UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

FACULTAD DE CIENCIAS

INTRODUCCIÓN A LA ESTIMACIÓN EN REGRESIÓN NO LINEAL

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARIA

P R E S E N T A :

EDNA GABRIELA LÓPEZ ESTRADA

FACULTAD DE CIENCIAS
UNAM

TUTOR: DRA. SILVIA RUIZ VELASCO ACOSTA

2007

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Agradecimientos

Con profundo agradecimiento a mis padres: Margarita Estrada y Dante López por quererme, cuidarme, haberme permitido tomar mis propias decisiones y por enseñarme que con esfuerzo y trabajo se pueden cumplir las metas.

A mis maravillosas hermanas: Estefany Karen y Karla Lisset

Por que en el fondo oscuro del cielo arden como pupilas de fuego.

Al amigo incondicional: Miguel Angel Ruelas Negrete

A la Dra. Silvia Ruiz Velasco por su paciencia al revisar y corregir este trabajo.

A mis sinodales:

Mat. Margarita Elvira
Dra. Ruth Selene
M. en C. Jesica
M. en C. Monica

por haber dedicado parte de su tiempo en la lectura de este trabajo y por sus comentarios.

A Juan por su comprensión y cariño, a Raúl por su amistad sincera, a Saraí, Reyna, Guillermo, Gonzalo y a todas aquellas personas que contribuyeron de alguna u otra forma al mejoramiento de este trabajo.

Índice general

Introducción

El trabajo pretende ser una introducción y una motivación para el estudio profundo de los Métodos Numéricos en materia de Estimación de Modelos No Lineales por mínimos cuadrados.

El estudio de las relaciones en problemas de toda índole como económicas, en la medicina, en la ingeniería etc. frecuentemente, es no lineal y el utilizar estimación de parámetros por el Método de Mínimos Cuadrados suele arrojar problemas de resolución en virtud de que el sistema de ecuaciones normales es, por lo general, no lineal en los parámetros.

Previamente, conviene dividir a los modelos no lineales en: intrínsicamente no lineales y los no intrínsicamente lineales. Estos últimos surgen cuando, en presencia de un modelo no lineal en los parámetros, existe alguna transformación de los datos que permita transformarlo en lineal; cuando dicha transformación no existe estamos ante la presencia de un modelo intrínsicamente no lineal. Este trabajo se basa en el estudio de los modelos intrínsicamente no lineales.

La no linealidad del sistema de ecuaciones normales que arrojan tales modelos no lineales suelen presentar problemas de resolución cuando no existe una manera algebraica de resolverlos. Ante tales circunstancias se hace necesario hallar métodos numéricos que faciliten la resolución de los mismos. En este trabajo se describirán y aplicarán los Métodos de Estimación basados en la minimización de la función de suma de cuadrados de los residuos: el algoritmo de **Gauss-Newton** y el algoritmo de **Newton-Ramphson.**

El trabajo está estructurado como sigue:

En el capítulo 1 se explica la importancia del Análisis de Regresión y la construcción de modelos de regresión simple y múltiple. También da las diferencias entre la Regresión lineal y la Regresión no lineal.

En el capítulo 2 se calculan los estimadores de regresión lineal por mínimos cuadrados y por la función de máxima verosimilitud, los cuales coinciden; se discuten sus propiedades y su geometría. Se explica la complicación de estimar por mínimos cuadrados un modelo de regresión no lineal y se describen dos métodos numéricos. El primer método es el algoritmo de Gauss-Newton,

de este se describe su geometría en la función suma de cuadrados de los residuales. El segundo método es el algoritmo de Newton-Ramphson. Al final del capítulo se hace inferencia estadística al encontrar los intervalos de confianza asintóticos basados en una estimación numérica.

En el capítulo 3 se explican las condiciones que debe cumplir una variable para llevar a cabo una regresión no lineal.

Finalmente, en el capítulo 4, se muestra una aplicación médica de la regresión no lineal ajustando un modelo a la relación que hay entre el plomo en la sangre de una persona y el coeficiente intelectual (IQ). Para ajustar este modelo se utiliza el programa STATISTICA 6.0.

Capítulo 1

Regresión Lineal y Regresión No Lineal

El Análisis de Regresión es una técnica estadística que sirve para investigar y modelar la relación entre variables. Son numerosas las aplicaciones de los modelos de regresión y las hay casi en cualquier campo.

Para modelar los datos usualmente se necesita la caracterización de la relación entre una variable que mida la respuesta o variable dependiente $\mathbf{Y}$, y el factor independiente o variable predictora o explicativa $\mathbf{X}$. La ecuación de una recta que relaciona dos variables es:

$$y = \beta_0 + \beta_1 x \tag{1.1}$$

Donde β_0 es la ordenada al origen y β_1 es la pendiente. Ahora bien, aunque exista una relación lineal los datos no caen exactamente sobre una recta si todos lo estuvieran no habría ninguna diferencia entre el valor observado y el valor de predicción, se debe tomar en cuenta un error aleatorio ϵ. Sea la diferencia entre el valor observado de y y el de la recta $\beta_0 + \beta_1 x$ un error ϵ, esto es, una variable aleatoria que explica por qué el modelo no ajusta exactamente los datos.

En los casos reales, las predicciones perfectas son prácticamente imposibles ya que existen causas externas que en ocasiones se pueden medir; sin embargo, la mayoría de las veces no son medibles.

Este error puede estar formado por los efectos de otras variables. Se supone que los errores tienen $E(\epsilon) = 0, Var(\epsilon) = \sigma^2$ desconocida, además, se suele suponer que los errores no están correlacionados o que tiene alguna distribución simétrica por ejemplo Normal. Así un modelo más plausible para los datos es:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{1.2}$$

La ecuación (1.2) se llama modelo de regresión lineal simple. Como la ecuación sólo tiene una variable regresora, se llama modelo de regresión lineal simple. A los parámetros β_0 y β_1 se les suele llamar **coeficientes de regresión**. La pendiente β_1 representa el cambio en la media de la distribución de y producido por un cambio unitario en x. Si el rango de valores incluye a $x = 0$, entonces la ordenada al origen β_0, es la media de la distribución de la respuesta y cuando x =0, si no, no tiene interptretación.

En general, la variable de respuesta Y se puede relacionar con k regresores $X_1, X_2, ..., X_k$ de modo que el modelo es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

al que se le llama modelo de regresión lineal múltiple, ya que implica a más de una regresora. El adjetivo lineal es para indicar que el modelo es lineal respecto a los parámetros $\beta_0, \beta_1, \beta_2, ..., \beta_k$ y no por que y sea una función lineal de las x. De hecho, el modelo de regresión lineal se podría escribir como sigue:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + ... + \beta_k z_k + \epsilon \tag{1.3}$$

en donde z_i representa cualquier función de las regresoras originales $x_1, x_2, ..., x_k$, incluyendo transformaciones como $exp(x_i)$, $sen(x_i)$. En la figura (1.1) la variable x sí es lineal con respecto a y, aunque podemos obtener comportamientos no lineales como el de la figura (1.2) donde la variable lnx no es lineal con respecto a y pero los parámetros β_0 y β_1 si lo son.

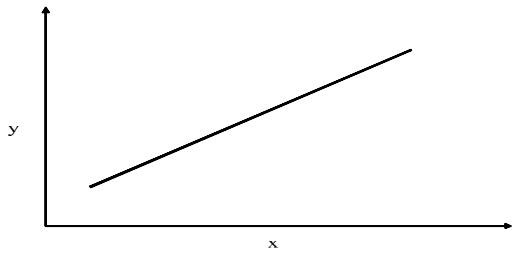


Figura 1.1: $y = \beta_0 + \beta_1 x$

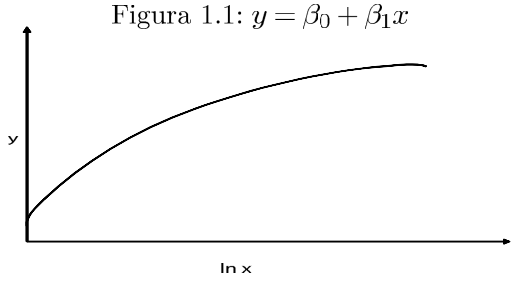


Figura 1.2: $y = \beta_0 + \beta_1 lnx$

Los parámetros β_j, $j = 0, 1 \ldots, k$, se llaman **coeficientes de regresión**. Este modelo describe a un hiperplano en el espacio de k dimensiones de las variables regresoras x_j. El parámetro β_j representa el cambio esperado en la respuesta y y por cambio unitario en x_j **cuando todas las demás variables regresoras** x_j $(i \neq j)$ **se mantienen constantes**. Por esta razón, a los parámetros β_j $j = 1 \ldots, k$ se les llama con frecuencia **coeficientes de regresión parcial.**

El modelo de regresión lineal (1.3) para n individuos se puede escribir en su forma matricial como:

$$\mathbf{y} = \mathbf{X}'\boldsymbol{\beta} + \epsilon$$

donde

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \tag{1.4}$$

cada $x_j' = (x_{1j}, x_{2j}, ..., x_{nj})$, $\mathbf{y}$ es un vector de dimensión n×1, $\mathbf{X}$ es una matriz de n×p, β es un vector de p×1, y ϵ es un vector de n×1. Como el valor esperado de los errores del modelo es cero, el valor esperado de la variable de respuesta es:

$$E(y) = E[f(x, \beta) + \epsilon] = f(x, \beta)$$

Se acostumbra llamar **función del valor esperado** a $f(x, \beta) = \mathbf{X}'\beta$ para el modelo.

Sin embargo, los modelos de regresión lineal no son adecuados a todas las situaciones, hay problemas en donde la variable de respuesta y la variable predictora se relacionan a través de una función **no lineal en los parámetros** conocida. Esto conduce a un modelo de **regresión no lineal**. La relación real entre la respuesta y las regresoras puede ser una ecuación diferencial o su solución.

Por ejemplo, el modelo

$$y = \theta_1 e^{\theta_2 x} + \epsilon$$

es no lineal en los parámetros desconocidos θ_1 y θ_2. Aquí se usará el símbolo θ para representar un parámetro en un modelo no lineal, para subrayar la diferencia entre el caso lineal y el no lineal.

En general, se escribirá el modelo de regresión no lineal en la forma

$$y = f(x, \theta) + \epsilon$$

en donde θ es un vector p x 1, de parámetros desconocidos, y ϵ es un error aleatorio con E(ϵ)=0, y Var(ϵ)=σ^2, no correlacionado, esto es, que la Cov(ϵ_i, ϵ_j) = 0 para todo $i \neq j$, también se supondrá que los errores tienen distribución normal, como en la regresión. Como

$$E(y) = E(f(x,\theta) + \epsilon) = f(x,\theta)$$

a la función $f(x,\theta)$ se le llama función de valor esperado. Esto se parece mucho al caso de la regresión lineal, excepto que ahora la función del valor esperado es una función **no lineal** de los parámetros.

En un modelo de regresión no lineal, al menos una de las derivadas de la función de valor esperado con respecto a los parámetros depende del cuando menos, uno de los parámetros. Por ejemplo, para el caso lineal

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

la función de valor esperado es $f(x,\beta)$=$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$ y sus derivadas parciales

$$\frac{\partial f(x,\beta)}{\partial \beta_j} = x_j, \qquad j = 0, 1, 2, ..., k$$

siendo $x_0 \equiv 1$. Obsérvese que en este caso lineal las derivadas no son funciones de las β.

