma cuadrática que da el factor de corrección es

$$\frac{1}{n}(\underline{1}\,\underline{1}') = \frac{1}{n}\begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix} = \frac{1}{n}J$$

La matriz $\left(\frac{1}{n}\right)J$ es idempotente con rango igual a $tr\left(\frac{1}{n}J\right) = 1$, es decir, el factor de corrección tiene un grado de libertad.

La suma de cuadrados adicional que es atribuida a las variables independientes en un modelo es

$$\begin{aligned} \mathbf{SC}_{reg} &= \mathbf{SC}_{mod} - \mathbf{SC}_{\mu} \\ &= \underline{Y}'\mathbf{H}\underline{Y} - \underline{Y}'\left(\frac{J}{n}\right)\underline{Y} \\ &= \underline{Y}'\left(\mathbf{H} - \frac{J}{n}\right)\underline{Y} \end{aligned}$$

Es decir, la matriz de definición para $\mathbf{SC}_{reg}$ es $\left(\mathbf{H} - \frac{J}{n}\right)$. La matriz $\frac{J}{n}$ es ortogonal a $\left(\mathbf{H} - \frac{J}{n}\right)$ y a $(\mathbf{I} - \mathbf{H})$, así que la suma total de cuadrados se ha descompuesto en las tres siguientes componentes ortogonales:

$$\begin{aligned} \underline{Y}'\underline{Y} &= \underline{Y}'\left(\frac{J}{n}\right)\underline{Y} + \underline{Y}'\left(\mathbf{H} - \frac{J}{n}\right)\underline{Y} + \underline{Y}'(\mathbf{I} - \mathbf{H})\underline{Y} \\ &= \mathbf{SC}_{\mu} + \mathbf{SC}_{reg} + \mathbf{SC}_{res} \end{aligned}$$

con 1, $(k' - 1) = k$ y $(n - k')$ grados de libertad, respectivamente.

Usualmente, $\mathbf{SC}_{\mu}$ es sustraída de $\underline{Y}'\underline{Y}$ y solamente la suma de cuadrados corregida, descompuesta en $\mathbf{SC}_{reg}$ y $\mathbf{SC}_{res}$ son reportadas.

Tabla 3: Resumen del análisis de varianza en regresión lineal

(Prueba de significancia global)

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrado Medio*
Total (no corregida)	n	$\underline{Y}'\underline{Y}$	$\frac{\underline{Y}'\underline{Y}}{n}$
Media	1	$n\bar{Y}^2$	$n\bar{Y}^2$
Total (corregida)	$(n-1)$	$\underline{Y}'\underline{Y} - n\bar{Y}^2$	$\frac{\underline{Y}'\underline{Y} - n\bar{Y}^2}{(n-1)}$
Regresión	k	$\underline{\hat{\beta}}'\mathbf{X}'\underline{Y} - n\bar{Y}^2$	$\frac{\underline{\hat{\beta}}'\mathbf{X}'\underline{Y} - n\bar{Y}^2}{k}$
Residuales	$(n-k')$	$\underline{Y}'\underline{Y} - \underline{\hat{\beta}}'\mathbf{X}'\underline{Y}$	$\frac{\underline{Y}'\underline{Y} - \underline{\hat{\beta}}'\mathbf{X}'\underline{Y}}{(n-k')}$

Ejemplo 9.- En los datos de Heagle:

$$\mathbf{SC}_{\mu} = \frac{911^2}{4} = 207,480.25$$

de esta forma

$$\mathbf{SC}_{reg} = 208,279.39 - 207,480.25 = 799.14$$

La tabla de análisis de varianza para los datos de Heagle se muestra a continuación

Tabla 4

Fuente de Variación	*Grados de Libertad*	*Suma de Cuadrados*	*Cuadrado Medio*
Total (no corregida)	4	208,279.39	52,069.85
Media	1	207,480.25	207,480.25
Total (corregida)	3	1,014.75	338.25
Regresión	1	799.14	799.14
Residuales	2	215.61	107.81

4.8 Esperanza de formas cuadráticas

Cada una de las formas cuadráticas calculadas en el análisis de varianza de $\underline{Y}$ estima alguna función sobre los parámetros del modelo. Las esperanzas de estas formas cuadráticas deben conocerse si deseamos hacer un uso apropiado de las sumas de cuadrados y de sus cuadrados medios. En el teorema 1 de este capítulo mostramos la forma general de la esperanza de formas cuadráticas. Bajo las suposiciones ordinarias del modelo, $E(\underline{Y}) = \mathbf{X}\underline{\beta}$ y $Var(\underline{Y}) = \mathbf{I}\sigma^2$, la esperanza de las formas cuadráticas es:

$$E(\underline{Y}'A\underline{Y}) = \sigma^2 tr(A) + \underline{\beta}'\mathbf{X}'A\mathbf{X}\underline{\beta}$$

Las esperanzas de las formas cuadráticas en el análisis de varianza se obtienen a partir de esta forma general, reemplazando A por la matriz de definición apropiada. Cuando A es idempotente, el coeficiente en σ^2 son los grados de libertad para la forma cuadrática.

$$\begin{aligned} E(\mathbf{SC}_{mod}) &= E(\underline{Y}'\mathbf{H}\underline{Y}) = \sigma^2 tr(\mathbf{H}) + \underline{\beta}'\mathbf{X}'\mathbf{H}\mathbf{X}\underline{\beta} \\ &= k'\sigma^2 + \underline{\beta}'\mathbf{X}'\mathbf{X}\underline{\beta} \end{aligned}$$

dado que $tr(\mathbf{H}) = k'$ y $\mathbf{HX} = \mathbf{X}$. Note que el segundo término en la ecuación anterior es una forma cuadrática en $\underline{\beta}$.

$$\begin{aligned} E(\mathbf{SC}_{reg}) &= E\left[\underline{Y}'\left(\mathbf{H}-\frac{J}{n}\right)\underline{Y}\right] \\ &= \sigma^2 tr\left(\mathbf{H}-\frac{J}{n}\right)+\underline{\beta}'\mathbf{X}'\left(\mathbf{H}-\frac{J}{n}\right)\mathbf{X}\underline{\beta} \\ &= k\sigma^2+\underline{\beta}'\mathbf{X}'\left(\mathbf{H}-\frac{J}{n}\right)\mathbf{X}\underline{\beta} \end{aligned}$$

dado que $\mathbf{X}'\mathbf{H} = \mathbf{X}'$. Esta forma cuadrática en $\underline{\beta}$ difiere de la anterior en que $\mathbf{X}'\left(\mathbf{H}-\frac{J}{n}\right)\mathbf{X}$ es una matriz de sumas de cuadrados corregidas y de productos de las X_j. Dado que la primera columna de $\mathbf{X}$ es constante, las sumas de cuadrados y productos que involucran a la primera columna son cero. Así, la primera columna y el primer renglón de $\mathbf{X}'\left(\mathbf{H}-\frac{J}{n}\right)\mathbf{X}$ contienen sólo ceros, lo cual remueve β_0 de la expresión cuadrática. Sólo los coeficientes de regresión de las variables independientes aparecen en la esperanza de la suma de cuadrados debida a la regresión.

$$\begin{aligned} E(\mathbf{SC}_{res}) &= E[\underline{Y}'(\mathbf{I}-\mathbf{H})\underline{Y}] \\ &= \sigma^2 tr(\mathbf{I}-\mathbf{H})+\underline{\beta}'\mathbf{X}'(\mathbf{I}-\mathbf{H})\mathbf{X}\underline{\beta} \\ &= \sigma^2(n-k')+\underline{\beta}'\mathbf{X}'(\mathbf{X}-\mathbf{X})\underline{\beta} \\ &= \sigma^2(n-k') \end{aligned}$$

El coeficiente de σ^2 en cada esperanza son los grados de libertad de la suma de cuadrados. Después de dividir cada una de ellas por sus grados de libertad para convertir las sumas de cuadrados en cuadrados medios, el coeficiente en σ^2 será uno en cada caso:

$$\begin{aligned} E(\mathbf{CM}_{reg}) &= \sigma^2+\frac{\underline{\beta}'\mathbf{X}'\left(\mathbf{H}-\frac{J}{n}\right)\mathbf{X}\underline{\beta}}{k} \\ E(\mathbf{CM}_{res}) &= \sigma^2 \end{aligned}$$

Esto muestra que el cuadrado medio de los residuales, $\mathbf{CM}_{res}$, es un estimador insesgado para σ^2. El cuadrado medio de la regresión es un estimador de σ^2 más una forma cuadrática para todas las β_j, exceptuando β_0. La comparación de $\mathbf{CM}_{res}$ y $\mathbf{CM}_{reg}$, por lo tanto, provee una base para juzgar la importancia de los coeficientes de regresión

o, equivalentemente, de las variables independientes. Dado que el segundo término en $E(\mathbf{CM}_{reg})$ es una forma cuadrática en $\underline{\beta}$, misma que no puede ser negativa, se sigue que cualquier contribución proveniente de las variables independientes a la predicción de Y_i hace que la esperanza de $\mathbf{CM}_{reg}$ sea mayor que la esperanza de $\mathbf{CM}_{res}$.

Las esperanzas suponen que el modelo usado en el análisis de varianza es, de hecho, el modelo correcto. Esta hipótesis es impuesta cuando se sustituye $E(\underline{Y})$ por $\mathbf{X}\underline{\beta}$. Si ha sido usado un modelo incorrecto, $E(\underline{Y}) \neq \mathbf{X}\underline{\beta}$ y el segundo término de $E(\mathbf{SC}_{res})$ no se hace cero. En lugar de ello, permanece una función cuadrática de los coeficientes de regresión de cualesquiera otras variables independientes -de importancia- que hubieran sido omitidas por error en el modelo. En este caso, $\mathbf{CM}_{res}$ será seguramente un estimador sesgado de σ^2.

Ejemplo 10.- Usando los datos de Heagle, el estimador de σ^2 obtenido a partir de $\mathbf{CM}_{res}$ es $\hat{\sigma}^2 = 107.81$ (tabla 3).

Este es un estimador muy pobre de σ^2, pues tiene solamente 2 grados de libertad. Sin embargo, este estimador de σ^2 es el que será usado en adelante.

4.8.1 Varianzas estimadas o estimación de varianzas

Hasta ahora, las matrices de varianzas y covarianzas para $\underline{\hat{\beta}}$, $\underline{\hat{Y}}$ y $\underline{e}$ han sido expresadas en términos de la verdadera varianza, σ^2. Los estimadores de las matrices de varianzas y covarianzas de estos vectores se obtienen al sustituir algún estimador de σ^2 en la fórmula correspondiente.

Ejemplo 11.- En el ejemplo del ozono y usando $\mathbf{CM}_{res} = \hat{\sigma}^2 = 107.81$ como un estimador de σ^2 se tiene:

$$\begin{aligned}
\widehat{Var}\left(\underline{\hat{\beta}}\right) &= (\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}^2 \\
&= \begin{pmatrix} 1.0755 & -9.4340 \\ -9.4340 & 107.8167 \end{pmatrix} 107.81 \\
&= \begin{pmatrix} 115.94 & -1,017.0 \\ -1,017.0 & 11,623 \end{pmatrix}
\end{aligned}$$

De esta manera:

$$\widehat{Var}\left(\hat{\beta}_0\right) = (1.0755)(107.81) = 115.94$$

$$\widehat{Var}\left(\hat{\beta}_1\right) = (107.8167)(107.81) = 11,623$$

$$\widehat{Cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) = (-9.4340)(107.81) = -1.017.0$$

En cada caso, el primer número en el producto es el coeficiente apropiado a partir de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$; el segundo término es $\hat{\sigma}^2$ (es coincidencia que el elemento inferior derecho sea tan parecido a $\hat{\sigma}^2$). Las matrices estimadas de varianzas y covarianzas para $\underline{\hat{Y}}$ y $\underline{e}$ se encuentran de manera similar, sustituyendo $\hat{\sigma}^2$ en las expresiones correspondientes.

4.9 Distribución de formas cuadráticas

La distribución probabilística de las formas cuadráticas nos da una base para desarrollar pruebas de hipótesis paramétricas. Es en este punto y al construir intervalos de confianza que las suposiciones básicas del modelo se vuelven cruciales. En esta sección se probarán y enunciarán algunos resultados que serán de suma utilidad en adelante.

Teorema 3.- Sea $\underline{Y}$ un vector aleatorio de tamaño n, distribuido $N\left(\mathbf{X}\underline{\beta}, \sigma^2\mathbf{I}_n\right)$, donde X es una matriz de tamaño $(n \times k')$, de rango k'. Entonces:

i) $\underline{Y} \sim N\left(\mathbf{X}\underline{\beta}, \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$

ii) $\left(\underline{\hat{\beta}} - \underline{\beta}\right)'\left(\mathbf{X}'\mathbf{X}\right)\left(\underline{\hat{\beta}} - \underline{\beta}\right)/\sigma^2 \sim \chi^2_{(k')}$

iii) $\underline{\hat{\beta}}$ es independiente de $\dfrac{\underline{e}'\underline{e}}{(n-k')} = \hat{\sigma}^2$

iv) $(n-k')\dfrac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k')}$

Demostración.- i) fue probado con anterioridad. Para probar ii) expresemos primero a

$$\left(\underline{\hat{\beta}} - \underline{\beta}\right)'\left(\mathbf{X}'\mathbf{X}\right)\left(\underline{\hat{\beta}} - \underline{\beta}\right)/\sigma^2$$

como

$$\left(\underline{\hat{\beta}} - \underline{\beta}\right)' \left(Var\left(\underline{\hat{\beta}}\right)^{-1}\right) \left(\underline{\hat{\beta}} - \underline{\beta}\right)$$

una forma cuadrática. Para continuar con la demostración, necesitamos probar el si - guiente resultado auxiliar:

Sea $\underline{Y}$ un vector aleatorio de tamaño n tal que $\underline{Y} \sim N(\underline{\theta}, \Sigma)$, con Σ definida positiva; entonces la forma cuadrática Q es tal que

$$Q = (\underline{Y} - \underline{\theta})' \Sigma^{-1} (\underline{Y} - \underline{\theta}) \sim \chi^2_{(n)}$$

En efecto, si Σ es definida positiva, entonces existe una matriz ortogonal T tal que

$$T'\Sigma T = diag(\lambda_1, \lambda_2, ..., \lambda_n) = \Lambda$$

donde $(\lambda_1, \lambda_2, ..., \lambda_n)$ son los vectores propios de Σ.

De esta forma se tiene

$$T'\Sigma T = \Lambda \Rightarrow \Sigma T = T\Lambda \Rightarrow \Sigma = T\Lambda T'$$

Definamos la transformación $\underline{Y} - \underline{\theta} = T\underline{X}$ donde $\underline{X} \sim N(\underline{0}, \Lambda)$, de aquí se obtiene $E(\underline{Y} - \underline{\theta}) = \underline{0}$ ó $E(\underline{Y}) = \underline{\theta}$. Así,

$$\begin{aligned} Var(\underline{Y}) &= E\left[(\underline{Y} - \underline{\theta})(\underline{Y} - \underline{\theta})'\right] \\ &= E\left[(T\underline{X})(T\underline{X})'\right] = E\left[T\underline{X}\underline{X}'T'\right] \\ &= T'E(\underline{X}\underline{X})'T' \\ &= T'Var(\underline{X})'T' \\ &= T\Lambda T' \\ &= \Sigma \end{aligned}$$

Por lo tanto, la forma cuadrática Q puede ser escrita de la siguiente forma:

$$Q = \underline{X}'T'\Sigma^{-1}T\underline{X}$$

donde $T'\Sigma^{-1}T = \Lambda^{-1}$. Así

$$Q = \underline{X}'\Lambda^{-1}\underline{X} = \sum_{i=1}^{n} \frac{X_i}{\lambda_i}$$

Sin embargo, $X_i \sim N(\underline{0}, \sigma_i^2)$ con $\sigma_i^2 = \lambda_i$. De esta forma, dado que el cuadrado de variables aleatorias normales con media cero se distribuye χ^2, $\frac{X_i}{\lambda_i} \sim \chi^2_{(1)}$. Así, la suma de las n variables aleatorias χ^2,

$$Q = \underline{X}'\Lambda^{-1}\underline{X} = \sum_{i=1}^{n} \frac{X_i}{\lambda_i} \sim \chi^2_{(n)}.$$

Usando este resultado y haciendo las siguientes sustituciones:

$$(\underline{Y} - \underline{\theta}) = \left(\underline{\hat{\beta}} - \underline{\beta}\right) \quad \text{y} \quad Var\left(\underline{\hat{\beta}}\right) = \Sigma$$

se prueba que

$$\left(\underline{\hat{\beta}} - \underline{\beta}\right)' \left(Var\left(\underline{\hat{\beta}}\right)^{-1}\right) \left(\underline{\hat{\beta}} - \underline{\beta}\right) \sim \chi^2_{(k')}$$

iii)Bajo las hipótesis usuales del modelo, basta probar $Cov\left(\underline{\hat{\beta}}, \hat{\sigma}^2\right) = 0$.

Para ello, basta probar

$$Cov\left(\underline{\hat{\beta}}, \underline{Y} - \mathbf{X}\underline{\hat{\beta}}\right) = 0$$

En general, la covarianza entre vectores aleatorios preserva las propiedades de la covarianza entre variables aleatorias reales. De esta forma

$$\begin{aligned} Cov\left(\underline{\hat{\beta}}, \underline{Y} - \mathbf{X}\underline{\hat{\beta}}\right) &= Cov\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\underline{Y}, (\mathbf{I}_n - \mathbf{H})\underline{Y}\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' Cov\left(\underline{Y}, \underline{Y}\right)(\mathbf{I}_n - \mathbf{H}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' Var\left(\underline{Y}\right)(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = 0 \end{aligned}$$

es decir, $\underline{\hat{\beta}}$ y $\underline{e}$ son independientes.

iv)Para demostrar este inciso usaremos los siguientes resultados:

Resultado a).- Sea $\underline{Y}$ un vector aleatorio de tamaño n tal que $\underline{Y} \sim N(\underline{\theta}, \sigma^2\mathbf{I}_n)$; sean $\underline{U} = A\underline{Y}$ y $\underline{V} = B\underline{Y}$, donde A y B son matrices cuadradas de orden n. Denotemos por A_1 a la matriz compuesta por los renglones linealmente independientes de A y por $\underline{U}_1$ a $A_1\underline{Y}$. Entonces, si $Cov(\underline{U}, \underline{V}) = 0$, se tiene:

i) U_1 es independiente de $\underline{V}'\underline{U}_1$

ii) $\underline{U}'\underline{U}$ y $\underline{V}'\underline{V}$ son independientes

Resultado b).-Sean $\underline{Y}$ un vector aleatorio de tamaño n tal que $\underline{Y} \sim N(\underline{\theta}, \sigma^2\mathbf{I}_n)$, P una matriz simétrica de rango r. Entonces, la forma cuadrática $Q = (\underline{Y} - \underline{\theta})' P (\underline{Y} - \underline{\theta}) \sim \chi^2_{(r)}$ si y sólo si P es una matriz cuadrada e idempotente.

Debemos demostrar $(n - k')\dfrac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k')}$.

Sean $\underline{U}_1 = \underline{\hat{\beta}}$ y $\underline{V} = \underline{Y} - \mathbf{X}\underline{\hat{\beta}}$, se tiene que $\underline{\hat{\beta}}$ es independiente de $\underline{e}'\underline{e}$. Note que

$$\begin{aligned} \underline{e}'\underline{e} &= \underline{Y}'(\mathbf{I}_n - \mathbf{H})\underline{Y} \\ &= (\underline{Y} - \mathbf{X}\underline{\beta})'(\mathbf{I}_n - \mathbf{H})(\underline{Y} - \mathbf{X}\underline{\beta}) \\ &= \underline{\varepsilon}'(\mathbf{I}_n - \mathbf{H})\underline{\varepsilon} \end{aligned}$$

dado que $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$. Como $(\mathbf{I}_n - \mathbf{H})$ es simétrica e idempotente y $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2\mathbf{I}_n)$, se tiene que $\underline{e}'\underline{e} \sim \chi^2_{(n-k')}$ y, por lo tanto, $(n - k')\dfrac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k')}$.

Teorema 4.- Sea $\underline{Z}$ un vector aleatorio tal que $\underline{Z} \sim N(\underline{\mu}, \Sigma)$, donde Σ tiene rango n. La forma cuadrática $U = \underline{Y}'A\underline{Y}$ se distribuye $\chi^2_{(k',l)}$, donde $l = \dfrac{1}{2}\underline{\mu}'A\underline{\mu}$ si y sólo si cualesquiera de las siguientes tres condiciones se satisface:

i) $A\Sigma$ es una matriz idempotente de rango k'

ii) ΣA es una matriz idempotente de rango k'

iii) Σ es la c-inversa de A y A tiene rango k'

Teorema 5.- Sea $\underline{Y}$ un vector aleatorio tal que $\underline{Y} \sim N(\underline{\mu}, \Sigma)$, donde Σ tiene rango n. Si $A\Sigma B = 0$, entonces las formas cuadráticas $\underline{Y}'A\underline{Y}$ y $\underline{Y}'B\underline{Y}$ son independientes.

(Una demostración de de los teoremas 4 y 5 puede hallarse en *Graybill, G. A., Theory and application of the linear model, cap. 4, USA, 1976*)

4.10 Pruebas de hipótesis: La hipótesis lineal general

En esta sección desarrollaremos y discutiremos una prueba para la hipótesis $C\underline{\beta} = \underline{\gamma}$ para el modelo lineal general. En secciones subsecuentes se discutirán algunos casos especiales de esta prueba.

Las pruebas que son generalmente de interés son aquellas que involucran funciones lineales de β_i, donde β_i es desconocida. Ilustraremos algunos de estos ejemplos.

Ejemplo 12.- Considere el modelo lineal simple

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, 3, ..., n$$

Los siguientes son ejemplos de pruebas que pueden resultar de interés:

1. $\beta_0 = 0$ (o $\beta_0 = b_0$, donde b_0 es una constante dada); es decir, probar si la intercepción es igual a cero (o a una constante dada b_0).

2. $\beta_1 = 1$ (o $\beta_1 = b_1$, donde b_1 es una constante dada); es decir, probar si la pendiente de la recta es igual a uno (o igual a una constante dada b_1).

Ejemplo 13.- Considere el modelo lineal general

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i \qquad i = 1, 2, 3, ..., n$$

Los siguientes son ejemplos de pruebas que pueden resultar de interés:

1. $\beta_1 = \beta_2$; probar si β_1 es igual a β_2, ignorando los valores de los parámetros restantes.

2. $\beta_1 = \beta_2$ y $\beta_3 = \beta_4$; probar si β_1 es igual a β_2 y β_3 es igual a β_4, ignorando el valor de β_0.

3. $\beta_1 = \beta_2 = 6$. Esta es una prueba diferente a la primera, puesto que en este caso la hipótesis es que β_1 y β_2 son ambas iguales a 6, mientras que en la otra sólo se plantea que $\beta_1 = \beta_2$ sin especificar cuál es el valor común de los parámetros.

4. $\beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$; probar si todos los parámetros son iguales a cero en forma simultánea.

5. $\beta_1 = \beta_2 = \beta_3 = \beta_4$; probar si todas las β_j son iguales entre sí pero sin especificar cuál es el valor común.

6. $$\left.\begin{array}{rcl} \beta_1 - 2\beta_2 &=& 4\beta_3 \\ \beta_1 + 2\beta_2 &=& 6 \end{array}\right\}$$

una prueba acerca de dos relaciones lineales específicas entre los parámetros.

Nótese que cada una de las hipótesis que se describieron en los dos ejemplos anteriores son casos especiales de

$$C\underline{\beta} = \underline{\gamma}$$

donde C es una matriz dada de tamaño $r \times k'$ y $\underline{\gamma}$ es un vector conocido de orden $r \times 1$. El vector $\underline{\beta}$ es el vector de los parámetros del modelo. A continuación describiremos C y $\underline{\gamma}$ para cada uno de las hipótesis del ejemplo 13.

1. $C = \begin{pmatrix} 0, & 1, & -1, & 0, & 0 \end{pmatrix}$; $\underline{\gamma} = 0$

2. $C = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$; $\underline{\gamma} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

3. $C = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$; $\underline{\gamma} = \begin{pmatrix} 6 \\ 6 \end{pmatrix}$

4. $C = \mathbf{I}_5$; $\underline{\gamma} = 0$

5. $C = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix}$; $\underline{\gamma} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$

6. $C = \begin{pmatrix} 0 & 1 & -2 & -4 & 0 \\ 0 & 1 & -2 & 0 & 0 \end{pmatrix}$; $\underline{\gamma} = \begin{pmatrix} 0 \\ 6 \end{pmatrix}$

El lector puede verificar cada hipótesis simplemente efectuando el producto y evaluando $C\underline{\beta} = \underline{\gamma}$ para cada caso.

Nótese que C y $\underline{\gamma}$ no son necesariamente únicos. Por ejemplo, considere la hipótesis 5 del ejemplo 13; ésta puede ser escrita como $C_1\underline{\beta} = \underline{\gamma}_1$ ó $C_2\underline{\beta} = \underline{\gamma}_2$, donde C_1 y C_2 están dadas abajo y $\underline{\gamma}_1 = \underline{\gamma}_2 = 0$

$$C_1 = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 1 & -2 & 0 \\ 0 & 1 & 1 & 1 & -3 \end{pmatrix} \qquad C_2 = \begin{pmatrix} 0 & -1 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

El lector puede verificar fácilmente que la hipótesis es cierta si y sólo si $C_1\underline{\beta} = \underline{\gamma}_1$ y si y sólo si $C_2\underline{\beta} = \underline{\gamma}_2$.

A continuación derivaremos una prueba para $H_0 : C\underline{\beta} = \underline{\gamma}$ vs $H_a : C\underline{\beta} \neq \underline{\gamma}$ para el modelo lineal general $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$, $\underline{\varepsilon} \sim N(\underline{\varepsilon}, \sigma^2\mathbf{I}_n)$ con las siguientes restricciones:

1. $C\underline{\beta} = \underline{\gamma}$ tiene solución.

2. La matriz de $r \times k'$, C, tiene rango r (esta restricción se requiere para que el producto $C'C$ sea de rango completo).

Se usará la prueba de razón de verosimilitudes generalizada. Definamos primero los espacios paramétricos:

$$\Theta_{H_0} = \left\{(\underline{\beta}, \sigma^2) \mid \underline{\beta} \in \Re^{k'},\ C\underline{\beta} = \underline{\gamma},\ 0 < \sigma^2 < \infty\right\}$$

y

$$\Theta = \left\{(\underline{\beta}, \sigma^2) \mid \underline{\beta} \in \Re^{k'},\ 0 < \sigma^2 < \infty\right\}$$

$L(\Theta_{H_0})$ y $L(\Theta)$ denotan a las funciones de verosimilitud en los espacios paramétricos correspondientes y $L\left(\widehat{\Theta_{H_0}}\right)$ y $L\left(\widehat{\Theta}\right)$ denotan a las funciones evaluadas en el punto donde alcanzan su valor máximo. De esta forma, la región crítica, utilizando cociente de verosimilitudes es

$$R_c = \left\{ \underline{Y} \mid \frac{L\left(\widehat{\Theta_{H_0}}\right)}{L\left(\widehat{\Theta}\right)} \leq q \right\}$$

El denominador del cociente de verosimilitudes se obtiene de manera directa, puesto que el problema se resolvió cuando se hallaron los estimadores máximo verosímiles para $\underline{\beta}$ y σ^2. Recuérdese que

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

y

$$\hat{\sigma}^2 = \frac{1}{n}\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)'\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)$$

de esta manera

$$\begin{aligned}
L(\Theta) &= \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left\{ -\frac{1}{2\sigma^2}\left[(Yi - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}))^2 \right] \right\} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Yi - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}))^2 \right\} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(\underline{Y} - \mathbf{X}\underline{\beta}\right)'\left(\underline{Y} - \mathbf{X}\underline{\beta}\right) \right\} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} n\sigma^2 \right\} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}} \\
&= (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\frac{n}{2}}
\end{aligned}$$

y $L\left(\widehat{\Theta}\right) = (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} e^{-\frac{n}{2}}$

Para obtener el numerador, $L\left(\widehat{\Theta_{H_0}}\right)$, se tiene que maximizar $L\left(\underline{\beta}, \sigma^2 \mid \underline{Y}\right)$ sujeto a la restricción $C\underline{\beta} = \underline{\gamma}$. Para ello, se procedería a obtener las derivadas parciales de $L\left(\underline{\beta}, \sigma^2 \mid \underline{Y}\right)$ con respecto a $\underline{\beta}$ y σ^2. Sin embargo, al derivar con respecto a $\underline{\beta}$ tenemos que imponer la restricción $C\underline{\beta} = \underline{\gamma}$. Por otra parte $L\left(\underline{\beta}, \sigma^2 \mid \underline{Y}\right)$ alcanza su valor máximo con respecto a $\underline{\beta}$ cuando $\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)'\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)$ es mínimo, por lo cual el problema es mi-

nimizar $(\underline{Y} - \mathbf{X}\underline{\beta})' (\underline{Y} - \mathbf{X}\underline{\beta})$ sujeto a $C\underline{\beta} = \underline{\gamma}$. Para ello, utilizaremos el procedimiento de multiplicadores de Lagrange.

Sea

$$G(\underline{\beta}, \underline{\lambda}) = (\underline{Y} - \mathbf{X}\underline{\beta})' (\underline{Y} - \mathbf{X}\underline{\beta}) + 2\underline{\lambda}' (C\underline{\beta} - \underline{\gamma})$$

donde

$$\underline{\lambda} \in \Re^{k'}$$

se tiene

$$\frac{\partial G}{\partial \underline{\beta}} = \frac{\partial}{\partial \underline{\beta}} \left[\underline{Y}'\underline{Y} - 2\underline{\beta}'\mathbf{X}'\underline{Y} + \underline{\beta}'\mathbf{X}'\mathbf{X}\underline{\beta} + 2\underline{\lambda}' (C\underline{\beta} - \underline{\gamma})\right]$$

donde $2\underline{\lambda}' (C\underline{\beta} - \underline{\gamma}) = 2 (\underline{\beta}'C' - \underline{\gamma}') \underline{\lambda}$

De esta manera

$$\begin{aligned}
\frac{\partial G}{\partial \underline{\beta}} &= -2\mathbf{X}'\underline{Y} + 2\mathbf{X}'\mathbf{X}\underline{\beta} + 2C'\underline{\lambda} \\
\frac{\partial G}{\partial \underline{\lambda}} &= 2 (C\underline{\beta} - \underline{\gamma}) \\
\frac{\partial G}{\partial \underline{\beta}} &= \underline{0} \Leftrightarrow \mathbf{X}'\mathbf{X}\tilde{\underline{\beta}} + C'\tilde{\underline{\lambda}} = \mathbf{X}'\underline{Y} \quad ...(1) \\
\frac{\partial G}{\partial \underline{\lambda}} &= \underline{0} \Leftrightarrow C\tilde{\underline{\beta}} = \underline{\gamma} \qquad ...(2)
\end{aligned}$$

De (1) obtenemos

$$\begin{aligned}
\tilde{\underline{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1} \left(\mathbf{X}'\underline{Y} - C'\tilde{\underline{\lambda}}\right) \\
&= \hat{\underline{\beta}} - (\mathbf{X}'\mathbf{X})^{-1} C'\tilde{\underline{\lambda}} \quad ...(3)
\end{aligned}$$

Sustituyendo $\tilde{\underline{\beta}}$ en (2)

$$\begin{aligned}
C\tilde{\underline{\beta}} &= C\hat{\underline{\beta}} - C(\mathbf{X}'\mathbf{X})^{-1} C'\tilde{\underline{\lambda}} = \underline{\gamma} \\
\Rightarrow \tilde{\underline{\lambda}} &= \left[C(\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1} \left(C\hat{\underline{\beta}} - \underline{\gamma}\right)
\end{aligned}$$

Sustituyendo en la ecuación (3) anterior

$$\tilde{\underline{\beta}} = \hat{\underline{\beta}} - (\mathbf{X}'\mathbf{X})^{-1} C' \left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1} \left(C\hat{\underline{\beta}} - \underline{\gamma}\right) \quad ...(4)$$

Entonces $\tilde{\underline{\beta}}$ minimiza $L\left(\underline{\beta}, \sigma^2 \mid \underline{Y}\right)$ sujeto a la restricción $C\underline{\beta} = \underline{\gamma}$. Aun falta derivar $L\left(\underline{\beta}, \sigma^2 \mid \underline{Y}\right)$ con respecto a σ^2 sujeto a la restricción $C\underline{\beta} = \underline{\gamma}$. Por facilidad, sea $l = \ln\left(L\left(\underline{\beta}, \sigma^2 \mid \underline{Y}\right)\right)$; maximizar l es equivalente a maximizar $L\left(\underline{\beta}, \sigma^2 \mid \underline{Y}\right)$, ya que la función ln es una función monótona.

$$\begin{aligned} l &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)'\left(\underline{Y} - \mathbf{X}\underline{\beta}\right) \\ \frac{\partial l}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\left(\underline{Y} - \mathbf{X}\underline{\beta}\right)'\left(\underline{Y} - \mathbf{X}\underline{\beta}\right) \\ \frac{\partial l}{\partial \sigma^2} &= 0 \Leftrightarrow \tilde{\sigma}^2 = \frac{1}{n}\left(\underline{Y} - \mathbf{X}\tilde{\underline{\beta}}\right)'\left(\underline{Y} - \mathbf{X}\tilde{\underline{\beta}}\right) \end{aligned}$$

con $\tilde{\underline{\beta}}$ arriba obtenida. Entonces

$$L\left(\widehat{\Theta_{H_0}}\right) = \left(2\pi\tilde{\sigma}^2\right)^{-n/2} e^{-n/2}$$

De esta forma, el cociente de verosimilitudes es:

$$\Lambda = \frac{L\left(\widehat{\Theta_{H_0}}\right)}{L\left(\widehat{\Theta}\right)} = \frac{\left(2\pi\tilde{\sigma}^2\right)^{-n/2} e^{-n/2}}{\left(2\pi\hat{\sigma}^2\right)^{-n/2} e^{-n/2}}$$

Se rechaza H_0 si y sólo si $\Lambda \leq q$,

$$\begin{aligned} \Lambda \leq q &\Leftrightarrow \left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}\right)^{-n/2} \leq q \\ &\Leftrightarrow \frac{\hat{\sigma}^2}{\tilde{\sigma}^2} \leq q_1 \\ &\Leftrightarrow \frac{\left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right)'\left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right)}{\left(\underline{Y} - \mathbf{X}\tilde{\underline{\beta}}\right)'\left(\underline{Y} - \mathbf{X}\tilde{\underline{\beta}}\right)} \leq q_1 \end{aligned}$$

Sea $\mathbf{SCE}_{MC} = \left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right)'\left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right)$. Esta suma se obtuvo sin ninguna restricción, es decir, considerando el modelo $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$, al cual llamaremos modelo completo.