Ahora se considera el modelo no lineal

$$y = f(x,\theta) + \epsilon = \theta_1 e^{\theta_2 x} + \epsilon$$

Las derivadas de la función de valor esperado con respecto a θ_1 y θ_2 son

$$\frac{\partial f(x,\beta)}{\partial \theta_1} = e^{\theta_2 x}$$

y

$$\frac{\partial f(x,\beta)}{\partial \theta_2} = \theta_1 x e^{\theta_2 x}$$

Como las derivadas son función de los parámetros desconocidos θ_1 y θ_2, el modelo no es lineal.

A veces es útil considerar una **transformación** que induzca la linealidad en la función de valor esperado del modelo. Por ejemplo, en el caso anterior

$$E(y) = f(x,\theta) = \theta_1 e^{\theta_2 x} \tag{1.5}$$

se puede linealizar la función del valor esperado sacando logaritmos

$$\begin{aligned} lnE(y) &= ln\theta_1 + \theta_2 x \\ E(lny) &= ln\theta_1 + \theta_2 x \end{aligned}$$

y como el logaritmo es una función lineal puede entrar a la esperanza, se reformula el modelo como sigue:

$$y* = \beta_0 + \beta_1 x + \epsilon \tag{1.6}$$

con $\beta_0 = ln\theta_1$ y $\beta_1 = \theta_2$. Y se puede utilizar la regresión **lineal simple** para estimar β_0 y β_0.

Los estimadores por mínimos cuadrados de los parámetros de la ecuación (1.6) no serán, en general, equivalentes a los estimadores no lineales de los parámetros en el modelo original de la ecuación (1.5). La razón es que en el **modelo no lineal original**, los mínimos cuadrados implican la minimización de la suma de los residuos al cuadrado respecto a y, mientras que en el **modelo transformado**, se está minimizando la suma de los de residuos al cuadrado respecto a ln y. También es necesario tomar en cuenta la naturaleza del error, en este caso podemos escribir:

$$\begin{aligned} exp(y*) &= exp(\beta_0) * exp(\beta_1 x) * exp(\epsilon) \\ y &= \theta_1 * exp(\theta_2 x) * \epsilon^* \\ E(y) &= E(\theta_1 * exp(\theta_2 x) * \epsilon^*) \\ E(y) &= \theta_1 * exp(\theta_2 x) * E(\epsilon^*) \end{aligned}$$

por lo que de la definición (1.5) $E(\epsilon^*) = 1$. Entonces el error es proporcional a la magnitud esperada en y, pero independiente de x. Hay que tomar en cuenta que al transformar estamos cambiando los supuestos sobre el error. La decisión de transformar o no, depende mucho de la naturaleza del error. Hay tres razones para transformar un modelo: la primera es trabajar con un modelo lineal; la segunda, es obtener errores que siguen una distribución aproximadamente simétrica; y la tercera, es obtener varianza constante en los errores.

A modelos que en un principio no tienen una expresión lineal, pero pueden obtenerla mediante transformaciones convenientes, se les conoce como **modelos intrínsicamente lineales.**

Capítulo 2

Estimación

2.1. Críterio de Mínimos Cuadrados para el modelo lineal

Supongáse que se tienen k pares de datos, los cuales se obtuvieron de manera experimental,

$$P_1(x_1, y_1), P_2(x_2, y_2), ..., P_n(x_n, y_n) \tag{2.1}$$

donde la variable y es una función de x, pero se desconoce la relación exacta que expresa a y en términos de x. Un examen de los puntos (2.1) en el plano xy puede guiar a la selección de una función continua suavizadora $f(x)$ tal que

$$f(x) = \text{la dependencia funcional de } y \text{ con respecto a } x.$$

Lo que se quiere es encontrar una función que aproxime a todas las observaciones, dado que los puntos $P_i(x_i, y_i)$ en (2.1) se obtuvieron de manera experimental, es poco probable que estén en la curva exacta (pero desconocida) de y con respecto a x; sin embargo, representa todo lo que sabemos de la curva. Así, cualquier indicador cuantitativo de qué tan bien $f(x)$ se ajusta a los datos dados debe estar basado en los n números:

$$\epsilon_i = y_i - f(x_i), \quad i = 1, 2, .., n \tag{2.2}$$

Geométricamente ϵ_i mide la distancia vertical desde el punto $P_i(x_i, y_i)$ a la gráfica de $f(x)$, como se muestra en la figura (2.1). Lo que se quiere es que la suma de los cuadrados de las diferencias entre las observaciones y_i y la función $f(x)$ al cuadrado. Si $f(x)$ tiene la forma funcional correcta, entonces al minimizar la suma de los errores al cuadrado se aproximará a la función real de y con respecto a x cuando n $\to \infty$. El mejor ajuste se logrará cuando minimicemos

$$SSE(f) = \sum_{i=1}^{n} [y_i - f(x_i)]^2$$

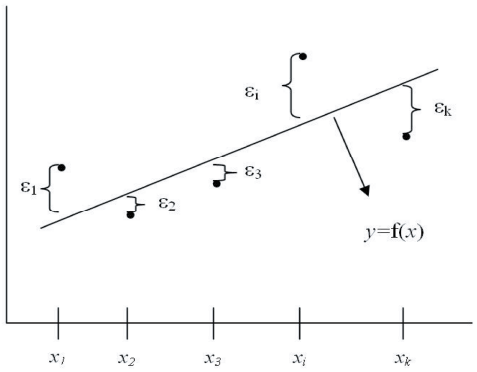


Figura 2.1: Diferencias entre las observaciones y_i y la función $f(x)$

A SSE se le conoce como la suma de cuadrados de los errores. A este procedimiento se le conoce como el *principio de mínimos cuadrados*.

Si $f(x)$ tiene k parámetros $\beta_1, \beta_2, ..., \beta_k$, entonces SSE($f$)=S($\boldsymbol{\beta}$) se puede renombrar dado que está en función de los parámetros y puede visualizarse como una función que suponemos es diferenciable, de k variables. Sabemos por cálculo que en ausencia de otras restricciones, el valor mínimo de S($\boldsymbol{\beta}$) ocurre cuando todas las derivadas parciales son simultáneamente cero, es decir, cuando

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \beta_1} = 0, \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_2} = 0, ..., \frac{\partial S(\boldsymbol{\beta})}{\partial \beta_k} = 0, \tag{2.3}$$

Las k ecuaciones con k incógnitas son las **ecuaciones normales** para $f(x)$; los valores obtenidos al minimizar los parámetros en (2.3) se llamarán **estimadores de los mínimos cuadrados** y se denotan por $\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$.

2.2. Estimación por Mínimos Cuadrados en Regresión Lineal

El modelo de regresión lineal simple establece que la verdadera media de la variable dependiente cambia en razón constante cuando el valor de la variable independiente crece o decrece. La desviación de las observaciones con respecto a su media se toma en cuenta sumándole un error aleatorio, de tal forma que la función que se utiliza es:

$$y = \beta_0 + \beta_1 x + \epsilon \tag{2.4}$$

Según la ecuación (2.4), se puede escribir

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, 3, ..., n. \tag{2.5}$$

Se considera que la ecuación (2.4) es un **modelo poblacional de regresión**, mientras que la ecuación (2.5) es un **modelo muestral de regresión**, escrito en términos de los n pares de datos (y_i, x_i). Los parámetros β_0 y β_1 son desconocidos y lineales, y deben estimarse con los datos de la muestra. Se supone que $E(\epsilon_i) = 0$, 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, $Var(\epsilon_i) = \sigma^2$, 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.

Aplicando el criterio de mínimos cuadrados (2.5) tenemos

$$\begin{aligned} \textit{minimizar} \quad S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \epsilon_i^2 \\ &= \sum_{i=1}^{n} [y_i - f(x_i)]^2 \\ &= \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i]^2 \end{aligned}$$

Los estimadores, por mínimos cuadrados deben satisfacer:

$$\begin{aligned} \frac{\partial S(\boldsymbol{\beta})}{\partial \hat{\beta}_0} &= -2 \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] = 0 \\ \frac{\partial S(\boldsymbol{\beta})}{\partial \hat{\beta}_1} &= -2 \sum_{i=1}^{n} [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i] x_i = 0 \end{aligned}$$

Simplificando obtenemos **las ecuaciones normales de mínimos cuadrados**

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \tag{2.6}$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \tag{2.7}$$

Despejando a $\hat{\beta}_0$ de la ecuación (2.6)

$$\begin{aligned} n\hat{\beta}_0 &= \sum_{i=1}^{n} y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \end{aligned}$$

Sustituimos $\hat{\beta}_0$ en la ecuación (2.7) para despejar $\hat{\beta}_1$

$$\begin{aligned} n\hat{\beta}_0\bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \\ n\bar{x}(\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \\ n\bar{y}\bar{x} - n\bar{x}^2\hat{\beta}_1 + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \end{aligned}$$

Entonces los **estimadores por mínimos cuadrados** son:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \tag{2.8}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{2.9}$$

Se puede simplificar $\hat{\beta}_1$ de la siguiente manera:

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \\ &= S_{xx} \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^{n} y_i(x_i - \bar{x}) &= \sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i \\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\ &= S_{xy} \end{aligned}$$

$$\Rightarrow \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{2.10}$$

Ejemplo 2.1. Carroll y Spiegelman (The Effects of Ignoring Small Measurement Errors in Precision Instrument Calibration.*Journal of Quality Technology*, **18**, 170-173,1986) examinan la relación entre la presión en un tanque y el volumen de líquido. El cuadro (2.1) muestra los datos.

Se quiere ajustar un modelo de regresión.

Volumen	Presión	Volumen	Presión	Volumen	Presión
2084	4599	2842	6380	3789	8599
2084	4600	3030	6818	3789	8600
2273	5044	3031	6817	3979	9048
2273	5043	3031	6818	3979	9048
2273	5044	3221	7266	4167	9484
2463	5488	3221	7268	4168	9487
2463	5487	3409	7709	4168	9487
2651	5931	3410	7710	4358	9936
2652	5932	3600	8156	4358	9938
2652	5932	3600	8156	4546	10377
2842	6380	3788	8597	4547	10379

Cuadro 2.1: Datos de la presión de un tanque y el volumen del líquido

Para escoger el tipo de regresión que se va a realizar graficamos los datos para saber cómo se correlacionan. En la figura (2.2) se observa que los datos siguen un comportamiento lineal y que la correlación es positiva, es decir, que conforme aumenta el volumen del líquido aumenta la presión del tanque. El coeficiente de correlación lineal de Pearson es de $\rho = 0{,}9835$ lo que confirma lo anterior.

Este coeficiente de correlación se calcula de la siguiente manera:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

y cuantifica la intensidad de la relación lineal entre dos variables. Su valor oscila entre -1 y $+1$; se aproxima a $+1$ cuando la relación tiende a ser lineal directa y se aproxima a -1 cuando la relación tiende a ser lineal inversa.

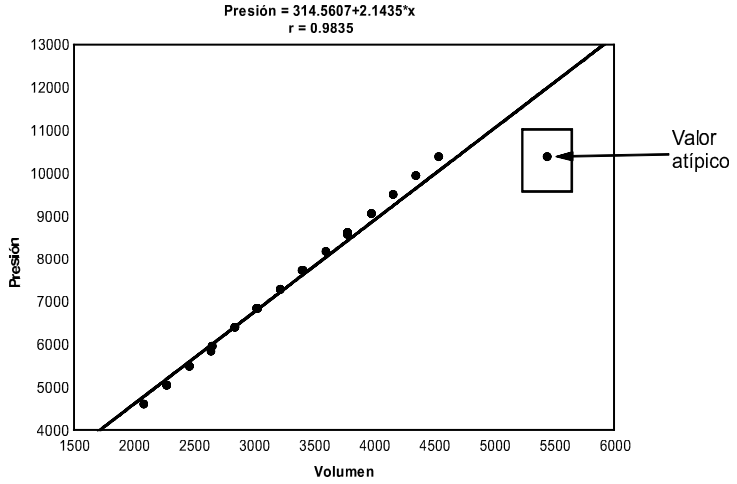


Figura 2.2: Modelo Lineal

Para estimar los parámetros del modelo se calculan primero:

$$\bar{x} = 3322{,}45,\ \bar{y} = 7436{,}30,\ \sum_{i=1}^{33} x_i y_i = 862306361,\ \sum_{i=1}^{33} x_i^2 = 386195717,$$

Según las ecuaciones 2.8 y 2.9

$$\begin{aligned} \hat{\beta}_1 &= \frac{862306361 - (33 * 7436{,}30 * 3322{,}45)}{386195717 - 33 * (3322{,}45)^2} = 2{,}1435 \\ \hat{\beta}_0 &= 7436{,}30 - (2{,}1435 * 3322{,}45) = 314{,}5607 \end{aligned}$$

El ajuste por mínimos cuadrados es:

$$\hat{y} = 314{,}5607 + 2{,}1435x$$

Se puede interpretar que por cada unidad que aumenta el volumen del líquido, la presión del tanque aumenta en promedio 2,1435 unidades. Como el límite inferior de las x no está cerca del origen, $\hat{\beta}_0$ no tiene interpretación práctica, en esté caso se tendría que buscar el estimador por mínimos cuadrados para una regresión en que la recta pasa por el origen.

2.2.1. Propiedades de los Estimadores por Mínimos Cuadrados en Regresión Lineal

Los estimadores por mínimos cuadrados $\hat{\beta}_0$ y $\hat{\beta}_1$ tienen varias propiedades estadisticas importantes. Primero, son combinaciones lineales de las $y_i's$. Definimos

$$C_i = \frac{(x_i - \bar{x})}{S_{xx}} \qquad i = 1, 2, 3, ..., n$$

$$\Rightarrow \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sum_{i=1}^{n} C_i y_i$$

Además

$$\begin{aligned}
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} &= \bar{y} - \bar{x} \sum_{i=1}^{n} C_i y_i \\
&= \frac{\sum_{i=1}^{n} y_i}{n} - \bar{x} \sum_{i=1}^{n} C_i y_i \\
&= \sum_{i=1}^{n} (\frac{y_i}{n} - \bar{x} C_i y_i) \\
&= \sum_{i=1}^{n} (\frac{1}{n} - \bar{x} C_i) y_i \\
&= \sum_{i=1}^{n} d_i y_i
\end{aligned}$$

donde

$$d_i = (\frac{1}{n} - \bar{x} C_i)$$

Tenemos que $\hat{\beta}_0$ y $\hat{\beta}_1$ son combinacines lineales de las observaciones $y_i's$. Aprovechando la notación anterior examinemos la propiedad de insesgamiento

$$\begin{aligned}
E(\hat{\beta}_1) &= E(\sum_{i=1}^{n} C_i y_i) = \sum_{i=1}^{n} C_i E(y_i) \\
&= \sum_{i=1}^{n} C_i E(\beta_0 + \beta_1 x_i + \epsilon_i) \\
&= \sum_{i=1}^{n} C_i (\beta_0 + \beta_1 x_i + E(\epsilon_i)) \\
&= \beta_0 \sum_{i=1}^{n} C_i + \beta_1 \sum_{i=1}^{n} C_i x_i \\
&= \beta_1
\end{aligned}$$

Se puede verificar que $\sum_{i=1}^{n} C_i = 0$ y $\sum_{i=1}^{n} C_i x_i = 1$

$$\begin{aligned}\sum_{i=1}^{n} C_i = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{S_{xx}} &= \frac{\sum_{i=1}^{n} x_i - n\bar{x}}{S_{xx}} \\ &= \frac{n\bar{x} - n\bar{x}}{S_{xx}} \\ &= 0\end{aligned}$$

y

$$\begin{aligned}\sum_{i=1}^{n} C_i x_i &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})x_i}{S_{xx}} \\ &= \frac{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}{S_{xx}} \\ &= \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{S_{xx}} \\ &= \frac{S_{xx}}{S_{xx}} \\ &= 1\end{aligned}$$

$$\begin{aligned}E(\hat{\beta}_0) &= E(\bar{y} - \hat{\beta}_1\bar{x}) \\ &= \frac{\sum_{i=1}^{n} E(y_i)}{n} - \bar{x}E(\hat{\beta}_1) \\ &= \frac{\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)}{n} - \bar{x}\beta_1 \\ &= \frac{n\beta_0 + \beta_1\sum_{i=1}^{n} x_i}{n} - \bar{x}\beta_1 \\ &= \beta_0 + \beta_1\bar{x_i} - \bar{x}\beta_1 \\ &= \beta_0\end{aligned}$$

Por lo tanto, $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores insesgados. Calculamos el Error Cuadrático Medio (ECM) para la consistencia.

$$ECM(\hat{\beta}_1) = Var(\hat{\beta}_1) + sesgo^2(\hat{\beta}_1)$$

Como $\hat{\beta}_1$ es insesgado y y_i es independiente de y_j $\forall i \neq j$ tenemos

$$\begin{aligned} ECM(\hat{\beta}_1) = Var(\hat{\beta}_1) = Var(\sum_{i=1}^{n} C_i y_i) &= \sum_{i=1}^{n} Var(C_i y_i) \\ &= \sum_{i=1}^{n} C_i^2 Var(y_i) \\ &= \sum_{i=1}^{n} C_i^2 Var(\beta_0 + \beta_1 x_i + \epsilon_i) \\ &= \sum_{i=1}^{n} C_i^2 Var(\epsilon_i) = \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} \\ &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{aligned}$$

Calculando el límite

$$lim_{n\to\infty} ECM(\hat{\beta}_1) = lim_{n\to\infty} \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

siempre existe, pero no se va a cero a menos que $\sum_{i=1}^{n}(x_i - \bar{x})^2 \longrightarrow \infty$ Por lo tanto $\hat{\beta}_1$ es inconsistente para β_1.

$$\begin{aligned} ECM(\hat{\beta}_0) &= Var(\hat{\beta}_0) + sesgo^2(\hat{\beta}_0) \\ &= Var(\hat{\beta}_0) = Var\left(\sum_{i=1}^{n} d_i y_i\right) = \sum_{i=1}^{n} d_i^2 Var(y_i) \\ &= \sigma^2\left[\frac{1}{n} - \frac{\bar{x}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2}\right] \\ &= \sigma^2\left[\frac{1}{n} - \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\ &= \sigma^2\left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \end{aligned}$$

Finalmente

$$lim_{n\to\infty} ECM(\hat{\beta}_0) = lim_{n\to\infty}\left[\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\sigma^2 = 0$$

Por lo tanto, $\hat{\beta}_0$ es consistente para β_0. Calculamos la covarianza para saber cómo se comportan $\hat{\beta}_0$ y $\hat{\beta}_1$ conjuntamente. Utilizando el supuesto de linealidad con y_i y sabiendo que y_i es independiente de y_j $\forall i \neq j$.

$$\begin{aligned}
Cov(\hat{\beta}_0,\hat{\beta}_0) &= E(\hat{\beta}_0 - E(\hat{\beta}_0))E(\hat{\beta}_1 - E(\hat{\beta}_1)) \\
&= E\left(\sum_{i=1}^{n} d_i y_i - \sum_{i=1}^{n} d_i E(y_i)\right)E\left(\sum_{i=1}^{n} C_i y_i - \sum_{i=1}^{n} C_i E(y_i)\right) \\
&= E\left(\sum_{i=1}^{n} d_i (y_i - E(y_i))\right)E\left(\sum_{i=1}^{n} C_i (y_i - E(y_i))\right) \\
&= \sum_{i=1}^{n} d_i C_i Var(y_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} d_i C_i Cov(y_i, y_j) \\
&= \sum_{i=1}^{n} d_i C_i Var(y_i) \\
&= \sum_{i=1}^{n}\left(\frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\left(\frac{1}{n} - \frac{\bar{x}(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)\sigma^2 \\
&= \sigma^2\left(-\frac{\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right) \neq 0
\end{aligned}$$

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 $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ $i = 1, 2, \ldots, n$; $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$(cte), ϵ_i independiente a ϵ_j $\forall$ $i \neq j$, los estimadores por mínimos cuadrados de β_0 y β_1 son, dentro de los estimadores lineales, los de menor varianza. Bajo la hipótesis de que $\epsilon_i \sim N(0, \sigma^2)$ se tiene que $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ también tiene distribución normal

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Con esto se puede verificar 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.

La función de verosimilitud para $\vec{Y} = y_i, y_2, \ldots, y_n$ es la siguiente:

$$\begin{aligned}\mathcal{L}(\beta_0, \beta_1, \sigma^2; \vec{Y}) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\left\{-\sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2}\right\}\end{aligned}$$

aplicando la función logaritmo

$$\ln \mathcal{L}(\beta_0, \beta_1, \sigma^2; \vec{Y}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Para maximizar la función obtenemos las derivadas parciales de $\ln \mathcal{L}$

$$\begin{aligned}\frac{\partial \ln \mathcal{L}}{\partial \beta_0} &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \Rightarrow \sum_{i=1}^{n} y_i &= n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \qquad (2.11) \\ \frac{\partial \ln \mathcal{L}}{\partial \beta_1} &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \\ \Rightarrow \sum_{i=1}^{n} y_i x_i &= \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \qquad (2.12) \\ \frac{\partial \ln \mathcal{L}}{\partial \sigma^2} &= \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^3} - \frac{n}{\sigma} = 0 \qquad (2.13)\end{aligned}$$

Observe que las ecuaciones (2.11) y (2.12) coinciden con las ecuaciones normales. La varianza del error, σ^2, es un parámetro adicional desconocido, cuyo estimador máximo verosímil es:

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

Este estimador no es insesgado, como veremos a continuación.