Sea $\mathbf{SCE}_{MR} = \left(\underline{Y} - \mathbf{X}\tilde{\underline{\beta}}\right)'\left(\underline{Y} - \mathbf{X}\tilde{\underline{\beta}}\right)$. Esta suma de cuadrados se obtuvo al imponer la restricción $C\underline{\beta} = \underline{\gamma}$. El modelo se reduce dependiendo de la forma de C y $\underline{\gamma}$. Entonces

$$\Lambda = \frac{\mathbf{SCE}_{MC}}{\mathbf{SCE}_{MR}}$$

Dada la partición fundamental de las sumas de cuadrados, se tiene que

$$\mathbf{SCE}_{MR} = \mathbf{SCE}_{MC} + \mathbf{SCE}_{H_0}$$

Se rechaza H_0 si

$$\frac{\mathbf{SCE}_{MC}}{\mathbf{SCE}_{MR}} \leq q_1 \Leftrightarrow \frac{\mathbf{SCE}_{MR}}{\mathbf{SCE}_{MC}} \geq q_2$$

$$\Leftrightarrow \frac{\mathbf{SCE}_{MC} + \mathbf{SCE}_{H_0}}{\mathbf{SCE}_{MC}} \geq q_2$$

$$\Leftrightarrow 1 + \frac{\mathbf{SCE}_{H_0}}{\mathbf{SCE}_{MC}} \geq q_2 \Leftrightarrow \frac{\mathbf{SCE}_{H_0}}{\mathbf{SCE}_{MC}} \geq q_3$$

De esta manera la región crítica es

$$R_c = \left\{ \underline{Y} \mid \frac{\mathbf{SCE}_{H_0}}{\mathbf{SCE}_{MC}} \geq q_3 \right\}$$

con q_3 tal que

$$P\left[\frac{\mathbf{SCE}_{H_0}}{\mathbf{SCE}_{MC}} \geq q_3 \mid H_0 \text{ es cierta} \right] = \alpha$$

Para obtener la expresión de la $\mathbf{SCE}_{H_0}$ se procederá como sigue,

$$\begin{aligned} \mathbf{SCE}_{MR} &= \left(\underline{Y} - \mathbf{X}\tilde{\underline{\beta}}\right)' \left(\underline{Y} - \mathbf{X}\tilde{\underline{\beta}}\right) \\ &= \left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}} + \mathbf{X}\hat{\underline{\beta}} - \mathbf{X}\tilde{\underline{\beta}}\right)' + \left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}} + \mathbf{X}\hat{\underline{\beta}} - \mathbf{X}\tilde{\underline{\beta}}\right) \\ &= \left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right)' \left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right) + \left(\hat{\underline{\beta}} - \tilde{\underline{\beta}}\right)' \mathbf{X}'\mathbf{X} \left(\hat{\underline{\beta}} - \tilde{\underline{\beta}}\right) \\ &\quad +2 \left(\hat{\underline{\beta}} - \tilde{\underline{\beta}}\right)' \mathbf{X}' \left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right) \end{aligned}$$

donde

$$\mathbf{X}' \left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right) = \mathbf{X}'\underline{Y} - \mathbf{X}'\mathbf{X}\hat{\underline{\beta}} = \mathbf{X}'\underline{Y} - \underbrace{\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}}_{\mathbf{I}}\hat{\underline{\beta}} = 0$$

entonces

$$\mathbf{SCE}_{MR} = \underbrace{\left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right)' \left(\underline{Y} - \mathbf{X}\hat{\underline{\beta}}\right)}_{\mathbf{SCE}_{MC}} + \underbrace{\left(\hat{\underline{\beta}} - \tilde{\underline{\beta}}\right)' \mathbf{X}'\mathbf{X} \left(\hat{\underline{\beta}} - \tilde{\underline{\beta}}\right)}_{\mathbf{SCE}_{H_0}}$$

Por otro lado, usando la ecuación 4,

$$\underline{\hat{\beta}} - \underline{\tilde{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} C' \left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1} \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)$$

lo cual implica

$$\begin{aligned}
\mathbf{SCE}_{H_0} &= \left(\underline{\hat{\beta}} - \underline{\tilde{\beta}}\right)' \mathbf{X}'\mathbf{X} \left(\underline{\hat{\beta}} - \underline{\tilde{\beta}}\right) \\
&= \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)' \left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1} C (\mathbf{X}'\mathbf{X})^{-1} \underbrace{\mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}}_{\mathbf{I}} \\
&\quad C' \left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1} \left(C\underline{\hat{\beta}} - \underline{\gamma}\right) \\
&= \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)' \left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1} \left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right] \left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1} \left(C\underline{\hat{\beta}} - \underline{\gamma}\right) \\
&= \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)' \left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1} \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)
\end{aligned}$$

A continuación necesitamos determinar la distribución de la estadística de prueba, $\dfrac{\mathbf{SCE}_{H_0}}{\mathbf{SCE}_{MC}}$. Recuérdese que

$$\frac{\mathbf{SCE}_{MC}}{\sigma^2} = \frac{\underline{Y}' \left(\mathbf{I} - \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\right) \underline{Y}}{\sigma^2} \sim \chi^2_{(n-k')}$$

Considere a continuación $\dfrac{\mathbf{SCE}_{H_0}}{\sigma^2}$ y recuerde que

$$\begin{aligned}
\underline{\hat{\beta}} &\sim N\left(\underline{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right) \\
\Rightarrow C\underline{\hat{\beta}} &\sim N\left(C\underline{\beta}, \sigma^2 C (\mathbf{X}'\mathbf{X})^{-1} C'\right) \\
\Rightarrow C\underline{\hat{\beta}} - \underline{\gamma} &\sim N\left(C\underline{\beta} - \underline{\gamma}, \sigma^2 C (\mathbf{X}'\mathbf{X})^{-1} C'\right)
\end{aligned}$$

$$\frac{\mathbf{SCE}_{H_0}}{\sigma^2} = \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)' \frac{\left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1}}{\sigma^2} \left(C\underline{\hat{\beta}} - \underline{\gamma}\right) = Q$$

Como $\left(C\underline{\hat{\beta}} - \underline{\gamma}\right) \sim N\left(C\underline{\beta} - \underline{\gamma}, \sigma^2 C (\mathbf{X}'\mathbf{X})^{-1} C'\right)$ y utilizando el teorema 4, con $\underline{Z} = \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)$, para hallar la distribución de Q, basta probar que $A\Sigma$ es idempotente. Como

$$\underbrace{\frac{\left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1}}{\sigma^2}}_{A} \underbrace{\sigma^2 C (\mathbf{X}'\mathbf{X})^{-1} C'}_{\Sigma} = \mathbf{I}$$

entonces $A\Sigma$ es una matriz idempotente. Entonces, $Q = \dfrac{\mathbf{SCE}_{H_0}}{\sigma^2} \sim \chi^2_{(r,l)}$; se tienen r grados de libertad, ya que $A = \dfrac{\left[C (\mathbf{X}'\mathbf{X})^{-1} C'\right]^{-1}}{\sigma^2}$ es de rango r. Sea $\underline{\mu} = C\underline{\beta} - \underline{\gamma}$; $l = \dfrac{1}{2}\underline{\mu}' A \underline{\mu}$, en este caso,

$$l = \frac{1}{2\sigma^2} \left(C\underline{\beta} - \underline{\gamma}\right)' \left[C \left(\mathbf{X}'\mathbf{X}\right)^{-1} C'\right]^{-1} \left(C\underline{\beta} - \underline{\gamma}\right)$$

pero bajo H_0, $C\underline{\beta} = \underline{\gamma}$, de donde $l = 0$, por lo que $Q = \dfrac{\mathbf{SCE}_{H_0}}{\sigma^2} \sim \chi^2_{(r)}$ bajo H_0. En resumen, se tiene que

$$\frac{\mathbf{SCE}_{MC}}{\sigma^2} = \frac{\underline{Y}' \left(\mathbf{I} - \mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\right) \underline{Y}}{\sigma^2} \sim \chi^2_{(n-k')}$$

y

$$\frac{\mathbf{SCE}_{H_0}}{\sigma^2} = \frac{1}{\sigma^2} \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)' \left[C \left(\mathbf{X}'\mathbf{X}\right)^{-1} C'\right]^{-1} \left(C\underline{\hat{\beta}} - \underline{\gamma}\right) \sim \chi^2_{(r)}$$

bajo H_0.

Si se demuestra que estas dos variables aleatorias son independientes entre sí, entonces su cociente (afectado por ciertas constantes) llevaría a una distribución F.

Nótese que

$$\mathbf{SCE}_{H_0} = \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)' \left[C \left(\mathbf{X}'\mathbf{X}\right)^{-1} C'\right]^{-1} \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)$$

con $\underline{\hat{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\underline{Y}$

entonces

$$\mathbf{SCE}_{H_0} = \left(C \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\underline{Y} - \underline{\gamma}\right)' \left[C \left(\mathbf{X}'\mathbf{X}\right)^{-1} C'\right]^{-1} \left(C \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\underline{Y} - \underline{\gamma}\right)$$

Ahora

$$\left[C \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\right] \mathbf{X}C' = CC'$$

lo cual implica

$$\left[C \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\right] \mathbf{X}C' \left(CC'\right)^{-1} = \mathbf{I}$$

en donde $(CC')^{-1}$ existe pues C es de rango completo. La $\mathbf{SCE}_{H_0}$ puede entonces expresarse como

$$\begin{aligned}\mathbf{SCE}_{H_0} &= \left(C\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left[\underline{Y}-\mathbf{X}C'\left(CC'\right)^{-1}\underline{\gamma}\right]\right)'\left[C\left(\mathbf{X}'\mathbf{X}\right)^{-1}C'\right]^{-1} \\ &\quad\cdot\left(C\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left[\underline{Y}-\mathbf{X}C'\left(CC'\right)^{-1}\underline{\gamma}\right]\right) \\ &= \left[\underline{Y}-\mathbf{X}C'\left(CC'\right)^{-1}\underline{\gamma}\right]'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}C'\left[C\left(\mathbf{X}'\mathbf{X}\right)^{-1}C'\right]^{-1}C \\ &\quad\cdot\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left[\underline{Y}-\mathbf{X}C'\left(CC'\right)^{-1}\underline{\gamma}\right]\end{aligned}$$

Sea $\underline{Y}_* = \underline{Y} - \mathbf{X}C'\left(CC'\right)^{-1}\underline{\gamma}$. Dado que

$$\underline{Y} \sim N\left(\mathbf{X}\underline{\beta}, \sigma^2\mathbf{I}\right)$$

se sigue que

$$\underline{Y}_* \sim N\left(\mathbf{X}\underline{\beta} - \mathbf{X}C'\left(CC'\right)^{-1}\underline{\gamma}, \sigma^2\mathbf{I}\right)$$

Sea $\underline{\theta} = \mathbf{X}\underline{\beta} - \mathbf{X}C'\left(CC'\right)^{-1}\underline{\gamma}$, entonces $\underline{Y}_* \sim N\left(\underline{\theta}, \sigma^2\mathbf{I}\right)$

Por lo anterior, $\mathbf{SCE}_{H_0} = \underline{Y}_*'\underbrace{\left[\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}C'\left[C\left(\mathbf{X}'\mathbf{X}\right)^{-1}C'\right]^{-1}C\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right]}_{A}\underline{Y}_*$

Con respecto a $\mathbf{SCE}_{MC}$ se tiene que

$$\mathbf{SCE}_{MC} = \underline{Y}'\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\underline{Y}$$

además

$$\mathbf{X}'\left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right) = \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\mathbf{X} = 0$$

La $\mathbf{SCE}_{MC}$ se puede expresar también como:

$$\begin{aligned}\mathbf{SCE}_{MC} &= \left(\underline{Y}-\mathbf{X}C'\left(CC'\right)^{-1}\underline{\gamma}\right)'\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)\left(\underline{Y}-\mathbf{X}C'\left(CC'\right)^{-1}\underline{\gamma}\right) \\ &= \underline{Y}_*'\underbrace{\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)}_{B}\underline{Y}_*\end{aligned}$$

Deseamos demostrar que $\underline{Y}_*'A\underline{Y}_*$ $(\mathbf{SCE}_{H_0})$ es independiente de $\underline{Y}_*B\underline{Y}_*(\mathbf{SCE}_{MC})$, donde $\underline{Y}_* \sim N\left(\underline{\theta}, \sigma^2\mathbf{I}\right)$. De acuerdo al teorema 5, se debe demostrar que $A\Sigma B = 0$. Sustituyendo se tiene

$$\sigma^2 \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} C'\left[C\left(\mathbf{X}'\mathbf{X}\right)^{-1} C'\right]^{-1} C\left(\mathbf{X}'\mathbf{X}\right)^{-1} \underbrace{\mathbf{X}'\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\right)}_{\mathbf{0}} = 0$$

Por lo tanto, $\mathbf{SCE}_{H_0}$ y $\mathbf{SCE}_{MC}$ son independientes.

Dado que

$$\frac{\mathbf{SCE}_{H_0}}{\sigma^2} \sim \chi^2_{(r)}$$

bajo H_0 y

$$\frac{\mathbf{SCE}_{MC}}{\sigma^2} \sim \chi^2_{(n-k')}$$

independientes entre sí, se tiene que

$$\frac{\dfrac{\mathbf{SCE}_{H_0}}{r\sigma^2}}{\dfrac{\mathbf{SCE}_{MC}}{(n-k')\,\sigma^2}} \sim F_{(r,n-k')}$$

bajo H_0, lo cual implica

$$\frac{\mathbf{SCE}_{H_0}/r}{\mathbf{SCE}_{MC}/(n-k')} \sim F_{(r,n-k')}$$

bajo H_0.

Denotemos por $\dfrac{\mathbf{SCE}_{H_0}}{r} = \mathbf{CME}_{H_0}$ y $\dfrac{\mathbf{SCE}_{MC}}{(n-k')} = \mathbf{CME}_{MC}$

La región crítica de esta prueba es

$$R_c = \left\{\underline{Y} \mid \frac{\mathbf{CME}_{H_0}}{\mathbf{CME}_{MC}} \geq F^{1-\alpha}_{(r,n-k')}\right\}$$

donde $F^{1-\alpha}_{(r,n-k')}$ denota al cuantil $(1-\alpha)$ de una distribución F con r y $(n-k')$ grados de libertad.

Tabla 5

Tabla de Análisis de Varianza para probar las hipótesis $H_0 : C\underline{\beta} = \underline{\gamma}$ vs. $H_a : C\underline{\beta} \neq \underline{\gamma}$, en el modelo $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2\mathbf{I})$, con $C_{r\times k'}$, de rango r, matriz de constantes conocidas y $\underline{\gamma}_{r\times 1}$, vector de constantes conocidas.

Fuente de variación	G L	*Sumas de cuadrados*	$C\ M$	F
H_0	r	$\left(C\underline{\hat{\beta}} - \underline{\gamma}\right)' \left[C\left(\mathbf{X}'\mathbf{X}\right)^{-1} C'\right]^{-1} \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)$	$\frac{\mathbf{SCE}_{H_0}}{r}$	$\frac{\mathbf{CME}_{H_0}}{\mathbf{CME}_{MC}}$
Modelo Completo	$n - k'$	$\left(\underline{Y} - \mathbf{X}\underline{\hat{\beta}}\right)' \left(\underline{Y} - \mathbf{X}\underline{\hat{\beta}}\right) = \underline{Y}'\underline{Y} - \underline{\hat{\beta}}'\mathbf{X}'\underline{Y}$	$\frac{\mathbf{SCE}_{MC}}{(n-k')}$	
Modelo Reducido	$n - k' + r$	$\left(\underline{Y} - \mathbf{X}\underline{\tilde{\beta}}\right)' \left(\underline{Y} - \mathbf{X}\underline{\tilde{\beta}}\right)$		

Nótese que

$$\begin{aligned}
\mathbf{SCE}_{MC} &= \left(\underline{Y} - \mathbf{X}\underline{\hat{\beta}}\right)' \left(\underline{Y} - \mathbf{X}\underline{\hat{\beta}}\right) \\
&= \underline{Y}'\underline{Y} - \underline{Y}'\mathbf{X}\underline{\hat{\beta}} - \underline{\hat{\beta}}'\mathbf{X}'\underline{Y} + \underline{\hat{\beta}}'\mathbf{X}'\mathbf{X}\underline{\hat{\beta}} \\
&= \underline{Y}'\underline{Y} - 2\underline{\hat{\beta}}'\mathbf{X}'\underline{Y} + \underline{\hat{\beta}}'\mathbf{X}'\mathbf{X}\underline{\hat{\beta}} \\
&= \underline{Y}'\underline{Y} - 2\underline{\hat{\beta}}'\mathbf{X}'\underline{Y} + \underline{\hat{\beta}}'\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\underline{Y} \\
&= \underline{Y}'\underline{Y} - 2\underline{\hat{\beta}}'\mathbf{X}'\underline{Y} + \underline{\hat{\beta}}'\mathbf{X}'\underline{Y} \\
&= \underline{Y}'\underline{Y} - \underline{\hat{\beta}}'\mathbf{X}'\underline{Y}
\end{aligned}$$

4.10.1 Casos especiales de la prueba lineal general

En esta sección discutiremos algunos casos especiales de la prueba lineal general y determinaremos cómo se modifica la estadística de prueba en cada uno de ellos.

Caso 1: $H_0 : \underline{\beta} = \underline{0}$ **vs.** $H_a : \underline{\beta} \neq \underline{0}$

Supóngase que se desea probar

$$H_0 : \underline{\beta} = \underline{0} \text{ vs. } H_a : \underline{\beta} \neq \underline{0}$$

En este caso, $C = \mathbf{I}_{k'}$, de rango k', $\underline{\gamma} = \underline{0}$ y el modelo reducido por la hipótesis nula es $Y_i = \varepsilon_i$. Sustituyendo esta elección particular de C y $\underline{\gamma}$ en la expresión de la $\mathbf{SCE}_{H_0}$ se tiene

$$\mathbf{SCE}_{H_0} = \hat{\underline{\beta}}' (\mathbf{X}'\mathbf{X}) \hat{\underline{\beta}}$$

Obviamente el modelo completo es $\underline{Y} = X\underline{\beta} + \underline{\varepsilon}$, por lo que la $\mathbf{SCE}_{MC}$ es la reportada en la tabla anterior. Entonces:

$$\frac{\hat{\underline{\beta}}' (\mathbf{X}'\mathbf{X}) \hat{\underline{\beta}}/k'}{\mathbf{CME}_{MC}} \sim F_{(k', n-k')}$$

Caso 2: $H_0 : \underline{\beta} = \underline{\beta}^*$ **vs.** $H_a : \underline{\beta} \neq \underline{\beta}^*$

Supóngase que se desea probar

$$H_0 : \underline{\beta} = \underline{\beta}^* \text{ vs. } H_a : \underline{\beta} \neq \underline{\beta}^*$$

con $\underline{\beta}^*_{k' \times 1}$ de constantes conocidas. En este caso, $C = \mathbf{I}_{k'}$, de rango k', $\underline{\gamma} = \underline{\beta}^*$. Entonces

$$\mathbf{SCE}_{H_0} = \left(\hat{\underline{\beta}} - \underline{\beta}^*\right)' (\mathbf{X}'\mathbf{X}) \left(\hat{\underline{\beta}} - \underline{\beta}^*\right)$$

y

$$F = \frac{\left(\hat{\underline{\beta}} - \underline{\beta}^*\right)' (\mathbf{X}'\mathbf{X}) \left(\hat{\underline{\beta}} - \underline{\beta}^*\right)/k'}{\mathbf{CME}_{MC}} \sim F_{(k', n-k')}$$

Caso 3: $H_0 : \underline{\lambda}'\underline{\beta} = m$ **vs.** $H_a : \underline{\lambda}'\underline{\beta} \neq m$

Supóngase que se desea probar

$$H_0 : \underline{\lambda}'\underline{\beta} = m \text{ vs. } H_a : \underline{\lambda}'\underline{\beta} \neq m$$

donde $\underline{\lambda} \in \Re^{k'}$, un vector de constantes conocidas y $m \in \Re$ una constante conocida.

En este caso, $C = \underline{\lambda}'_{1 \times k'}$, de rango 1, $\underline{\gamma} = m$. Así,

$$\mathbf{SCE}_{H_0} = \left(\underline{\lambda}'\underline{\hat{\beta}} - m\right)' \left(\underline{\lambda}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \underline{\lambda}\right)^{-1} \left(\underline{\lambda}'\underline{\hat{\beta}} - m\right) = \frac{\left(\underline{\lambda}'\underline{\hat{\beta}} - m\right)^2}{\left(\underline{\lambda}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \underline{\lambda}\right)}$$

y

$$F = \frac{\left(\underline{\lambda}'\underline{\hat{\beta}} - m\right)^2}{\left(\underline{\lambda}' \left(\mathbf{X}'\mathbf{X}\right)^{-1} \underline{\lambda}\right) \mathbf{CME}_{MC}} \sim F_{(1,n-k')}$$

Caso 4: Prueba de significancia global

$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ vs. H_a : No H_0

$$C = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

con rango de C igual a k y $\underline{\gamma}_{k\times 1} = \underline{0}$

Tenemos entonces que el modelo reducido es $Y_i = \beta_0 + \varepsilon_i$. La $\mathbf{SCE}_{MC}$ es

$$\mathbf{SCE}_{MC} = \underline{Y}'\underline{Y} - \underline{\hat{\beta}}'\mathbf{X}'\underline{Y}$$

La del modelo reducido:

$$\mathbf{SCE}_{MR} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \underline{Y}'\underline{Y} - n\bar{Y}^2$$

Y la $\mathbf{SCE}_{H_0}$

$$\mathbf{SCE}_{H_0} = \underline{\hat{\beta}}'\mathbf{X}'\underline{Y} - n\bar{Y}^2$$

De esta manera

$$F = \frac{\left(\underline{\hat{\beta}}'\mathbf{X}'\underline{Y} - n\bar{Y}^2\right)/k}{\mathbf{CME}_{MC}} \sim F_{(k,n-k')}$$

Ejemplo 14.- Para ilustrar el uso de la hipótesis lineal general, se usarán datos provenientes de un programa de acondicionamiento físico realizado en North Carolina State University.

Se tomaron medidas sobre una muestra de $n = 31$ hombres inscritos en el programa. Las variables medidas fueron: edad en años, peso en kilogramos, oxígeno aspirado (ml por kg de peso por minuto), tiempo para correr 1.5 millas en minutos, frecuencia cardiaca en reposo, frecuencia cardiaca durante la carrera (al mismo tiempo se midió el oxígeno aspirado) y la frecuencia cardiaca mínima y máxima durante la carrera. Los datos se presentan en la tabla 6.

Los resultados que se analizarán serán sobre el oxígeno aspirado, Y y sobre las cuatro variables independientes X_1, X_2, X_3, y X_4.

El modelo es $\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$, donde $\underline{\beta}' = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$, donde cada subíndice indica a la variable independiente correspondiente (mencionadas arriba). El modelo ajustado es:

$$\begin{aligned} \underline{Y} &= \mathbf{X}\underline{\hat{\beta}} \\ &= \mathbf{X}\begin{pmatrix} 84.26902 \\ -3.06981 \\ 0.00799 \\ -0.11671 \\ 0.08518 \end{pmatrix} \end{aligned}$$

El $\mathbf{CME}_{MC}$ es

$$\mathbf{CME}_{MC} = \frac{\underline{e}'\underline{e}}{n-k'} = \frac{193.117}{31-5} = \frac{193.1178}{26} = 7.4276$$

Tabla 6: Medidas sobre la condición física de 31 hombres en un programa de acondicionamiento físico en la North Carolina State University.

				Frecuencia Cardiaca		
Edad (*años*)	*Peso* (*kg*)	*Demanda de Oxígeno*	*Tiempo* (min)	*Descansando*	*Corriendo*	*Máximo*
44	89.47	44.609	11.37	62	178	182
40	75.07	45.313	10.07	62	185	185
44	85.84	54.297	8.65	45	156	184
42	68.15	59.571	8.17	40	166	172
38	89.02	49.874	9.22	55	178	180
47	77.45	44.811	11.63	58	176	176
40	75.98	45.681	11.95	70	176	180
43	81.19	49.091	10.85	64	162	170
44	81.42	39.442	13.08	63	174	176
38	81.87	60.055	8.63	48	170	186
44	73.03	50.541	10.13	45	168	168
45	87.66	37.388	14.03	56	186	192
45	66.45	44.754	11.12	51	176	176
47	79.15	47.273	10.60	47	162	164
54	83.12	51.855	10.33	50	166	170
49	81.42	49.156	8.95	44	180	185
51	69.63	40.836	10.95	57	168	172
51	77.91	46.672	10.00	48	162	168
48	91.63	46.774	10.25	48	162	164

Tabla 6, continuación

				Frecuencia Cardiaca		
Edad (*años*)	*Peso* (*kg*)	*Demanda de Oxígeno*	*Tiempo* (min)	*Descansando*	*Corriendo*	*Máximo*
49	73.37	50.388	10.08	67	168	168
57	73.37	39.407	12.63	58	174	176
54	79.38	46.080	11.17	62	156	176
52	76.32	45.441	9.63	48	164	166
50	70.87	54.625	8.92	48	146	186
51	67.25	45.118	11.08	48	172	172
54	91.63	39.203	12.88	44	168	172
51	73.71	45.790	10.47	59	186	188
57	59.08	50.545	9.93	49	148	160
49	76.32	48.673	9.40	56	186	188
48	61.24	47.920	11.50	52	170	176
52	82.78	47.467	10.50	53	170	172

Las pruebas de hipótesis requieren el cálculo de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$, que en nuestro caso es:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 17.423095 & -0.1596195 & 0.0072675 & -0.0140450 & -0.0779657 \\ -0.1596195 & 0.0236857 & -0.0016967 & -0.0009852 & 0.0009482 \\ 0.0072675 & -0.0016967 & 0.0007776 & -0.0000935 & -0.0000854 \\ -0.0140450 & -0.0009852 & -0.0000935 & 0.0005428 & -0.0003562 \\ -0.0779657 & 0.0009482 & -0.0000854 & -0.0003562 & 0.0007560 \end{pmatrix}$$

Primeramente probaremos la hipótesis $H_0 : \beta_2 = \beta_4 = 0$, con $\alpha = 0.05$. La hipótesis alternativa es que al menos una de ellas es diferente de cero. Esta hipótesis nula se puede escribir, en la forma lineal general, como sigue:

$$H_0 : C\underline{\beta} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Note que el rango de C es dos. Así, la estadística de prueba es

$$F = \frac{\left(\frac{1}{r}\right)\left(C\underline{\hat{\beta}} - \underline{\gamma}\right)' \left[C\left(\mathbf{X}'\mathbf{X}\right)^{-1} C'\right]^{-1} \left(C\underline{\hat{\beta}} - \underline{\gamma}\right)}{\mathbf{CME}_{MC}}$$

donde

$$\left[C\left(\mathbf{X}'\mathbf{X}\right)^{-1} C'\right]^{-1} = \begin{pmatrix} 0.0007776 & -0.0000854 \\ -0.0000854 & 0.0007560 \end{pmatrix}^{-1} = \begin{pmatrix} 1302.223 & 147.101 \\ 147.101 & 1339.322 \end{pmatrix}$$

y

$$C\underline{\hat{\beta}} - \underline{\gamma} = \begin{bmatrix} 0.00799 \\ 0.08518 \end{bmatrix}$$

es decir

$$F = \frac{\left(\frac{1}{2}\right) \begin{pmatrix} 0.00799 \\ 0.08518 \end{pmatrix}' \begin{pmatrix} 1302.223 & 147.101 \\ 147.101 & 1339.322 \end{pmatrix} \begin{pmatrix} 0.00799 \\ 0.08518 \end{pmatrix}}{7.4276}$$

$$= \frac{\left(\frac{1}{2}\right) 10.0016}{7.4276} = 0.673$$

El valor calculado de F es mucho menor que $F^{0.95}_{(2,26)} = 3.37$, es decir, la prueba no es significativa, por lo cual no existe razón para rechazar la hipótesis nula.

La tabla de análisis de varianza correspondiente a esta prueba es,

Tabla 7: Tabla de análisis de varianza ($H_0 : \beta_2 = \beta_4 = 0$)

Fuente de variación	*G L*	*Sumas de cuadrados*	*C M*	*F*
H_0	2	10.0016	5.0008	
Modelo Completo	26	193.17	7.4276	0.673
Modelo Reducido	28	203.1716		

La segunda hipótesis que se probará (con $\alpha = 0.05$) ilustra el caso donde $\underline{\gamma} \neq \underline{0}$. Suponga que información previa sugiere que la intercepción β_0 para un grupo de hombres de esta edad y peso debe ser 90. Entonces, la hipótesis nula de interés es $H_0 : \beta_0 = 90$ y, para ilustrar los cálculos, se construirá una hipótesis compuesta añadiendo a las restricciones las dos condiciones en la hipótesis nula anterior.

La hipótesis nula es

$H_0 : C\underline{\beta} - \underline{\gamma} = \underline{0}$

donde

$$C\underline{\beta} - \underline{\gamma} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = \begin{pmatrix} 90 \\ 0 \\ 0 \end{pmatrix}$$

Para esta hipótesis

$$C\hat{\underline{\beta}} - \underline{\gamma} = \begin{pmatrix} -5.73098 \\ 0.00799 \\ 0.08518 \end{pmatrix}$$

y

$$
\begin{aligned}
\left[C\left(\mathbf{X}'\mathbf{X}\right)^{-1}C'\right]^{-1} &= \begin{pmatrix} 17.423095 & 0.0072675 & -0.0779657 \\ 0.0072675 & 0.0007776 & -0.0000854 \\ -0.0779657 & -0.0000854 & 0.0007560 \end{pmatrix}^{-1} \\
&= \begin{pmatrix} 0.106613 & 0.213749 & 11.018666 \\ 0.213749 & 1302.65230 & 169.19219 \\ 11.018666 & 169.19219 & 2478.12557 \end{pmatrix}
\end{aligned}
$$

de tal forma que

$$
\begin{aligned}
F &= \frac{\left(\frac{1}{r}\right)\left(C\underline{\hat{\beta}}-\underline{\gamma}\right)'\left[C\left(\mathbf{X}'\mathbf{X}\right)^{-1}C'\right]^{-1}\left(C\underline{\hat{\beta}}-\underline{\gamma}\right)}{\mathbf{CME}_{MC}} \\
&= \frac{\left(\frac{1}{3}\right)11.0187}{7.4276} = 0.494
\end{aligned}
$$

la cual es, nuevamente, mucho menor que $F^{0.95}_{(3,26)} = 2.98$, es decir, la prueba no es significativa, por lo que no existe razón para rechazar la hipótesis nula.

Tabla 8

Tabla de análisis de varianza para probar $H_0 : \beta_0 = 90$

Fuente de variación	*G L*	*Sumas de cuadrados*	*C M*	*F*
H_0	3	11.0187	3.6729	0.494
Modelo Completo	26	193.17	7.4276	
Modelo Reducido	29	204.1887		

4.11 Intervalos de confianza

Los intervalos de confianza sobre los parámetros proveen de más información al analista que aquella que pudiera darle solamente la estimación puntual de los mismos. En esta

sección se desarrollan los intervalos de confianza que revisten más interés en el modelo lineal general.

4.11.1 Intervalo de confianza para σ^2

A continuación obtendremos un intervalo de confianza para σ^2. Dado que

$$\frac{(n-k')\,\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k')}$$

con

$$\hat{\sigma}^2 = \frac{\underline{e}'\underline{e}}{n-k'}$$

se tiene

$$P\left[\chi^{\alpha/2}_{(n-k')} < \frac{(n-k')\,\hat{\sigma}^2}{\sigma^2} < \chi^{1-\alpha/2}_{(n-k')}\right] = 1-\alpha$$

Por lo tanto, un intervalo del $(1-\alpha)\times 100\%$ de confianza para σ^2 está dado por

$$\left(\frac{(n-k')\,\hat{\sigma}^2}{\chi^{1-\alpha/2}_{(n-k')}}, \frac{(n-k')\,\hat{\sigma}^2}{\chi^{\alpha/2}_{(n-k')}}\right)$$

donde $\chi^{\alpha/2}_{(n-k')}$ y $\chi^{1-\alpha/2}_{(n-k')}$ denotan los cuantiles $\frac{\alpha}{2}$ y $1-\frac{\alpha}{2}$ de una distribución χ^2 con $(n-k')$ grados de libertad.