Tomando en cuenta que:

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

Los grados de libertad corresponden a n, el número de observaciones o tamaño de la muestra, se pierde un grado de libertad por cada parámetro estimado, en este caso es $\bar{x}$, por eso es $n-1$. En la ecuación (2.13) $\hat{\beta}_0$ y $\hat{\beta}_1$ son los parámetros a estimar, por lo tanto, se pierden 2 grados de libertad.

$$\Rightarrow \frac{\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

$$\Rightarrow E\left(\frac{\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2}{\sigma^2}\right) = (n-2)$$

Por lo que

$$E(\hat{\sigma}^2) = E\left(\frac{\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2}{n}\right) = \frac{(n-2)\sigma^2}{n} \neq \sigma^2$$

Por lo tanto $\hat{\sigma}^2$ es un estimador sesgado. Sin embargo, es posible construir un estimador insesgado a partir de la última expresión obtenida de la siguente manera:

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

2.3. Forma Matricial del Modelo de Regresión Lineal

La siguiente notación nos permitirá continuar con la construcción de modelos más generales.

Tenemos el siguiente modelo: $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$; $i = 1, 2, \ldots, n$ con $\epsilon \sim N(0, \sigma^2)$ y $\epsilon_i \perp \epsilon_j$. Las ecuaciones para las n observaciones son:

$$\begin{aligned}
y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \epsilon_1 \\
y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \epsilon_2 \\
&\vdots \\
y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \epsilon_n
\end{aligned}$$

en forma matricial quedan:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_k x_{1k} + \epsilon_1 \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_k x_{2k} + \epsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_k x_{nk} + \epsilon_n \end{pmatrix}$$

Al descomponer el miembro derecho

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ & & \vdots & & \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

se tiene la siguiente ecuación de la recta ajustada

$$\mathbf{y} = \mathbf{X}\beta + \epsilon \tag{2.14}$$

donde $\mathbf{y}$ es un vector de dimensión n×1, $\mathbf{X}$ es una matriz de n×p, β es un vector de p×1, y ϵ es un vector de n×1. Notamos que al obtener las derivadas parciales para cada uno de los parámetros desconocidos de la regresión se tendrán p−k×1 ecuaciones normales. La suma de cuadrados de los errores puede ser expresada como

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \epsilon_i^2 = \epsilon'\epsilon$$

y desarrollandola queda

$$\begin{aligned} S(\boldsymbol{\beta}) &= \mathbf{y}'\mathbf{y} - \beta'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta \\ &= \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta \end{aligned}$$

donde $\beta'\mathbf{X}'\mathbf{y}$ es un matriz 1×1, o un escalar, y su transpuesta $(\beta'\mathbf{X}'\mathbf{y})' = \mathbf{y}'\mathbf{X}'\beta$ es el mismo escalar. El estimador mínimos cuadrados debe satisfacer

$$\left.\frac{\partial S(\boldsymbol{\beta})}{\partial \beta}\right|_{\hat{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\beta = 0$$

simplificando se tiene

$$\mathbf{X}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{y} \tag{2.15}$$

El sistema (2.15) son las ecuaciones normales de mínimos cuadrados. Asumiendo $\mathbf{X}$ de rango completo y multiplicando por la inversa de $\mathbf{X}'\mathbf{X}$ obtenemos el estimador mínimos cuadrados

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.16}$$

Se prueba en el anexo [A.3] que $E(\hat{\beta}) = \beta$ y $Var(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.

2.4. Mínimos cuadrados en Regresión no Lineal

La notación para el modelo de regresión no lineal es diferente al modelo de regresión lineal.

$$\mathbf{y_i} = f(x_i, \boldsymbol{\theta}) + \epsilon_i \qquad i = 1, 2, 3, ..., n \tag{2.17}$$

donde $\boldsymbol{\theta}$ puede ser un vector. La ecuación anterior se descompone de la siguiente manera

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n\times 1} = \begin{pmatrix} f(x_1, \theta) \\ f(x_2, \theta) \\ \vdots \\ f(x_n, \theta) \end{pmatrix}_{n\times 1} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{n\times 1}$$

Con los supuestos de que $\epsilon \sim N(\mathbf{0}, \sigma^2\boldsymbol{I})$ y $\epsilon_i \perp \epsilon_j$, por lo que se conservan los mismos supuestos que en regresión lineal. El estimador por mínimos cuadrados de $\boldsymbol{\theta}$, denotado por $\hat{\boldsymbol{\theta}}$ es el que minimiza la suma de cuadrados de los errores

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} [y_i - f(x_i, \boldsymbol{\theta})]^2 \qquad i = 1, 2, 3, ..., n \tag{2.18}$$

que en forma vectorial es:

$$S(\boldsymbol{\theta}) = [y - f(\mathbf{x}, \boldsymbol{\theta})]'[y - f(\mathbf{x}, \boldsymbol{\theta})] = \| \, y - f(\mathbf{x_i}) \, \|^2$$

Al hacer los supuestos de normalidad sobre ϵ_i el estimador máximo verosimil de $\hat{\boldsymbol{\theta}}$ es también el estimador por mínimos cuadrados. Esto es por que la función de verosimilitud es

$$\mathcal{L}(\boldsymbol{\theta}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} exp\left\{-\frac{S(\boldsymbol{\theta})}{2\sigma^2}\right\}$$

Así que si σ^2 es conocida, maximizar $L(\boldsymbol{\theta}, \sigma^2)$ con respecto a $\boldsymbol{\theta}$ es equivalente a maximizar $exp(-S(\boldsymbol{\theta})/2\sigma^2)$, y por las caracteristicas de la función exponencial y dado que $S(\boldsymbol{\theta}) \geq 0$ lo anterior es equivalente a minimizar la función $S(\boldsymbol{\theta})$ con respecto a $\boldsymbol{\theta}$; es decir, los estimadores conciden.

Para encontrar el estimador por mínimos cuadrados $\hat{\boldsymbol{\theta}}$ se necesita diferenciar la expresión (2.18) con respecto a $\boldsymbol{\theta}$. La ecuación normal toma la forma:

$$\sum_{i=1}^{n} \{y_i - f(x_i, \boldsymbol{\theta})\} \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_r}\bigg|_{\theta=\hat{\theta}} = 0 \qquad r = 1, 2, 3, ..., p \tag{2.19}$$

Ejemplo: Consideramos el siguiente modelo

$$\begin{aligned} y_i &= \theta_1 x_i^{\theta_2} + \epsilon_i \qquad i = 1, 2, 3, ..., n \\ \frac{\partial f}{\partial \theta_1} &= x_i^{\theta_2} \\ \frac{\partial f}{\partial \theta_2} &= \theta_1 x_i^{\theta_2} log x_i \end{aligned}$$

Sustituyendo este resultado en (2.19), obtenemos las ecuaciones normales:

$$\sum_{i=1}^{n}[y_i - \theta_1 x_i^{\theta_2}][x_i^{\theta_2}] = 0 \tag{2.20}$$

$$\sum_{i=1}^{n}[y_i - \theta_1 x_i^{\theta_2}][\theta_1 x_i^{\theta_2} log x_i] = 0 \tag{2.21}$$

Hasta aquí, la teoría desarrollada es idéntica que en el caso lineal; sin embargo, al querer resolver las ecuaciónes normales el grado de dificultad que se tendría para tratar de despejar a θ_1 y θ_2 en la ecuación (2.21) y (2.21) es alto y esta dificultad se acentúa cuando el modelo involucra más parámetros; es decir, no hay una forma cerrada. A esto se debe añadir el hecho de que en el caso no lineal, la función puede tener varios mínimos locales o varios mínimos globales.

Ha sido necesario entonces, que se busquen alternativas para la estimación no lineal de los parámetros. Por lo que se han desarrollado métodos iterativos como alternativas para la estimación puntual. Nos referimos a un método iterativo como aquel que se repite tantas veces sea necesario hasta que, con base en un criterio definido, se considere que los resultados finalmente encontrados convergen a una solución o divergen.

Es necesario mencionar que los métodos que se van a enunciar tienen la característica de utilizar un punto inicial $\boldsymbol{\theta}^0$, sobre cuya elección e importancia se tratará al final de exponer el primer método.

2.4.1. Geometría de mínimos cuadrados no lineales

Para comprender las complicaciones que introduce un modelo no lineal lo examinaremos gráficamente.

El objetivo de la suma de errores al cuadrado $S(\boldsymbol{\theta})$ sigue siendo la misma que en modelos lineales; minimizar la norma[1] de la distancia del vector respuesta $y = (y_1, y_2, \ldots, y_n)$ al vector sobre el espacio de estimación $\theta = \theta_1, \theta_1, \ldots, \theta_p$. Para determinada muestra, la función suma de cuadrados de los residuos, $S(\boldsymbol{\theta})$, sólo depende de los parámetros del modelo θ. Así, en el espacio de parámetros, se puede representar la función $S(\boldsymbol{\theta})$ con una gráfica de curvas de nivel, en la que cada curva de nivel en la superficie es una línea de suma constante de residuos al cuadrado.

Primero se tiene que en el modelo lineal, los parámetros están representados por $\boldsymbol{\beta}$ y la suma de cuadrados de los residuales es $S(\boldsymbol{\beta})$. En la figura (2.3) observamos que las curvas de nivel son elipsoides anidadas que van convergiendo al mínimo, el cual es único y corresponde a $\hat{\boldsymbol{\beta}}$.

[1]Se refiere a la norma Euclidiana.

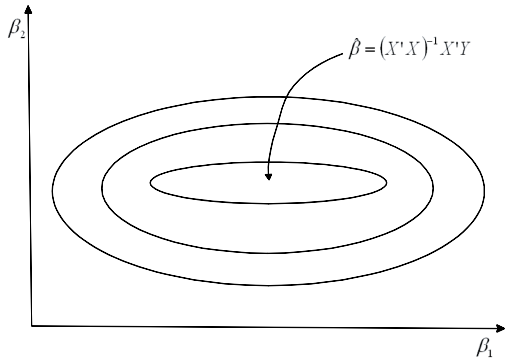


Figura 2.3: Modelo Lineal

Para el caso no lineal la gráfica varía de acuerdo con la función y los datos que se hayan obtenido. Está gráfica ya no está bien definida y suele ser muy irregular como en la figura (2.4) donde las curvas de nivel se muestran aplastadas y más estiradas; o en el caso de la figura (2.5) en donde encontramos un mínimo local y uno global, aunque podemos encontrar el caso en donde existan varios mínimos locales y quizás más de un mínimo global.