Ejemplo 15.- Calculemos un intervalo del 95% de confianza para σ^2 usando los datos del ejemplo anterior.

$$\begin{aligned}\left(\frac{(n-k')\,\hat{\sigma}^2}{\chi^{1-\alpha/2}_{(n-k')}}, \frac{(n-k')\,\hat{\sigma}^2}{\chi^{\alpha/2}_{(n-k')}}\right) &= \left(\frac{(26)\,7.4276}{\chi^{0.975}_{(26)}}, \frac{(26)\,7.4276}{\chi^{0.025}_{(26)}}\right)\\ &= \left(\frac{(26)\,7.4276}{41.9}, \frac{(26)\,7.4276}{13.8}\right)\\ &= (4.609, 13.994)\end{aligned}$$

4.11.2 Intervalo de confianza para $\underline{\lambda}'\underline{\beta}$

Frecuentemente el analista está interesado en construir intervalos de confianza para alguna función de las $\beta_i's$. Aquí presentamos un intervalo de confianza para uno de los casos más importantes, el de una función lineal de las $\beta_i's$. Sea $\underline{\lambda}$ un vector de constantes conocidas, de tamaño k',

$$\underline{\lambda}'\hat{\underline{\beta}} \sim N\left(\underline{\lambda}'\underline{\beta}, \sigma^2\underline{\lambda}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underline{\lambda}\right)$$

En efecto, se sabe que una combinación lineal de variables aleatorias normales se distribuye normal. Por otro lado, los parámetros de esta distribución en particular son:

$$E\left(\underline{\lambda}'\hat{\underline{\beta}}\right) = \underline{\lambda}'E\left(\hat{\underline{\beta}}\right) = \underline{\lambda}'\underline{\beta}$$

y

$$\begin{aligned} Var\left(\underline{\lambda}'\hat{\underline{\beta}}\right) &= E\left[\left(\underline{\lambda}'\hat{\underline{\beta}} - \underline{\lambda}'\underline{\beta}\right)'\left(\underline{\lambda}'\hat{\underline{\beta}} - \underline{\lambda}'\underline{\beta}\right)\right] \\ &= E\left[\underline{\lambda}'\left(\hat{\underline{\beta}} - \underline{\beta}\right)'\left(\hat{\underline{\beta}} - \underline{\beta}\right)\underline{\lambda}'\right] \\ &= \underline{\lambda}'Var\left(\hat{\underline{\beta}}\right)\underline{\lambda} \\ &= \sigma^2\underline{\lambda}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underline{\lambda} \end{aligned}$$

Por lo anterior:

$$\frac{\underline{\lambda}'\hat{\underline{\beta}} - \underline{\lambda}'\underline{\beta}}{\sqrt{\sigma^2\underline{\lambda}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underline{\lambda}}} \sim N(0,1)$$

de lo cual se obtiene

$$\frac{\underline{\lambda}'\hat{\underline{\beta}} - \underline{\lambda}'\underline{\beta}}{\hat{\sigma}\sqrt{\underline{\lambda}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underline{\lambda}}} \sim t_{(n-k')}$$

donde

$$\hat{\sigma}^2 = \frac{\underline{e}'\underline{e}}{n-k'}$$

De esta manera, tenemos

$$P\left[-t_{(n-k')}^{1-\frac{\alpha}{2}} \leq \frac{\underline{\lambda}'\hat{\underline{\beta}} - \underline{\lambda}'\underline{\beta}}{\hat{\sigma}\sqrt{\underline{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\underline{\lambda}}} \leq t_{(n-k')}^{1-\frac{\alpha}{2}}\right] = 1-\alpha$$

y el intervalo de $(1-\alpha)\times 100\%$ de confianza para $\underline{\lambda}'\underline{\beta}$ está dado por

$$\underline{\lambda}'\hat{\underline{\beta}} \pm t_{(n-k')}^{1-\frac{\alpha}{2}}\hat{\sigma}\sqrt{\underline{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\underline{\lambda}}$$

donde $t_{(n-k')}^{1-\frac{\alpha}{2}}$ denota al cuantil $1-\dfrac{\alpha}{2}$ de una distribución t con $(n-k')$ grados de libertad.

Observe que esta estadística nos permite tener un intervalo de confianza para cualquier β_i. Por ejemplo, si deseamos construir un intervalo de confianza para β_1, se construye

$$\underline{\lambda}' = (0,1,0,...,0)$$

de tal forma que

$$\underline{\lambda}'\underline{\beta} = (0,1,0,...,0)\begin{pmatrix}\beta_0\\ \beta_1\\ \vdots\\ \beta_k\end{pmatrix} = \beta_1$$

Ejemplo 16.- Calcularemos un intervalo del 95% de confianza para $\beta_2+\beta_4=0$. En este caso

$$\underline{\lambda}'\underline{\beta} = (0,0,1,0,1)\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix} = \beta_2+\beta_4$$

$$\underline{\lambda}'\hat{\underline{\beta}} = (0,0,1,0,1)\begin{pmatrix}84.26902\\ -3.06981\\ 0.00799\\ -0.11671\\ 0.08518\end{pmatrix} = 0.00799+0.08518 = 0.09317$$

$$\underline{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\underline{\lambda} = 0.001363$$

y

$$t^{1-\frac{\alpha}{2}}_{(n-k')} = t^{.975}_{(26)} = 2.056$$

Por lo tanto, el intervalo del 95% de confianza para $\underline{\lambda}'\underline{\beta}$ está dado por:

$$\begin{aligned}\underline{\lambda}'\hat{\underline{\beta}} \pm t^{1-\frac{\alpha}{2}}_{(n-k')}\hat{\sigma}\sqrt{\underline{\lambda}'(\mathbf{X}'\mathbf{X})^{-1}\underline{\lambda}} &= 0.09317 \pm (2.056)\,7.4276\sqrt{0.001363} \\ &= 0.09317 \pm (2.056)\,7.4276\,(0.100609) \\ &= 0.09317 \pm 1.53641\end{aligned}$$

que es

$$(-1.4432,1\ ,\ 1.6296)$$

4.11.3 Intervalo de confianza para $E(Y_0 \mid \underline{X}_0)$

Si las β_i fueran conocidas, entonces, para una observación particular de las variables independientes $\mathbf{X}$, $\underline{X}_0 = (X_{01}, X_{02}, ..., X_{0k})$, la cantidad $E(Y_0)$ puede calcularse con toda precisión, dado que $E(Y_0) = \sum_{i=1}^{k} X_{0i}\beta_i$. Sin embargo, como el vector de parámetros es desconocido, debe ser estimado a partir de una muestra de tamaño n de las variables independientes $\mathbf{X}$ y con ello estimar $E(Y_0)$, la respuesta media de Y para un valor particular de las variables regresoras $\mathbf{X}$. Sea $\underline{X}_0$ el nivel de esta variable para el cual deseamos estimar la respuesta media, es decir $\widehat{E(Y_0 \mid \underline{X}_0)}$. Suponemos que $\underline{X}_0$ es cualquier observación particular sobre el rango de datos originales X usado para ajustar el modelo.

Un estimador puntual de $E(Y_0)$ es

$$\widehat{E(Y_0)} = \hat{Y}_0 = \underline{X}_0\hat{\underline{\beta}}$$

Este es un estimador insesgado de $\hat{Y}_0$, en efecto:

$$E\left(\hat{Y}_0\right) = E\left(\underline{X}_0\hat{\underline{\beta}}\right) = \underline{X}_0 E\left(\hat{\underline{\beta}}\right) = \underline{X}_0\underline{\beta}$$

La varianza de este estimador está dada por:

$$Var\left(\hat{Y}_0\right) = Var\left(\underline{X}_0\hat{\underline{\beta}}\right) = \underline{X}_0' Var\left(\hat{\underline{\beta}}\right)\underline{X}_0 = \underline{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\sigma^2\underline{X}_0$$

De lo anterior obtenemos

$$\hat{Y}_0 \sim N\left(\underline{X}_0\underline{\beta}, \underline{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\sigma^2\underline{X}_0\right)$$

por lo cual

$$\frac{\hat{Y}_0 - \underline{X}_0\hat{\underline{\beta}}}{\sigma\sqrt{\underline{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underline{X}_0}} \sim N(0,1)$$

dado que $\dfrac{(\underline{e}'\underline{e})/(n-k')}{\sigma^2} \sim \chi^2_{(n-k')}$ y es independiente de $\hat{\underline{\beta}}$.

Obsérvese que lo anterior implica

$$\frac{\dfrac{\hat{Y}_0 - \underline{X}_0\hat{\underline{\beta}}}{\sigma\sqrt{\underline{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underline{X}_0}}}{\sqrt{\dfrac{(\underline{e}'\underline{e})/(n-k')}{\sigma^2}}} = \frac{\hat{Y}_0 - \underline{X}_0\hat{\underline{\beta}}}{\sqrt{\dfrac{(\underline{e}'\underline{e})}{(n-k')}}\sqrt{\underline{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underline{X}_0}} \sim t_{(n-k')}$$

De esta forma, podemos plantear

$$P\left[-t^{1-\frac{\alpha}{2}}_{(n-k')} \leq \frac{\hat{Y}_0 - \underline{X}_0\hat{\underline{\beta}}}{\hat{\sigma}\sqrt{\underline{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underline{X}_0}} \leq t^{1-\frac{\alpha}{2}}_{(n-k')}\right] = 1-\alpha$$

Así, el intervalo del $(1-\alpha)\times 100\%$ de confianza para el valor medio de Y cuando se observa $\underline{X}_0$está dado por:

$$\hat{Y}_0 \pm t^{1-\frac{\alpha}{2}}_{(n-k')}\hat{\sigma}\sqrt{\underline{X}_0'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\underline{X}_0}$$

Note que éste es un caso especial del descrito en la sección anterior

4.11.4 Intervalo de predicción para la respuesta media de futuras observaciones

Suponga que una muestra aleatoria de m valores de Y es descrita por la función de distribución

$$f(Y; X_{01}, X_{02}, ..., X_{0k'})$$

donde cada uno de los m valores de Y es seleccionado a partir de la misma función de distribución, es decir, a partir de los mismos valores de X_i.

Denotemos a estos m valores de Y por $Y_{01}, Y_{02}, ..., Y_{0m}$, y a la media de ellos por $\bar{Y}_0$, es decir

$$\bar{Y}_0 = \frac{1}{m} \sum_{i=1}^{m} Y_{0i}$$

Suponga que se desea construir un intervalo de confianza para la media $\bar{Y}_0$ de estos m valores de Y. Es decir, suponga que se construye un intervalo tal que la probabilidad de que la media $\bar{Y}_0$ esté en él sea igual a un valor predeterminado $1 - \alpha$. El intervalo será llamado un intervalo de predicción del $(1 - \alpha) \times 100\%$ de confianza, puesto que la probabilidad de que la media $\bar{Y}_0$ de una muestra futura de valores de Y seleccionada aleatoriamente a partir de la función de densidad $f(Y; X_{01}, X_{02}, ..., X_{0k'})$ esté en él es igual a $1 - \alpha$.

Dado que $\underline{\beta}$ y σ^2 son desconocidas, se deben seleccionar n valores muestrales

$$A_i' = (Y_i; X_{1i}, X_{2i}, ..., X_{ji})$$

donde

$$j = 1, 2, ..., n$$

Se deberán usar estos valores muestrales para estimar $\underline{\beta}$ y σ^2 y con ellos construir el intervalo de confianza deseado. Para hacer esto, observe que, dado que $\bar{Y}_0$ y $\hat{\underline{\beta}}$ son independientes, la cantidad

$$z = \bar{Y}_0 - \hat{\underline{\beta}}' \underline{X}_0 \sim N\left(0, \left(\frac{1}{m} + \underline{X}_0' (\mathbf{X}'\mathbf{X})^{-1} \underline{X}_0\right) \sigma^2\right)$$

donde $\underline{X}_0$ es un vector de tamaño k' cuyo i-ésimo elemento es X_{0i}. De esta expresión obtenemos

$$w = \frac{\bar{Y}_0 - \underline{\hat{\beta}}' \underline{X}_0}{\sigma \sqrt{\frac{1}{m} + \underline{X}_0' (\mathbf{X}'\mathbf{X})^{-1} \underline{X}_0}} \sim N(0,1)$$

Como $v = \frac{(n-k')\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-k')}$, donde $\hat{\sigma}^2 = \frac{\underline{e}'\underline{e}}{n-k'}$ y es independiente de w, se tiene

$$u = \frac{w\sqrt{n-k'}}{\sqrt{v}} \sim t_{(n-k')}$$

Por lo tanto

$$P\left[-t^{\frac{\alpha}{2}}_{(n-k')} \leq u \leq t^{\frac{\alpha}{2}}_{(n-k')}\right] = 1 - \alpha$$

Sustituyendo u se tiene:

$$P\left[-t^{\frac{\alpha}{2}}_{(n-k')} \leq \frac{\bar{Y}_0 - \underline{\hat{\beta}}' \underline{X}_0}{\hat{\sigma} \sqrt{\frac{1}{m} + \underline{X}_0' (\mathbf{X}'\mathbf{X})^{-1} \underline{X}_0}} \leq t^{\frac{\alpha}{2}}_{(n-k')}\right] = 1 - \alpha$$

Con lo cual se llega al intervalo de predicción del $(1-\alpha) \times 100\%$ de confianza para la respuesta media de futuras observaciones:

$$\underline{\hat{\beta}}' \underline{X}_0 \pm t^{\frac{\alpha}{2}}_{(n-k')} \hat{\sigma} \sqrt{\frac{1}{m} + \underline{X}_0' (\mathbf{X}'\mathbf{X})^{-1} \underline{X}_0}$$

4.11.5 Inferencia simultánea en regresión lineal múltiple

Considérese el vector de parámetros $\underline{\beta}$. Los intervalos de confianza para $\underline{\beta}$ definirían regiones en $\Re^{k'}$, donde, con el $(1-\alpha) \times 100\%$ de confianza $\beta_0, \beta_1, ..., \beta_k$ estarían simultáneamente. Para construir este intervalo, considérese

$$\underline{\hat{\beta}} \sim N\left(\underline{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$$

por lo cual

$$\underline{\hat{\beta}} - \underline{\beta} \sim N\left(\underline{0}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$$

Por otro lado, el producto

$$\frac{1}{\sigma^2}\left(\underline{\hat{\beta}}-\underline{\beta}\right)'(\mathbf{X}'\mathbf{X})\left(\underline{\hat{\beta}}-\underline{\beta}\right)\sim\chi^2_{k'}$$

Sabemos también que $\underline{\hat{\beta}}$ es independiente de $\frac{\underline{e}'\underline{e}}{n-k'}=\mathbf{CME}=\frac{\mathbf{SCE}}{n-k'}$ y que $\frac{\underline{e}'\underline{e}}{\sigma^2}\sim\chi^2_{(n-k')}$

Por lo anterior se tiene

$$\frac{\left(\underline{\hat{\beta}}-\underline{\beta}\right)'(\mathbf{X}'\mathbf{X})\left(\underline{\hat{\beta}}-\underline{\beta}\right)}{\frac{\underline{e}'\underline{e}}{n-k'}}\sim F_{(k',n-k')}$$

con lo cual se puede plantear

$$P\left[\frac{n-k'}{k'}\left(\underline{\hat{\beta}}-\underline{\beta}\right)'(\mathbf{X}'\mathbf{X})\left(\underline{\hat{\beta}}-\underline{\beta}\right)\leq F^{1-\alpha}_{(k',n-k')}\right]=1-\alpha$$

Haciendo las sustituciones pertinentes en la desigualdad anterior, se obtiene un elipsoide k'-dimensional que es el intervalo de confianza para todos los parámetros del modelo. La pendiente de los ejes y la excentricidad del elipsoide muestran la dirección y la magnitud, respectivamente, de las correlaciones entre los estimadores de los parámetros.

Hasta ahora hemos supuesto que cada una de las variables regresoras incluidas en el modelo contiene información que influye en éste, y que esta información no es aportada por ninguna otra de las variables. Es decir, hemos supuesto que no existe información redundante entre las variables regresoras.

En la práctica esto casi no ocurre. La necesidad de contar con modelos más eficientes nos lleva a analizar cuáles de las variables de las que se dispone son más significativas, descartando aquellas que no contribuyen con información al modelo si las primeras están incluidas.

En el siguiente capítulo se presentan algunas técnicas y criterios para construir un modelo eficiente.

Capítulo 5

Selección de variables: El problema de definición del modelo

En los capítulos previos hemos supuesto que se sabe que todas las variables regresoras influyen en el modelo. Hasta ahora nuestro objetivo ha sido desarrollar las técnicas que aseguren que la forma funcional del modelo es correcta y que las hipótesis del mismo no son violadas. En algunas aplicaciones, consideraciones teóricas o experiencia previa pueden ser útiles al seleccionar las variables regresoras que van a ser usadas en el modelo. Sin embargo, en la mayoría de los problemas prácticos, el analista tiene una variedad de variables candidatas que podrían incluir factores de influencia, pero el subconjunto real de variables regresoras que deberían ser usadas en el modelo necesita ser determinado. Encontrar un subconjunto adecuado de variables regresoras es llamado *el problema de selección de variables.*

Construir un modelo de regresión que incluya sólo un subconjunto de las variables regresoras disponibles involucra dos objetivos conflictivos:

1. Se desea que el modelo incluya el mayor número posible de variables regresoras, de tal forma que la información contenida en estos factores influya en el valor pronosticado de la variable respuesta, Y.

2. Se desea también que el modelo incluya el menor número posible de variables regresoras, puesto que la varianza de $\hat{Y}$ crece cuando el número de aquellas aumenta. También debemos tomar en consideración que cuantas más variables regresoras contenga la ecuación de regresión, mayor será el costo de recolección de los datos y de mantenimiento del modelo.

El proceso de búsqueda de un modelo que combine adecuadamente estos dos objetivos es llamado *selección de la mejor ecuación de regresión*.

Desafortunadamente no existe una única definición de *la mejor ecuación de regresión*, como se verá más adelante. Además, los distintos algoritmos disponibles para la selección de variables a menudo especifican diferentes subconjuntos del conjunto de las candidatas a variables regresoras como "el mejor".

El problema de selección de variables es discutido frecuentemente en un escenario idealizado; usualmente se supone que la especificación funcional correcta de las variables regresoras es conocida y que no están presentes outliers u observaciones que influyan demasiado en el modelo. En la práctica, estas suposiciones raramente son conocidas. El análisis de los residuales (discutido en el capítulo 2), es útil para señalar posibles variables regresoras, revelar formas funcionales de éstas que deberían ser investigadas, e identificar defectos en los datos tales como los outliers. El efecto de observaciones con un alto nivel de influencia en el modelo también debe ser analizado.

El diagnóstico de qué tan adecuado es el modelo está ligado al problema de selección de variables. Aunque idealmente estos problemas deberían ser resueltos simultáneamente, usualmente se utiliza un proceso iterativo en el cual:

1. Se emplea una estrategia de selección de variables en particular y
2. El modelo resultante se revisa para verificar que las especificaciones funcionales son correctas y que no existen outliers ni observaciones con una alta influencia. Esto podría indicar que el paso1 debe ser repetido.

Es posible que se requieran varias iteraciones para obtener un modelo adecuado.

Ninguno de los procedimientos de selección de variables descrito en este capítulo garantiza que mediante él se produzca la mejor ecuación de regresión para un determinado conjunto de datos. De hecho, usualmente no existe una única mejor ecuación de regresión sino varias igualmente buenas. Debido a que los algoritmos de selección de variables dependen casi totalmente del uso de una computadora, algunas veces el analista se siente tentado a confiar demasiado en los resultados de un procedimiento en particular. Esta tentación debe evitarse. La experiencia previa del analista, juicios personales acerca del fenómeno en estudio y consideraciones subjetivas acerca del mismo, todas ellas, entran en el problema de selección de la ecuación de regresión. Los procedimientos de selección de variables, deben ser usados por el analista como métodos que le permitan explorar la estructura de los datos.

5.1 Consecuencias de la definición incorrecta del modelo

Con el fin de favorecer el interés en seleccionar adecuadamente las variables del modelo, revisaremos brevemente las consecuencias de la selección incorrecta de las variables del mismo.

Supóngase que existen k candidatas a variables regresoras, $X_1, X_2, ..., X_k$, con $n > k+1$ observaciones sobre estas variables y sobre la variable respuesta Y. El modelo completo, que incluye a las k variables regresoras es:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \varepsilon_i, \ \ i = 1, 2, ..., n$$

lo cual es equivalente a

$$\underline{Y} = \mathbf{X}\underline{\beta} + \underline{\varepsilon}$$

Se supone que la lista de variables regresoras contiene a todas las variables que influyen en el modelo. Note que la ecuación anterior contiene un término β_0; mientras que β_0

podría ser una variable candidata en el proceso de selección, normalmente es forzada a ser incluida en el modelo: Se supone que todas las ecuaciones incluyen un término β_0. Sea r el número de variables regresoras que son removidas del modelo. De este modo, el número de variables que son retenidas es $p = k + 1 - r$. Dado que la intercepción es incluida, el modelo formado por este subconjunto tiene $p - 1 = k - r$ de las variables regresoras originales.

De esta forma, la ecuación anterior puede escribirse como

$$\underline{Y} = \mathbf{X}_p\underline{\beta}_p + \mathbf{X}_r\underline{\beta}_r + \underline{\varepsilon}$$

donde la matriz $\mathbf{X}$ ha sido particionada en $\mathbf{X}_p$, una matriz de $n \times p$ cuyas columnas representan la intercepción y las $p - 1$ variables regresoras que serán retenidas en el modelo formado por el subconjunto de variables elegido (en adelante, llamaremos a éste *el modelo incompleto*), y $\mathbf{X}_r$, una matriz de $n \times r$ cuyas columnas representan las variables regresoras que se removerán del modelo completo.

Supóngase que $\underline{\beta}$ se particiona en $\underline{\beta}_p$ y $\underline{\beta}_r$. Para el modelo completo, el estimador por mínimos cuadrados de $\underline{\beta}$ es:

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

y un estimador de la varianza de los residuales, σ^2, es

$$\hat{\sigma}^2_* = \frac{\underline{Y}'\underline{Y} - \hat{\underline{\beta}}^*\mathbf{X}'\underline{Y}}{n-k-1} = \frac{\underline{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\underline{Y}}{n-k-1}$$

Los componentes de $\hat{\underline{\beta}}^*$ se denotan por $\hat{\underline{\beta}}^*_p$ y $\hat{\underline{\beta}}^*_r$; los valores $\hat{Y}_i^*$ son los valores ajustados. Para el modelo incompleto:

$$\underline{Y} = \mathbf{X}_p\underline{\beta}_p + \underline{\varepsilon}$$

El estimador por mínimos cuadrados para $\underline{\beta}_p$ es:

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

El estimador de la varianza de los residuales es

$$\hat{\sigma}^2 = \frac{\underline{Y}'\underline{Y} - \underline{\hat{\beta}}_p \mathbf{X}_p'\underline{Y}}{n-p} = \frac{\underline{Y}'\left(\mathbf{I} - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\right)\underline{Y}}{n-p}$$

y los valores ajustados son $\hat{Y}_i$.

Las propiedades de los estimadores $\underline{\hat{\beta}}_p$ y $\hat{\sigma}^2$ han sido investigadas por varios autores y se pueden resumir como sigue:

1. El valor esperado de $\underline{\hat{\beta}}_p$ es

$$\begin{aligned} E(\underline{\hat{\beta}}_p) &= \underline{\beta}_p + (\mathbf{X}_p'\mathbf{X}_p)^{-1}\left(\mathbf{X}_p'\mathbf{X}_r\underline{\beta}_r\right) \\ &= \underline{\beta}_p + A\underline{\beta}_r \end{aligned}$$

donde $A = (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{X}_r$ (algunas veces A es llamada la matriz *alias*). Así, $\underline{\hat{\beta}}_p$ es un estimador sesgado de $\underline{\beta}_p$, a menos que los coeficientes de regresión correspondientes a las variables removidas ($\underline{\beta}_r$) sean cero o que las variables conservadas sean ortogonales a las variables removidas ($\mathbf{X}_p'\mathbf{X}_r = 0$).

2. Las varianzas de $\underline{\hat{\beta}}_p$ y $\underline{\hat{\beta}}^*$ son $Var(\underline{\hat{\beta}}_p) = \sigma^2(\mathbf{X}_p'\mathbf{X}_p)^{-1}$ y $Var(\underline{\hat{\beta}}^*) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, respectivamente. Además, la matriz $Var(\underline{\hat{\beta}}^*) - Var(\underline{\hat{\beta}}_p)$ es definida semipositiva, esto es, las varianzas de los estimadores por mínimos cuadrados de los parámetros del modelo completo son mayores o iguales que las varianzas de los parámetros correspondientes del modelo incompleto. Consecuentemente, remover variables nunca aumenta las varianzas de los estimadores de los parámetros restantes.

3. Puesto que $\underline{\hat{\beta}}_p$ es un estimador sesgado de $\underline{\beta}_p$ y $\underline{\hat{\beta}}_p^*$ no lo es, parece más razonable comparar la precisión de los estimadores de los parámetros del modelo completo y del modelo incompleto en términos del cuadrado medio del error (Si $\hat{\theta}$ es un estimador de θ, el cuadrado medio del error de $\hat{\theta}$ es $\mathbf{CME}(\hat{\theta}) = Var(\hat{\theta}) + \left(E(\hat{\theta}) - \theta\right)^2$). El cuadrado medio del error de $\underline{\hat{\beta}}_p$ es:

$$\mathbf{CME}(\underline{\hat{\beta}}_p) = \sigma^2(\mathbf{X}_p'\mathbf{X}_p)^{-1} + A\underline{\beta}_r\underline{\beta}_r'A'$$

Si la matriz $Var(\hat{\underline{\beta}}^*) - \underline{\beta}_r \underline{\beta}_r'$ es definida semipositiva, la matriz $Var(\hat{\underline{\beta}}^*) - CME(\hat{\underline{\beta}}_p)$ también lo es. Esto significa que, si las variables regresoras removidas tienen coeficientes de regresión menores que el error estándar de sus estimadores en el modelo completo, entonces los estimadores por mínimos cuadrados de los parámetros en el modelo incompleto tienen un cuadrado medio del error menor que el correspondiente a los estimadores en el modelo completo.

4. El parámetro $\hat{\sigma}_*^2$ del modelo completo es un estimador insesgado de σ^2. Sin embargo, para el modelo incompleto

$$E(\hat{\sigma}^2) = \frac{\sigma^2 + \underline{\beta}_r' \mathbf{X}_r' \left(\mathbf{I} - \mathbf{X}_p(\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\right) \mathbf{X}_r \underline{\beta}_r}{n-p}$$

Esto es, generalmente $\hat{\sigma}^2$ es un estimador sesgado (hacia arriba) de σ^2.

5. Suponga que se desea predecir la respuesta en el punto $\underline{X} = (\underline{X}_p', \underline{X}_r')$. Si se usa el modelo completo, el valor ajustado de éste es $\hat{Y}^* = \underline{X}' \hat{\underline{\beta}}^*$, con media $\underline{X}'\underline{\beta}$ y varianza del valor ajustado:

$$Var(\hat{Y}^*) = \sigma^2 \left(1 + \underline{X}' (\mathbf{X}'\mathbf{X})^{-1} \underline{X}\right)$$

Sin embargo, si se usa el modelo incompleto, $\hat{Y} = \underline{X}_p' \hat{\underline{\beta}}_p$ con media

$$E(\hat{Y}) = \underline{X}_p' \underline{\beta}_p + \underline{X}_p' A \underline{\beta}_r$$

y un cuadrado medio del error del valor ajustado

$$\mathbf{CME}(\hat{Y}) = \sigma^2 \left(1 + \underline{X}_p' (\mathbf{X}_p'\mathbf{X}_p)^{-1} \underline{X}_p\right) + \left(\underline{X}_p' A \underline{\beta}_r - \underline{X}_r' \underline{\beta}_r\right)^2$$

Note que $\hat{Y}$ es un estimador sesgado de Y a menos que $\underline{X}_p' A \underline{\beta}_r = 0$, lo cual ocurre en general sólo si $\mathbf{X}_p'\mathbf{X}_r\underline{\beta}_r = 0$. Más aún, la varianza de $\hat{Y}^*$ del modelo completo no es menor que la varianza del modelo incompleto. En términos del cuadrado medio del error se puede probar que

$$Var(\hat{Y}^*) \geq \mathbf{CME}(\hat{Y})$$

toda vez que la matriz $Var(\hat{\underline{\beta}}^*) - \underline{\beta}_r \underline{\beta}_r'$ sea definida semipositiva.

Nuestra motivación para la selección de variables se puede resumir como sigue: Removiendo variables del modelo puede mejorarse la precisión de los estimadores de los parámetros de las variables que se conservan, aun cuando algunas de las variables no sean despreciables. Esto aplica también en el caso de la varianza de la respuesta pronosticada. Remover variables del modelo introduce (potencialmente) sesgo en los estimadores de los coeficientes de las variables retenidas y de la variable respuesta; sin embargo, si las variables removidas tienen efectos pequeños en el modelo, el cuadrado medio del error de los estimadores sesgados será menor que la varianza de los estimadores insesgados, esto es, el sesgo que se introduce es menor que la reducción en la varianza. Existe también el riesgo de conservar variables despreciables, esto es, variables con coeficientes menores que cero o menores que los errores estándar correspondientes en el modelo completo. Este riesgo es tal que las varianzas de los estimadores de los parámetros y de los valores ajustados se incrementan.

Existe todavía una consideración final que debe tomarse en cuenta. Frecuentemente, los modelos de regresión se construyen a partir de datos que se conocen por su nombre en inglés: *happenstance*, los cuales son datos extraídos de registros históricos. Los datos *happenstance* a menudo están saturados de defectos (incluidos los outliers) e inconsistencias provenientes de cambios en la organización de la recolección de los datos. Estos defectos de los datos pueden tener un impacto importante en el proceso de selección de las variables y llevar a una especificación del modelo incorrecta. Un problema muy común en los datos *happenstance*, es encontrarse con que algunas de las variables regresoras han estado tan controladas que varían sobre un rango muy limitado; frecuentemente, estas variables son las que más influyen en el modelo y debido a ello estuvieron tan controladas (con el fin de mantener la respuesta dentro de límites aceptables). Precisamente por lo limitado del rango de los datos, la variable regresora puede parecer carente de importancia en el ajuste por mínimos cuadrados. Este es un problema serio durante la especificación del modelo y solamente el conocimiento no estadístico del analista acerca del contexto y

ambiente del modelo puede prevenirlo. Cuando el rango de las variables que se piensa son importantes es controlado muy estrechamente, probablemente se requiera recolectar nuevos datos para construir el modelo.

5.2 Criterios para evaluar modelos de regresión incompletos

Los dos aspectos fundamentales del problema de selección de variables son: generar un modelo incompleto y decidir si éste es mejor que otro formado por un subconjunto distinto de variables regresoras. A continuación se presentan criterios para evaluar y comparar modelos de regresión incompletos.

5.2.1 Coeficiente de determinación múltiple

Una medida de qué tan adecuado es un modelo de regresión que es usado ampliamente es el coeficiente de determinación múltiple, R^2, definido como

$$R^2 = \frac{\mathbf{SCR}}{S_{yy}} = 1 - \frac{\mathbf{SCE}}{S_{yy}}$$

Denotemos por R_p^2 al coeficiente de determinación múltiple de un modelo de regresión incompleto con p términos, es decir, $p-1$ variables regresoras y un término β_0. Se tiene entonces

$$R^2 = \frac{\mathbf{SCR}(p)}{S_{yy}} = 1 - \frac{\mathbf{SCE}(p)}{S_{yy}}$$

donde $\mathbf{SCR}(p)$ y $\mathbf{SCE}(p)$ denotan la suma de cuadrados de la regresión y de los residuales, respectivamente, para un modelo incompleto de regresión con p términos. Note que hay $\begin{pmatrix} k \\ p-1 \end{pmatrix}$ valores de R_p^2 para cada valor de p, uno por cada posible modelo de tamaño p.

R_p^2 crece cuando p crece y su valor máximo lo alcanza cuando $p = k + 1$. Por ello, este criterio puede ser usado añadiendo variables regresoras al modelo hasta el punto en el cual una variable adicional ya no sea útil, puesto que el incremento que aporta a R_p^2 es muy pequeño. La aproximación general es ilustrada en la figura siguiente, la cual presenta una gráfica hipotética del valor máximo de R_p^2 para cada modelo incompleto de tamaño p contra p.

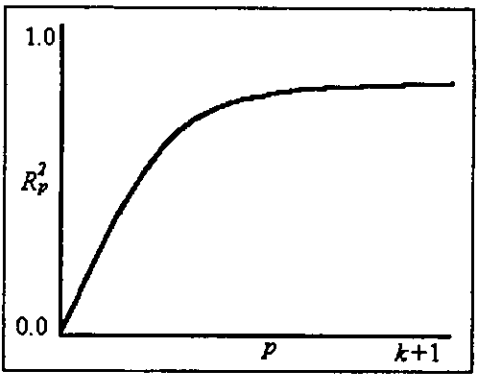


Figura 1: R_p^2 vs. p

En la práctica, se examina una gráfica de este tipo y se selecciona el número de variables regresoras para el modelo final como el punto en el cual la rodilla de la curva se vuelve aparente. Es claro que esto requiere experiencia por parte del analista.

Dado que no se puede hallar un valor óptimo de R^2 para un modelo de regresión incompleto, se debe entonces buscar un valor satisfactorio de éste. Aitkin propuso una solución a este problema diseñando una prueba por medio de la cual se pueda identificar a cualquier modelo incompleto cuyo valor de R^2 no sea sustancialmente distinto del valor de R^2 para el modelo completo. Sea

$$R_0^2 = 1 - \left(1 - R_{k+1}^2\right)\left(1 + d_{\alpha,n,k}\right)$$

donde

$$d_{\alpha,n,k} = \frac{k \cdot F^{\alpha}_{(n,n-k-1)}}{n-k-1}$$

y R^2_{k+1} es el valor de R^2 para el modelo completo. Cualquier subconjunto de variables regresoras que produzca un valor de R^2 mayor que R^2_0 será llamado un conjunto *alfa-adecuado.*

Generalmente R^2 no se puede usar directamente como criterio para escoger el número de regresores para ser incluidos en el modelo. De cualquier forma, para un número fijo de variables p, R^2_p puede ser usado para comparar los $\binom{k}{p-1}$ modelos incompletos generados. Se prefiere a los modelos con los valores de R^2_p mayores.

5.2.2 R^2 ajustado

Para evitar las dificultades en la interpretación de R^2, algunos analistas prefieren usar la estadística $\bar{R}^2_p$ (R^2 ajustado), definida por la ecuación

$$\bar{R}^2_p = 1 - \left(\frac{n-1}{n-p}\right)(1 - R^2_p)$$

La estadística $\bar{R}^2_p$ no necesariamente crece cuando se agregan variables regresoras adicionales al modelo. De hecho, puede probarse que si se añaden s variables regresoras al modelo, $\bar{R}^2_{p+s}$ excederá a $\bar{R}^2_p$ si y sólo si la estadística F-parcial para la prueba de significancia de las s variables regresoras adicionales es mayor que 1. Consecuentemente, un criterio para la selección de un modelo incompleto óptimo es elegir el modelo que posea el valor máximo de $\bar{R}^2_p$. De cualquier forma, este criterio es equivalente al criterio del cuadrado medio de los residuales, mismo que se presenta a continuación.