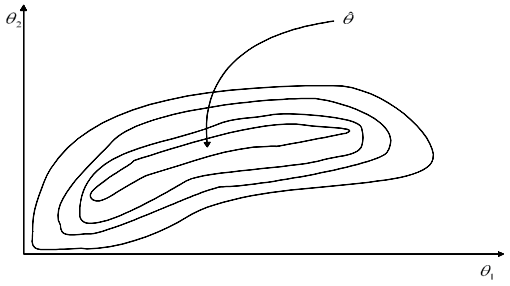


Figura 2.4: Modelo No lineal

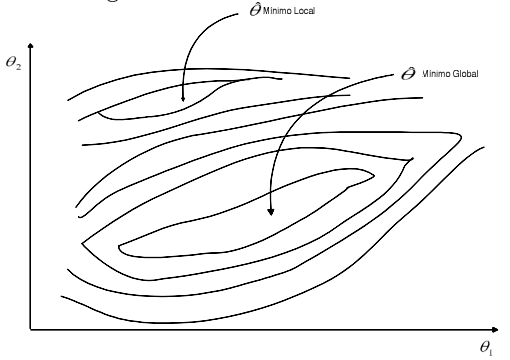


Figura 2.5: Modelo No Lineal Varios Mínimos

2.4.2. Método Gauss-Newton

En este método se usa una aproximación lineal para la función suavizadora para encontrar el estimador $\boldsymbol{\theta}$, mejorando el punto inicial $\boldsymbol{\theta}^0$ mediante varias iteraciones hasta que el estimador no cambie bajo cierta tolerancia y precisión, porque puede haber decimales que sigan cambiando, por ejemplo 3,8542254210 y 3,8542254214 si cambia pero la precisión ya puede ser suficiente y ya no se siguen con las iteraciones. Sea $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_p^0)$ la expansión en serie de Taylor de orden uno de $f(x_i, \boldsymbol{\theta})$ alrededor del punto $\boldsymbol{\theta}^0$ es:

$$f(x_i, \boldsymbol{\theta}) \approx f(x_i, \boldsymbol{\theta}^0) + \sum_{r=1}^{p} \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_r}\bigg|_{\theta=\theta^0} (\theta_r - \theta_r^0) \tag{2.22}$$

Si renombramos

$$\begin{aligned} \mathbf{f}(\boldsymbol{\theta}) &= \big(f(x_1, \boldsymbol{\theta}), f(x_2, \boldsymbol{\theta}) \ldots, f(x_n, \boldsymbol{\theta})\big)' \\ \mathbf{F_i} &= \frac{\partial f(x_i, \boldsymbol{\theta})}{\partial \theta_r} \end{aligned}$$

Agrupando todos los términos podemos escribir la ecuación (2.22) de la siguiente manera

$$\mathbf{f}(\boldsymbol{\theta}) \approx \mathbf{f}(\boldsymbol{\theta}^0) + \mathbf{F} * (\boldsymbol{\theta} - \boldsymbol{\theta^0}) \tag{2.23}$$

La suma de errores al cuadrado queda

$$\begin{aligned} S(\boldsymbol{\theta}) &= \| \mathbf{y} - \mathbf{f}(\boldsymbol{\theta}) \|^2 \\ &\approx \| \mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^0) - \mathbf{F}(\boldsymbol{\theta} - \boldsymbol{\theta}^0) \|^2 \\ &= \| \mathbf{z} - \mathbf{F}\boldsymbol{\beta} \|^2 \end{aligned} \tag{2.24}$$

donde $\mathbf{z} = \mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^0) = \epsilon$ y $\boldsymbol{\beta} = (\boldsymbol{\theta} - \boldsymbol{\theta}^0)$. Ahora tenemos una aproximación a un modelo lineal y podemos aplicar el resultado (2.16) donde $S(\boldsymbol{\theta})$ es minimizada cuando $\boldsymbol{\beta}$ está dada por

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{z} = \boldsymbol{\delta}^0$$

Ahora como $\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\theta}^0 = \boldsymbol{\delta}^0$ podemos usar a $\boldsymbol{\theta}^1 = \boldsymbol{\theta}^0 + \boldsymbol{\delta}^0$ como un mejor estimador para el parámetro desconocido $\boldsymbol{\theta}$. Para la siguiente iteración se reemplaza $\boldsymbol{\theta}^0$ por $\boldsymbol{\theta}^1$ en la ecuación (2.22), para producir otro estimador corregido, una nueva matriz de derivadas y un nuevo incremento, que sería un $\boldsymbol{\theta}^2$, y así sucesivamente. En general, en la a-ésima iteración se tiene

$$\hat{\boldsymbol{\theta}}^{a+1} = \boldsymbol{\theta}^a + \boldsymbol{\delta}^a = \boldsymbol{\theta}^a + (\mathbf{F}'\mathbf{F})^{-1}\mathbf{F}'\mathbf{z} \tag{2.25}$$

donde los residuos son $\mathbf{z} = \mathbf{y} - \mathbf{f}(\boldsymbol{\theta}^{a+1}) = \epsilon$. Usualmente se le conoce a $\boldsymbol{\delta}^0$ como el vector incremento.

Este proceso se repite hasta que el incremento sea tan pequeño que no haya un cambio radical en el vector de parámetros. Típicamente este criterio de convergencia está basado en

$$\left|\frac{\boldsymbol{\theta}^{a+1} - \boldsymbol{\theta}^{a}}{\boldsymbol{\theta}^{a}}\right| < \gamma$$

donde γ puede ser un número tan pequeño como 10^{-6}.

Una vez alcanzado el valor notable por medio del proceso iterativo, restará por determinar si dicho valor corresponde realmente a un mínimo o si el valor es un máximo de cáracter global o local. Para tratar de maximizar las posibilidades de que se trate de un mínimo absoluto y no tan solo de un mínimo local una de las prácticas habituales consiste en utilizar el algoritmo para diferentes valores iniciales de $\boldsymbol{\theta}$. Para los distintos valores iniciales, podemos obtener diferentes mínimos de la función, el mínimo correspondiente con la menor suma de cuadrados de los errores será el estimador por mínimos cuadrados no lineales.

Por otra parte, el propio procedimiento no puede conducir a un máximo porque F'F será siempre positivo debido a que es siempre una función cuadrática. Si se comienza el procedimiento con un valor inicial de θ, la pendiente de la función a minimizar, $S(\boldsymbol{\theta})$, será positiva, por lo cual el algoritmo conducirá en la dirección correcta y obtendremos valores de $\boldsymbol{\theta}$ menores que $\boldsymbol{\theta}^0$; por lo tanto, nos estaremos moviendo hacia un mínimo (absoluto o local). Así mismo, si se comienza con un valor inicial de $\boldsymbol{\theta}$ situado a la izquierda de un mínimo, la pendiente de $S(\boldsymbol{\theta})$ será negativa, por lo cual el cambio en $\boldsymbol{\theta}$ será positivo y nuevamente nos moveremos hacia un mínimo.

Un problema adicional que puede ocurrir es que los sucesivos cambios producidos en las estimaciones iterativas de θ sean demasiado grandes y no puedan localizar el mínimo. Puede ser conveniente introducir una variable de **longitud de paso** $\boldsymbol{\lambda}$ de manera que disminuya la posibilidad de que el cambio en θ sea demasiado grande (es decir, se pase de un valor de θ a la izquierda del θ que minimiza la función, a otro a la derecha o viceversa), lo cual puede conducir a que el número de iteraciones necesarias para arribar al valor crítico sea muy grande o que jamás converja. Considerando esta posibilidad, el algoritmo se transforma en:

$$\hat{\boldsymbol{\theta}}^{a+1} = \boldsymbol{\theta}^{a} + \lambda \boldsymbol{\delta}^{a}$$

donde λ es tal que

$$S(\boldsymbol{\theta}^1) < S(\boldsymbol{\theta}^0) \tag{2.26}$$

Un método común de seleccionar a λ es empezar con $\lambda = 1$ y ver si la condición (2.26) se satisface, aunque generalmente, la variable se calcula a traves de un proceso de prueba y error.

2.4.3. Geometría del método de Gauss-Newton

Ya se ha mencionado que las curvas de nivel de un modelo no lineal son irregulares, aplastadas y muy estiradas. El método de Gauss-Newton convierte el problema no lineal en uno lineal que comienza en el punto $\boldsymbol{\theta}^0$.

La primera iteración cambia las curvas irregulares de nivel por un conjunto de curvas elípticas. Las curvas irregulares de $S(\boldsymbol{\theta})$ pasan exactamente a través del punto inicial $\boldsymbol{\theta}^0$, en la siguiente iteración sólo se repite el proceso, comenzando en la nueva solución $\boldsymbol{\theta}^1$ como se ve en la figura (2.6). La evolución definitiva de la linealización es una secuencia de problemas lineales para los cuales las soluciones se acercan hacia un mínimo global de la función no lineal, esto se ve en la figura (2.7).

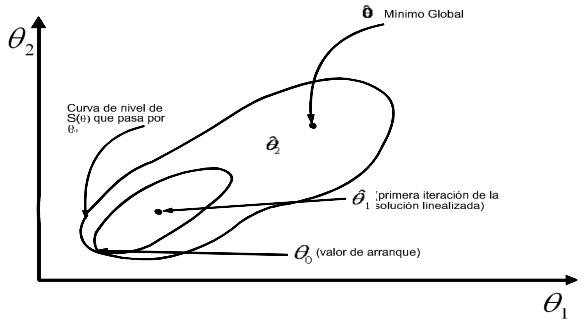


Figura 2.6: La primera iteración

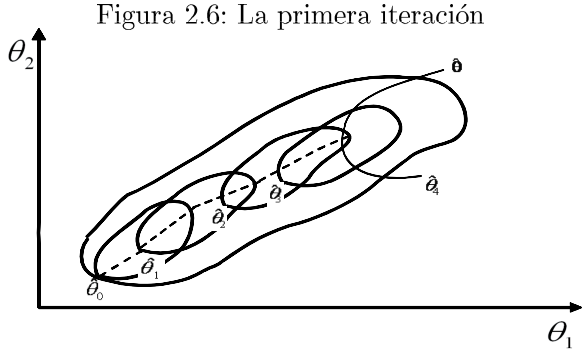


Figura 2.7: Evolución de sucesivas iteraciones de linealización

2.4.4. Método de Newton-Raphson

Se parte nuevamente del modelo en notación matricial

$$\mathbf{y_i} = f(x_i, \boldsymbol{\theta}) + \epsilon_i \qquad i = 1, 2, 3, ..., n$$

y de los supuestos $\epsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ y $\epsilon_i \perp \epsilon_j$. En este caso, en vez de aproximar a la función $f(x_i, \boldsymbol{\theta})$ por un desarrollo en serie de Taylor de primer orden, se aproximará a la función $S(\boldsymbol{\theta})$ por un desarrollo en serie de Taylor de segundo orden alrededor de un punto inicial de $\boldsymbol{\theta}$.

Primero se parte de un modelo general con un único parámetro

$$\mathbf{y} = f(x, \theta) + \epsilon$$

donde θ es un escalar, se toma θ^0 como el valor inicial, la expansión en serie de Taylor de segundo grado de S(θ) alrededor del punto θ^0 es:

$$S(\theta) \approx S(\theta^0) + \frac{dS(\theta)}{d\theta}\bigg|_{\theta=\theta^0}(\theta - \theta^0) + \frac{1}{2}\frac{d^2S(\theta)}{d\theta^2}\bigg|_{\theta=\theta^0}(\theta - \theta^0)^2$$

Para resolver el ploblema de minimización derivamos con respecto a θ (el polinomio de Taylor de segundo grado)

$$\frac{dS(\theta)}{d\theta} \approx \frac{dS(\theta)}{d\theta}\bigg|_{\theta=\theta^0} + \frac{d^2S(\theta)}{d\theta^2}\bigg|_{\theta=\theta^0}(\theta - \theta^0) \tag{2.27}$$

se nombra

$$H(\theta^0) = \frac{d^2S(\theta)}{d\theta^2}\bigg|_{\theta=\theta^0}$$

se reemplaza $H(\theta^0)$, se iguala a cero la ecuación (2.27) y se resuleve para θ

$$\begin{aligned}
\frac{dS(\theta)}{d\theta} &\approx \frac{dS(\theta)}{d\theta}\bigg|_{\theta=\theta^0} + H(\theta^0)(\theta - \theta^0) = 0 \\
\Longrightarrow H(\theta^0)\theta - H(\theta^0)\theta^0 &= -\frac{dS(\theta)}{d\theta}\bigg|_{\theta=\theta^0} \qquad (2.28) \\
H(\theta^0)\theta &= H(\theta^0)\theta^0 - \frac{dS(\theta)}{d\theta}\bigg|_{\theta=\theta^0} \qquad (2.29) \\
\theta &= \theta^0 - H^{-1}(\theta^0)\frac{dS(\theta)}{d\theta}\bigg|_{\theta=\theta^0} \qquad (2.30)
\end{aligned}$$

Se toma a θ como la siguiente estimación y se denomina θ^1. Continuando el proceso en iteraciones sucesivas, se obtiene la iteración a-ésima del algoritmo que estará dada por:

$$\theta^{a+1} = \theta^a + H(\theta^a)\frac{dS(\theta)}{d\theta}\bigg|_{\theta=\theta^a} \tag{2.31}$$

A continuación se hace extensivo el resultado para el caso de un modelo generalizado no lineal con k parámetros a estimar. Se usa la notación

$$\mathbf{g}(\boldsymbol{\theta}) = \frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \frac{\partial S(\theta)}{\theta_1} \\ \frac{\partial S(\theta)}{\theta_2} \\ \vdots \\ \frac{\partial S(\theta)}{\theta_p} \end{pmatrix}_{p\times 1}$$

y

$$H(\boldsymbol{\theta}^0) = \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \begin{pmatrix} \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_1^2} & \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_p} \\ \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_2^2} & \cdots & \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_2 \partial \theta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_1} & \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_2} & \cdots & \frac{\partial^2 S(\boldsymbol{\theta})}{\partial \theta_p^2} \end{pmatrix}_{p \times p}$$

denotan, respectivamente, el vector gradiente y la matriz Hessiana de S($\boldsymbol{\theta}$).

Sea $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_p^0)'$ el vector de valor inicial de dimensión p×1, la expansión en serie de Taylor de segundo grado de S($\boldsymbol{\theta}$) alrededor del punto $\boldsymbol{\theta}^0$ es:

$$S(\boldsymbol{\theta}) \approx S(\boldsymbol{\theta}^0) + \mathbf{g}'(\boldsymbol{\theta^0})(\boldsymbol{\theta} - \boldsymbol{\theta^0}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta^0})' H(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta^0})$$

Se deriva con respecto a $\boldsymbol{\theta}$

$$\frac{\partial S(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \approx \mathbf{g}'(\boldsymbol{\theta^0}) + H(\boldsymbol{\theta}^0)(\boldsymbol{\theta} - \boldsymbol{\theta^0})$$

generalizando los resultados obtenidos para el caso de un parámetro, se encuentra el valor $\boldsymbol{\theta}$ que minimiza $S(\boldsymbol{\theta})$ es

$$\boldsymbol{\theta} = \boldsymbol{\theta}_0 - \mathbf{H}(\boldsymbol{\theta}^0)^{-1}\mathbf{g}(\boldsymbol{\theta^0}) \tag{2.32}$$

La iteración a-ésima del algoritmo está dada por

$$\hat{\boldsymbol{\theta}}^{a+1} = \boldsymbol{\theta}_a - H(\boldsymbol{\theta}^a)^{-1}\mathbf{g}(\boldsymbol{\theta^a}) \tag{2.33}$$

Al igual que el algoritmo de Gauss-Newton este algoritmo continuará hasta que se se obtenga un estimador el cual la precisión sea suficiente como no continuar con las iteraciones.

Surge la interrogante de saber si el algoritmo de Newton-Raphson converge a un mínimo y de ser así, si es local o global. Debido a que en el entorno de un mínimo el término $H(\theta^0)$ será positivo, el procedimiento conducirá a un mínimo si el valor inicial de θ está lo suficientemente cerca del mínimo. Sin embargo, si θ^0 no se encuentra cercano a un mínimo sino a un máximo, la derivada de la función será negativa y $H(\theta^0)$ será también negativo por lo cual el procedimiento conducirá en dirección de un máximo. Por lo tanto, los resultados pueden ser tanto mínimos como máximos e incluso pueden darse situaciones extrañas como puntos sillas en caso de que la segunda derivada también sea 0. Está característica del algoritmo de Newton-Raphson hace necesario ser cuidadoso a la hora de aplicar el método y no tomar un único valor del estimador como definitivo, por lo que es muy recomendable realizar la estimación para varios valores iniciales de θ^0. Como en el procedimiento de Gauss–Newton, existe la posibilidad de que se pase por alto el valor del mínimo, por lo cual se recomienda, como se hizo antes, la utilización de una

variable de longitud de paso $\boldsymbol{\lambda}$ para disminuir la probabilidad de que esto ocurra. La forma más general del algoritmo queda establecida como:

$$\hat{\boldsymbol{\theta}}^{a+1} = \boldsymbol{\theta}_a - \lambda H(\theta^0)^{-1} \frac{\partial S(\boldsymbol{\theta})}{\partial \theta}\bigg|_{\theta=\theta^0}$$

Para cada iteración λ es especificado de tal forma que $S(\boldsymbol{\theta}^{a+1}) < S(\boldsymbol{\theta}^a)$.

2.5. Estimación de la varianza

Es de destacar que la varianza puede ser estimada como:

$$\sigma^2 = \frac{\sum_{r=1}^{p}(y_i - \hat{y}_i)^2}{n-p} \tag{2.34}$$

si se ve como

$$\sigma^2 = \frac{\sum_{r=1}^{p}(y_i - f(x_i, \hat{\theta}))^2}{n-p}$$

recordando que p es el número de parámetros en el modelo de regresión no lineal. Se puede estimar el valor de la matriz de covarianza asintótica (de muestras grandes) a través del vector de parámetros $\hat{\theta}$ que sería:

$$Var(\hat{\theta}) = \sigma^2(F'F)^{-1}$$

donde X es compuesta por las derivadas parciales ya definidas y evaluadas en el estimador de θ.

2.6. La importancia de buenos valores iniciales

Uno de los aspectos que puede hacer que se tenga éxito en el análisis no lineal es elegir valores iniciales para los parámetros, que sean cercanos a los valores verdaderos de los parámetros, para minimizar las dificultades de convergencia. Una mala elección de arranque podría causar la convergencia hacia un mínimo local de la función, o que pase completamente desapercibido el que se haya obtenido una solución subóptima. Estos simples principios pueden ser usados para determinar valores iniciales:

1. Interpretar la función de valor esperado. Una de las ventajas de la regresión no lineal es que los parámetros de la función suavizadora tienen usualmente algún significado para los científicos o investigadores. Este significado puede ser gráfico, físico, biologico, químico, etc. y esto puede ser de mucha ayuda para determinar los valores iniciales. Los estimadores iniciales pueden obtenerse mediante un experimento. También graficar la función de valor esperado usando varios valores de los parámetros es un ejercicio extremadamente benéfico,

por que con este camino uno puede familiarizarse con la función y con los parámetros que realmente la afectan.

2. Transformar la función de valor esperado para obtener los valores de arranque. Los mínimos cuadrados lineales se pueden usar con los datos recíprocos para producir estimados de los parámetros lineales.

3. Reducir la dimensión sustituyendo en algunos parámetros valores específicos. Con está técnica se van estimando los parámetros sucesivamente. Como ejemplo tenemos la función $f = \theta_1 + \theta_2 e^{-\theta_3 x}$, donde θ_3 es positivo. El límite de la función cuando $x \to \infty$ es θ_1 y el valor de f cuando $x = 0$ es $\theta_1 + \theta_2$. Dependiendo de donde se incrementan o decrecen los datos se tomar y_{max} o y_{min} como valor inicial de θ_1^0, y se usa la diferencia $y(0) - \theta_1^0$ para encontrar θ_2^0. Se puede diseñar una regresión lineal (sin el término constante) para $ln[(y - \theta_1^0)/\theta_2^0]$ y obtener θ_1^0.

2.7. Intervalos de Confianza Asintóticos

2.7.1. Intervalos de Confianza Asintóticos para θ

Sea

$$\mathbf{y_i} = f(\mathbf{x_i}, \boldsymbol{\theta}) + \epsilon_i \qquad i = 1, 2, 3, ..., n \tag{2.35}$$

donde $\epsilon \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$ y $\epsilon_i \perp \epsilon_j$. Denotamos el verdadero valor del vector parámetro por $\boldsymbol{\theta}$. Si es de interés la construcción de intervalos de confianza para una combinación lineal de $\boldsymbol{a'\theta}$ se puede aplicar el resultado del apéndice [A.8](sólo que se cambia $\mathbf{X}$ por $\mathbf{F}$). En particular tenemos un resultado asintótico

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}, \sigma^2 \mathbf{C^{-1}}) \qquad \mathbf{C} = \mathbf{F'F} \tag{2.36}$$

De lo anterior se tiene que asintóticamente $\boldsymbol{a'\hat{\theta}} \sim N(\mathbf{a'\theta}, \sigma^2 \mathbf{a'C^{-1}a})$ y es independiente de $S^2 = \| \mathbf{y} - \mathbf{f}(\hat{\boldsymbol{\theta}}) \|^2 / (n - p)$, el cuál es un estimador de σ^2. Se tiene que

$$\frac{\mathbf{a'\hat{\theta}} - \mathbf{a'\theta}}{\sigma \sqrt{\mathbf{a'C^{-1}a}}} \sim N(0, 1)$$

y que un estimador insesgado de σ^2 es:

$$\frac{(n - p)S^2}{\sigma^2} \sim \chi^2_{(n-p)}$$

De aquí que para n grande, tenemos que aproximadamente

$$T = \frac{\mathbf{a}'\hat{\boldsymbol{\theta}} - \mathbf{a}'\boldsymbol{\theta}}{S\sqrt{\mathbf{a}'\mathbf{C}^{-1}\mathbf{a}}} \sim t_{(n-p)} \tag{2.37}$$

donde $t_{(n-p)}$ es la distribución t con n-p grados de libertad. Un intervalo de confianza aproximado al $100 * (1-\alpha)\,\%$ para $\boldsymbol{a}'\hat{\boldsymbol{\theta}}$ es entonces

$$\mathbf{a}'\hat{\boldsymbol{\theta}} \pm t^{\alpha/2}_{(n-p)} S\sqrt{(\mathbf{a}'\mathbf{C}^{-1}\mathbf{a})}$$