5.2.3 Cuadrado medio de los residuales

El cuadrado medio de los residuales de un modelo de regresión incompleto, por ejemplo,

$$\mathbf{CME}(p) = \frac{\mathbf{SCE}(p)}{n-p}$$

puede ser usado como un criterio de evaluación del modelo. El comportamiento general de $\mathbf{CME}(p)$ cuando p crece se ilustra en la figura que sigue.

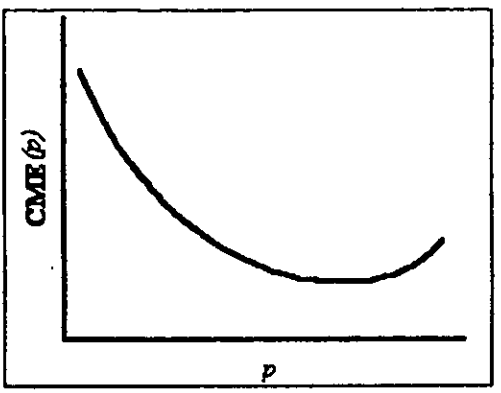


Figura 2: $\mathbf{CME}(p)$ *vs.* p

Dado que la $\mathbf{SCE}(p)$ siempre decrece cuando p crece, $\mathbf{CME}(p)$ inicialmente decrece, luego se estabiliza y eventualmente puede crecer. Esto último ocurre cuando la reducción de $\mathbf{SCE}(p)$ al añadir una variable regresora al modelo no es suficiente para compensar la pérdida de un grado de libertad en el denominador de $\mathbf{CME}(p) = \dfrac{\mathbf{SCE}(p)}{n-p}$. Es decir, añadir una variable regresora a un modelo con p términos, ocasionará que $\mathbf{CME}(p) > \mathbf{CME}(p+1)$ si la reducción en la suma de cuadrados de los residuales es menor que $\mathbf{CME}(p)$. Si se desea utilizar este criterio en la selección del tamaño del modelo, se deberá hacer una gráfica como la anterior y basar la decisión en:

1. El valor mínimo de $\mathbf{CME}(p)$
2. El valor de p tal que $\mathbf{CME}(p)$ sea aproximadamente igual al valor de $\mathbf{CME}$ del modelo completo, o
3. Un valor de p cercano al punto en el cual el valor mínimo de $\mathbf{CME}(p)$ comience a crecer.

El modelo reducido de regresión que minimice a $\mathbf{CME}(p)$ también minimizará a $\bar{R}_p^2$. Para ver esto, note que

$$\begin{aligned}\bar{R}_p^2 &= 1 - \frac{n-1}{n-p}\left(1 - R_p^2\right) \\ &= 1 - \frac{n-1}{n-p}\frac{\mathbf{SCE}(p)}{S_{yy}} \\ &= 1 - \frac{n-1}{S_{yy}}\frac{\mathbf{SCE}(p)}{n-p} \\ &= 1 - \frac{n-1}{S_{yy}}\mathbf{CME}(p)\end{aligned}$$

Es decir, el criterio del mínimo $\mathbf{CME}(p)$ y el máximo de $\bar{R}_p^2$ son equivalentes.

5.2.4 Estadística C_p de Mallows

Mallows propuso un criterio que está relacionado con el cuadrado medio del error del valor ajustado $\hat{Y}_i$, esto es:

$$E\left(\hat{Y}_i - E\left(Y_i\right)\right)^2 = \left(E\left(Y_i\right) - E(\hat{Y}_i)\right)^2 + Var(\hat{Y}_i)$$

Note que $E\left(Y_i\right)$ es la respuesta esperada de X_i en el verdadero modelo de regresión y $E\left(\hat{Y}_i\right)$ es la respuesta esperada del modelo incompleto (con p términos). De esta forma, $E\left(Y_i\right) - E(\hat{Y}_i)$ es el sesgo en el i-ésimo punto. Consecuentemente, los dos términos en el lado derecho de la ecuación anterior son el sesgo al cuadrado y la varianza, respectivamente, del cuadrado medio del error. Denotemos por $SCS(p)$ al sesgo cuadrado total de una ecuación de regresión con p términos,

$$SCS(p) = \sum_{i=1}^{n}\left(E\left(Y_i\right) - E(\hat{Y}_i)\right)^2$$

y definamos el cuadrado medio del error total estandarizado, Γ_p, como

$$\begin{aligned}\Gamma_p &= \frac{1}{\sigma^2}\left[\sum_{i=1}^{n}\left(E\left(Y_i\right) - E(\hat{Y}_i)\right)^2 + \sum_{i=1}^{n} Var(\hat{Y}_i)\right] \\ &= \frac{SCS(p)}{\sigma^2} + \frac{1}{\sigma^2}\sum_{i=1}^{n} Var(\hat{Y}_i)\end{aligned}$$

Se puede probar que

$$\sum_{i=1}^{n} Var(\hat{Y}_i) = p\sigma^2$$

y que el valor estimado de la suma de cuadrados de los residuales de una ecuación con p términos es

$$E\left(\mathbf{SCE}(p)\right) = SCS(p) + (n-p)\sigma^2$$

Sustituyendo $\sum_{i=1}^{n} Var(\hat{Y}_i)$ y $SCS(p)$ en Γ_p se tiene

$$\begin{aligned}\Gamma_p &= \frac{1}{\sigma^2}\left[E\left(\mathbf{SCE}(p)\right) - (n-p)\sigma^2 + p\sigma^2\right] \\ &= \frac{E\left(\mathbf{SCE}(p)\right)}{\sigma^2} - n + 2p\end{aligned}$$

Supóngase que $\hat{\sigma}^2$ es un buen estimador de σ^2. Reemplazando $E\left(\mathbf{SCE}(p)\right)$ por el valor observado $\mathbf{SCE}(p)$ se tiene un estimador de Γ_p, a saber

$$C_p = \frac{\mathbf{SCE}(p)}{\hat{\sigma}^2} - n + 2p$$

Si el modelo de p términos tiene un sesgo despreciable, entonces $SCS(p) = 0$. Consecuentemente, $E\left(\mathbf{SCE}(p)\right) = (n-p)\sigma^2$ y

$$E\left(C_p \mid sesgo = 0\right) = \frac{(n-p)\sigma^2}{\sigma^2} - n + 2p = p$$

Cuando se usa el criterio de C_p, es útil construir una gráfica de C_p como función de p para cada ecuación de regresión como la que mostramos en la siguiente página.

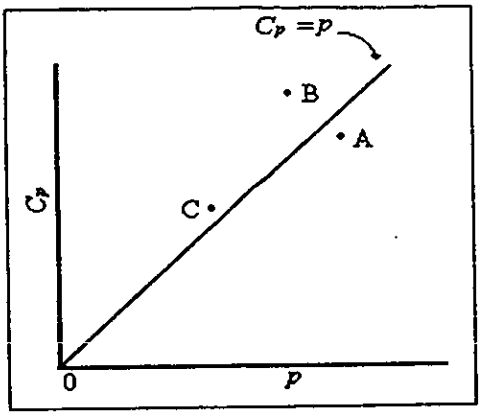


Figura 3: C_p vs. p

Las ecuaciones de regresión cuyo sesgo sea pequeño, tendrán valores de C_p que sean cercanos a línea $C_p = p$ (punto A en la figura anterior), mientras que aquellas ecuaciones con un sesgo considerable se hallarán por encima de ésta (punto B). Generalmente se desea que el valor de C_p sea pequeño. Por ejemplo: aunque el punto C en la figura está por arriba de la línea $Cp = p$, está abajo del punto A y por lo tanto, representa un modelo cuyo error total es menor: Puede ser preferible aceptar cierto sesgo en la ecuación para reducir el error promedio en la predicción.

Para calcular C_p se necesita un estimador insesgado de σ^2. Frecuentemente se usa el cuadrado medio del error del modelo completo para este propósito. Sin embargo, esto ocasiona que $C_p = p = k + 1$ para el modelo completo.

Al usar **CME**$(k + 1)$ como un estimador de σ^2 se está suponiendo que el sesgo del modelo completo es despreciable. Si el modelo completo posee varias variables regresoras que no contribuyan de manera sustancial al modelo (variables cuyos coeficientes de regresión sean iguales o cercanos a cero), entonces **CME**$(k+1)$ frecuentemente sobreestimará a σ^2 y, consecuentemente, los valores de C_p serán pequeños. Si se desea que la estadística C_p funcione adecuadamente, debe usarse un buen estimador de σ^2 .

5.3 Usos del modelo de regresión y criterios para evaluarlo

El criterio que se use para la selección del modelo de regresión debe estar relacionado con el uso que se le dará al modelo. Existen varios usos posibles del modelo, incluyendo descripción de los datos, predicción y estimación, estimación de los parámetros y control.

Si el objetivo del modelo es obtener una buena descripción de un proceso determinado o modelar un sistema complejo, está indicada la búsqueda de ecuaciones de regresión cuyas sumas de cuadrados de los residuales sean pequeñas. Dado que **SCE** se minimiza al usar todas las variables candidatas, usualmemente se decide eliminar variables sólo si el incremento en **SCE** es pequeño. En general, se desea describir el sistema con el menor número posible de variables regresoras, explicando al mismo tiempo la mayor parte de la variabilidad en Y.

Frecuentemente, las ecuaciones de regresión se usan para la predicción de futuras observaciones o para la estimación de la respuesta media. En estos casos se desea seleccionar las variables regresoras de forma tal que el cuadrado medio del error de la predicción sea mínimo. Esto implica, usualmente, que todas las variables regresoras cuyo efecto en el modelo no sea muy grande, deban ser removidas de éste.

Si se está interesado en la estimación de los parámetros, entonces se deben considerar tanto el sesgo que resulte al remover variables, como las varianzas de los estimadores de los coeficientes. Cuando las variables regresoras son altamente multicolineales, los estimadores por mínimos cuadrados de los coeficientes de regresión pueden ser extremadamente pobres.

Cuando un modelo de regresión es usado para control, es importante que los estimadores de los parámetros sean precisos. Esto significa que los errores estándar de los coeficientes de regresión deben ser pequeños. Además, dado que los ajustes hechos sobre las X's para controlar la respuesta Y son proporcionales a las $\hat{\beta}$'s, los coeficientes de regresión deberían reflejar de una forma bastante clara los efectos de las variables

regresoras. Si las variables regresoras son altamente multicolineales, las $\hat{\beta}$'s pueden ser estimadores bastante pobres de los efectos de las variables regresoras individuales.

5.4 Técnicas de cómputo para la selección de variables

Se ha visto que es deseable considerar modelos de regresión que empleen un subconjunto de las candidatas a variables regresoras. Para hallar el subconjunto de variables que vaya a ser usado en la ecuación de regresión definitiva, es natural que se considere ajustar modelos con distintas combinaciones de las variables regresoras candidatas. En esta sección se discutirán algunas técnicas de cómputo para generar modelos de regresión incompletos e ilustrar los criterios de evaluación de estos modelos.

5.4.1 Todas las regresiones posibles

Este procedimiento requiere que el analista ajuste todas las ecuaciones de regresión que contengan una variable regresora candidata, dos variables regresoras candidatas y así sucesivamente. Estas ecuaciones se evalúan de acuerdo a algún criterio apropiado y se selecciona el "mejor" modelo de regresión.

Si suponemos que la intercepción β_0 se incluye en todas las ecuaciones y que existen k variables candidatas, se tienen 2^k ecuaciones para ser ajustadas y examinadas. Por ejemplo, si $k = 4$, entonces existen $2^4 = 16$ ecuaciones posibles, mientras que si $k = 10$, entonces hay $2^{10} = 1024$ ecuaciones de regresión posibles; claramente, el número de ecuaciones que serán examinadas crece rápidamente cuando el número de variables candidatas crece.

Ejemplo 1.- Hald presenta datos relativos al calor desprendido, medido en calorías por gramo de cemento (Y), como una función de cada uno de cuatro ingredientes en la mezcla: aluminato tricálcico (X_1), silicato tricálcico (X_2), alúmino férrico tetracálcico (X_3) y silicato dicálcico (X_4). Los datos se muestran en la tabla 1.

Tabla1: Datos de Hald

Obs. i	Y_i	X_{i1}	X_{i2}	X_{i3}	X_{i4}
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Estos datos serán usados para ilustrar el método de todas las regresiones posibles para la selección de variables.

Puesto que hay 4 variables regresoras candidatas, existen 16 posibles ecuaciones de regresión si siempre se incluye la intercepción β_0. Los resultados de ajustar estas 16 ecuaciones se muestran en la tabla 2. También se incluyen en ella las estadísticas R_p^2, $\bar{R}_p^2$, **SCE**(p), **CME**(p) y C_p.

Tabla 2: Resultados de ajustar todas las regresiones posibles en los Datos de Hald

Número de Variables en el modelo	p	Variables incluidas	SCE(p)	R_p^2	$\bar{R}_p^2$	CME(p)	C_p
Ninguna	1	Ninguna	2,715.7635	0	0	226.3136	442.92
1	2	X_1	1,265.6867	0.53395	0.49158	115.0624	202.55
1	2	X_2	906.3363	0.66627	0.63593	82.3942	142.49
1	2	X_3	1,939.4005	0.28587	0.22095	176.3092	315.16
1	2	X_4	883.8669	0.67459	0.64495	80.3515	138.73
2	3	X_1X_2	57.9045	0.97868	0.97441	5.7904	2.68
2	3	X_1X_3	1,227.0721	0.54817	0.45780	122.7073	198.10
2	3	X_1X_4	74.7621	0.97247	0.96697	7.4762	5.50
2	3	X_2X_3	415.4427	0.84703	0.81644	41.5443	62.44
2	3	X_2X_4	868.8801	0.68006	0.61607	86.8880	138.23
2	3	X_3X_4	175.7380	0.93529	0.92235	17.5738	22.37
3	4	$X_1X_2X_3$	48.1106	0.98228	0.97638	5.3456	3.04
3	4	$X_1X_2X_4$	47.9727	0.98234	0.97645	5.3303	3.02
3	4	$X_1X_3X_4$	50.8361	0.98128	0.97504	5.6485	3.50
3	4	$X_2X_3X_4$	73.8145	0.97282	0.96376	8.2017	7.34
4	5	$X_1X_2X_3X_4$	47.8636	0.98238	0.97356	5.9829	5.00

La tabla 3 contiene los estimadores por mínimos cuadrados de los coeficientes de regresión. La naturaleza parcial de estos estimadores resulta evidente al examinarla. Considérese por ejemplo a X_2. Cuando el modelo está formado solamente por X_2, el estimador por mínimos cuadrados del efecto de éste es 0.789. Si X_4 se incluye también en el modelo,el efecto de X_2 es 0.311, una reducción mayor al 100%. Cuando además se agrega X_3, el efecto cambia a -0.923. Claramente, el estimador por mínimos cuadrados de cada coeficiente de regresión depende fuertemente de las otras variables regresoras

incluidas en el modelo. Estos cambios tan marcados indican que existe una correlación sustancial entre las cuatro variables regresoras.

Tabla 3: Estimadores por mínimos cuadrados de los coeficientes de regresión en los Datos de Hald

Variables incluidas	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
X_1	81.479	1.869			
X_2	57.424		0.789		
X_3	110.203			-1.256	
X_4	117.568				-0.738
X_1X_2	52.577	1.468	0.662		
X_1X_3	72.349	2.312		0.494	
X_1X_4	103.097	1.440			-0.614
X_2X_3	72.075		0.731	-1.008	
X_2X_4	94.160		0.311		-0.457
X_3X_4	131.282			-1.200	-0.724
$X_1X_2X_3$	48.194	1.696	0.657	0.250	
$X_1X_2X_4$	71.648	1.452	0.416		-0.237
$X_2X_3X_4$	203.642		-0.923	-1.448	-1.557
$X_1X_3X_4$	111.684	1.052		-0.410	-0.643
$X_1X_2X_3X_4$	62.405	1.551	0.510	0.102	-0.144

Evaluaremos los modelos incompletos generados usando el criterio de R_p^2. La figura siguiente muestra una gráfica de R_p^2 contra p. Al examinarla resulta claro que cuando el número de variables es mayor que 2, se gana muy poco en términos de R^2 al introducir variables adicionales.

Los modelos (X_1, X_2) y (X_1, X_4) tienen casi el mismo valor de R^2 y, en términos de este criterio, hay poca diferencia si se selecciona uno u otro modelo como el definitivo.

Si examinamos los modelos con una sola variable independiente, veremos que, a la luz del criterio R^2, el mejor modelo lo constituye X_4; tomando esto en consideración, podría decirse que es preferible usar (X_1, X_4).

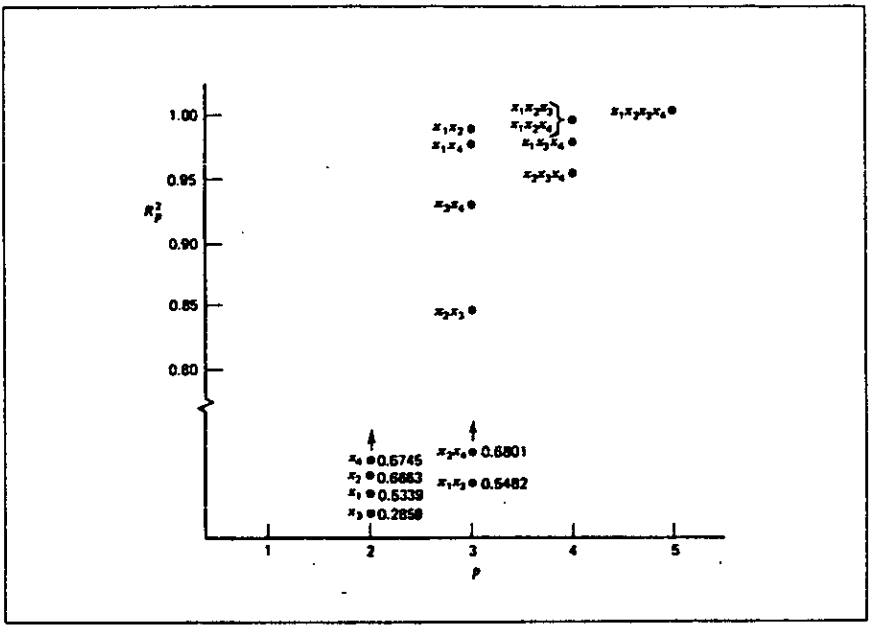


Figura 4: R_p^2 vs. p en los datos de Hald

Usemos la estadística

$$R_0^2 = 1 - \left(1 - R_{k+1}^2\right)\left(1 + d_{\alpha,n,k}\right)$$

donde

$$d_{\alpha,n,k} = \frac{k \cdot F_{(n,n-k-1)}^{\alpha}}{n-k-1}$$

Con $\alpha = 0.05$ se tiene:

$$
\begin{aligned}
R_0^2 &= 1-(1-R_5^2)\left(1+\frac{4\cdot F_{(4,8)}^{.05}}{8}\right) \\
&= 1-(1-0.98238)\left(1+\frac{4(3.84)}{8}\right) \\
&= 0.94855
\end{aligned}
$$

Por lo tanto, cualquier modelo incompleto para el cual $R_p^2 > R_o^2 = 0.94855$ es R^2-adecuado, es decir, su R^2 no es sustancialmente distinta de R_{k+1}^2. A partir de la tabla 2 puede verse que varios modelos satisfacen este criterio y, por ello, la elección del modelo final aun no es clara.

Es ilustrativo observar la correlación existente entre X_i y X_j, asi como entre X_i y Y. Los valores de estas correlaciones simples se muestran en la tabla que sigue:

Tabla 4: Correlaciones simples entre las variables de los Datos de Hald

	X_1	X_2	X_3	X_4	Y
X_1	1.0				
X_2	0.229	1.0			
X_3	-0.824	-0.139	1.0		
X_4	-0.245	-0.973	0.030	1.0	
Y	0.731	0.816	-0.535	-0.821	1.0

Note que los pares de variables (X_1, X_3) y (X_2, X_4) están altamente correlacionados, puesto que $r_{13} = -0.824$ y $r_{24} = -0.973$. Consecuentemente, añadir alguna otra variable regresora cuando X_1 y X_2 ó X_1 y X_4 ya están incluidas en el modelo, será de poca utilidad, puesto que la información contenida en las variables excluidas está esencialmente presente en las variables que ya están incluidas en el modelo. Esta estructura correlativa es responsable en parte de los cambios en los coeficientes de regresión que se observan en la tabla 3.

La figura siguiente muestra una gráfica de **CME**(p) contra p.

El modelo que minimiza la suma de cuadrados de los residuales es (X_1, X_2, X_4), con $\mathbf{CME}(4) = 5.3303$. Note que, como se esperaba, el modelo que minimiza $\mathbf{CME}(p)$ también maximiza el valor ajustado $\bar{R}_p^2$. De cualquier forma, 2 de los otros modelos formados por 3 variables,(X_1, X_2, X_3) y (X_1, X_3, X_4), tienen valores similares del cuadrado medio de los residuales.

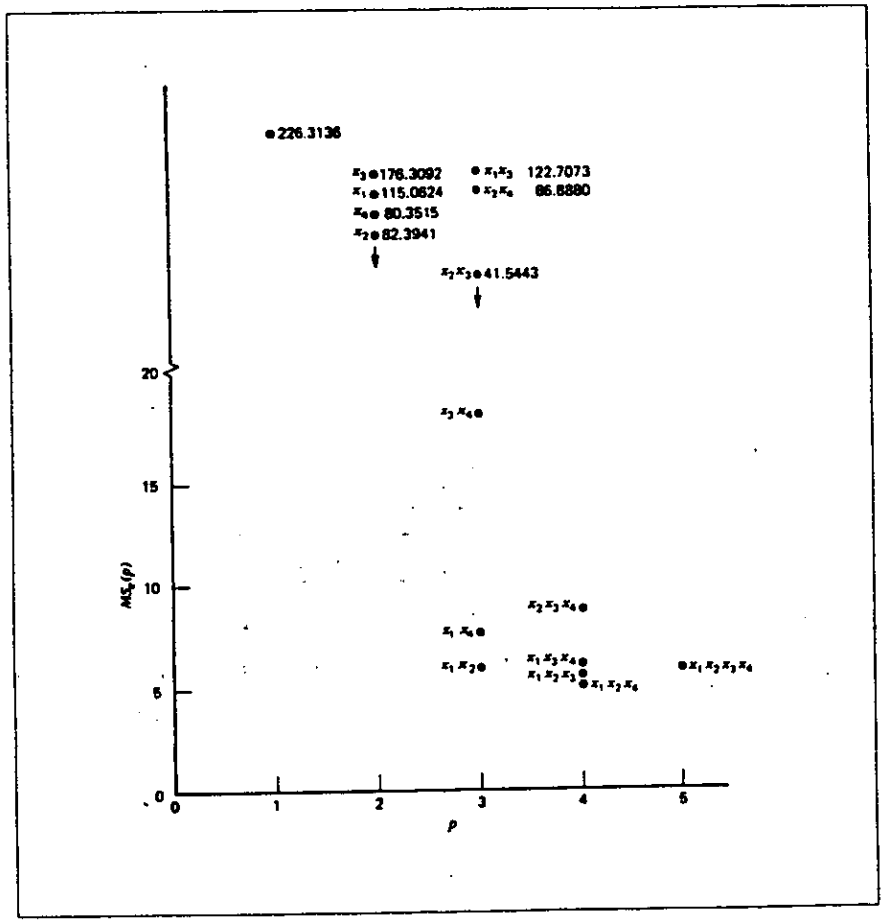


Figura 5: $\mathbf{CME}(p)$ *vs.* p *en los Datos de Hald*

Si (X_1, X_2) ó (X_1, X_4) se encuentran en el modelo, la reducción en el cuadrado medio de los residuales es muy poca cuando se añaden otras variables. El modelo (X_1, X_2) podría ser más apropiado que (X_1, X_4) puesto que tiene un valor menor del cuadrado medio de los residuales.

La figura 6 muestra el comportamiento de C_p contra p.

Para ilustrar los cálculos, suponga que se toma $\hat{\sigma} = 5.9829$ (**CME** del modelo completo) y calculamos C_3 para el modelo (X_1, X_4). Se tiene

$$\begin{aligned} C_3 &= \frac{\mathbf{SCE}(3)}{\hat{\sigma}^2} - n + 2p \\ &= \frac{74.7621}{5.9829} - 13 + 2(3) \\ &= 5.496 \end{aligned}$$

Al examinar la gráfica anterior hallamos que existen 4 modelos que podrían ser aceptables: (X_1, X_2), (X_1, X_2, X_3), (X_1, X_2, X_4) y (X_1, X_3, X_4). Sin considerar factores adicionales como información técnica acerca de las variables o el costo de recolección de los datos, sería apropiado escoger al modelo más simple, (X_1, X_2), como el modelo final, dado que éste tiene el valor más pequeño de C_p.

Con este ejemplo se ilustra todo el procedimiento de cálculo asociado a la definición del modelo con el método de *todas las regresiones posibles.* Nótese que no hay una elección clara de la mejor ecuación de regresión. Frecuentemente encontramos que criterios diferentes sugieren ecuaciones diferentes; por ejemplo: El valor mínimo de C_p lo tiene la ecuación formada por el conjunto (X_1, X_2), mientras que (X_1, X_2, X_4) minimiza a **CME**. Todos los modelos que parezcan apropiados para ser elegidos como el definitivo, deben ser sometidos a las pruebas usuales para verificar qué tan apropiados son, incluyendo la búsqueda de outliers, puntos con demasiada influencia en la ecuación y multicolinealidad.

Generación eficiente de todas las regresiones posibles

Existen varios algoritmos disponibles para la generación eficiente de todas las regresiones posibles. La idea básica de estos algoritmos es efectuar los cálculos para los 2^k modelos

de regresión posibles, de tal forma que los modelos secuenciales difieran solamente por una variable. Esto permite que algunos métodos numéricos muy eficientes sean usados en los cálculos. Algunos de estos algoritmos están incluidos en software comercial.

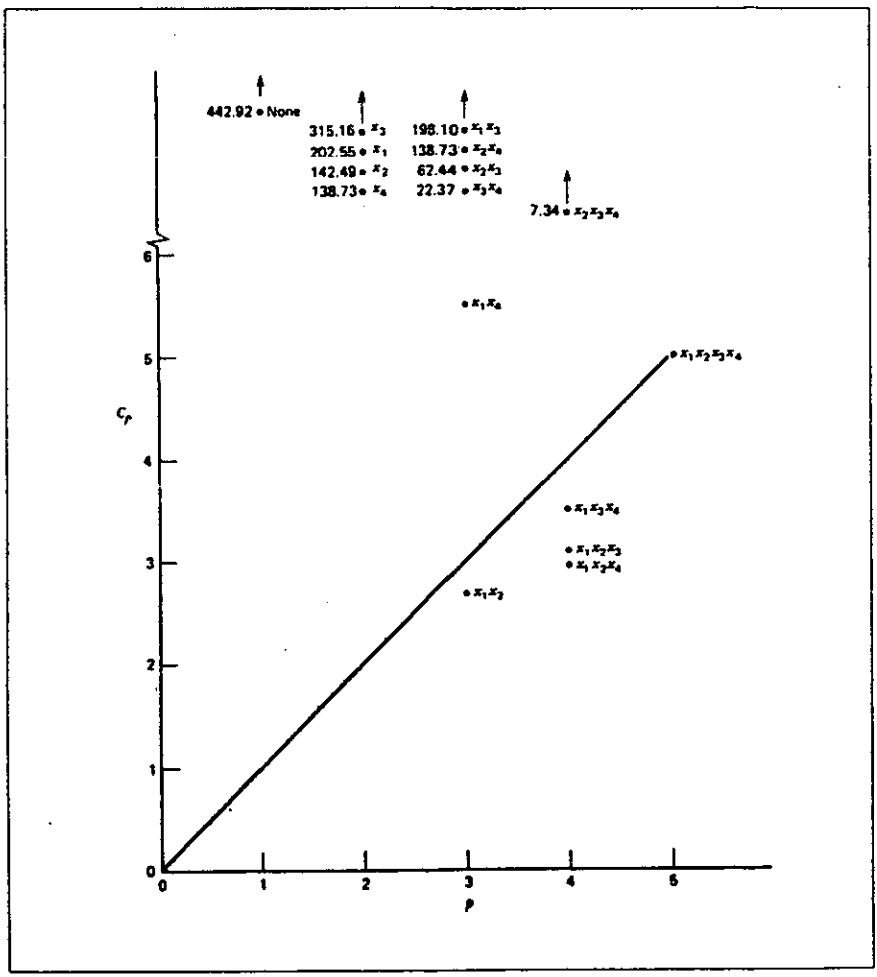


Figura 6: C_p vs. p en los Datos de Hald

5.4.2 Métodos de regresión stepwise

Debido a que los cálculos que se originan de evaluar todas las regresiones posibles puede ser gravoso, se han desarrollado varios métodos para evaluar sólo un número reducido de modelos de regresión, ya sea añadiendo o removiendo variables regresoras de una en una. Estos métodos son conocidos normalmente como *métodos stepwise* y pueden ser clasificados en tres categorías generales:

1. Selección "hacia adelante" (forward)
2. Eliminación "hacia atrás" (backward)
3. Regresión stepwise (que es una combinación de los métodos 1 y 2)

Forward

Este procedimiento inicia con la suposición de que no existe otra variable regresora en el modelo que no sea la intercepción. Los esfuerzos se concentran en encontrar una ecuación "óptima" añadiendo variables al modelo de una en una. La primera variable regresora introducida en el modelo es aquella cuya correlación simple con la variable respuesta Y sea la mayor. Supongamos que esta variable es X_1, X_1 también posee el mayor valor de la estadística F en la prueba de significancia. Esta variable se introduce en el modelo si el valor que produzca de F excede a un cierto valor F_{in} que haya sido previamente seleccionado (F_{in} también puede llamarse F de "entrada"; por conveniencia se dejará a F el subíndice "in"). La segunda variable elegida para entrar al modelo será aquella que, después de ajustar el efecto de la primera variable introducida al modelo en Y, tenga el mayor valor de correlación con ésta. Estas correlaciones son conocidas también como *correlaciones parciales* y son las correlaciones simples entre los residuales de la ecuación de regresión $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_0 X_1$ y los residuales de las ecuaciones de regresión de las otras variables regresoras candidatas en X_1, digamos $\hat{X}_j = \hat{\alpha}_{0j} + \hat{\alpha}_{1j} X_1$, $j = 2, 3, ..., k$.

Supóngase que en el segundo paso, la variable regresora con correlación parcial más alta es X_2. Esto implica que el valor más alto de la estadística F-parcial es

$$F = \frac{\mathbf{SCR}(X_1 \mid X_2)}{\mathbf{CME}(X_1, X_2)}$$

Si $F > F_{in}$, entonces X_2 se agrega al modelo. En general, en cada paso, la variable regresora que tenga la correlación parcial más alta con Y (o, de forma equivalente, aquella cuya F- parcial, dado que las otras variables están incluidas en el modelo, sea la mayor) se incluye en el modelo si su F-parcial es mayor que F_{in}. El proceso termina cuando la mayor de las estadísticas F-parciales no excede a F_{in} o cuando la última variable regresora es incluida en el modelo.

Ejemplo 2.- Se aplicará el método de selección forward a los datos de Hald (tabla 1). La figura 7 muestra los resultados obtenidos cuando el análisis fue efectuado usando el programa SAS. En este programa el usuario especifica el valor de F_{in} eligiendo el nivel de significancia α; el algoritmo toma entonces a $F_{in} = F^{\alpha}_{(1,n-p)}$. En este ejemplo seleccionaremos $\alpha = 0.10$ para determinar F_{in}. Algunos programas requieren un valor numérico determinado para F_{in}. Normalmente se elige un valor que se encuentre entre 2 y 4.

En la tabla 4 se puede ver que la variable regresora más altamente correlacionada con Y es X_4 ($r_{4y} = -0.821$) y, dado que la estadística F-parcial asociada a esa variable es $F = 2.80 > F^{0.10}_{(1,11)} = 3.23$, X_4 se incluye en el modelo. En el segundo paso, la variable regresora con la mayor correlación parcial con Y es X_1, y, como la F-parcial de esta variable es

$$F = \frac{\mathbf{SCR}(X_1 \mid X_4)}{\mathbf{CME}(X_1, X_4)} = \frac{809.1048}{7.4762} = 108.22 > F^{0.10}_{(1,10)} = 3.29$$

X_1 se añade al modelo. En el tercer paso, X_2 muestra la mayor correlación parcial con Y. La estadística F-parcial es

$$F = \frac{\mathbf{SCR}(X_2 \mid X_1, X_4)}{\mathbf{CME}(X_1, X_2, X_4)} = \frac{26.7894}{5.3303} = 5.03 > F^{0.10}_{(1,9)} = 3.36$$

X_2 se añade al modelo. En este punto sólo queda fuera del modelo X_4, cuya F-parcial es menor a $F^{0.10}_{(1,8)} = 3.46$, por lo cual termina el procedimiento de selección con la ecuación

$$\hat{Y} = 71.6483 + 1.4519\,X_1 + 0.4161\,X_2 - 0.2365\,X_4$$

como la ecuación final.

SAS

FORWARD SELECTION PROCEDURE FOR DEPENDENT VARIABLE Y

STEP 1 VARIABLE X4 ENTERED R SQUARE = 0.67454196 C(P) = 138.73083349

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	1	1831.89616002	1831.89616002	22.80	0.0006
ERROR	11	883.86691690	80.35153790		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	117.56793118				
X4	−0.73816181	0.15459600	1831.89616002	22.80	0.0006

BOUNDS ON CONDITION NUMBER: 1. 1

STEP 2 VARIABLE X1 ENTERED R SQUARE = 0.97247105 C(P) = 5.49585082

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	2	2641.00096477	1320.50048238	176.63	0.0001
ERROR	10	74.76211216	7.47621122		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	103.09738164				
X1	1.43995828	0.13841664	809.10480474	108.22	0.0001
X4	−0.61395363	0.04864455	1190.92463664	159.30	0.0001

BOUNDS ON CONDITION NUMBER: 1.064105, 4.256421

STEP 3 VARIABLE X2 ENTERED R SQUARE = 0.98233545 C(P) = 3.01823347

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	3	2667.79034752	889.26344917	166.83	0.0001
ERROR	9	47.97272940	5.33030327		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	71.64830697				
X1	1.45193796	0.11699759	820.90740153	154.01	0.0001
X2	0.41610976	0.18561049	26.78938276	5.03	0.0517
X4	−0.23654022	0.17328779	9.93175378	1.86	0.2054

BOUNDS ON CONDITION NUMBER: 18.94008, 116.3601

NO OTHER VARIABLES MET THE 0.1000 SIGNIFICANCE LEVEL FOR ENTRY INTO THE MODEL.

SUMMARY OF FORWARD SELECTION PROCEDURE FOR DEPENDENT VARIABLE Y

STEP	VARIABLE ENTERED	NUMBER IN	PARTIAL R**2	MODEL R**2	C(P)	F	PROB > F
1	X4	1	0.6745	0.6745	138.731	22.7985	0.0006
2	X1	2	0.2979	0.9725	5.496	108.2239	0.0001
3	X2	3	0.0099	0.9823	3.018	5.0259	0.0517

Figura 7: Resultados obtenidos al aplicar el procedimiento de selección forward de SAS a los datos de Hald

Backward

El método forward empieza con cero variables regresoras en el modelo y va seleccionando variables hasta que se obtiene un modelo adecuado. El método backward intenta generar un buen modelo trabajando en la dirección opuesta, esto es, se inicia el procedimiento con un modelo que contenga las k variables candidatas. A continuación, la estadística F-parcial de cada variable se calcula, como si ésta fuera la última variable a incluir en el modelo; la menor de todas las F-parciales es comparada con F_{out}, un valor "de salida" previamente seleccionado. Si esta F-parcial es menor que F_{out}, la variable regresora asociada con ella es removida del modelo. A continuación se ajusta una ecuación de regresión con las $k-1$ variables restantes, se calculan las F-parciales y el proceso se repite.

El procedimiento backward termina cuando la menor de las F-parciales es mayor o igual que F_{out}. Éste suele ser un algoritmo bastante eficiente y tiene además la ventaja de que permite apreciar el efecto de incluir todas las variables candidatas, de tal forma que nada que parezca ser obvio sea omitido.

Ejemplo 3.- El método backward se ilustra utilizando los datos de Hald. La figura 8 muestra los resultados obtenidos al aplicar el algoritmo incluido en SAS en estos datos. Se selecciona $\alpha = 0.10$ para que el programa calcule F_{out} como $F^{\alpha}_{(1,n-p)}$. En el paso 0 se muestran los resultados obtenidos al ajustar la ecuación de regresión del modelo completo.

La menor de las F-parciales es $F = 0.02$ y está asociada a X_3. Puesto que $F = 0.02 < F^{0.10}_{(1,8)} = 3.46$, X_3 es removida del modelo. En el paso 1, se muestra la ecuación de regresión ajustada con las variables X_1, X_2 y X_4. La menor de las estadísticas parciales está asociada a X_4 y es igual a 1.86. Se tiene $F = 1.86 < F^{0.10}_{(1,9)} = 3.36$, por lo tanto, X_4 se remueve del modelo. En el paso 2 se observa la ecuación de la recta ajustada del modelo que contiene las variables X_1 y X_2. La menor de las estadísticas parciales es $F = 145.52$, asociada a X_1 y puesto que ésta excede a $F_{out} = F^{0.10}_{(1,10)} = 3.29$, el procedimiento de selección se detiene en este punto. La ecuación de regresión final es

$$\hat{Y} = 52.5773 + 1.4683\,X_1 + 0.6623\,X_2$$

Note que este modelo es diferente del obtenido por el método forward. Más aun, el método de todas las regresiones posibles sugiere este mismo modelo como el mejor.

SAS

BACKWARD ELIMINATION PROCEDURE FOR DEPENDENT VARIABLE Y

STEP 0 ALL VARIABLES ENTERED R SQUARE = 0.98237562 C(P) = 5.00000000

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	4	2667.89943757	666.97485939	111.48	0.0001
ERROR	8	47.86363935	5.98295492		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	62.40536930				
X1	1.55110265	0.74476987	25.95091138	4.34	0.0708
X2	0.51016758	0.72378800	2.97247824	0.50	0.5009
X3	0.10190940	0.75470905	0.10909005	0.02	0.8959
X4	−0.14406103	0.70905206	0.24697472	0.04	0.8441

BOUNDS ON CONDITION NUMBER: 282.5129, 2489.203

STEP 1 VARIABLE X3 REMOVED R SQUARE = 0.98233545 C(P) = 3.01823347

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	3	2667.79034752	889.26344917	166.83	0.0001
ERROR	9	47.97272940	5.33030327		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	71.64830697				
X1	1.45193796	0.11699759	820.90740153	154.01	0.0001
X2	0.41610976	0.18561049	26.78938276	5.03	0.0517
X4	−0.23654022	0.17328779	9.93175378	1.86	0.2054

BOUNDS ON CONDITION NUMBER: 18.94008, 116.3601

STEP 2 VARIABLE X4 REMOVED R SQUARE = 0.97867837 C(P) = 2.67824160

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	2	2657.85859375	1328.92929687	229.50	0.0001
ERROR	10	57.90448318	5.79044832		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	52.57734888				
X1	1.46830574	0.12130092	848.43186034	146.52	0.0001
X2	0.66225049	0.04585472	1207.78226562	208.58	0.0001

BOUNDS ON CONDITION NUMBER: 1.055129, 4.220516

ALL VARIABLES IN THE MODEL ARE SIGNIFICANT AT THE 0.1000 LEVEL.

SUMMARY OF BACKWARD ELIMINATION PROCEDURE FOR DEPENDENT VARIABLE Y

STEP	VARIABLE REMOVED	NUMBER IN	PARTIAL R**2	MODEL R**2	C(P)	F	PROB > F
1	X3	3	0.0000	0.9823	3.018	0.0182	0.8959
2	X4	2	0.0037	0.9787	2.678	1.8633	0.2054

Figura 8: Resultados obtenidos al aplicar el procedimiento de selección backward de SAS a los datos de Hald

Regresión Stepwise

Los dos procedimientos antes descritos sugieren la existencia de varias combinaciones posibles entre ellos. Una de las más populares es el algoritmo de regresión stepwise de Efroymsom. Este algoritmo es una modificación del método forward, en el cual, en cada paso, todas las variables regresoras previamente incluidas en el modelo son re-evaluadas a partir de sus estadísticas F-parciales. Una variable incluida en el modelo en uno de los primeros pasos podría en un paso subsecuente resultar redundante, debido a la relación que guarda con las demás variables incluidas en ese momento en la ecuación de regresión.

El método stepwise requiere de dos valores límite: F_{in} y F_{out}. Algunas personas gustan de seleccionarlos de forma tal que $F_{in} = F_{out}$, aunque esto no es necesario. Normalmente se elige $F_{in} > F_{out}$, haciendo relativamente más difícil añadir una variable que removerla. Ejemplo 4.- En la figura 9 se presentan los resultados obtenidos al usar el algoritmo de regresión stepwise de SAS en los datos de Hald. Se ha elegido $F_{in} = F_{out} = F^{(\alpha)}_{(1,n-p)}$ con $\alpha = 0.10$. En el paso 1, el procedimiento empieza con cero variables en el modelo y trata de incluir a X_4. Puesto que la F-parcial asociada excede a $F_{in} = 3.23$, X_4 se añade al modelo.

En el paso 2, X_2 es añadido al modelo; en este modelo, si la F-parcial de X_4 fuera menor que $F_{out} = F^{0.1}_{(1,10)} = 3.29$, X_4 tendría que ser removida del modelo. Sin embargo, la F-parcial de X_4 en este paso es $F = 159.30$, así que X_4 permanece en el modelo. En el paso 3, las F-parciales de X_1 y X_4 se comparan con $F_{out} = F^{0.1}_{(1,10)} = 3.36$. Dado que para X_4, $F = 1.86 < 3.36$, X_4,es removida del modelo

El paso 4 muestra los resultados obtenidos después de remover X_4 del modelo. En este punto la única variable regresora candidata es X_3, misma que no puede ser incluida en el modelo puesto que su F-parcial no es mayor que F_{in}. Así, el procedimiento de selección termina con el modelo

$$\hat{Y} = 52.5773 + 1.4683\,X_1 + 0.6623\,X_2$$

que es la mismo que el identificado como el mejor por el procedimiento backward.

SAS

STEPWISE REGRESSION PROCEDURE FOR DEPENDENT VARIABLE Y

NOTE: SLENTRY AND SLSTAY HAVE BEEN SET TO .10 FOR THE STEPWISE TECHNIQUE.

STEP 1 VARIABLE X4 ENTERED R SQUARE = 0.67454196 C(P) = 138.73083349

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	1	1831.89616002	1831.89616002	22.80	0.0006
ERROR	11	883.86691690	80.35153790		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	117.56793118				
X4	−0.73816181	0.15459600	1831.89616002	22.80	0.0006

BOUNDS ON CONDITION NUMBER: 1, 1

STEP 2 VARIABLE X1 ENTERED R SQUARE = 0.97247105 C(P) = 5.49585082

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	2	2641.00096477	1320.50048238	176.63	0.0001
ERROR	10	74.76211216	7.47621122		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	103.09738164				
X1	1.43995828	0.13841664	809.10480474	108.22	0.0001
X4	−0.61395363	0.04864455	1190.92463664	159.30	0.0001

BOUNDS ON CONDITION NUMBER: 1.064105, 4.256421

STEP 3 VARIABLE X2 ENTERED R SQUARE = 0.98233545 C(P) = 3.01823347

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	3	2667.79034752	889.26344917	166.83	0.0001
ERROR	9	47.97272940	5.33030327		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	71.64830697				
X1	1.45193796	0.11699759	820.90740153	154.01	0.0001
X2	0.41610976	0.18561049	26.78938276	5.03	0.0517
X4	−0.23654022	0.17328779	9.93175378	1.86	0.2054

BOUNDS ON CONDITION NUMBER: 18.94008, 116.3601

STEP 4 VARIABLE X4 REMOVED R SQUARE = 0.97867837 C(P) = 2.67824160

	DF	SUM OF SQUARES	MEAN SQUARE	F	PROB > F
REGRESSION	2	2657.85859375	1328.92929687	229.50	0.0001
ERROR	10	57.90448318	5.79044832		
TOTAL	12	2715.76307692			

	B VALUE	STD ERROR	TYPE II SS	F	PROB > F
INTERCEPT	52.57734888				
X1	1.46830574	0.12130092	848.43186034	146.52	0.0001
X2	0.66225049	0.04585472	1207.78226562	208.58	0.0001

BOUNDS ON CONDITION NUMBER: 1.055129, 4.220516

NO OTHER VARIABLES MET THE 0.1000 SIGNIFICANCE LEVEL FOR ENTRY INTO THE MODEL.

SUMMARY OF STEPWISE REGRESSION PROCEDURE FOR DEPENDENT VARIABLE Y

STEP	VARIABLE ENTERED	VARIABLE REMOVED	NUMBER IN	PARTIAL R**2	MODEL R**2	C(P)	F	PROB > F
1	X4		1	0.6745	0.6745	138.731	22.7985	0.0006
2	X1		2	0.2979	0.9725	5.496	108.2239	0.0001
3	X2		3	0.0099	0.9823	3.018	5.0259	0.0517
4		X4	2	0.0037	0.9787	2.678	1.8633	0.2054

Figura 9: Resultados obtenidos al aplicar el procedimiento de selección stepwise de SAS a los datos de Hald

5.4.3 Comentarios generales acerca de los métodos de regresión stepwise

Los algoritmos de regresión antes descritos, han sido criticados usando varios argumentos; el más común es el hecho de que ninguno de ellos garantiza que el mejor subconjunto de variables regresoras sea identificado. Más aun, puesto que todos estos procedimientos terminan con una ecuación final, usuarios con poca experiencia podrían concluir que han encontrado un modelo que es óptimo en algún sentido. Parte del problema es que, aparentemente, no existe un mejor modelo, sino varios igualmente buenos.

El analista también debe tomar en consideración que el orden en el cual las variables regresoras son agregadas al modelo, no necesariamente refleja el orden de importancia de ellas. No es inusual encontrar que una variable regresora agregada al modelo en los primeros pasos de un procedimiento se vuelva despreciable en una etapa subsecuente del mismo (Esto es evidente en los datos de Hald: Al ser aplicado el método forward en esos datos, la primera variable elegida para entar en el modelo es X_4. Sin embargo, cuando X_2 se incluye en éste, X_4 ya no es requerida debido a la alta correlación que muestra tener con X_2. De hecho, este es un problema general del método forward: una vez que una variable es incluida en el modelo, ya no puede ser removida).

Note que los tres métodos que hemos discutido (forward, backward y stepwise) no necesariamente llevan a la misma ecuación de regresión. La correlación entre las variables regresoras afecta el orden de entrada y remoción. Por ejemplo, usando los datos de Hald, encontramos que las variables elegidas por cada uno de los procedimientos fueron como sigue:

Forward: X_1, X_2, X_4

Backward: X_1, X_2

Stepwise: X_1, X_2

Algunos usuarios recomiendan que todos los procedimientos de selección de variables sean aplicados a los datos, con la esperanza ya sea de observar algún resultado común o de descubrir alguna característica de la estructura de los datos que deba ser examinada

usando sólo uno de los procedimientos de selección. Más aun, no necesariamente debe existir consistencia entre los resultados obtenidos mediante el método de todas las regresiones posibles y los métodos de regresión stepwise. Sin embargo, el método forward tiende a ser consistente con el método de todas las regresiones posibles para subconjuntos pequeños de variables, pero no es así para subconjuntos grandes.

Por estas razones, los métodos de selección stepwise deben ser usados con cautela.

5.4.4 Criterios de decisión para detener un procedimiento de selección de variables

Elegir los valores límite F_{in} y F_{out} en los procedimientos de selección stepwise, puede ser visto como una regla específica para detener el algoritmo. Algunos programas permiten al analista elegir estos valores directamente, mientras que otros requieren una tasa de ocurrencia del error tipo I para generarlos. Sin embargo, dado que la estadística F-parcial examinada en cada paso es el máximo de varias variables correlacionadas, considerar a α como un nivel de significancia o como una tasa de ocurrencia del error tipo I es engañoso. Varios autores han investigado este problema y muy poco se ha avanzado en los intentos, ya sea de encontrar las condiciones bajo las cuales el referido nivel de significancia en F tiene significado o de desarrollar la distribución exacta de las estadísticas F_{in} y F_{out}.

Algunos usuarios prefieren elegir valores relativamente pequeños de F_{in} y F_{out}, de tal forma que variables regresoras que de otra forma hubieran sido rechazadas por valores más convencionales de F deban ser investigadas. En una situación extrema, se deberían elegir valores de F_{in} y F_{out} tales que todas las variables regresoras sean elegidas por el procedimiento forward o rechazadas por el método backward, revelando de esta forma un modelo tentativo para cada tamaño del modelo p. Así, cada uno de estos modelos deberá ser evaluado con criterios tales como C_p ó **CME** para determinar el modelo final. Si se elige esta estrategia, debe tenerse en mente que el modelo final no es necesariamente óptimo y que es incluso probable que el mejor modelo haya sido pasado por alto.

Otra posibilidad es ejecutar varias veces el mismo procedimiento de selección usan-

do diferentes valores de F_{in} y F_{out} para observar el efecto de cada uno de ellos en los subconjuntos obtenidos.

La elección de F_{in} y F_{out} recae fundamentalmente en la preferencia personal del analista y suele tomarse considerable libertad en esta área.

5.5 Algunas consideraciones finales

En este capítulo hemos discutido varios procedimientos para la selección de variables, tres de los cuales son del tipo stepwise. La principal ventaja de ellos es que son rápidos y que están disponibles en forma de software comercial, la mayor parte del cual puede implementarse en computadoras personales. Sus desventajas son que no necesariamente generan modelos finales que sean los mejores con respecto a algún criterio estándar y, más aún, dado que están diseñados para producir una única ecuación final, un usuario con poca experiencia podría creer que ese modelo es óptimo en algún sentido.

Por otro lado, el procedimiento de todas las regresiones posibles, identifica los subconjuntos de variables que forman el mejor modelo con respecto a cualquier criterio que el analista imponga. Para problemas que tengan entre 20 y 30 variables regresoras candidatas, el costo (en tiempo) de aplicar cualquiera de los procedimientos de regresión aquí descritos es más o menos el mismo. Sin embargo, los métodos stepwise están más extendidos entre los programas estadísticos que el de todas las regresiones posibles.

Cuando el número de variables candidatas es demasiado grande, es posible usar el método de todas las regresiones posibles si se usa una estrategia en dos etapas: Algún método stepwise puede aplicarse para visualizar a las variables regresoras, eliminando aquellas que tengan efectos despreciables en la variable de respuesta, de tal forma que se pueda llegar a una lista menor de variables candidatas y el subconjunto que resulte pueda ser analizado con el método de todas las regresiones posibles.

Durante el análisis, siempre deben utilizarse los conocimientos disponibles acerca del contexto en el que ocurre el fenómeno que va a ser investigado, así como el sentido común

para evaluar a las variables disponibles. El tiempo invertido en evaluar teóricamente el problema antes de acudir a la computadora, casi siempre es redituable. Frecuentemente se encuentra que algunas variables pueden ser eliminadas por razones lógicas inherentes al problema.

Se han discutido diversos criterios para evaluar modelos incompletos, tales como la estadística C_p de Mallow y el CME. Sin embargo, la elección del modelo de regresión final no tiene un criterio concreto. Además de los criterios de evaluación formales, es importante hacerse las siguientes preguntas:

1. ¿La ecuación de regresión parece razonable?, es decir, ¿Las variables regresoras incluidas en el modelo tienen sentido en el contexto del problema?

2. ¿El modelo puede ser usado para su propósito original?. Por ejemplo, un modelo cuyo propósito original sea la predicción y que contenga una variable regresora que no pueda ser observada en el momento en el que la predicción es requerida, es inútil. Si el costo de recolección de los datos de una de las variables independientes es demasiado alto, el modelo es inútil también.

3. ¿Los coeficientes de regresión tienen valores razonables?. Esto es, ¿Los signos y magnitudes de los coeficientes de regresión son realistas y los errores estándar relativamente pequeños?

4. Cuando se examina qué tan adecuado es el modelo, ¿Se observan resultados satisfactorios?. Por ejemplo, ¿En las gráficas de residuales se observan outliers o puntos con demasiada influencia en el modelo que pudieran estar controlando la recta ajustada?

Si estas cuatro preguntas se toman con seriedad y las respuestas son aplicadas estrictamente, en algunos (quizá en la mayoría) de los casos puede no haber una ecuación de regresión definitiva que sea totalmente satisfactoria. Es claro entonces que se requiere de experiencia y conocimientos acerca del contexto del fenómeno y del propósito del modelo.

Finalmente, aunque la ecuación de regresión se ajuste adecuadamente al modelo, no existe la seguridad de que prediga observaciones nuevas con precisión. Se recomienda que la capacidad de predicción del modelo sea evaluada observando su comportamiento sobre datos nuevos que no hayan sido utilizados en la construcción de éste. Si esto no puede llevarse a cabo fácilmente, entonces algunos de los datos originales pueden ser apartados para luego ser usados con este propósito.

Capítulo 6

Elementos de álgebra lineal

6.1 Matrices

6.1.1 Definiciones básicas

Una *matriz* se define como un arreglo rectangular de elementos ordenados en renglones o en columnas. En general, una matriz A se denota como

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

En el presente texto generalmente usamos mayúsculas para denotar a las matrices. Los subíndices en cada uno de los elementos denotan el renglón y la columna, respectivamente, en la cual se encuentran ubicados; así, el (ij)-ésimo elemento a_{ij} está en el renglón i columna j. Se dice que dos matrices A y B son iguales si $a_{ij} = b_{ij}$, $\forall\ i, j$. El *orden* de una matriz es su tamaño en renglones y columnas. A es una matriz de orden m por n; en adelante diremos que A es de orden $m \times n$ o de tamaño $m \times n$. Algunas veces denotaremos esta condición mediante un subíndice, es decir $A_{m\times n}$.

El *rango* de una matriz se define como el número de columnas linealmente independientes en la matriz. Los elementos de un subconjunto de columnas de una matriz son

linealmente independientes si ninguna de las columnas puede ser expresada como combinación lineal de las demás. Si no existe ninguna dependencia lineal entre las columnas de una matriz, se dice que ésta es de *rango completo* o bien, *no singular*.

Ejemplo 1.- La matriz

$$A = \begin{pmatrix} 1 & 2 & 4 \\ 3 & 0 & 6 \\ 5 & 3 & 13 \end{pmatrix}$$

es una matriz de orden 3×3; el elemento a_{23} es igual a 6. El rango de A, denotado por $rango(A)$ es 2. En efecto, obsérvese que

$$2(1, 3, 5) + (2, 0, 3) = (4, 6, 13)$$

la cual es la última columna de A. De hecho, cualquier columna de A puede obtenerse a partir de una combinación lineal de las otras dos. Por otro lado, cualquier par de columnas de A es linealmente independiente, puesto que ninguna de ellas es un múltiplo de cualquier otra. Así, $rango(A) = 2$.

6.1.2 Tipos especiales de matrices.

Un *vector* es una matriz que tiene sólo un renglón o una sola columna. Por ejemplo,

$$\underline{a} = \begin{pmatrix} 1 \\ 6 \\ 5 \\ 2 \end{pmatrix}$$

es un vector columna de orden 4×1

$$\underline{\varepsilon} = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$$

es un vector renglón de orden 1×3.

En general, los vectores los definiremos como vectores columna y serán denotados indistintamente por letras mayúsculas y minúsculas subrayadas.

A un número le llamaremos *escalar*.

Una *matriz cuadrada* es una matriz cuyo número de columnas es igual al número de renglones; en el ejemplo anterior, A es una matriz cuadrada de 3×3. En algunas

ocasiones, para denotar una matriz cuadrada de orden $n \times n$ usaremos un subíndice, es decir A_n, también diremos que la matriz cuadrada A es de orden n.

Una *matriz diagonal* es una matriz cuadrada en la cual todos los elementos son cero, excepto alguno de los que se hallan en la diagonal principal, es decir, los elementos de la forma a_{ij}, $i = j$. Las matrices

$$A = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 4 \end{pmatrix} \text{ y } B = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 13 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -16 \end{pmatrix}$$

son matrices diagonales.

Una *matriz identidad* es una matriz diagonal cuyos elementos en la diagonal principal son todos iguales a uno. Estas matrices serán denotadas por $\mathbf{I}$ o bien por $\mathbf{I}_n$, donde el subíndice n denota el orden de la matriz. Así

$$\mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

es una matriz identidad de 2×2.

Una *matriz simétrica* es una matriz cuadrada en la cual el elemento a_{ij} es igual al elemento a_{ji}, para toda i, j.

$$A = \begin{pmatrix} 9 & 2 & -6 \\ 2 & 5 & -2 \\ -6 & -2 & 15 \end{pmatrix}$$

es una matriz simétrica. Note que la primera columna es igual al primer renglón, la segunda columna es igual al segundo renglón y la tercera columna es igual al tercer renglón.

6.1.3 Operaciones con matrices

Si A es una matriz cuadrada de orden n, la *traza* de A, denotada por $tr(A)$, es la suma de los elementos en su diagonal principal, es decir:

$$tr(A) = \sum_{i=1}^{n} a_{ii}$$

Por ejemplo, si

$$A = \begin{pmatrix} -4 & 0 & 6 \\ 0 & 1 & -3 \\ 6 & -3 & 5 \end{pmatrix}$$

se tiene

$$tr(A) = \sum_{i=1}^{n} a_{ii} = -4 + 1 + 5 = -2$$

La *transpuesta* de una matriz A, denotada por A', es el resultado de intercambiar los renglones y las columnas; es decir, el primer renglón de A se convierte en la primera columna de la transpuesta de A, A', el segundo renglón de A en la segunda columna de A' y así sucesivamente. En general, el (i,j)-ésimo elemento de A es el (j,i)-ésimo elemento de A'.

Ejemplo 2.- Si

$$A = \begin{pmatrix} 4 & 1 \\ 2 & 6 \\ 5 & 7 \\ 8 & 3 \end{pmatrix}$$

la transpuesta de A es

$$A' = \begin{pmatrix} 4 & 2 & 5 & 8 \\ 1 & 6 & 7 & 3 \end{pmatrix}$$

A partir de este operador podemos dar una definición alternativa de una matriz simétrica; una matriz simétrica A es una matriz idéntica a su transpuesta, es decir, es una matriz tal que $A = A'$.

La *suma* de dos matrices está definida si y sólo si ambas son del mismo orden. Si A y B son del mismo orden, se define $A + B = C$, C del mismo orden, una nueva matriz en la cual $c_{ij} = a_{ij} + b_{ij}$, $\forall i, j$. Esto es, se suman los elementos correspondientes de A y B para obtener los elementos de C. Por ejemplo, si

$$A = \begin{pmatrix} 2 & 0 \\ -5 & 6 \end{pmatrix} \text{ y } B = \begin{pmatrix} -3 & 6 \\ 4 & 1 \end{pmatrix}, C = A + B = \begin{pmatrix} -1 & 6 \\ -1 & 7 \end{pmatrix}$$

La adición de matrices es conmutativa $A + B = B + A$ y asociativa $(A + B) + C = A + (B + C)$.

Si λ es un escalar se define a la multiplicación escalar de λ por A a la matriz λA cuyos elementos son el resultado de multiplicar λ por los elementos correspondientes de A. Por ejemplo, si

$$A = \begin{pmatrix} 2 & 0 \\ -5 & 6 \end{pmatrix} \text{ y } \lambda = -5, \text{ entonces } \lambda A = \begin{pmatrix} -10 & 0 \\ 25 & -30 \end{pmatrix}$$

La propiedad distributiva se cumple en el caso de la multiplicación escalar, es decir, $\lambda(A+B) = \lambda A + \lambda B$ y $(\lambda + \mu)A = \lambda A + \mu A$.

A partir de las reglas de la adición de matrices y la multiplicación escalar se deduce que la *sustracción* $A - B = C$ es la matriz cuyos elementos son $c_{ij} = a_{ij} - b_{ij}$.

La multiplicación de dos matrices está definida si y solamente si el número de columnas de la primera matriz es igual al número de renglones de la segunda. Si A es de orden $m \times n$ y B es de orden $r \times p$ el producto AB existe si y sólo si $n = r$, el producto BA existe si y sólo si $p = m$.

Este operador puede definirse de un modo más sencillo si consideramos primero la multiplicación entre vectores. Sean $\underline{a}' = (a_1, a_2, a_3)$ y $\underline{b} = (b_1, b_2, b_3)$. El producto

$$\underline{a}'\underline{b} = (a_1, a_2, a_3) \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = a_1b_1 + a_2b_2 + a_3b_3$$

El resultado es un escalar igual a la suma de los productos de los elementos correspondientes. Sean $\underline{a}' = (3, 6, 1)$ y $\underline{b}' = (2, 4, 8)$

$$\underline{a}'\,\underline{b} = (3, 6, 1) \begin{pmatrix} 2 \\ 4 \\ 8 \end{pmatrix} = 2 \cdot 3 + 6 \cdot 4 + 1 \cdot 8 = 6 + 24 + 8 = 38$$

La multiplicación de matrices se define como una secuencia de multiplicaciones vectoriales. Sean

$\underline{a}_1' = (a_{11}, a_{12}, a_{13})$ y $\underline{a}_2' = (a_{21}, a_{22}, a_{23})$. La matriz

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$$

donde $\underline{a}_1'$ y $\underline{a}_2'$ son los vectores renglón de 1×3 de A, puede ser escrita como

$$A = \begin{pmatrix} \underline{a}_1' \\ \underline{a}_2' \end{pmatrix}$$

Análogamente escribamos

$$B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} \text{ como } B = (\underline{b}_1, \underline{b}_2)$$

donde $\underline{b}_1$ y $\underline{b}_2$ son los vectores columna de 3×1 de B.

De esta manera, el producto AB es la matriz de 2×2

$$AB = C = \begin{pmatrix} \underline{a}_1'\underline{b}_1 & \underline{a}_1'\underline{b}_2 \\ \underline{a}_2'\underline{b}_1 & \underline{a}_2'\underline{b}_2 \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

donde

$$\begin{pmatrix} c_{11} = \underline{a}_1'\underline{b}_1 = \sum_{j=1}^{3} a_{1j}b_{j1} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} \\ c_{12} = \underline{a}_1'\underline{b}_2 = \sum_{j=1}^{3} a_{1j}b_{j2} = a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ c_{21} = \underline{a}_2'\underline{b}_1 = \sum_{j=1}^{3} a_{2j}b_{j1} = a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \\ c_{22} = \underline{a}_2'\underline{b}_1 = \sum_{j=1}^{3} a_{2j}b_{j2} = a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{pmatrix}$$

En general, si A es de orden $m \times n$ y B es de orden $n \times p$ el producto AB es una matriz C de orden $m \times p$ cuyo (i,j)-ésimo elemento es $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$.

Ejemplo 3.- Sean

$$A = \begin{pmatrix} 1 & 2 \\ 4 & 5 \\ 3 & 0 \end{pmatrix} \text{ y } B = \begin{pmatrix} -1 \\ 3 \end{pmatrix}$$

$$AB = \begin{pmatrix} 1 & 2 \\ 4 & 5 \\ 3 & 0 \end{pmatrix} \begin{pmatrix} -1 \\ 3 \end{pmatrix} = \begin{pmatrix} 1(-1) + 2(3) \\ 4(-1) + 5(3) \\ 3(-1) + 0(3) \end{pmatrix} = \begin{pmatrix} 5 \\ 11 \\ -3 \end{pmatrix}$$

El producto BA no está definido.. En general $AB \neq BA$ excepto para un tipo muy especial de matrices cuadradas; es decir, el producto de matrices no es conmutativo. Si las matrices son de orden $m \times n$ y $n \times m$, entonces ambos productos están definidos, pero son de diferentes órdenes y, por lo tanto, diferentes. Si ambas matrices son cuadradas,

entonces ambos productos existen y son del mismo orden, pero no necesariamente iguales, como se muestra en el siguiente ejemplo.

Sean $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ y $B = \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix}$ se tiene que $AB = \begin{pmatrix} 7 & 2 \\ 4 & 2 \end{pmatrix}$, $BA = \begin{pmatrix} 6 & 3 \\ 4 & 3 \end{pmatrix}$

La propiedad asociativa del producto se mantiene en el caso de las matrices, es decir, $(AB)C = A(BC)$. Este es el caso también para la propiedad distributiva: $A(B+C) = AB + AC$ y $(B+C)A = BA + CA$.

La definición de las reglas para la multiplicación de matrices nos permiten introducir el concepto de *matriz idempotente.* Una matriz es idempotente si no cambia al ser multiplicada por si misma, es decir, A es idempotente si $AA = A^2 = A$.

Ejemplo 4.- La matriz

$$A = \frac{1}{6}\begin{pmatrix} 5 & 2 & -2 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix}$$ es idempotente puesto que

$$AA = \frac{1}{6}\begin{pmatrix} 5 & 2 & -2 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix}\frac{1}{6}\begin{pmatrix} 5 & 2 & -2 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix} = A^2 = \frac{1}{6}\begin{pmatrix} 5 & 2 & -2 \\ 2 & 2 & 2 \\ -1 & 2 & 5 \end{pmatrix} = A$$

6.1.4 Teoremas acerca de la transposición de una matriz

1. $(A')' = A$, es decir, la transpuesta de la transpuesta de una matriz es la matriz original

2. $(A+B)' = A' + B'$; la transpuesta de la suma de dos matrices es la suma de las transpuestas de cada una de ellas.

3. $(AB)' = B'A'$; la transpuesta del producto de dos matrices es el producto de las transpuestas de cada una en el orden inverso.

Las primeras dos propiedades se deducen directamente a partir de la definición del operador transposición. Para probar 3, note que

$$\begin{aligned} el\ (ji) - ésimo\ elemento\ en\ BA &= j - ésimo\ renglón\ de\ B'\ por\ la\ i - ésima\ columna\ de\ A' \\ &= j - ésima\ columna\ de\ B\ \ por\ el\ \ i - ésimo\ renglón\ de\ A \\ &= el\ (ij) - ésimo\ elemento\ de\ AB \end{aligned}$$

Lo anterior implica que $(ABC)' = ((AB)C)' = C'(AB)' = C'B'A'$.

Los siguientes ejemplos ilustran algunas operaciones con matrices y sirven también para introducir algunas expresiones algebraicas que se usan en el presente texto.

Ejemplo 5.- El conjunto de ecuaciones simultáneas

$$a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n = h_1$$

$$a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n = h_2$$

$$\vdots$$

$$a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = h_n$$

puede ser escrito como

$$A\underline{X} = \underline{h}$$

donde

A es una matriz de orden $n \times n$ de los coeficientes a_{ij}

$\underline{X}$ es un vector columna de las n incógnitas

$\underline{h}$ es un vector columna de los n términos independientes

El producto de la matriz A y el vector $\underline{X}$ da por resultado un vector columna con n elementos que se igualan uno a uno con los elementos del vector $\underline{h}$. El primer elemento de $A\underline{X}$ es $a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n$, que se iguala con h_1.