Aquí $\mathbf{C}$ puede ser estimada por $\hat{\mathbf{C}} = \hat{\mathbf{F}}'\hat{\mathbf{F}}$. Tomando $\mathbf{a}' = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$, donde el r-ésimo elemento de $\mathbf{a}$ es uno y el resto son cero, un intervalo de confianza para el r-ésimo elemento de $\boldsymbol{\theta}$, θ_r, es

$$\theta_r \pm t^{\alpha/2}_{(n-p)} S\sqrt{\mathbf{C}^{-1}} \tag{2.38}$$

2.7.2. Intervalos de Confianza Asintóticos de Predicción

Usando la linealización asintótica de (2.35) se puede predecir nuevas observaciones de y y que correspondan a un nivel especificado de la variable regresora x. Si $\mathbf{x_0}$ es el valor de interés de la variable regresora, entonces

$$\hat{y}_0 = f(\mathbf{x_0}, \hat{\boldsymbol{\theta}})$$

es el estimador puntual del nuevo valor de la respuesta y_0. A continuación se obtendrá un estimado para la esta observación futura y_0. Entonces para n grande, $\hat{\boldsymbol{\theta}}$ cercano al verdadero valor de $\boldsymbol{\theta}$, tenemos la usual expansión de Taylor

$$f(x_0, \hat{\boldsymbol{\theta}}) \approx f(\mathbf{x_0}, \hat{\boldsymbol{\theta}}) + \mathbf{f}'_0 * (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

donde $\mathbf{f}'_0$

$$\mathbf{f}'_0 = (\frac{\partial f(\mathbf{x_0}, \boldsymbol{\theta})}{\partial \theta_1}, \frac{\partial f(\mathbf{x_0}, \boldsymbol{\theta})}{\partial \theta_2}, \ldots, \frac{\partial f(\mathbf{x_0}, \boldsymbol{\theta})}{\partial \theta_p})$$

y de aquí

$$y_0 - \hat{y}_0 \approx y_0 - f(\mathbf{x_0}, \boldsymbol{\theta}) - \mathbf{f}'_0(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \epsilon_0 - \mathbf{f}'_0(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

De(2.36) y de la independencia estadística de $\hat{\boldsymbol{\theta}}$ y ϵ

$$\begin{aligned} E[y_0 - \hat{y}_0] &\approx E[\epsilon_0] - \mathbf{f}'_0 E[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}| \approx 0 \\ Var[y_0 - \hat{y}_0] &\approx Var[\epsilon_0] + Var[\mathbf{f}'_0(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] \\ &\approx \sigma^2 + \sigma^2 \mathbf{f}'_0(\mathbf{F}'\mathbf{F})^{-1}\mathbf{f_0} \\ &= \sigma^2(1 + v_0) \end{aligned}$$

donde $v_0 = f_0' * (\mathbf{F}'\mathbf{F})^{-1} * f_0$.

Nótese que la variable $y_0 - \hat{y}_0$ se distribuye asintóticamente $N(0, \sigma^2(1 + v_0))$. Ahora S^2 es independiente de y_0 y es asintóticamente independiente de $\hat{\boldsymbol{\theta}}$ de aquí que:

$$\frac{y_0 - \hat{y}_0}{S\sqrt{1 + v_0}} \sim t_{(n-p)}$$

y un intervalo de confianza asintótico de predicción para y_0 está dado por

$$\hat{y}_0 \pm t_{(n-p)}^{\alpha/2} S|1 + \mathbf{f_0'}(\mathbf{F}'\mathbf{F})^{-1}\mathbf{f_0}]^{1/2} \tag{2.39}$$

Se toma en cuenta que $\mathbf{f_0}$ y $\mathbf{F}$ son funciones de $\boldsymbol{\theta}$, v_0 es una función de $\boldsymbol{\theta}$ y puede ser estimada remplazando $\boldsymbol{\theta}$ por $\hat{\boldsymbol{\theta}}$.

2.8. Pruebas de Hipótesis

2.8.1. Pruebas de Hipótesis concernientes a una sola θ

Se desea probar la hipótesis de que θ_r es igual a una costante θ_0. Las hipótesis correspondientes son:

$$\begin{aligned} H_0 : \theta_r &= \theta_0 \\ H_1 : \theta_r &\neq \theta_0 \end{aligned}$$

Tomamos como estadístico de prueba, cuando n es razonablemente grande, a la ecuación (2.37)

$$T = \frac{\mathbf{a}'\hat{\boldsymbol{\theta}} - \mathbf{a}'\boldsymbol{\theta}}{S\sqrt{\mathbf{a}'\mathbf{C}^{-1}\mathbf{a})}} \sim t_{(n-p)}$$

donde $t_{(n-p)}$ es la distribución t con n-p grados de libertad.
La regla de decisión a nivel de significancia α es:

$$\text{Rechaza } H_0 \text{ si } |T| > t_{(1-\alpha/2;n-p)}$$

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

Que se rechace la hipótesis nula $\theta_0 = 0$, implica que el regresor θ_r contribuye al modelo en forma significativa.

2.8.2. Prueba de Hipótesis para varias θ

Para la prueba de varias θ:

$$\begin{aligned} H_0 &: \boldsymbol{\theta} = \theta_0 \\ H_1 &: \textit{existe alguna } \theta_i \neq \theta_{i0} \end{aligned}$$

Para el estadístico de prueba se utiliza el hecho demostrado en el apéndice [A.11] de que

$$F^* = \frac{MS_R}{MS_Res}$$

Tiene una distribución no centrada $F_{(k,n-p,\lambda)}$, con parámetro de no centralidad λ para un n grande, cuando H_0 cs cicrta. Sc calcula cl cstadístico dc prucba F^* y se rechaza H_0 si:

$$F^* > F_{(k,n-p,\lambda)}$$

El procedimiento de prueba se puede resumir en una **Tabla de Análisis de Varianza** como se muestra a continuación

Fuente de variación	**Suma de Cuadrados**	**Grados de libertad**	**Cuadrado Medio**	F_0
Regresión	SS_R	k	MS_R	MS_R/MS_{Res}
Residuos	SS_{Res}	n-k	MS_{Res}	
Total	SS_T	n-1		

Cuadro 2.2: Tabla dc Análisis dc Varianza

Capítulo 3

Construyendo El Modelo de Regresión No Lineal

Como se vio en el capítulo anterior, ajustar un modelo de regresión no lineal es demasiado laborioso, por lo que antes de comenzar el análisis es conveniente seleccionar un programa de cómputo para facilitar el trabajo. Sin embargo, como es bien sabido no basta "decirle" a la computadora que lo resuelva, es necesario desarrollar un código para el modelo y ello hay que escribirlo de manera conveniente. En este capítulo se explicará la construcción de un modelo de regresión no lineal y la forma de interpretar los resultados.

3.1. Preparando los datos para una regresión no lineal

La regresión no lineal se usa para construir un modelo apropiado en el cual se define Y en función de X. La meta es generar una curva estándar que se pueda usar para predecir valores desconocidos de Y. Se utiliza para variables cuantitativas dependientes, que se pueden medir en un intervalo como el tamaño de una persona, la presión sanguinea o la temperatura. Si los resultados pertenecen a una distribución binomial como el género, masculino o femenino, viable o no viable, la regresión lineal no es apropiada. En estos casos se necesita un análisis especial como la regresión logistica.

No debe usarse la regresión lineal sólo para evitar usar la regresión no lineal. Encontrar curvas con regresión no lineal no es difícil, y considerando todo el tiempo y el esfuerzo que se pone en la recolección de datos, es preferible utilizar el mejor análisis posible.

3.1.1. Transformando los valores de Y

Multiplicar o dividir todos los valores de Y por una constante no cambia la curva de mejor ajuste. Se obtendrán valores ajustados de los parámetros iguales o equivalentes con intervalos de confianza equivalentes. Puede ser una buena idea transformar, o cambiar las unidades, para evitar valores muy grandes o muy pequeños.

Se debe de notar que todos los parámetros están expresados en unidades de Y, por lo que si se cambian los valores de Y, también se cambiarán las unidades de los parámetros.

Restar una constante a todos los valores de Y no cambiará la distancia de los puntos a la mejor curva ajustada, por lo que no afectará la curva obtenida por la regresión no lineal.

Mientras que transformar linealmente los valores de Y (como dividir todos los valores por una constante o restar una constante) no cambia la naturaleza de la mejor curva suavizadora, en contraste, una transformación no lineal (como el logaritmo de todos los valores de Y, calcular la raíz cuadrada) cambia la posición relativa de los datos y da como resultado una curva diferente que minimiza la suma de cuadrados. Por lo que una transformación no lineal arroja diferentes valores para los estimadores de los parámetros. Dependiendo de problemas a estudiar, esto puede ser bueno o malo.

La regresión no lineal se basa en los supuestos de que los datos alrededor de la curva siguen una distribución Normal. Si los datos ya siguen una distribución Normal realizar una transformación no lineal puede hacer que el supuesto se viole. Sin embargo, si los datos no siguen una distribución normal, una transoformación no lineal puede ayudar a que este supuesto se cumpla y, en estos casos, una transformación no lineal es una buena opción.

3.1.2. Criterios para Remover las observaciones discrepantes (Outliers)

Las observaciones discrepantes están definidas como aquellas con un residuos grande, en ocasiones es equivalente hablar de observaciones con valores grandes. Estos valores son llamados *outliers*. Primeras preguntas antes de intentar borrar el outlier:

1. ¿Es el dato encontrado correcto? Puede haber un error de captura.
2. ¿Hubo algún problema experimental con el valor? Por ejemplo, si notamos que durante el experimento alguno de los aparatos no funciona correctamente, se puede pensar que el dato obtenido es un error del aparato y se tiene justificación para borrarlo del análisis.

3. ¿Puede el outlier ser causado por una diversidad biológica? Si cada valor es resultado de una persona o animal diferente, el outlier puede ser un valor correcto. Entonces el outlier no es resultado de un error, se debe a que el individuo es considerablemente diferente a los demás. Es interesante encontrar este tipo de datos.

Antes de decir NO a estas tres preguntas, hay que decidir qué hacer con el outlier. Hay dos posibilidades:

- Una posibilidad es que el outlier es correcto, en este caso el dato se conserva para el análisis. El valor tiene la misma distribución que los otros valores, entonces este debe de ser incluido.
- La otra posibilidad es que el valor es un error. Cuando se incluye un valor erróneo en el análisis los resultados pueden ser incorrectos, por lo que es preferible quitar el dato. En otras palabras, el valor es resultado de una población diferente a la de los otros y es engañoso.

El problema es que nunca se puede estar seguro cuál de estas posibilidades es correcta. Los estadísticos han desarrollado varios métodos para detectar la probabilidad de encontrarlos. En todos los métodos primero se cuantifica qué tan lejos está el outlier de cierto valor, como la media de todos los puntos, la diferencia entre el outlier y la media de los valores restantes, o la diferencia entre el outlier y el valor más cercano. Después estandarizamos este valor dividiendo por alguna medida de disipersión. Puede ser la desviación estándar de todos los valores, la desviación estándar de los valores restantes o el rango del dato. Finalmente se cálcula un valor "P" que responde a está pregunta: Si todos los valores se distribuyen de forma Normal ¿Cuál es la probabilidad de obtener un outlier? Si esa probabilidad es pequeña, entonces se puede concluir que el outlier es un valor erróneo, y se tiene la justificación de excluirlo del análisis.

Un punto de **influencia**, o **valor influeyente**, tiene un impacto notable sobre los coeficientes del modelo, porque "jala" al modelo de regresión en su dirección.