Ejemplo 6.- Suma de cuadrados

$$\begin{aligned} e_1^2 + e_2^2 + ... + e_n^2 &= \sum_{i=1}^{n} e_i^2 \\ &= \underline{e}'\underline{e} \end{aligned}$$

donde $\underline{e}' = (e_1, e_2, ..., e_n)$

Ejemplo 7.- Suma ponderada de cuadrados

$$a_{11}x_1^2 + a_{22}x_2^2 + ... + a_{nn}x_n^2 = \underline{X}'A\underline{X}$$

donde

A es una matriz diagonal de orden $n \times n$

$\underline{X}$ es un vector columna de n elementos

Este resultado puede verse como una aplicación directa de las reglas de multiplicación. Para ilustrarlo tomemos el caso $n = 3$:

$$\underline{X}'A\underline{X} = (x_1, x_2, x_3)\begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2$$

Ejemplo 8.- Formas cuadráticas, $\underline{X}'A\underline{X}$.

Una función muy importante en estadística se obtiene cuando consideramos la expresión general $\underline{X}'A\underline{X}$ sin la restricción sobre A de ser diagonal, sino simplemente simétrica.

Para el caso $n = 2$:

$$\underline{X}'A\underline{X} = (x_1, x_2)\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = a_{11}x_1^2 + a_{22}x_2^2 + 2x_1x_2a_{12}$$

Si $n = 3$

$$\begin{aligned} \underline{X}'A\underline{X} &= (x_1, x_2, x_3)\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \\ &= a_{11}x_1^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + a_{22}x_2^2 + 2a_{23}x_2x_3 + a_{33}x_3^2 \end{aligned}$$

La restricción de que A sea simétrica no es muy seria, puesto que si A no lo es, un producto cruzado típico sería $(a_{ij} + a_{ji})x_ix_j$. La forma cuadrática general puede ser escrita como sigue.

$$\begin{aligned} \underline{X}'A\underline{X} = {} & a_{11}x_1^2 + 2a_{12}x_1x_2 + \ldots + 2a_{1n}x_1x_n \\ & + a_{22}x_2^2 + \ldots + 2a_{2n}x_2x_n \\ & + \ldots \\ & \ddots \\ & + a_{nn}x_n^2 \end{aligned}$$

donde A es una matriz simétrica de orden $n \times n$ y $\underline{X}$ es un vector de n elementos.

Si, por ejemplo, tenemos la forma cuadrática

$x_1^2 + 3x_2^2 - 5x_3^2 + 2x_1x_2 - 8x_2x_3$

por inspección se obtiene

$$\underline{X} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \text{ y } A = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 3 & -4 \\ 0 & -4 & -5 \end{pmatrix}$$

puesto que los coeficientes de los términos x_i^2 son los elementos de la diagonal de A y los términos fuera de ella son simétricos y la mitad de los coeficientes de los términos x_{ij}.

6.1.5 Determinantes

El *determinante* de una matriz A es un escalar calculado a partir de los elementos de A de acuerdo a reglas bien definidas. Se define sólo para matrices cuadradas y se denota por $|A|$ o por $\det(A)$.

El determinante de una matriz de orden 1×1, $|\lambda|$, es el escalar mismo. El determinante de una matriz A de orden 2×2

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

se define como

$$|A| = a_{11}a_{22} - a_{21}a_{12}$$

El determinante de matrices de orden mayor se obtiene expandiéndolo como una función lineal de submatrices de orden 2. Es conveniente definir, en primer lugar, el *menor* y el *co-factor* de un elemento en una matriz.

Sea A una matriz cuadrada de orden n. Para cualquier elemento a_{rs} en A, una matriz cuadrada de orden $n-1$ se forma eliminando el renglón y la columna que contienen a a_{rs}. Denotemos por A_{rs} a esta matriz. El determinante de A_{rs}, $|A_{rs}|$, es llamado el menor del elemento a_{rs}. El producto $\theta_{rs} = (-1)^{r+s}|A_{rs}|$ es llamado el co-factor de a_{rs}.

El determinante de una matriz A de orden n es

$$|A| = \sum_{j=1}^{n} \theta_{ij} a_{ij}$$

donde cada θ_{ij} contiene un determinante de orden $n-1$.

De este modo, cada determinante de orden n se desarrolla como una función de determinantes de orden menor. Cada uno de estos determinantes se desarrolla a su

vez como función de determinantes de orden $n-2$. Esta sucesión de determinantes se continúa hasta que $|A|$ puede ser expresado en términos de determinantes de submatrices de orden 2 de A.

En el caso de $n = 3$

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Si usamos el i-ésimo renglón para el desarrollo de $|A|$

$$|A| = \sum_{j=1}^{3} \theta_{ij} a_{ij} = \theta_{i1}a_{i1} + \theta_{i2}a_{i2} + \theta_{i3}a_{i3}$$

donde

$$\theta_{11} = a_{22}a_{33} - a_{32}a_{23} = (-1)^{1+1} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}$$

$$\theta_{12} = a_{23}a_{31} - a_{21}a_{33} = (-1)^{1+2} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}$$

$$\theta_{13} = a_{21}a_{32} - a_{22}a_{31} = (-1)^{1+3} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

por lo cual

$$|A| = a_{11}(a_{22}a_{33} - a_{32}a_{23}) + a_{12}(a_{23}a_{31} - a_{21}a_{33}) + a_{13}(a_{21}a_{32} - a_{22}a_{31})$$

Ejemplo 9.- Sea

$$A = \begin{pmatrix} 2 & 4 & 6 \\ 1 & 2 & 3 \\ 5 & 7 & 9 \end{pmatrix}$$

Usando el primer renglón de A para el desarrollo de $|A|$

$$\theta_{11} = (-1)^2 \begin{vmatrix} 2 & 3 \\ 9 & 7 \end{vmatrix} = 18 - 21 = -3$$

$$\theta_{12} = (-1)^3 \begin{vmatrix} 1 & 3 \\ 5 & 9 \end{vmatrix} = -(9 - 15) = 6$$

$$\theta_{13} = (-1)^4 \begin{vmatrix} 1 & 2 \\ 5 & 7 \end{vmatrix} = 7 - 10 = -3$$

El determinante de A es:

$2(-3) + 4(6) + 6(-3) = 0$

Si el determinante de una matriz es cero se dice de ésta que es una matriz *singular*; si $|A| \neq 0$, se dice que es *no singular* o *de rango completo*. Note que en nuestro ejemplo el primer renglón es igual al doble del segundo renglón, por lo cual no es de rango completo.

El determinante de matrices de orden mayor es difícil de calcular rápidamente en forma manual, por lo cual se recomienda el uso de una computadora; la mayoría de los paquetes estadísticos ya están programados para calcular el determinante de una matriz.

Propiedades de los determinantes

1. $|A'| = |A|$, es decir, el determinante de la matriz transpuesta es igual al determinante de la matriz original. Para probar este resultado hagamos $B = A'$. Necesitamos hallar el determinante de B en términos del determinante de A; ilustraremos la demostración considerando una matriz cuadrada de orden 3. En ella el término $b_{13}b_{21}b_{32}$ en la expansión de B es el producto de los tres elementos a_{31}, a_{12} y a_{23}. Su signo en el determinante de B está dado por el número de inversiones en 3, 1, 2 (segundos subíndices de las expresiones en b), por lo cual es positivo. El signo de $a_{31}a_{12}a_{23}$ en el determinante de A puede hallarse cuando estos términos son reordenados con el primer subíndice en el orden natural. Este reordenamiento, sin embargo, dará tantas inversiones en el segundo subíndice como el número de inversiones que sean removidas del primer subíndice, por ejemplo, como el primer elemento es llevado a la tercera posición, se remueven dos inversiones del primer subíndice puesto que el entero 3 va después del 1 y el 2, pero se crean dos inversiones en el segundo subíndice puesto que ahora el entero 1 se mueve a la derecha del 2 y del 3. De este modo, $a_{31}a_{12}a_{23}$ tiene el mismo signo en la expansión del determinante de A y en la de B. Esto puede ser aplicado a cualquier término, y

así, se prueba el resultado.

2. Intercambiar dos columnas o renglones de A cambia el signo del determinante de A. En efecto: intercambiar dos columnas de A significa que el primer subíndice de todos los términos de $|A|$ no cambia, pero el signo de cada término sí, con lo cual el signo de A cambia. Sea B la matriz obtenida al intercambiar las columnas j y k de A; hemos probado$|B| = -|A|$. Entonces B' es la matriz obtenida al intercambiar los renglones j y k de A', por el resultado anterior

$$|B'| = |B| = -|A| = -|A'|.$$

3. El determinante de una matriz con dos columnas o renglones iguales es cero. En efecto, suponga que las columnas j y k de una matriz A son idénticas. Al intercambiarlas la matriz permanece inalterada; por el resultado anterior se tiene $|A| = -|A|$, es decir, $|A| = 0$

4. Si cada elemento de un renglón o columna de una matriz A es multiplicado por un escalar λ para dar una nueva matriz B, entonces $|B| = \lambda |A|$

5. Si cada elemento de una matriz A de orden n es multiplicado por un escalar λ, entonces $|\lambda A| = \lambda^n |A|$

Estas dos últimas propiedades son un resultado directo de que cada término en el desarrollo del determinante contiene uno y sólo un elemento de cada renglón (o columna) de la matriz.

6. El determinante de una matriz A permanece inalterado en su valor cuando cualquier columna o renglón es sumado al múltiplo de cualquier otra columna o renglón. Este resultado es útil en la evaluación numérica de los determinantes; ilustraremos la demostración para el caso $n = 3$.

Considere

$$\Delta = \begin{vmatrix} a_{11}+\lambda\alpha & a_{12}+\lambda\beta & a_{13}+\lambda\gamma \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

Usando el primer renglón del determinante para desarrollarlo por menores se tiene:

$$\Delta = (a_{11}+\lambda\alpha)\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - (a_{12}+\lambda\beta)\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + (a_{13}+\lambda\gamma)\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

Note que

$$\Delta = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} + \lambda \begin{vmatrix} \alpha & \beta & \gamma \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

Si $(\alpha, \beta, \gamma) = (a_{21}, a_{22}, a_{23})$ o $(\alpha, \beta, \gamma) = (a_{31}, a_{32}, a_{33})$. El último determinante en el lado derecho se vuelve cero, puesto que tiene dos renglones idénticos; así, el lado derecho se reduce al valor del determinante original. Es sencillo notar que de este resultado se deriva que el mismo resultado se cumpla si cualquier renglón o columna es sumada a una combinación lineal de otros renglones o columnas.

7. $|AB| = |A|\,|B|$

8. Expansiones en términos de co-factores distintos son idénticamente cero.

Considere

$a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + ... + a_{in}\theta_{jn},\ i \neq j$

Los elementos son aquellos del renglón i y los co-factores son los del renglón j. Esta es la expresión que debemos obtener para el determinante de una matriz en la cual los renglones i y j son idénticos, mismo que, como hemos visto, es cero. De esta manera:

$$\sum_{r=1}^{n} a_{ir}\theta_{jr} = 0\ ,\ i \neq j$$

$$\sum_{s=1}^{n} a_{sk}\theta_{sl} = 0,\ k \neq l$$

6.1.6 Inversa (o recíproca) de una matriz

En álgebra escalar tenemos

$xx^{-1} = x^{-1}x = 1$

Ello sugiere que, en el álgebra de matrices, podemos preguntarnos si, para alguna matriz A existe una matriz A^{-1} tal que $AA^{-1} = A^{-1}A = \mathbf{I}$, la matriz identidad. A^{-1}es llamada la *inversa* o *recíproca* de A. Los pasos para su construcción son:

1. A partir de una matriz A de orden $n \times n$ forme una nueva matriz, donde a_{ij} sea reemplazado por su correspondiente co-factor θ_{ij} y transponga la nueva matriz. La matriz que resulta se llama adjunta de A $(adj(A))$:

$$adj(A) = \begin{pmatrix} \theta_{11} & \theta_{21} & .... & \theta_{n1} \\ \theta_{12} & \theta_{22} & ... & \theta_{n2} \\ \vdots & & & \vdots \\ \theta_{1n} & \theta_{2n} & & \theta_{nn} \end{pmatrix}$$

Usando el último resultado citado de las propiedades de los determinantes

$$A(adj(A)) = (adj(A))A = \begin{pmatrix} |A| & 0 & ... & 0 \\ 0 & |A| & & 0 \\ \vdots & & & \vdots \\ 0 & 0 & ... & |A| \end{pmatrix} = |A|\mathbf{I}$$

2. Defina A^{-1} como:

$$A^{-1} = \frac{1}{|A|}(adj(A)) = \begin{pmatrix} \frac{\theta_{11}}{|A|} & \frac{\theta_{21}}{|A|} & .... & \frac{\theta_{n1}}{|A|} \\ \frac{\theta_{12}}{|A|} & \frac{\theta_{22}}{|A|} & ... & \frac{\theta_{n2}}{|A|} \\ \vdots & & & \vdots \\ \frac{\theta_{1n}}{|A|} & \frac{\theta_{2n}}{|A|} & & \frac{\theta_{nn}}{|A|} \end{pmatrix}$$

Este último paso es posible sólo si $|A| \neq 0$, es decir, si A es no singular.

A^{-1} es única. En efecto, suponga que existe una matriz B tal que $AB = \mathbf{I}$, entonces $A^{-1} = A^{-1}\mathbf{I} = A^{-1}AB = B$.

Por otro lado, suponga que existe una matriz C tal que $CA = \mathbf{I}$, entonces $A^{-1} = \mathbf{I}A^{-1} = A^{-1}AC = C$

Ejemplo10, Regla de Cramèr.- Si $A\underline{X} = \underline{h}$, donde A es una matriz no singular, multiplicando por el lado izquierdo ambos miembros de la ecuación por A^{-1} se obtiene:

$\underline{X} = A^{-1}\underline{h}$

En el caso $n = 3$ se tiene:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \frac{1}{|A|} \begin{pmatrix} \theta_{11} & \theta_{21} & .... & \theta_{31} \\ \theta_{12} & \theta_{22} & ... & \theta_{32} \\ \vdots & & & \vdots \\ \theta_{13} & \theta_{23} & & \theta_{33} \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix}$$

donde las θ's denotan los co-factores. De lo anterior resulta:

$$x_1 = \frac{h_1\theta_{11} + h_2\theta_{21} + h_3\theta_{31}}{|A|}$$

El numerador de esta expresión para x_1 es la expresión del determinante de una matriz sobre los elementos de su primera columna, donde esta columna es $\underline{h}$ y las columnas restantes son la segunda y tercera columna de A. Así, en un sistema de ecuaciones simultáneas, x_i se halla calculando la razón de dos determinantes, siendo el primero el de la matriz de los coeficientes, en la cual el vector $\underline{h}$ reemplaza a la i-ésima columna; el segundo determinante es el de la matriz de los coeficientes:

$$x_1 = \frac{\begin{pmatrix} h_1 & a_{12} & a_{13} \\ h_2 & a_{22} & a_{23} \\ h_3 & a_{32} & a_{33} \end{pmatrix}}{|A|}, \quad x_2 = \frac{\begin{pmatrix} a_{11} & h_1 & a_{13} \\ a_{21} & h_2 & a_{23} \\ a_{31} & h_3 & a_{33} \end{pmatrix}}{|A|}, \quad x_3 = \frac{\begin{pmatrix} a_{11} & a_{12} & h_1 \\ a_{21} & a_{22} & h_2 \\ a_{31} & a_{32} & h_3 \end{pmatrix}}{|A|}$$

Propiedades de la inversa de una matriz

1. $(AB)^{-1} = B^{-1}A^{-1}$; la inversa del producto de dos matrices es el producto de las inversas en el orden opuesto.

Demostración:

$$\begin{aligned} (AB)(B^{-1}A^{-1}) &= A(BB^{-1})A^{-1} \\ &= AIA^{-1} \\ &= AA^{-1} \\ &= \mathbf{I} \end{aligned}$$

de manera similar tenemos:

$(B^{-1}A^{-1})(AB) = \mathbf{I}$

Este resultado puede ser extendido fácilmente para obtener:

$(ABC)^{-1} = C^{-1}B^{-1}A^{-1}$

2. $(A^{-1})^{-1} = A$; la inversa de la inversa de una matriz es la matriz original

Demostración:

Por la definición de inversa:

$(A^{-1})(A^{-1})^{-1} = \mathbf{I}$

Si multiplicamos por la izquierda ambos lados de la ecuación anterior por A se obtiene:

$$\begin{aligned} A(A^{-1})(A^{-1})^{-1} &= A\mathbf{I} \\ \mathbf{I}(A^{-1})^{-1} &= A \\ (A^{-1})^{-1} &= A \end{aligned}$$

3. $(A')^{-1} = (A^{-1})'$; la inversa de la transpuesta de una matriz A es la transpuesta de la inversa.

Demostración:

$AA^{-1} = \mathbf{I}$

transponiendo

$(A^{-1})'A' = \mathbf{I}$

multiplicando por la derecha ambos lados de la ecuación anterior por $(A')^{-1}$

$(A^{-1})' \underbrace{A'(A')^{-1}}_{\mathbf{I}} = \mathbf{I}(A')^{-1}$

De esta manera:

$(A^{-1})' = (A')^{-1}$

4. $|A^{-1}| = \dfrac{1}{|A|}$; el determinante de la inversa de una matriz A es el inverso del determinante.

Demostración:

Recuérdese que $|AB| = |A|\,|B|$; sabemos que $A^{-1}A = \mathrm{I}$, por lo cual

$|\mathrm{I}| = 1 = |A^{-1}A| = |A^{-1}|\,|A|$

Por lo anterior:

$$|A^{-1}| = \frac{1}{|A|}$$

6.1.7 Matrices particionadas

Dado que una matriz es un arreglo rectangular de elementos, ésta puede ser dividida por medio de líneas verticales u horizontales en arreglos más pequeños o en submatrices. Por ejemplo, sea

$$A = \left(\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ \hline a_{31} & a_{32} & a_{33} & a_{34} \end{array}\right),$$

puede ser particionada a través de las dos líneas mostradas para dar las siguientes cuatro submatrices:

$$A_{11} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}, A_{12} = \begin{pmatrix} a_{14} \\ a_{24} \end{pmatrix}$$

$$A_{21} = \begin{pmatrix} a_{31} & a_{32} & a_{33} \end{pmatrix}, A_{22} = \begin{pmatrix} a_{34} \end{pmatrix}$$

y A puede ser reescrita como

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

Note que las líneas a través de las cuales se hace la partición deben extenderse a todo lo largo y ancho de la matriz original. Las operaciones básicas de suma y multiplicación siguen aplicándose igual en el caso de las matrices particionadas, siempre que las matrices que participan en la operación estén particionadas de manera similar. Por ejemplo, si B es también de orden 3×4 y es particionada

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

donde B_{ij} es del mismo orden que A_{ij}; de esta manera, la suma $A+B$ puede ser escrita como:

$$A+B=\begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ A_{21}+B_{21} & A_{22}+B_{22} \end{pmatrix}$$

Para considerar el caso de la multiplicación, supóngase una matriz B de orden 4×2 :

$$B=\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \\ b_{41} & b_{42} \end{pmatrix}$$

El producto AB está definido y es de orden 3×2. Para que el producto pueda ser expresado en términos de matrices particionadas, la única condición es que la partición de los renglones de B sea similar a la partición de las columnas de A. Por ejemplo, particionemos a B como sigue:

$$B=\begin{pmatrix} B_{11} \\ B_{21} \end{pmatrix}$$

donde

$$B_{11}=\begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix} \text{ y } B_{21}=\begin{pmatrix} b_{41} & b_{42} \end{pmatrix}$$

Si tratamos a las submatrices como elementos ordinarios podemos escribir el producto AB como:

$$AB=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\begin{pmatrix} B_{11} \\ B_{21} \end{pmatrix}=\begin{pmatrix} A_{11}B_{11}+A_{12}B_{21} \\ A_{21}B_{11}+A_{22}B_{21} \end{pmatrix}$$

Por inspección podemos verificar que el resultado es el mismo que si hubiéramos efectuado la multiplicación sobre las matrices originales. Así, las submatrices pueden ser tratadas como elementos ordinarios, siempre que las matrices hayan sido particionadas de manera similar.

La condición de partición similar también pudo haber sido satisfecha por:

$$B=\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

donde

$$B_{11} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}, \; B_{12} = \begin{pmatrix} b_{12} \\ b_{22} \\ b_{32} \end{pmatrix}, \; B_{21} = (b_{41}) \text{ y } B_{22} = (b_{42})$$

y el producto AB aparecer como

$$AB = \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix}$$

en términos de las nuevas submatrices.

6.1.8 Inversa de una matriz particionada

Frecuentemente podemos necesitar la inversa de una matriz particionada y, desde luego, el cálculo de la inversa puede simplificarse mediante una partición adecuada de la matriz original. Sea A una matriz no singular, la cual es particionada como sigue

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

donde A_{11} y A_{22} son ambas cuadradas y no singulares. Particionemos a la inversa de A, A^{-1} de manera similar y escribámosla como

$$A^{-1} = \begin{pmatrix} E & F \\ G & H \end{pmatrix}$$

Usando las ecuaciones

$$AA^{-1} = A^{-1}A = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{pmatrix},$$

donde la matriz identidad ha sido también particionada similarmente, es posible probar que la inversa A^{-1} puede ser expresada como

$$A^{-1} = \begin{pmatrix} E & -EA_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}E & A_{22}^{-1} + A_{22}^{-1}A_{21}EA_{12}A_{22}^{-1} \end{pmatrix}$$

donde $E = \left(A_{11} - A_{12}A_{22}^{-1}A_{21}\right)^{-1}$.

Esta es sólo una de las muchas formas en que la inversa de una matriz particionada puede ser expresada. Es útil, con fines de cálculo, que A_{22}^{-1} sea particularmente simple. Una forma alternativa puede obtenerse basándose en A_{11}^{-1}.

Una aplicación útil del resultado anterior es el cálculo de la inversa de la matriz

$$A = \begin{pmatrix} X'X & X'D \\ D'X & D'D \end{pmatrix}$$

donde X es una matriz de tamaño $n \times k$ cuyos elementos son observaciones sobre ciertas variables y D es una matriz de $n \times s$ de variables indicadoras (*dummy*) convenientemente especificadas.

Aplicando sobre A el resultado anterior se obtiene

$$\begin{pmatrix} E & -E(X'D)(D'D)^{-1} \\ -(D'D)^{-1}(D'X)E & (D'D)^{-1} + (D'D)^{-1}(D'X)E(X'D)(D'D)^{-1} \end{pmatrix}$$

donde $E = (X'MX)^{-1}$ y $M = \mathbf{I} - D(D'D)^{-1}D'$

De este resultado también se desprende que si A es una matriz no singular y de la forma

$$A = \begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}$$

entonces

$$A^{-1} = \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{pmatrix}$$

Esta clase de matrices son llamadas *diagonal por bloques* y la propiedad se mantiene para matrices que puedan ser particionadas de forma tal que muestren más de dos submatrices a lo largo de la diagonal principal.

6.1.9 Determinantes de una matriz particionada

Algunas veces necesitamos encontrar los determinantes de matrices particionadas. Para empezar, note que

$$\begin{vmatrix} A_{11} & 0 \\ 0 & \mathbf{I} \end{vmatrix} = |A_{11}|$$

En efecto, si evaluamos el determinante del miembro del lado izquierdo expandiendo en términos de los elementos del último renglón, el único término distinto de cero es el último, mismo que está unitariamente multiplicado por un determinante de la misma forma, sólo que el orden de $\mathbf{I}$ ha sido reducido en uno. De aquí se sigue que el determinante de una matriz diagonal por bloques es:

$$\begin{vmatrix} A_{11} & 0 \\ 0 & A_{22} \end{vmatrix} = \begin{vmatrix} A_{11} & 0 \\ 0 & \mathbf{I} \end{vmatrix} \begin{vmatrix} \mathbf{I} & 0 \\ 0 & A_{22} \end{vmatrix} = |A_{11}||A_{22}|$$

Considérese a continuación el determinante

$$\begin{vmatrix} A_{11} & A_{12} \\ 0 & \mathbf{I} \end{vmatrix}$$

De forma análoga al razonamiento inicial se tiene

$$\begin{vmatrix} A_{11} & A_{12} \\ 0 & \mathbf{I} \end{vmatrix} = |A_{11}|$$

Esto nos permite calcular el determinante de una matriz triangular por bloques:

$$\begin{vmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{vmatrix} = \begin{vmatrix} \mathbf{I} & 0 \\ 0 & A_{22} \end{vmatrix} \begin{vmatrix} A_{11} & A_{12} \\ 0 & \mathbf{I} \end{vmatrix} = |A_{11}||A_{22}|$$

Esta es una generalización matricial del resultado que dice que el determinante de una matriz triangular es el producto de los elementos en la diagonal. Finalmente, para hallar el determinante de una matriz particionada cualquiera, sea:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

Como antes y suponiendo que A_{22} es no singular, definamos

$$B_1 = \begin{pmatrix} \mathbf{I} & -A_{12}A_{22}^{-1} \\ 0 & \mathbf{I} \end{pmatrix}, \; B_2 = \begin{pmatrix} \mathbf{I} & 0 \\ -A_{22}^{-1}A_{21} & \mathbf{I} \end{pmatrix}$$

entonces

$$B_1AB_2 = \begin{pmatrix} A_{11}-A_{12}A_{22}^{-1}A_{21} & 0 \\ 0 & A_{22} \end{pmatrix}$$

y dado que $|B_1| = |B_2| = 1$

$$|A| = |A_{22}|\left|A_{11} - A_{12}A_{22}^{-1}A_{21}\right|$$

De forma similar, si A_{11} es no singular, se puede probar que

$$|A| = |A_{11}|\left|A_{22} - A_{21}A_{11}^{-1}A_{12}\right|$$

6.2 Dependencia lineal, rango y solución de ecuaciones homogéneas

Considere el conjunto de ecuaciones homogéneas

$$A\underline{X} = \underline{0}$$

donde A es un matriz de orden $m \times n$, cuyos elementos son constantes conocidas y $\underline{X}$ es un vector columna de n incógnitas. Si denotamos las columnas de A por $\underline{a}_1, \underline{a}_2, ..., \underline{a}_n$, la expresión anterior puede ser escrita como

$$x_1\underline{a}_1 + x_2\underline{a}_2 + ... + x_n\underline{a}_n = 0$$

donde el lado izquierdo denota una combinación lineal de las columnas de A. Si la única solución de esta ecuación es la trivial, $\underline{X} = \underline{0}$, es decir, si cada uno de los elementos de $\underline{X}$ es igual a cero, se dice que los vectores $\underline{a}_1, \underline{a}_2, ..., \underline{a}_n$ son linealmente independientes. Si existe algún vector $\underline{X} \neq \underline{0}$ (es decir, al menos uno de los elementos de $\underline{X}$ es distinto de cero) que satisfaga la ecuación, entonces se dice que las columnas de A son linealmente dependientes.

Recuérdese que el rango de una matriz A se define como el número de columnas (o renglones) linealmente independientes de A.

Sea A una matriz de orden $n \times m$ y r el rango de A. Se puede ver que r no puede exceder a m. En efecto, supongamos que $r > m$ y que las primeras r columnas de A son linealmente independientes, las primeras m columnas de A son entonces linealmente independientes y si A_1 denota la matriz cuadrada de orden m, constituida por las primeras m columnas de A, tenemos $|A_1| \neq 0$. Así, para cualquiera de los vectores linealmente independientes restantes $\underline{a}_i$ $(i = m+1, m+2, ..., r)$ la ecuación

$$A_1\underline{b} = \underline{a}_i$$

tiene solución distinta de cero:

$$\underline{b} = A_1^{-1}\underline{a}_i$$

lo cual contradice la suposición de que hay más de m linealmente independientes. Así,

$$rango(A) < \min\{m, n\}$$

Si $rango(A) = r$ y si suponemos nuevamente que las columnas de A están ordenadas de tal forma que las primeras r son linealmente independientes, la única solución para

$$x_1\underline{a}_1 + x_2\underline{a}_2 + ... + x_n\underline{a}_n = \underline{0}$$

es

$x_1 = x_2 = ... = x_n = 0$

Si $r = m$, la matriz cuadrada formada por los vectores $\underline{a}_1$, $\underline{a}_2$, ..., $\underline{a}_r$ tiene un determinante distinto de cero. Si $r < m$, podemos descartar $(m - r)$ ecuaciones linealmente dependientes para obtener

$x_1\underline{a}_1^* + x_2\underline{a}_2^* + ... + x_r\underline{a}_r^* = \underline{0}$

donde las $\underline{a}_i^*$ denotan vectores columna de r elementos. La única solución para la ecuación anterior sigue siendo

$x_1 = x_2 = ... = x_n = 0$

y existe al menos un menor de orden r de A que es distinto de cero. Dado que cualquier conjunto de $r + 1$ columnas de A es linealmente dependiente, el conjunto de todos los menores de orden $r + 1$ deben ser cero; lo mismo ocurre en los menores de orden $r + 2$, $r + 3$, etc. También, dado que las implicaciones de un menor distinto de cero para la independencia lineal de renglones y columnas son idénticas, se sigue que el rango de una matriz puede ser definido equivalentemente como el máximo número de renglones linealmente independientes, el máximo número de columnas linealmente independientes o el máximo orden de los menores distintos de cero.

Por ejemplo, el rango de

$$A = \begin{pmatrix} 4 & 8 & 2 \\ 2 & 4 & 4 \\ 2 & 4 & -2 \end{pmatrix}$$

es 2, $|A| = 0$ y aunque algunos menores de orden 2 son cero, otros son distintos de cero. Es claro, a partir de la definición del rango, que $rango(\mathbf{I}_n) = n$ y que el rango de una matriz diagonal es igual al número de elementos distintos de cero en la diagonal.

Otro teorema importante es el siguiente:

$rango(AB) \leq \min\{rango(A), rango(B)\}$

Este teorema establece que el rango del producto AB no puede exceder al más pequeño de los rangos de A y B. De éste se desprende un corolario importante:

El producto, ya sea por la derecha o por la izquierda de A por una matriz no singular, da como resultado una matriz cuyo rango es igual al rango de A. Sea A una matriz de orden

$m \times n$ y $rango\,(A) = r$. Sea B una matriz no singular de orden n tal que $rango\,(B) = n$. Sea $rango(AB) = k$, entonces, por el teorema antes enunciado se tiene

$$k \leq \min\{r, n\}$$

esto es, $k \leq r$, puesto que $r \leq n$, pero

$$A = (AB)\,B^{-1}$$

por lo cual

$$r \leq \min\{k, n\}$$

es decir $r \leq k$, por lo cual $k = r$ y se tiene $rango\,(AB) = rango\,(A)$ si B es no singular. De manera análoga se prueba el otro caso.

Considere el conjunto de ecuaciones

$$\begin{pmatrix} 1 & 2 & -3 \\ 2 & 0 & 2 \\ 4 & 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Note que $rango\,(A) = 2$ y que cualquier renglón puede ser expresado como combinación lineal de los otros dos; así, cualquiera de las tres ecuaciones puede ser descartada sin perder información. Quitemos la tercera ecuación y reescribamos las dos primeras en la forma:

$$\begin{pmatrix} 1 & 2 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = - \begin{pmatrix} -3 \\ 2 \end{pmatrix} x_3$$

Resolviendo se obtiene

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -x_3 \\ 2x_3 \end{pmatrix}$$

El vector solución para este conjunto de ecuaciones es

$$\underline{X} = (-x_3, 2x_3, x_3) = (-1, 2, 1)x_3.$$

El valor de x_3 es arbitrario y, por ello, existe un número infinito de soluciones, pero, una vez que x_3 ha sido elegido, los valores de x_1 y x_2 son un cierto múltiplo de él. En otras palabras, las proporciones de los elementos en el vector solución están determinadas de forma única y todas las soluciones posibles caen en una única línea recta por el origen en el espacio tridimensional. Geométricamente se dice que la dimensión del subespacio de las soluciones de nuestro ejemplo es uno y notamos que en él, la dimensión es igual a

la diferencia entre el número de incógnitas ($n = 3$) y el rango de A ($rango(A) = 2$).

Consideremos ahora el conjunto de ecuaciones

$$\begin{pmatrix} 1 & 2 & 6 \\ 2 & 4 & 12 \\ 4 & 8 & 24 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$rango(A) = 1$, es decir, sólo hay una ecuación linealmente independiente, a la cual podemos escribir como:

$$x_1 = -2x_2 - 6x_3$$

En este caso dos elementos del vector solución son arbitrarios y el conjunto de soluciones para esta ecuación es un plano que pasa por el origen en el espacio tridimensional. También en este ejemplo, la dimensión del subespacio de soluciones es igual a la diferencia entre el número de incógnitas y el rango de A. Estos dos últimos ejemplos ilustran y sugieren el resultado general:

Si A es una matriz de orden $m \times n$ con rango igual a r, entonces la dimensión del espacio de soluciones para $A\underline{X} = \underline{0}$ es $(n - r)$. Este resultado puede probarse como sigue:

Particionemos a A de la siguiente manera:

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} = \underline{0}$$

donde A_{11} es cuadrada y no singular de orden r y A_{12} es de orden $r \times (n - r)$. Dado que los últimos $(m - r)$ renglones de la expresión anterior son redundantes, tenemos

$$A_{11}\underline{X}_1 + A_{12}\underline{X}_2 = \underline{0}$$

de lo cual se obtiene

$$\underline{X}_1 = -A_{11}^{-1}A_{12}\underline{X}_2$$

De este modo, todas las soluciones de $A\underline{X} = \underline{0}$ son de la forma

$$\underline{X} = \begin{pmatrix} -A_{11}^{-1}A_{12}\underline{X}_2 \\ \underline{X}_2 \end{pmatrix} = \begin{pmatrix} -A_{11}^{-1}A_{12} \\ \mathrm{I}_{(n-r)} \end{pmatrix} \underline{X}_2$$

La matriz en el lado derecho de esta expresión es de orden $n \times (n - r)$, tiene rango igual a $(n - r)$ y $\underline{X}_2$ denota un vector arbitrario de $(n - r)$ elementos. De esta manera, todos los vectores en el espacio solución son combinaciones lineales de $(n - r)$ vectores

linealmente independientes, es decir, la dimensión del espacio solución de $A\underline{X} = \underline{0}$ es $n - rango(A)$.

6.3 Raíces y vectores característicos

El problema de los vectores característicos se define como encontrar los valores de un escalar λ y un vector asociado $\underline{X} \neq \underline{0}$ que satisfagan la ecuación

$$A\underline{X} = \lambda\underline{X}$$

donde A es una matriz cuadrada de tamaño n; λ es llamada una *raíz característica* de A y $\underline{X}$ un *vector característico*. Nombres alternativos son *raíces* y *vectores propios* y *eigenvalores* y *eigenvectores*. Si tomamos el caso $n = 2$ para ilustrar este problema se tiene

$$(a_{11} - \lambda)\, x_1 + a_{12}x_2 = 0$$

$$a_{21}x_1 + (a_{22} - \lambda)\, x_2 = 0$$

que puede ser escrito nuevamente en forma matricial como sigue

$$(A - \lambda\mathbf{I})\, \underline{X} = \underline{0}$$

Esta ecuación tiene solución distinta de la trivial sólo si $(A - \lambda\mathbf{I})$ es singular, es decir, sólo si $|A - \lambda\mathbf{I}| = 0$. Esta expresión forma un polinomio de grado n en λ, que puede resolverse para esta incógnita y, con ello, hallar los vectores característicos. Por ejemplo, en el caso $n = 2$ se tiene

$$(a_{11} - \lambda)\,(a_{22} - \lambda) - a_{12}a_{21} = 0$$

es decir,

$$\lambda^2 - (a_{11} + a_{22})\,\lambda + (a_{11}a_{22} - a_{12}a_{21}) = 0$$

cuyas raíces son

$$\lambda_{\frac{1}{2}} = \frac{1}{2}\left((a_{11} + a_{22}) \pm \sqrt{(a_{11} + a_{22})^2 - 4\,(a_{11}a_{22} - a_{12}a_{21})}\right)$$

En el caso de que A sea simétrica, es decir, $a_{12} = a_{21}$ las raíces son

$$\lambda_{\frac{1}{2}} = \frac{1}{2}\left((a_{11} + a_{22}) \pm \sqrt{(a_{11} - a_{22})^2 + 4a_{12}^2}\right)$$

y dado que el argumento de la raíz cuadrada es la suma de dos cuadrados, las raíces λ_1

y λ_2 son necesariamente reales en el caso de una matriz simétrica.

Ejemplo 10: Considere la matriz

$$A = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}$$

Sustituyendo en la última ecuación se obtiene $\lambda_1 = 5$ y $\lambda_2 = 0$. Para hallar el vector característico asociado a cada raíz, sustituimos λ_i en $(A - \lambda \mathbf{I})\,\underline{X} = \underline{0}$, de lo cual se obtiene, para $\lambda_1 = 5$

$$\begin{aligned} -x_1 + 2x_2 &= 0 \\ 2x_1 - 4x_2 &= 0 \end{aligned}$$

es decir

$$x_1 = 2x_2$$

Así, un elemento del vector característico es arbitrario y, por ello, si $\underline{X}$ satisface $(A - \lambda \mathbf{I})\,\underline{X} = \underline{0}$ para una determinada λ, entonces también $k\underline{X}$, $k \in \Re$. Se debe normalizar al vector haciendo su longitud unitaria, es decir,

$$x_1^2 + x_2^2 = 1$$

la cual, si $x_1 = 2x_2$ da

$$x_1 = \frac{2}{\sqrt{5}} \qquad x_2 = \frac{1}{\sqrt{5}}$$

Así, el vector característico asociado a $\lambda_1 = 5$ es

$$\underline{X}_1 = \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right)$$

De manera análoga se puede verificar que el vector asociado a $\lambda_2 = 0$ es

$$\underline{X}_2 = \left(\frac{1}{\sqrt{5}}, \frac{-2}{\sqrt{5}}\right)$$

Nótese que $\underline{X}_1'\underline{X}_2 = 0$, es decir, los vectores característicos de esta matriz simétrica son ortogonales. En general, sólo trabajaremos con matrices simétricas de elementos reales. Las dos últimas propiedades (raíces características reales y vectores característicos ortogonales) se preservan en el caso de matrices simétricas, con elementos reales, de orden n. Mas aun, si una raíz característica λ tiene multiplicidad k (es decir, se repite k veces), habrá k vectores ortogonales asociados a esta matriz.

La matriz simétrica A de orden n, tiene raíces características $\lambda_1, \lambda_2, ..., \lambda_n$ posiblemente no todas distintas; corresponde a estas raíces un conjunto de vectores ortogonales

$\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$, tales que $\underline{X}_i'\underline{X}_j = 0$, $i \neq j$, $i, j = 1, 2, ..., n$.

Podemos normalizar los vectores de tal forma que $\underline{X}_i'\underline{X}_i = 1$, para toda $i = 1, 2, ..., n$. Un conjunto de vectores ortogonales normalizados es llamado un conjunto ortonormal de vectores.

Ahora podemos escribir las condiciones para estos vectores, como

$$\underline{X}_i'\underline{X}_j = \delta_{ij}$$

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

donde δ_{ij} es llamada *la delta de Kronecker*. Si $\mathbf{X}$ denota la matriz de orden n cuyas columnas son los vectores $\underline{X}_1, \underline{X}_2, ..., \underline{X}_n$, entonces

$$\mathbf{X}'\mathbf{X} = \mathbf{I}_n$$

por lo cual

$$\mathbf{X}' = \mathbf{X}^{-1}$$

es decir, la transpuesta de $\mathbf{X}$ es igual a su inversa. De una matriz con tales características se dice que es una matriz ortogonal. También se sigue directamente que $\mathbf{X}\mathbf{X}' = \mathbf{I}_n$. Si ahora formamos el producto $\mathbf{X}'A\mathbf{X}$, el ij-ésimo elemento en la matriz (ordenada) que resulta es

$$\underline{X}_i'A\underline{X}_j = \lambda_i\underline{X}_i'\underline{X}_j$$

aplicando $A\underline{X} = \lambda\underline{X}$ y dado que $\underline{X}_j$ es el vector característico correspondiente a λ_j. De esta manera

$$\underline{X}_i'A\underline{X}_j = \lambda_i\delta_{ij}$$

usando la definición de la delta de Kronecker.

Por lo anterior

$$\mathbf{X}'A\mathbf{X} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

Esto es, ordenando los vectores característicos de A como las columnas de $\mathbf{X}$ y for-

mando el producto $\mathbf{X}'A\mathbf{X}$ se produce una matriz diagonal con las raíces características de A en la diagonal principal. La ecuación anterior es un ejemplo de la diagonalización de una matriz simétrica.

6.4 Formas cuadráticas y matrices definidas positivas

Recordemos que una forma cuadrática es una expresión de la forma $\underline{X}'A\underline{X}$, donde A es una matriz simétrica de orden n y $\underline{X}$ es un vector con n entradas.

Se dice que la forma cuadrática $\underline{X}'A\underline{X}$ y la matriz A son definidas positivas si y solamente si

$\underline{X}'A\underline{X} > 0$ para toda $\underline{X} \neq \underline{0}$

Se dice que la forma cuadrática y la matriz asociada son definidas semipositivas si $\underline{X}'A\underline{X} \geq 0$ para toda $\underline{X}$.

De aquí se sigue que si A es definida positiva debe ser no singular, puesto que si ésta fuera singular, la ecuación $A\underline{X} = \underline{0}$ tendría una solución distinta de la trivial y $\underline{X}'A\underline{X} = 0$ para $\underline{X} \neq \underline{0}$, lo cual contradice la suposición de que A es definida positiva. En el resto de este capítulo supondremos que A es una matriz simétrica, a menos que se establezca lo contrario, que es de orden n y definida positiva.

Teorema.- Si B denota una matriz de orden $n \times s$, $s < n$, con $rango(B) = s$, entonces $B'AB$ es definida positiva.

Demostración.- $B'AB$ es simétrica de orden s. Considere cualquier vector $\underline{Y} \neq \underline{0}$ de s elementos y sea $\underline{X} = B\underline{Y}$; se tiene entonces que $\underline{X} \neq 0$. Si $s < n$, podemos escribir esta ecuación como

$$\begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix} \underline{Y}$$

donde la partición es a través de los primeros s renglones y los restantes $(n - s)$. Dado que B tiene rango s, podemos suponer que B_1 es no singular y, por lo tanto, $\underline{Y} = B_1^{-1}\underline{X}_1$.

Si $\underline{X}$ fuera el vector cero, entonces $\underline{X}_1 = \underline{0}$ y se tendría $\underline{Y} = \underline{0}$, lo cual contradice la hipótesis inicial, así que $\underline{X} \neq \underline{0}$. Si $s = n$, se obtiene el mismo resultado; de este modo

$$\underline{Y}'B'AB\,\underline{Y} = \underline{X}'A\underline{X} > 0$$

es decir, $B'AB$ es definida positiva.

Se tiene un caso especial de $B'AB$ cuando $s = n$ y B es no singular. Las condiciones del teorema aun se satisfacen, así que $B'AB$ es definida positiva para una matriz no singular B.

A partir de la definición de matriz definida positiva, se sigue que la matriz identidad $\mathbf{I}$ es definida positiva. En efecto,

$$\underline{X}'\mathbf{I}\underline{X} = \underline{X}'\underline{X} > 0 \text{ para } \underline{X} \neq \underline{0}$$

Si B es no singular de rango s, $B'\mathbf{I}B = B'B$ es definida positiva y, por lo tanto, $B'B$ es no singular de rango s.

El siguiente resultado formula una importante propiedad de las matrices definidas positivas:

A es definida positiva si y sólo si sus raíces características son positivas.

Sea $\mathbf{X}$ la matriz ortogonal que diagonaliza a A, esto es

$$\mathbf{X}'A\mathbf{X} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \cdots & 0 \\ 0 & 0 & \lambda_3 & \cdots & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

donde las λ_i son las raíces características de A. Sea $\underline{Z}$ cualquier vector no nulo de n elementos. Definamos $\underline{Y} = \mathbf{X}'\underline{Z}$, entonces $\underline{Z} = \mathbf{X}\underline{Y}$, puesto que $\mathbf{X}$ es ortogonal. Así

$$\underline{Z}'A\underline{Z} = \underline{Y}'\mathbf{X}'A\mathbf{X}\underline{Y} = \sum_{i=1}^{n} \lambda_i Y_i^2$$

Para probar la condición de suficiencia, sea $\lambda_i > 0$ para toda $i = 1, 2, ..., n$, así $\underline{Z}'A\underline{Z} > 0$ y A es definida positiva.

Para probar la condición de necesidad, sea A definida positiva, esto significa

$$\underline{Z}'A\underline{Z} = \sum_{i=1}^{n} \lambda_i Y_i^2 > 0$$

Supongamos, por ejemplo que $\lambda_j \leq 0$; tomemos $\underline{Y} = (1, 0, ..., 0)$. Dado que $\underline{Z} = \mathbf{X}\underline{Y}$,

es claro que $\underline{Z} = \mathbf{X}\underline{Y}$ es no nulo, aun más, $\underline{Z}$ es la primera columna de $\mathbf{X}$, de este modo tenemos $\underline{Z}'A\underline{Z} = \lambda_1 \leq 0$, lo cual contradice la hipótesis de que A es definida positiva.

Dado que $\mathbf{X}'\mathbf{X} = \mathbf{I}$ se obtiene $|\mathbf{X}|^2 = 1$, para $|\mathbf{X}'| = |\mathbf{X}|$, y, por lo tanto $|\mathbf{X}| = \pm 1$, es decir, el determinante de una matriz ortogonal es igual a ± 1; por lo anterior,

$$|A| = \lambda_1 \cdot \lambda_2 \cdot \ldots \cdot \lambda_n$$

y, por lo tanto, $|A| > 0$.

Así pues, una matriz definida positiva es no singular, tiene raíces características positivas y determinante positivo. De esto se deduce también que todos los menores principales de una matriz definida positiva sean también positivos.

Por otro lado Λ puede ser adaptada para hallar un resultado muy importante. Dado que $\lambda_i > 0$ para toda $i = 1, 2, ..., n$, podemos definir una matriz diagonal D como

$$D = \begin{pmatrix} \frac{1}{\sqrt{\lambda_1}} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sqrt{\lambda_2}} & & \cdots & 0 \\ 0 & 0 & \frac{1}{\sqrt{\lambda_3}} & \cdots & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{1}{\sqrt{\lambda_n}} \end{pmatrix}$$

Haciendo $D\Lambda D$ se obtiene

$$(\mathbf{X}D)'A(\mathbf{X}D) = \mathbf{I}_n$$

o

$$Q'AQ = \mathbf{I}_n$$

donde $Q = \mathbf{X}D$

Como $\mathbf{X}$ y D son ambas no singulares, Q es no singular. Con una multiplicación adecuada, de $Q'AQ = \mathbf{I}_n$ se obtiene

$$A = \left(Q^{-1}\right)'Q^{-1}$$

De esta manera, si A es definida positiva, se puede hallar una matriz no singular $P = (Q^{-1})'$ tal que $A = PP'$.

Otro resultado útil que se obtiene directamente de Λ, es que la traza de A es la suma de sus raíces características; en efecto:

$\sum_{i=1}^{n} \lambda_i = tr(\mathbf{X}'A\mathbf{X}) = tr(A\mathbf{X}\mathbf{X}') = tr(A)$

Anteriormente hemos considerado la transformación $\underline{Y} = \mathbf{X}'\underline{Z}$ y probado que $\underline{Z}'A\underline{Z} = \sum_{i=1}^{n} \lambda_i Y_i^2$.

Cuando $\underline{Z}$ tiene por elementos variables aleatorias normalmente distribuidas y $\underline{Z}'A\underline{Z}$ es una forma cuadrática en variables normales, y refiriéndonos a la expresión general de las formas cuadráticas antes presentada, se verá que $\sum_{i=1}^{n} \lambda_i Y_i^2$ representa una considerable simplificación. El hecho de que la transformación $\mathbf{X}$ sea ortogonal, significa que la distribución de los elementos de $\underline{Y}$ puede hallarse fácilmente a partir de los elementos de $\underline{Z}$. Suponga que el vector $\underline{Z}$ conste de variables aleatorias independientes, distribuidas normalmente, con media cero y varianza constante, esto es

$E(Z_i) = 0, i = 1, 2, ..., n$

$E(Z_i^2) = \sigma^2, i = 1, 2, ..., n$

$E(Z_i Z_j) = 0, i \neq j$

en otras palabras,

$\underline{Z} \sim N(\underline{0}, \sigma^2\mathbf{I})$

Dado que $\underline{Y} = \mathbf{X}'\underline{Z}$, $\underline{Y}$ tiene también distribución normal multivariada con $E(\underline{Y}) = \mathbf{X}'E(\underline{Z}) = \underline{0}$. La matriz de varianzas y covarianzas para un vector $\underline{Y}$ con media cero se define como

$$E(\underline{Y}'\underline{Y}) = \begin{pmatrix} E(Y_1^2) & E(Y_1Y_2) & \dots & E(Y_1Y_n) \\ E(Y_2Y_1) & E(Y_2^2) & \dots & E(Y_2Y_n) \\ \vdots & & & \vdots \\ E(Y_nY_1) & E(Y_nY_2) & \dots & E(Y_n^2) \end{pmatrix}$$

donde los elementos en la diagonal principal son varianzas y los elementos fuera de ella covarianzas. De esta manera

$$\begin{aligned} E(\underline{Y}'\underline{Y}) &= E(\mathbf{X}'\underline{Z}\,\underline{Z}'\mathbf{X}) \\ &= \mathbf{X}'E(\underline{Z}\,\underline{Z}')\mathbf{X} \\ &= \sigma^2\mathbf{X}\mathbf{I}\mathbf{X} \\ &= \sigma^2\mathbf{X}'\mathbf{X} \\ &= \sigma^2 \end{aligned}$$

es decir,

$$\underline{Y} \sim N(\underline{0}, \sigma^2 \mathbf{I})$$

Se tiene un caso especial de $\underline{Z}A\underline{Z}$ cuando A es una matriz idempotente: Las raíces características de una matriz idempotente son cero o uno. En efecto,

$$A\underline{X} = \lambda \underline{X}$$

por lo cual

$$\begin{aligned} A^2\underline{X} &= \lambda A\underline{X} \\ &= \lambda^2 \underline{X} \end{aligned}$$

Pero

$$A^2\underline{X} = A\underline{X} = \lambda \underline{X}$$

por consiguiente

$$\lambda^2 \underline{X} = \lambda \underline{X}$$

donde $\underline{X} \neq 0$, por lo cual

$$\lambda^2 - \lambda = 0 \text{ y } \lambda = 0 \text{ ó } \lambda = 1$$

A partir de Λ se deduce también que cuando una matriz idempotente de orden n y rango $h - k$ $(0 \leq k < n)$ se diagonaliza, habrá $n - k$ unidades en la diagonal principal y k ceros para que el rango de la matriz en el lado derecho sea $n - k$, dado que la multiplicación de A por matrices no singulares no cambia el rango. Así, hemos llegado a un resultado muy importante: Si $\underline{X}'A\underline{X}$ es una forma cuadrática en n variables aleatorias independientes, normalmente distribuidas, con media cero y varianza común y constante σ^2 y si A es idempotente de rango $n - k$, entonces

$$\underline{X}'A\underline{X} = Y_1^2 + Y_2^2 + \ldots + Y_{n-k}^2$$

donde las Y's son también variables independientes, normalmente distribuidas, con media cero y varianza común y constante.

Ejemplo 11.- Considere

$$Z = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{pmatrix}$$

$$(Z'Z)^{-1} = \begin{pmatrix} \frac{14}{6} & -1 \\ -1 & \frac{3}{6} \end{pmatrix}$$

y

$$Z(Z'Z)^{-1}Z' = \begin{pmatrix} \frac{5}{6} & \frac{2}{6} & -\frac{1}{6} \\ \frac{2}{6} & \frac{2}{6} & \frac{2}{6} \\ -\frac{1}{6} & \frac{2}{6} & \frac{5}{6} \end{pmatrix}$$

Definamos una nueva matriz A como

$$A = \mathbf{I}_3 - Z(Z'Z)^{-1}Z'$$

$$A = \begin{pmatrix} \frac{1}{6} & -\frac{2}{6} & \frac{1}{6} \\ -\frac{2}{6} & \frac{4}{6} & -\frac{2}{6} \\ \frac{1}{6} & -\frac{2}{6} & \frac{1}{6} \end{pmatrix}$$

A es una matriz simétrica, idempotente y con rango 1. Deseamos hallar sus vectores y raíces característicos. Se puede verificar que la ecuación característica

$$|A - \lambda| = 0$$

nos da

$$\lambda^3 - \lambda^2 = 0$$

Esta ecuación tiene raíces $\lambda_1 = 1$ y $\lambda_2 = 0$, λ_2 con multiplicidad dos, lo cual ilustra el hecho de que las raíces características de una matriz idempotente son cero y uno, y que el número de raíces características unitarias es igual al rango de A. Para $\lambda = 1$:

$$(A - \lambda\mathbf{I}) = \begin{pmatrix} -\frac{5}{6} & -\frac{2}{6} & \frac{1}{6} \\ -\frac{2}{6} & -\frac{2}{6} & -\frac{2}{6} \\ \frac{1}{6} & -\frac{2}{6} & -\frac{5}{6} \end{pmatrix}$$

y la matriz ecuación $(A - \lambda \mathbf{I})\,\underline{X} = 0$ nos da

$$\begin{aligned} -5x_1 - 2x_2 - x_3 &= 0 \\ -2x_1 - 2x_2 - 2x_3 &= 0 \\ x_1 - 2x_2 - 5x_3 &= 0 \end{aligned}$$

Para estas tres ecuaciones fijamos

$$x_1^2 + x_2^2 + x_3^2 = 1$$

con el fin de normalizar al vector $\underline{X}$. De estas cuatro ecuaciones se obtiene

$$\underline{X} = \left(\frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right)$$

Para $\lambda = 0$, la ecuación $(A - \lambda \mathbf{I})\,\underline{X} = 0$ nos da una sola ecuación independiente

$$x_1 - 2x_2 + x_3 = 0$$

la cual, junto con la ecuación de normalización, significa que una de las x's puede ser tomada como un valor arbitrario. Por ejemplo, sea $x_3 = 0$, las condiciones arriba citadas dan

$$\underline{X}^* = \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}, 0\right)$$

que es un vector característico asociado a $\lambda = 0$. Dado que esta raíz tiene multiplicidad dos, esperamos tener otro vector característico $\underline{X}^{**}$ asociado a esta raíz y ortogonal a $\underline{X}^*$. La condición de ortogonalidad $(\underline{X}^*)'\,\underline{X}^{**} = 0$ y la necesidad de $\underline{X}^{**}$ de ser unitario y de satisfacer $x_1 - 2x_2 + x_3 = 0$ nos llevan a obtener

$$\underline{X}^{**} = \left(\frac{1}{\sqrt{30}}, \frac{-2}{\sqrt{30}}, \frac{-5}{\sqrt{30}}\right)$$

Definamos ahora la matriz $\mathbf{X}$, cuyas columnas son los tres vectores característicos de A,

$$\mathbf{X} = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{2}{\sqrt{5}} & \frac{1}{\sqrt{30}} \\ -\frac{2}{\sqrt{6}} & \frac{1}{\sqrt{5}} & -\frac{2}{\sqrt{30}} \\ \frac{1}{\sqrt{6}} & 0 & -\frac{5}{\sqrt{30}} \end{pmatrix}$$

Sea $\underline{X}_i$ la i-ésima columna de $\mathbf{X}$, notemos que $\underline{X}_i'\underline{X}_j = \begin{cases} 0 & si\ i \neq j \\ 1 & si\ i = j \end{cases}$

es decir, $\mathbf{X}$ es una matriz ortogonal y $\mathbf{X}' = \mathbf{X}^{-1}$. Si formamos el producto $\mathbf{X'AX}$ hallaremos

$$\mathbf{X}'A\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \mathbf{\Lambda}$$

Si consideramos un vector columna $\underline{Z}$ de 3 elementos y la transformación ortogonal $\underline{Y} = \mathbf{X}'\underline{Z}$, entonces

$\underline{Z}'A\underline{Z} = \underline{Y}\mathbf{X}'A\mathbf{X}\underline{Y} = Y_1^2$

La forma cuadrática general involucra términos en z_1^2, z_2^2, z_3^2 y los términos cruzados $z_i z_j$. sin embargo, en el caso en el que la matriz A de la forma cuadrática es idempotente de rango uno, hemos visto que es posible definir una transformación ortogonal $\underline{Y} = \mathbf{X}'\underline{Z}$ tal que $\underline{Z}'A\underline{Z} = Y_1^2$

6.5 Inversa generalizada e inversa condicional de una matriz

Previamente hemos definido la inversa de una matriz y discutido algunas de sus propiedades más importantes. Se dejó establecido que para que una matriz A tenga inversa, A debe ser cuadrada y su determinante diferente de cero. La teoría de los modelos lineales, la cual incluye una buena parte de estadística teórica y aplicada, involucra la solución de sistemas de ecuaciones lineales

$A\underline{X} = \underline{g}$

y funciones de las soluciones. Cuando la matriz A es invertible, la solución existe y es $\underline{X} = A^{-1}\underline{g}$. Sin embargo, en algunas situaciones A no es invertible y se necesita encontrar una solución al sistema de ecuaciones, en estas circunstancias, una teoría que permitiera hallarla sería de gran utilidad. Tal teoría involucra los conceptos de *inversa generalizada* e *inversa condicional* de una matriz.

6.5.1 Inversa generalizada de una matriz

Sea A una matriz de tamaño $m \times n$. Se define por *inversa generalizada de A ó g-inversa* de A a una matriz denotada por A^- de tamaño $n \times m$ y que satisface:

1. AA^- es simétrica
2. A^-A es simétrica
3. $AA^-A = A$
4. $A^-AA^- = A^-$

Para toda matriz A, existe A^-, es única y cumple las siguientes propiedades:

a) $(A')^- = (A^-)'$, la g-inversa de la transpuesta de una matriz es igual a la transpuesta de la g-inversa de la matriz

b) $(A^-)^- = A$, es decir, la g-inversa de la g-inversa de una matriz es la matriz original

c) $rango\,(A^-) = rango\,(A)$, el rango de la g-inversa de una matriz es igual al rango de la matriz original

d) Si A es una matriz simétrica, entonces A^- también lo es

e) Si A es una matriz no singular, entonces $A^{-1} = A^-$

f) Si A es una matriz idempotente, entonces $A^- = A$

g) Sea D una matriz diagonal de tamaño n cuyos elementos en la diagonal, d_{ii}, $i = 1, 2, ..., n$. La g-inversa de D, D^-, es una matriz diagonal cuyo i-ésimo elemento en la diagonal es igual a d_{ii}^{-1} si $d_{ii} \neq 0$ o igual a cero si $d_{ii} = 0$.

h) Sea B una matriz de tamaño $m \times r$ y rango $r > 0$; sea C una matriz de tamaño $r \times m$ y rango $m > 0$. Entonces $(BC)^- = C^-B^-$

i) $(A'A)' = A^-\,(A')^-$ para cualquier matriz A

j) $(AA^-)^- = AA^-$ y $(A^-A)^- = A^-A$ para cualquier matriz A

k) Sean P una matriz ortogonal de tamaño $m \times m$, Q una matriz ortogonal de tamaño $n \times n$ y A una matriz de orden $m \times n$. Entonces $(PAQ)^- = Q'A^-P'$

l) Sea A una matriz de orden $m \times n$, entonces $AB = AA^-$ si y sólo si B es tal que $ABA = A$ y AB es simétrica

m) Sea A una matriz de orden $m \times n$ y particionémosla como $A = [A_1, A_2]$, donde A_1 es de tamaño $m \times r$ y $r > 0$. Entonces

$$AA^- = A_1A_1^- + \left[\left(\mathbf{I} - A_1A_1^-\right)A_2\right]\left[\left(\mathbf{I} - A_1A_1^-\right)A_2\right]^-$$

6.5.2 Inversa condicional de una matriz

Sea A una matriz de orden $m \times n$. Se define como *inversa condicional* o *c-inversa* de una matriz, a la matriz denotada por A^c y que satisface $AA^cA = A$. La matriz condicional de una matriz A cumple las siguientes propiedades:

a) La inversa generalizada de A es también una inversa condicional, pero una inversa condicional de A no necesariamente es la inversa generalizada de A.

b) Para toda matriz A existe una matriz condicional, pero ésta no es necesariamente única

Sea X una matriz de tamaño $m \times n$ y rango $r > 0$, los siguientes resultados se cumplen:

c) $rango(X^c) \geq rango(X)$

d) X^cX y XX^c son matrices idempotentes

e) $rango(X^cX) = rango(XX^c) = r$

f) $X^cX = \mathbf{I}$ si y sólo si $rango(X) = n$

g) $XX^c = \mathbf{I}$ si y sólo si $rango(X) = m$

h) $tr(X^cX) = tr(XX^c) = r$

i) Si X^c es una matriz c-inversa de X, entonces $(X^c)'$ es una matriz c-inversa de X'

j) $X(X'X)^c X' = XX^-$

Sea $K = X(X'X)^c X'$, los siguientes resultados se cumplen

k) K es una matriz simétrica e idempotente

l) $rango(K) = tr(K) = rango(X) = r$

m) $KX = X$; $X'K = X'$

n) $(X'X)^c X'$ es una c-inversa de X para cualquier c-inversa de $X'X$

o) $X(X'X)^c$ es una c-inversa de X' para cualquier c-inversa de $X'X$

6.6 Cálculo diferencial en notación matricial

En el capítulo relativo al modelo múltiple nos es necesario diferenciar algunas expresiones simples que involucran vectores y matrices; por ello, estableceremos en esta sección algunos conceptos básicos.

Consideremos primero

$$\underline{a}'\underline{x} = (a_1, a_2, ..., a_n)\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = a_1x_1 + a_2x_2 + ... + a_nx_n$$

Si tomamos las derivadas parciales de $\underline{a}'\underline{x}$ con respecto a x_i tenemos

$$\frac{\partial \underline{a}'\underline{x}}{\partial x_1} = a_1$$

$$\frac{\partial \underline{a}'\underline{x}}{\partial x_2} = a_2$$

$$\frac{\partial \underline{a}'\underline{x}}{\partial x_3} = a_3 \cdot$$

es decir, las derivadas parciales son simplemente los elementos del vector $\underline{a}$. Así, si tomamos las n derivadas parciales en cuestión y las colocamos en un arreglo como el vector $\underline{a}$, podemos considerar el proceso como el de la diferenciación de un vector, definido como

$$\frac{\partial (\underline{a}'\underline{x})}{\partial \underline{x}} = \underline{a}$$

donde el lado izquierdo indica la operación de diferenciar con respecto a los elementos del vector $\underline{x}$.

Consideremos a continuación la forma cuadrática

$$\underline{X}'A\underline{X} = (x_1, x_2, ..., x_n) \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$\begin{aligned} \underline{X}'A\underline{X} = {} & a_{11}x_1^2 + 2a_{12}x_1x_2 + \dots + 2a_{1n}x_1x_n \\ & + a_{22}x_2^2 + \dots + 2a_{2n}x_2x_n \\ & + \dots \\ & \ddots \\ & + a_{nn}x_n^2 \end{aligned}$$

Tomando las derivadas parciales con respecto a los elementos de $\underline{X}$ se obtiene:

$$\frac{\partial}{\partial x_1}(\underline{X}'A\underline{X}) = 2(a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n)$$

$$\frac{\partial}{\partial x_2}(\underline{X}'A\underline{X}) = 2(a_{12}x_1 + a_{22}x_2 + \dots + a_{2n}x_n)$$

$$\vdots$$

$$\frac{\partial}{\partial x_n}(\underline{X}'A\underline{X}) = 2(a_{1n}x_1 + a_{2n}x_2 + a_{23}x_{3+} \dots + a_{nn}x_n)$$

Sin tomar en cuenta el factor 2, el lado derecho de las ecuaciones anteriores contiene los elementos del producto $A\underline{X}$, lo cual da un vector columna de n elementos. De manera alternativa, podemos ver este resultado como el producto $\underline{X}'A$, en cuyo caso se tiene un vector renglón de n elementos; así

$$\frac{\partial}{\partial \underline{X}}(\underline{X}'A\underline{X}) = 2A\underline{X} \text{ ó } \frac{\partial}{\partial \underline{X}}(\underline{X}'A\underline{X}) = 2\underline{X}'A$$

La elección entre alguno de estos dos resultados equivalentes, usualmente está determinado por el contexto en el cual la diferenciación tiene lugar. A continuación examinemos el caso en el que $\underline{Y}$ es un vector columna de n elementos, cada uno de los cuales es una función de los m elementos de $\underline{X}$, es decir

$$Y_i = f_i(X_1, X_2, ..., X_n)$$

Cada Y_i puede ser derivada parcialmente con respecto a cada X_j, dando esto por resultado mn derivadas parciales en total. Si ordenamos estas derivadas parciales en una matriz de tamaño $m \times n$, el resultado es el que sigue:

$$\frac{\partial \underline{Y}}{\partial \underline{X}} = \begin{pmatrix} \frac{\partial Y_1}{\partial X_1} & \frac{\partial Y_2}{\partial X_1} & \cdots & \frac{\partial Y_n}{\partial X_1} \\ \\ \frac{\partial Y_1}{\partial X_2} & \frac{\partial Y_2}{\partial X_2} & \cdots & \frac{\partial Y_n}{\partial X_2} \\ \vdots & \vdots & & \vdots \\ \frac{\partial Y_1}{\partial X_m} & \frac{\partial Y_2}{\partial X_m} & \cdots & \frac{\partial Y_n}{\partial X_m} \end{pmatrix}$$

Las derivadas de primer orden pueden ser usadas para localizar los valores críticos. Para distinguir entre posiciones máximas y mínimas necesitamos examinar las derivadas de segundo orden. Así, si

$$Y = f(X_1, X_2, ..., X_m)$$

haciendo

$$\frac{\partial Y}{\partial \underline{X}} = \underline{0}$$

se tiene un vector solución $\underline{X}_0$ en el cual Y es un valor crítico. Este valor crítico es un mínimo si

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2 Y}{\partial X_i \partial X_j}dX_i dX_j > 0$$

para cada conjunto de dX's, no todas cero, evaluando las derivadas en $\underline{X}_0$. El valor crítico es un máximo si el signo de la desigualdad anterior es opuesto. Para expresar la derivada de segundo orden en forma matricial, definimos

$$\frac{\partial^2 Y}{\partial \underline{X}^2} = \frac{\partial\,(\partial Y/\partial \underline{X})}{\partial \underline{X}} = \begin{pmatrix} \frac{\partial^2 Y}{\partial X_1^2} & \frac{\partial^2 Y}{\partial X_1 \partial X_2} & \cdots & \frac{\partial^2 Y}{\partial X_1 \partial X_n} \\ \frac{\partial^2 Y}{\partial X_2 \partial X_1} & \frac{\partial^2 Y}{\partial X_2^2} & \cdots & \frac{\partial^2 Y}{\partial X_2 \partial X_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 Y}{\partial X_n \partial X_1} & \frac{\partial^2 Y}{\partial X_n \partial X_2} & \cdots & \frac{\partial^2 Y}{\partial X_n^2} \end{pmatrix}$$

Capítulo 7

Algunos resultados básicos de probabilidad y estadística

7.1 Operador suma y operador producto

7.1.1 Operador suma

El operador suma, $\sum$, se define como sigue

$$\sum_{i=1}^{n} Y_i = Y_1 + Y_2 + ... + Y_n$$

Algunas propiedades importantes de este operador son

1. $\sum_{i=1}^{n} k = nk$, donde k es constante

2. $\sum_{i=1}^{n} (Y_i + Z_i) = \sum_{i=1}^{n} Y_i + \sum_{i=1}^{n} Z_i$

3. $\sum_{i=1}^{n} (a + cY_i) = na + c\sum_{i=1}^{n} Y_i$

El operador doble suma, $\sum\sum$, se define como sigue

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} &= \sum_{i=1}^{m} (Y_{i1} + Y_{i2} + ... + Y_{im}) \\ &= Y_{11} + Y_{12} + ... + Y_{1m} + Y_{21} + ... + Y_{2m} + ... + Y_{nm}\end{aligned}$$

Una propiedad importante del operador doble suma es

$$\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} Y_{ij}$$

7.1.2 Operador producto

El operador producto, $\prod$, se define como sigue

$$\prod_{i=1}^{n} Y_i = Y_1 \cdot Y_2 \cdot \ldots \cdot Y_n$$

7.2 Probabilidad

7.2.1 Teorema de la adición

Sean A_i y A_j dos eventos definidos en un espacio muestral. Entonces

$$P(A_i \cup A_j) = P(A_i) + P(A_j) - P(A_i \cap A_j)$$

donde $P(A_i \cup A_j)$ denota la probabilidad de que cualquiera, A_i ocurra ó A_j ocurra ó ambos, A_i y A_j ocurran; $P(A_i)$ y $P(A_j)$ denotan la probabilidad de A_i y de A_j, respectivamente y $P(A_i \cap A_j)$ denota la probabilidad de que ambos, A_i y A_j ocurran.

7.2.2 Teorema de la multiplicación (o Teorema de Bayes)

Sea $P(A_i \mid A_j)$ la probabilidad condicional de que A_i ocurra, dado que A_j ha ocurrido. Esta probabilidad condicional se define como sigue

$$P(A_i \mid A_j) = \frac{P(A_i \cap A_j)}{P(A_j)} \quad P(A_j) \neq 0$$

Según el teorema de la multiplicación :

$$\begin{aligned} P(A_i \cap A_j) &= P(A_i)\,P(A_j \mid A_i) \\ &= P(A_j)\,P(A_i \mid A_j) \end{aligned}$$

7.2.3 Evento complemento

El evento complemento de A_i, es denotado por $\overline{A_i}$. Los siguientes resultados de los eventos complemento son útiles:

1. $P\left(\overline{A_i}\right) = 1 - P(A_i)$

2. $P\left(\overline{A_i \cup A_j}\right) = P\left(\overline{A_i} \cap \overline{A_j}\right)$

7.3 Variables aleatorias

A lo largo de esta sección, a menos que especifiquemos lo contrario, supondremos que la variable aleatoria Y toma un número finito de valores.

7.3.1 Valor esperado

Hagamos a la variable aleatoria Y tomar los valores Y_1, Y_2, ..., Y_k, con probabilidades dadas por la función de probabilidad

$f(Y_s) = P(Y = Y_s) \quad s = 1, 2, ..., k$

El valor esperado de Y, denotado por $E(Y)$, se define como

$E(Y) = \sum_{s=1}^{k} Y_s f(Y_s)$

$E()$ es llamado el operador esperanza. Una propiedad importante del operador esperanza es

$E(a + cY) = a + cE(Y)$

donde a y c son constantes.

Algunos casos especiales de esta propiedad son

1. $E(a) = a$
2. $E(cY) = cE(Y)$
3. $E(a + Y) = a + E(Y)$

Si la variable aleatoria Y es continua, con función de densidad $f(Y)$, $E(Y)$ se define como

$E(Y) = \int_{-\infty}^{\infty} Y f(Y)\, dY$

7.3.2 Varianza

La varianza de la variable aleatoria Y denotada por $Var(Y)$ se define como sigue

$Var(Y) = E\{[Y - E(Y)]^2\}$

Una expresión equivalente es

$Var(Y) = E(Y^2) - (E(Y))^2$

$Var()$ es llamado el operador varianza.

La varianza de una función lineal de Y, $Var(a + cY)$ es

$Var(a + cY) = c^2 Var(Y)$

Algunos casos especiales de esta propiedad son

1. $Var(a + Y) = Var(Y)$
2. $Var(cY) = c^2 Var(Y)$

donde a y c son constantes.

Si Y es una variable aleatoria continua, $Var(Y)$ se define como sigue

$$Var(Y) = \int_{-\infty}^{\infty} (Y - E(Y))^2 f(Y)\, dY$$

7.3.3 Distribuciones de probabilidad conjunta, marginal y condicional

Denotemos a la función conjunta de probabilidad de dos variables aleatorias Y y X por $g(Y, Z)$:

$g(Y_s, Z_t) = P(Y = Y_s \cap Z = Z_t) \quad s = 1, 2, ..., k;\ t = 1, 2, ..., m$

La función marginal de probabilidad de Y, denotada por $f(Y)$ es

$$f(Y_s) = \sum_{t=1}^{m} g(Y_s, Z_t) \quad s = 1, 2, ..., k$$

y la función marginal de probabilidad de Z, denotada por $h(Z)$, es

$$h(Z_t) = \sum_{s=1}^{k} g(Y_s, Z_t) \quad t = 1, 2, ..., m$$

La función de probabilidad condicional de Y dado que $Z = Z_t$, es

$$f(Y_s \mid Z_t) = \frac{g(Y_s, Z_t)}{h(Z_t)} \quad h(Z_t) \neq 0;\ s = 1, 2, ..., k$$

La función de probabilidad condicional de Z dado que $Y = Y_s$, es

$$h(Z_t \mid Y_s) = \frac{g(Y_s, Z_t)}{f(Y_s)} \quad f(Y_s) \neq 0; \quad t = 1, 2, ..., m$$

7.3.4 Covarianza

La covarianza de Y y Z se denota por $Var(Y, Z)$ y se define como

$$Var(Y, Z) = E[(Y - E(Y))(Z - E(Z))]$$

Una expresión equivalente es

$$Var(Y, Z) = E(YZ) - [E(Y)E(Z)]$$

$Var\left(\ , \ \right)$, es llamado el operador covarianza.

La covarianza de $a_1 + c_1Y$ y $a_2 + c_2Z$ se denota por $Var(a_1 + c_1Y, a_2 + c_2Z)$, y se tiene

$$Var(a_1 + c_1Y, a_2 + c_2Z) = c_1c_2\,Var(Y, Z)$$

donde a_1, a_2, c_1 y c_2 son constantes.

Algunos casos especiales son

$$Var(c_1Y, c_2Z) = c_1c_2Var(Y, Z)$$

$$Var(a_1 + Y, a_2 + Z) = Var(Y, Z)$$

Por definición se tiene

$$Var(Y, Y) = Var(Y)$$

donde $Var(Y)$ es la varianza de Y.

7.3.5 Variables aleatorias independientes

Las variables aleatorias Y y Z son independientes si y sólo si

$$g(Y_s, Z_t) = f(Y_s)\,h(Z_t) \quad s = 1, 2, ..., k; \quad t = 1, 2, ..., m$$

Si Y y Z son variables aleatorias independientes

$$Var(Y, Z) = 0$$

(Cuando ocurre el caso especial de que Y y Z se distribuyen normales conjuntamente, $Var(Y, Z) = 0$ implica que Y y Z son independientes).

7.3.6 Funciones de variables aleatorias

Sean $(Y_1, Y_2, ..., Y_n)$ n variables aleatorias. Considere la función $\sum a_i Y_i$, donde a_1 son constantes. Se tiene entonces

$$E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i)$$

y

$$Var^2\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Var(Y_i, Y_j)$$

Específicamente, tenemos, para $n = 2$

$$E(a_1 Y_1 + a_2 Y_2) = a_1 E(Y_1) + a_2 E(Y_2)$$

$$Var(a_1 Y_1 + a_2 Y_2) = a_1^2 Var(Y_1) + a_2^2 Var(Y_2) + 2a_1 a_2 Var(Y_1, Y_2)$$

Si las variables aleatorias Y_i son independientes, tenemos

$$Var\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i{}^2 Var(Y_i)$$

7.3.7 Teorema del límite central

Si $Y_1, Y_2, ..., Y_n$ son observaciones aleatorias independientes de una población con función de probabilidad $f(Y)$ para la cual $Var(Y)$ es finita, la media muestral $\bar{Y}$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

se distribuye aproximadamente normal cuando el tamaño de la muestra n es razonablemente grande, con media $E(Y)$ y varianza $\frac{Var(Y)}{n}$.

7.4 Algunas distribuciones de probabilidad

7.4.1 Distribución normal

La función de densidad de una variable aleatoria normal Y es

$$f(Y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y-\mu}{\sigma}\right)^2\right] \quad -\infty < Y < +\infty$$

donde μ y σ son los dos parámetros de la distribución normal y $\exp(a)$ denota e^a.

La media y la varianza de una variable aleatoria normal son

$$E(Y) = \mu$$
$$Var(Y) = \sigma^2$$

Funciones de variables aleatorias

Una función lineal de una variable aleatoria normal tiene la siguiente propiedad

Si Y es una variable aleatoria normal, la variable transformada $Y' = a + cY$ (a y c son constantes) se distribuye normalmente, con media $a + cE(Y)$ y varianza $c^2 Var(Y)$.

Variable aleatoria normal estandarizada

La variable aleatoria normal estandarizada z

$$z = \frac{Y - \mu}{\sigma}$$

donde Y es una variable aleatoria normalmente distribuida. z se distribuye normalmente, con media 0 y varianza 1. Esta condición se denota como sigue,

$Y \sim N(0, 1)$

Las tablas contienen probabilidades acumulativas A para los cuantiles $z(A)$ donde

$P(z \leq z(A)) = A$

Dado que la distribución normal es simétrica en torno al cero, $z(-A) = 1 - z(A)$.

Funciones de variables aleatorias independientes

Sean $(Y_1, Y_2, ..., Y_n)$, n variables aleatorias independientes normalmente distribuidas. Se tiene que la combinación lineal $a_1Y_1 + a_2Y_2 + ... + a_nY_n$ se distribuye normalmente, con media $\sum_{i=1}^{n} a_i E(Y_i)$ y varianza $\sum_{i=1}^{n} a_i^2 Var(Y_i)$.

7.4.2 Distribución χ^2

Sean $(z_1, z_2, ..., z_v)$ v variables aleatorias independientes distribuidas normalmente, con media 0 y varianza 1 (normales estandarizadas). Se define

$\chi^2_{(v)} = z_1^2 + z_2^2 + ... + z_v^2$

donde las z_i son independientes. La distribución χ^2 tiene un parámetro, v, el cual es conocido como los grados de libertad. La media de la distribución χ^2 con v grados de libertad es

$$E\left(\chi^2_{(v)}\right) = v$$

7.4.3 Distribución t

Sean z y $\chi^2_{(v)}$ variables aleatorias independientes (distribuidas como normal estandarizada y χ^2, respectivamente). Se define,

$$t_{(v)} = \frac{z}{\left(\frac{\chi^2_{(v)}}{v}\right)^{\frac{1}{2}}}$$

La distribución t tiene un parámetro, los grados de libertad v. La media de una distribución t con v grados de libertad es

$$E\left(t_{(v)}\right) = 0$$

La distribución t es simétrica en torno al cero.

7.4.4 Distribución F

Sean $\chi^2_{(v_1)}$ y $\chi^2_{(v_2)}$ dos variables aleatorias independientes, distribuidas χ^2. Se define

$$F_{(v_1,v_2)} = \frac{\chi^2_{(v_1)}}{v_1} \div \frac{\chi^2_{(v_2)}}{v_2}$$

La distribución F tiene dos parámetros, los grados de libertad del numerador, v_1 y los grados de libertad del denominador, v_2.

Los percentiles por debajo de 50 pueden obtenerse usando la siguiente relación

$$F^{\alpha}_{(v_1,v_2)} = \frac{1}{F^{1-\alpha}_{(v_2,v_1)}}$$

La siguiente relación existe entre las variables aleatorias distribuidas t y las variables aleatorias distribuidas F

$$\left(t_{(v)}\right)^2 = F_{(1,v)}$$

y los percentiles de las distribuciones t y F se relacionan como sigue

$$\left(t^{.5+\alpha/2}_{(v)}\right)^2 = F^{\alpha}_{(1,v)}$$

7.5 Estimación estadística

7.5.1 Propiedades de los estimadores

1. Un estimador $\hat{\theta}$ del parámetro θ es insesgado si $E\left(\hat{\theta}\right) = \theta$
2. Un estimador $\hat{\theta}$ es un estimador consistente de θ si $\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| \geq \epsilon\right) = 0$ para cualquier $\epsilon > 0$
3. Un estimador $\hat{\theta}$ es un estimador suficiente de θ si la probabilidad condicional conjunta de las observaciones muestrales, dado $\hat{\theta}$, no depende del parámetro θ.
4. Un estimador $\hat{\theta}$ es un estimador de varianza mínima de θ si para cualquier otro estimador $\hat{\theta}^*$, $\left(Var\left(\hat{\theta}\right)\right)^2 \leq \left(Var\left(\hat{\theta}^*\right)\right)^2$, para toda $\hat{\theta}^*$.

7.5.2 Estimadores máximo verosímiles

El método de máxima verosimilitud es un método general para hallar estimadores. Suponga que se toma una muestra de una población cuya función de probabilidad $f(Y;\theta)$ involucra un parámetro, θ. Dadas las observaciones independientes $Y_1, Y_2, ..., Y_n$, la función de probabilidad conjunta de las observaciones muestrales es

$$g(Y_1, Y_2, ..., Y_n) = \prod_{i=1}^{n} f(Y_i;\theta)$$

Cuando esta función de probabilidad conjunta se ve como una función de θ, con las observaciones dadas, es llamada la función de verosimilitud, $L(\theta)$.

$$L(\theta) = \prod_{i=1}^{n} f(Y_i;\theta)$$

Maximizando $L(\theta)$ con respecto a θ se obtiene el estimador máximo verosímil de θ. Bajo condiciones generales, los estimadores máximo verosímiles son consistentes y suficientes.

7.5.3 Estimadores por mínimos cuadrados

El método de mínimos cuadrados es otro método general para hallar estimadores. Se supone que las observaciones muestrales son, para el caso de un solo parámetro θ, de la forma

$$Y_i = f_i(\theta) + \varepsilon_i \quad i = 1, 2, ..., n$$

donde $f_i(\theta)$ es una función conocida del parámetro θ y las ε_i son variables aleatorias; usualmente se supone que $E(\varepsilon_i) = 0$.

Con el método de mínimos cuadrados, para las observaciones muestrales dadas, la suma de cuadrados

$$Q = \sum_{i=1}^{n} [Y_i - f_i(\theta)]^2$$

se considera como una función de θ. El estimador por mínimos cuadrados de θ se obtiene minimizando Q con respecto a θ. En muchas circunstancias, los estimadores por mínimos cuadrados son insesgados y consistentes.

7.6 Inferencias acerca de la media de una población normal

Suponga que se tiene una muestra aleatoria de n observaciones $Y_1, Y_2, ..., Y_n$ de una población normal con media μ y desviación estándar σ^2. La media y la desviación estándar observacionales son

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

y

$$s = \left(\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} \right)^{\frac{1}{2}} = \left(\frac{\sum_{i=1}^{n} Y_i^2 - \left(\frac{\sum_{i=1}^{n} Y_i}{n} \right)^2}{n-1} \right)^{\frac{1}{2}}$$

respectivamente, y la desviación estándar de $\bar{Y}$, denotada por $S(\bar{Y})$, es

$$S(\bar{Y}) = \frac{s}{\sqrt{n}}$$

Así se tiene que $\frac{\bar{Y} - \mu}{S(\bar{Y})}$ se distribuye t con $(n-1)$ grados de libertad, es decir

$$\frac{\bar{Y} - \mu}{S(\bar{Y})} \sim t_{(n-1)}$$

7.6.1 Estimación por intervalos

El intervalo del $(1-\alpha) \times 100\%$ de confianza para μ, se obtiene a partir de la ecuación anterior y es

$$\bar{Y} \pm t_{(n-1)}^{1-\frac{\alpha}{2}} S(\bar{Y})$$

Ejemplo 1.- Obtenga un intervalo del 95% confianza para μ cuando

$n = 10 \quad \bar{Y} = 20 \quad s = 4$

Se tiene $S(\bar{Y}) = \frac{4}{\sqrt{10}} = 1.265 \quad t_{(9)}^{.975} = 2.262$

De esta forma, los límites del intervalo de confianza son $20 \pm 2.262\,(1.265)$. Así, el intervalo del 95% de confianza para μ es

$$17.1 \leq \mu \leq 22.9$$

7.6.2 Pruebas de hipótesis

Las pruebas de hipótesis concernientes a la media poblacional μ, se construyen a partir de la estadística de prueba

$$t^* = \frac{\bar{Y} - \mu_0}{S(\bar{Y})} \sim t_{(n-1)}$$

La tabla 1 muestra las reglas de decisión para cada uno de tres posibles casos, con la probabilidad α del error tipo I.

Tabla 1: Reglas de decisión para pruebas concernientes a la media μ de una población normal

Alternativas	*Regla de decisión*
	(a)
$H_0 : \mu = \mu_0$	$Si\ \|t^*\| \leq t^{1-\alpha/2}_{(n-1)}$, *concluya* H_0
$H_a : \mu \neq \mu_0$	$Si\ \|t^*\| > t^{1-\alpha/2}_{(n-1)}$, *concluya* H_a
	$donde : t^* = \dfrac{\bar{Y} - \mu_0}{S(\bar{Y})}$
	(b)
$H_0 : \mu \geq \mu_0$	$Si\ t^* \geq t^{\alpha}_{(n-1)}$, *concluya* H_0
$H_a : \mu < \mu_0$	$Si\ t^* < t^{\alpha}_{(n-1)}$, *concluya* H_a
	(c)
$H_0 : \mu \leq \mu_0$	$Si\ t^* \leq t^{1-\alpha}_{(n-1)}$, *concluya* H_0
$H_a : \mu > \mu_0$	$Si\ t^* > t^{1-\alpha}_{(n-1)}$, *concluya* H_a

Ejemplo 2.- Elija entre las alternativas $H_0 : \mu \leq 20$, $H_a : \mu > 20$; con $\alpha = 0.05$ y $n = 15$, $\bar{Y} = 24$ y $s = 6$.

Se tiene

$$S(\bar{Y}) = \frac{6}{\sqrt{15}} = 1.549$$

y

$$t^{95}_{(14)} = 1.761$$

La regla de decisión es

$Si\ t^* \leq 1.761$, *concluya* H_0
$Si\ t^* > 1.761$, *concluya* H_a

Dado que $t^* = \dfrac{(24-20)}{1.549} = 2.58 > 1.761$, se concluye H_a.

Ejemplo 3.- Elija entre las alternativas $H_0 : \mu = 10$, $H_a : \mu \neq 10$; con $\alpha = 0.02$ y $n = 25$, $\bar{Y} = 5.7$ y $s = 8$.

Se tiene

$$S\left(\bar{Y}\right) = \frac{8}{\sqrt{25}} = 1.6$$

y

$$t_{(24)}^{.99} = 2.492$$

La regla de decisión es

Si $|t^*| \leq 2.492$, *concluya* H_0

Si $|t^*| > 2.492$, *concluya* H_a

donde el símbolo $\left|\ \right|$ denota el valor absoluto. Dado que

$$|t^*| = |(5.7 - 10)/1.6| = |-2.69| = 2.69 > 2.492$$

se concluye H_a.

7.6.3 Relación entre pruebas de hipótesis e intervalos de confianza

Existe una relación directa entre las pruebas de hipótesis y los intervalos de confianza. Por ejemplo, los límites del intervalo para μ, $\bar{Y} \pm t_{(n-1)}^{1-\frac{\alpha}{2}} S\left(\bar{Y}\right)$, pueden ser usados para probar $H_0 : \mu = \mu_0$ vs. $H_a : \mu \neq \mu_0$. Si μ_0 está contenido en el intervalo del $(1 - \alpha) \times 100\%$ de confianza, la regla de decisión de la tabla 1 con nivel de significancia α llevará a concluir H_0, y viceversa.

Existen correspondencias similares entre los otros intervalos de confianza y las pruebas de hipótesis.

7.7 Comparación entre las medias de dos poblaciones normales

7.7.1 Muestras independientes

Suponga que se tienen dos poblaciones normales, con medias μ_1y μ_2, respectivamente, y con desviación estándar común, σ. Las medias μ_1y μ_2 serán comparadas a partir de

muestras aleatorias independientes para cada una de las dos poblaciones:

Muestra 1: $Y_1, Y_2, ..., Y_{n_1}$

Muestra 2: $Z_1, Z_2, ..., Z_{n_2}$

Los estimadores de las medias poblacionales, son las medias muestrales

$$\bar{Y} = \frac{\sum_i Y_i}{n_1} \quad \text{y} \quad \bar{Z} = \frac{\sum_i Z_i}{n_2}$$

y un estimador de $\mu_1 - \mu_2$ es $\bar{Y} - \bar{Z}$.

Un estimador de la varianza común, σ^2, es

$$s^2 = \frac{\sum_i \left(Y_i - \bar{Y}\right)^2 + \sum_i \left(Z_i - \bar{Z}\right)^2}{n_1 + n_2 - 2}$$

y un estimador de $Var\left(\bar{Y} - \bar{Z}\right)$ es

$$S^2\left(\bar{Y} - \bar{Z}\right) = s^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Se tiene

$$\frac{\left(\bar{Y} - \bar{Z}\right) - (\mu_1 - \mu_2)}{S\left(\bar{Y} - \bar{Z}\right)} \sim t_{(n_1+n_2-2)}$$

7.7.2 Estimación por intervalos

Los límites para el intervalo del $(1-\alpha) \times 100\%$ de confianza para $\mu_1 - \mu_2$ se obtienen a partir de la ecuación anterior y son

$$\left(\bar{Y} - \bar{Z}\right) \pm t^{1-\alpha/2}_{(n_1+n_2-2)} S\left(\bar{Y} - \bar{Z}\right)$$

Ejemplo 4.- Obtenga un intervalo del $(1-\alpha) \times 100\%$ de confianza para $\mu_1 - \mu_2$ cuando

$n_1 = 10 \quad \bar{Y} = 14 \quad \sum\left(Y_i - \bar{Y}\right)^2 = 105$

$n_2 = 20 \quad \bar{Z} = 8 \quad \sum\left(Z_i - \bar{Z}\right)^2 = 224$

Se tiene

$$s^2 = \frac{105 + 224}{10 + 20 - 2} = 11.75$$

$$S^2\left(\bar{Y} - \bar{Z}\right) = 11.75\left(\frac{1}{10} + \frac{1}{20}\right) = 1.7625$$

$$t^{.975}_{(28)} = 2.048$$

Con lo cual el intervalo buscado es

$$3.3 = (14 - 8) - 2.048\,(1.328) \leq \mu_1 - \mu_2 \leq (14 - 8) + 2.048\,(1.328) = 8.7$$

7.7.3 Pruebas de hipótesis

Las pruebas de hipótesis para $\mu_1 - \mu_2$ se construyen a partir de

$$\frac{(\bar{Y} - \bar{Z}) - (\mu_1 - \mu_2)}{S(\bar{Y} - \bar{Z})} \sim t_{(n_1+n_2-2)}$$

La tabla 2 muestra las reglas de decisión para cada uno de tres posibles casos, basado en la estadística de prueba

$$t^* = \frac{(\bar{Y} - \bar{Z})}{s(\bar{Y} - \bar{Z})}$$

con la probabilidad de error tipo I igual a α.

Tabla 2: Reglas de decisión para pruebas concernientes a μ_1 y μ_2 de dos poblaciones normales ($\sigma_1 = \sigma_2 = \sigma$). Muestras independientes

Alternativas	*Regla de decisión*
	(a)
$H_0 : \mu_1 = \mu_2$	$Si\ \|t^*\| \leq t^{1-\alpha/2}_{(n_1+n_2-2)}$, *concluya* H_0
$H_a : \mu_1 \neq \mu_2$	$Si\ \|t^*\| > t^{1-\alpha/2}_{(n_1+n_2-2)}$, *concluya* H_a
	$donde : t^* = \frac{\bar{Y} - \bar{Z}}{S(\bar{Y} - \bar{Z})}$
	(b)
$H_0 : \mu_1 \geq \mu_2$	$Si\ t^* \geq t^{\alpha}_{(n_1+n_2-2)}$, *concluya* H_0
$H_a : \mu_1 < \mu_2$	$Si\ t^* < t^{\alpha}_{(n_1+n_2-2)}$, *concluya* H_a
	(c)
$H_0 : \mu_1 \leq \mu_2$	$Si\ t^* \leq t^{1-\alpha}_{(n_1+n_2-2)}$, *concluya* H_0
$H_a : \mu_1 > \mu_2$	$Si\ t^* > t^{1-\alpha}_{(n_1+n_2-2)}$, *concluya* H_a

Ejemplo 5.- Elija entre las alternativas $H_0 : \mu_1 = \mu_2$, $H_a : \mu_1 \neq \mu_2$; donde $\alpha = 0.10$ y con los datos del ejemplo anterior. Se tiene

$t^{.95}_{(28)} = 1.701$

La regla de decisión es

$Si\ |t^*| \leq 1.701$, *concluya* H_0

$Si\ |t^*| > 1.701$, *concluya* H_a

Dado que $|t^*| = |(14 - 8)/1.328| = |4.52| = 4.52 > 1.701$, se concluye H_a.

7.7.4 Observaciones apareadas

Cuando las observaciones en dos muestras están apareadas (por ejemplo, Y_i y Z_i, resultados de la evaluación de desempeño para el i-ésimo empleado antes y después de un año de experiencia laboral), se usan las diferencias

$$W_i = Y_i - Z_i \quad i = 1, 2, ..., n$$

para trabajar con ellas como si provinieran de una muestra de una única población. Así, cuando las W_i pueden ser tratadas como observaciones de una población normal, se tiene

$$\frac{\bar{W} - (\mu_1 - \mu_2)}{S(\bar{W})} \sim t_{(n-1)}$$

donde $t_{(n-1)}$ denota una distribución t con $(n-1)$ grados de libertad. Se tiene también que

$$\bar{W} = \frac{\sum_i W_i}{n}$$

y

$$S^2(\bar{W}) = \left(\frac{\sum_i (W_i - \bar{W})^2}{n-1} \right) \div n$$

7.8 Inferencias acerca de la varianza de una población normal

Cuando se tiene una muestra aleatoria de una población normal, se cumple

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$$

donde $\chi^2_{(n-1)}$ denota una distribución χ^2 con $(n-1)$ grados de libertad.

7.8.1 Estimación por intervalos

El límite inferior L y el límite superior U del intervalo del $(1-\alpha) \times 100\%$ confianza para σ^2, la varianza poblacional, se obtienen a partir de la ecuación anterior y son

$$L = \frac{(n-1)s^2}{\chi^2_{(1-\alpha/2;n-1)}} \qquad U = \frac{(n-1)s^2}{\chi^2_{(\alpha/2;n-1)}}$$

donde $\chi^2_{(1-\alpha/2;n-1)}$ y $\chi^2_{(\alpha/2;n-1)}$ denotan los cuantiles $(1-\alpha/2)$ y $(\alpha/2)$, respectivamente, de una distribución χ^2 con $(n-1)$ grados de libertad.

Ejemplo 5.- Obtenga un intervalo del 98% de confianza para σ^2, usando los datos del ejemplo 1.

Se tiene $s^2 = 16 \quad \chi^2_{(.01;9)} = 21.67 \quad \chi^2_{(.99;9)} = 2.09$

Así, el intervalo del 98% de confianza para σ^2 está dado por

$$6.6 = \frac{9(16)}{21.67} \leq \sigma^2 \leq \frac{9(16)}{2.09} = 68.9$$

7.8.2 Pruebas de hipótesis

Las pruebas de hipótesis para σ^2 se construyen a partir de $\dfrac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$. La tabla 3 muestra las reglas de decisión para cada uno de tres casos, con la probabilidad de error tipo I igual a α.

Tabla 3: Reglas de decisión para pruebas concernientes a σ^2, la varianza de una población normal

Alternativas	*Regla de decisión*
	(a)
$H_0 : \sigma^2 = \sigma_0^2$	*Si* $\chi^2_{(\alpha/2;n-1)} \leq \dfrac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{(1-\alpha/2;n-1)}$, *concluya* H_0
$H_a : \sigma^2 \neq \sigma_0^2$	*En otro caso, concluya* H_a
	(b)
$H_0 : \sigma^2 \geq \sigma_0^2$	*Si* $\dfrac{(n-1)s^2}{\sigma_0^2} \geq \chi^2_{(\alpha;n-1)}$, *concluya* H_0
$H_a : \sigma^2 < \sigma_0^2$	*Si* $\dfrac{(n-1)s^2}{\sigma_0^2} < \chi^2_{(\alpha;n-1)}$, *concluya* H_a
	(c)
$H_0 : \sigma^2 \leq \sigma_0^2$	*Si* $\dfrac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{(1-\alpha;n-1)}$, *concluya* H_0
$H_a : \sigma^2 > \sigma_0^2$	*Si* $\dfrac{(n-1)s^2}{\sigma_0^2} > \chi^2_{(1-\alpha;n-1)}$, *concluya* H_a

7.9 Comparaciones entre las varianzas de dos poblaciones normales

Suponga que se seleccionan dos muestras independientes de dos poblaciones normales, con medias y varianzas μ_1, σ_1^2 , μ_2 y σ_2^2, respectivamente. Las varianzas muestrales son

$$s_1^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{n_1 - 1}$$

y

$$s_2^2 = \frac{\sum_i (Z_i - \bar{Z})^2}{n_2 - 1}$$

Se tiene

$$\frac{s_1^2}{\sigma_1^2} \div \frac{s_2^2}{\sigma_2^2} \sim F_{(n_1-1,n_2-2)}$$

donde $F_{(n1-1,n_2-2)}$ denota una distribución F con $(n_1 - 1)$ y $(n_2 - 1)$ grados de libertad.

7.9.1 Estimación por intervalos

Los límites inferior y superior, L y U, de un intervalo del $(1 - \alpha) \times 100\%$ de confianza para σ_1^2/σ_2^2 se obtienen a partir de la ecuación anterior y son:

$$L = \frac{s_1^2}{s_2^2}\left[\frac{1}{F^{1-\alpha/2}_{(n_1-1,n_2-1)}}\right] \quad U = \frac{s_1^2}{s_2^2}\left[\frac{1}{F^{\alpha/2}_{(n_1-1,n_2-1)}}\right]$$

Ejemplo 7.- Obtenga un intervalo del 90% de confianza para σ_1^2/σ_2^2 usando con los siguientes datos

$n_1 = 16 \qquad n_2 = 21$

$s_1^2 = 54.2 \qquad s_2^2 = 17.8$

Se tiene

$$F^{.05}_{(15,20)} = 1/F^{.95}_{(20,15)} = 1/2.33 = .429$$

$$F^{.95}_{(15,20)} = 2.33$$

con lo cual el intervalo buscado es

$$1.4 = \frac{54.2}{17.8}\left(\frac{1}{2.20}\right) \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{54.2}{17.8}\left(\frac{1}{.429}\right) = 7.1$$

7.9.2 Pruebas de hipótesis

Las pruebas de hipótesis para $\frac{\sigma_1^2}{\sigma_2^2}$ se construyen a partir de $\frac{s_1^2}{\sigma_1^2} \div \frac{s_2^2}{\sigma_2^2} \sim F_{(n_1-1,n_2-2)}$. La tabla 4 muestra las reglas de decisión para cada uno de tres posibles casos, con probabilidad de error tipo uno igual a α.

Tabla 4: Reglas de decisión para pruebas concernientes a σ^2, la varianza de una población normal

Alternativas	*Regla de decisión*
	(a)
$H_0 : \sigma_1^2 = \sigma_2^2$	*Si* $F^{\alpha/2}_{(n_1-1;n_2-1)} \leq \frac{s_1^2}{s_2^2} \leq F^{1-\alpha/2}_{(n_1-1;n_2-1)}$, *concluya* H_0
$H_a : \sigma_1^2 \neq \sigma_2^2$	*En otro caso, concluya* H_a
	(b)
$H_0 : \sigma_1^2 \geq \sigma_2^2$	*Si* $\frac{s_1^2}{s_2^2} \geq F^{\alpha}_{(n_1-1;n_2-1)}$, *concluya* H_0
$H_a : \sigma_1^2 < \sigma_2^2$	*Si* $\frac{s_1^2}{s_2^2} < F^{\alpha}_{(n_1-1;n_2-1)}$, *concluya* H_a
	(c)
$H_0 : \sigma_1^2 \leq \sigma_2^2$	*Si* $\frac{s_1^2}{s_2^2} \leq F^{1-\alpha}_{(n_1-1;n_2-1)}$, *concluya* H_0
$H_a : \sigma_1^2 > \sigma_2^2$	*Si* $\frac{s_1^2}{s_2^2} > F^{1-\alpha}_{(n_1-1;n_2-1)}$, *concluya* H_a

Ejemplo 8.- Elija entre las alternativas

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_a : \sigma_1^2 \neq \sigma_2^2$

con $\alpha = .02$ y usando los datos del ejemplo 7.

Se tiene

$F^{.01}_{(15,20)} = 1/F^{.99}_{(20,15)} = 1/3.37 = .297$

$F^{.99}_{(15,20)} = = 3.09$

La regla de decisión es: si $0.297 \leq \frac{s_1^2}{s_2^2} \leq 3.09$, concluya H_0, en otro caso, concluya H_a. Como $\frac{s_1^2}{s_2^2} = 54.2/17.8 = 3.04$, se concluye H_0.

Bibliografía

[1] Chatterjee, Samprit; *Regression Analysis by Example*; 1st Edition; John Wiley and Sons; USA, 1977

[2] Graybill, Franklin A; *An Introduction to Linear Statistics Models*;Vol I; 1st Edition; Mc Graw Hill; USA, 1961

[3] Graybill, Franklin A; *Theory and Application of the Linear model*; 1st Edition; Duxbury Press; USA, 1976

[4] Hoel, Paul G; *Introduction to Mathemathical statistics*; 3rd Edition; John Wiley and Sons; USA, 1966

[5] Hoffman, Kenneth, et. al; *Algebra Lineal* ;2a Edición; Prentice Hall Hispanoamericana; México, 1971

[6] Johnston, J; *Econometric Methods*; 2nd Edition; Mc Graw Hill; Tokio, 1972

[7] Montgomery, Douglas C, et. al.; *Introduction to Linear Regression Analysis*; 2nd Edition; John Wiley and sons; USA, 1992

[8] Neter, John, et. al; *Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Designs*; 3rd Edition; Irwin; USA, 1990

[9] Pfeiffer, Paul E; *Concepts of Probability Theory*; 2nd Edition; Dover, USA, 1978

[10] Rawlings, John, O.; *Applied Regresssion Analysis, a Research tool*; 1st Edition; Wadsworth & Brooks; USA, 1988