A veces se ve que un pequeño subconjunto de los datos ejerce una influencia desproporcionada sobre los coeficientes y las propiedades del modelo. En un caso extremo, los estimados de parámetro pueden depender más del subconjunto influyente de puntos que de la mayor parte de los datos. Estos modelos son indeseables, se prefiere un modelo de regresión que sea representativo de todas las observaciones en la muestra, y no sea solo de unas cuantas. En consecuencia, se desea localizar esos puntos influyentes y evaluar su impacto sobre el modelo. Si esos puntos son en realidad valores "malos", se deberían eliminar de la muestra. Por otro lado, puede que

no haya nada de malo en ellos, pero si controlan las propiedades clave del modelo, sería bueno conocerlo, porque podrían afectar el uso final del modelo de regresión. Existen métodos de diagnóstico para balanceo e influencia, se pueden consultar en "The geometry of case deletion and the assessment of influence in nonlinear regression", *The Canadian Journal of Statistics*, Vol. 15, No. 2, 91-103, 1987. y "Leverage, local influence and curvature in nolinear regression", *Biometrika*, **80**, 1, 99-106.

3.1.3. El coeficiente de determinación R^2

La cantidad

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = \frac{SS_R}{SS_T} \\ &= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} = 1 - \frac{SS_{Res}}{SS_T} \end{aligned}$$

se llama **coeficiente de determinación**. Donde SS_T significa Suma de cuadrados del Total, SS_R es la suma de cuadrados de la regresión y SS_{Res} es la suma de cuadrados de los residuos. Como SS_T es una medida de la variabilidad de y sin considerar el efecto de la variable regresora x y SS_{Res} es una medida de la variabilidad de y que queda después de haber tenido en consideración a x, R^2 se llama con frecuencia, la proporción de la variación explicada por el regresor x. Ya que $0 \leq SS_{Res} \leq SS_T$, entonces $0 \leq R^2 \leq 1$. Los valores de R^2 cercanos a 1 implican que la mayor parte de la variabilidad de y está explicada por el modelo de regresión.

El estadístico R^2 se debe usar con precaución, porque siempre es posible conseguir que R^2 sea grande agregando términos suficientes al modelo. Por ejemplo, si no hay puntos repetidos (más de un valor de y con el mismo valor de x), un polinomio de grado $n-1$ producirá un ajuste "perfecto", con $R^2 = 1$, de los n puntos de datos. Cuando hay puntos repetidos, R^2 nunca puede ser exactamente igual a 1, porque el modelo no puede explicar la variabilidad relacionada con el error "puro".

3.2. Las primeras cinco preguntas antes de los resultados de la regresión no lineal

1. ¿Los valores de los parámetros son plausibles?
   Cuando evaluamos los valores de los parámetros reportados por la regresión no lineal, lo primero que debemos hacer es checar que el resultado es científicamente plausible. El programa puede encontrar el mejor valor para los parámetros, pero no puede saber

qué significan, por lo que pueden no tener sentido en el contexto del problema. Si los valores ajustados no son científicamente sensatos, entonces el ajuste no es bueno. Se puede forzar al programa a buscar en un rango, e intentar el ajuste otra vez.

2. ¿La curva está cerca de los datos?
   Si el objetivo del análisis es simplemente encontrar una curva que ayude a interpolar valores desconocidos, puede ser suficiente sólo mirar la curva. En este contexto no son de impotancia los valores de los parámetros, aunque si lo es la construcción de intervalos de confianza y de predicción, sus errores estándar. Aunque, frecuentemente se necesitará ajustar curvas para los datos que ayuden a entender su sistema. Bajo está circunstancia necesitas mirar cuidadosamente los resultados de regresión no lineal.

3. ¿Qué tan precisos son los valores ajustados de los parámetros?
   No sólo se quiere conocer cual es el mejor ajuste, sino también se quiere saber cuál es la precisión de este valor. Los programas de regresión no lineal reportan un error estándar para cada parámetro, así como intervalos del 95 % de confianza. Algunos calculan en base a intervalos asinstóticos como los vistos en el capítulo 2. Si todos los supuestos de la regresión son ciertos, la probabilidad de que el intervalo contenga al valor verdadero del parámetro es del 95 %. Si el intervalo de confianza es razonablemente angosto, has alcanzado la meta, encontrar el mejor valor ajustado del parámetro con razonable certeza. Si el intervalo de confianza es realmente ancho, entonces se tienen problemas. El parámetro estaría dentro de un rango de valores muy amplio.

4. ¿Puede otro modelo ser apropiado?
   La regresión no lineal encuentra parámetros que hacen ajustar el modelo a los datos tanto como es posible (dados algunos supuestos). Esto no determina si un modelo podría trabajar mejor que otro.

   Aunque un modelo ajusta bien los datos, éste puede no ser el mejor[1] modelo o el más correcto, se debe de estar siempre alerta a la posibilidad de que un modelo diferente trabaje mejor. En algunos casos, se necesita recolectar datos en un rango más amplio de X para poder notar que modelo ajusta mejor los datos. En caso de ser un experimento se pueden recolectar los datos bajo diferentes condiciones experimentales. Y en caso de que los datos no se puedan recolectar, al ser un modelo una abstracción, en particular se puede tener una idea de lo que queremos y se podrían simular los datos.

[1]En realidad no existe el mejor modelo, lo que se busca es un modelo que ajuste correctamente a los datos según ciertos criterios

5. ¿Se ha violado alguno de los supuestos de la regresión no lineal?
   La regresión no lineal está basada en un conjunto de supuestos. Cuando se revisan los resultados, también se debe de revisar que los supuestos no hayan sido violados. Puede darse una violación de los supuestos; sin embargo, dependiendo del objetivo de la estimación, algunas violaciones de los supuestos podrían no requerir corrección.

   SUPUESTOS

- X es conocida. Todo el error está en Y.

- Los errores siguen una distribición normal.
  Pequeñas desviaciones respecto a la hipótesis de normalidad no afectan mucho al modelo, pero una no normalidad grande es potencialmente más seria, porque la inferencia estadística a través de los estadísticos t o F y los intervalos de confianza y de predicción dependen del supuesto de normalidad. Un método sencillo para comprobar la suposición de normalidad es trazar una gráfica de probabilidad normal de los residuos. Es una gráfica diseñada para que al desplegarse la distribución normal acumulada parezca una línea recta. Sean $e_1 < e_2 < \ldots < e_n$ los residuos ordenados en orden creciente. Si se grafican e_i en función de la probabilidad acumulada $P_i = (i - \frac{1}{2}), i = 1, 2, \ldots, n$, en papel de probabilidad normal, los puntos que resulten deberían estar aproximadamente sobre una línea recta. Esa recta se suele determinar en forma visual, con énfasis en los valores centrales y no en los extremos. Las diferencias apreciables respecto a la recta indican que la distribución no es normal. En la figura 3.1a se muestra la gráfica de probabilidad idealizada. Obsérvese que los puntos caen aproximadamente sobre una recta. Las partes b a d de la gráfica muestran otros problemas característicos. La parte b muestra una curva que va bruscamente hacia arriba y hacia abajo en los dos extremos, lo que indica que las colas de esta distribución son demasiado pesadas para poder considerarla como normal. Al contrario, la parte c muestra un aplanamiento en los extremos, que es un comportamiento característico de las muestras tomadas de una distribución con colas más ligera que la normal. La parte d de la gráfica muestra patrones asociados con asimetría positiva y negativa, respectivamente.

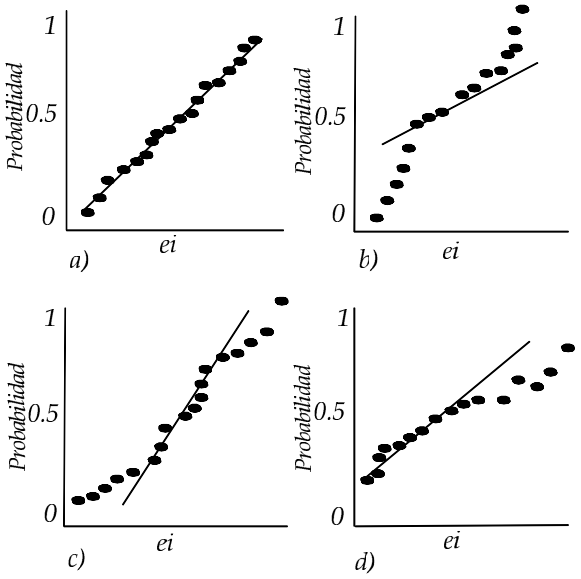


Figura 3.1: Gráficas de probabilidad normal

También existen pruebas no paramétricas[2] para probar la normalidad como la prueba Kolmogorov, la prueba χ^2 de bondad y ajuste, la prueba Anderson-Darling o la prueba Shapiro-Wilks.

- Se conoce como *homocedasticidad*[3] de los residuos al hecho de que la dispersión de la desviación estándar de los errores es la misma no importando que la curva crezca. Si este supuesto no se cumple se tiene una disminución de la eficiencia del estimador mínimo cuadrático, éste deja de ser el de varianza miníma entre todos los estimadores lineales e insesgados. La gráfica se construye poniendo a los residuos e_i en función de los valores correspondientes $\hat{y}_i$. Y sirve para poder corraborar el supuesto de varianza constante en los residuos.

  Si esta gráfica se parece a la de la figura 3.2a, que indica que los residuos se pueden encerrar en una banda horizontal, entonces no hay defectos obvios del modelo. Las gráficas de e_i en función de $\hat{y}_i$ que se parezcan a cualquiera de los patrones de las partes b a d son síntomas de deficiencias del modelo.

[2]Al realizar estás pruebas se debe tener en cuenta que los residuos no son independientes; esto es por la forma en que se calculan: $e = y - \hat{y}$. Al calcular los estimados se necesita tomar en cuenta toda la muestra, y esto hace que los residuos no sean independientes

[3]Varianza constante

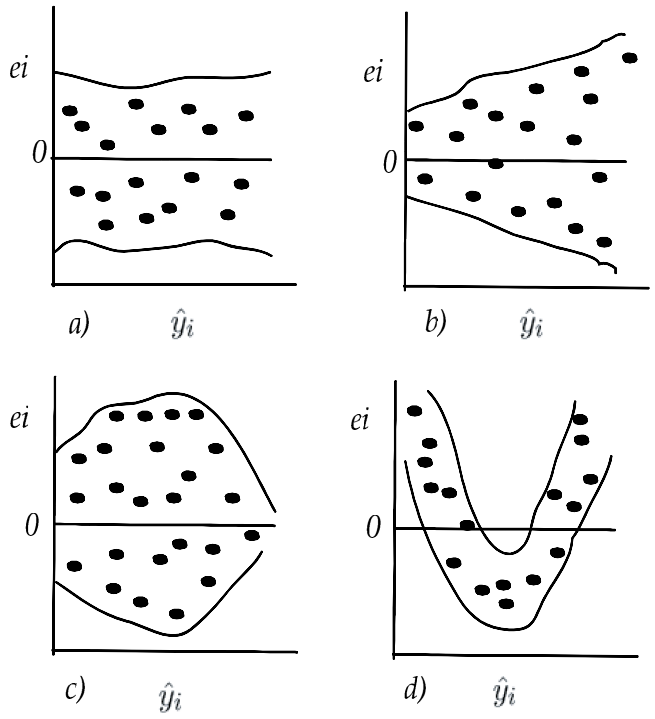


Figura 3.2: Patrones de gráfica de los residuales

Las pruebas que se pueden hacer para comprobar heterocedasticidad son: Prueba Bartlet, Prueba de White.

- Para poder observar si los residuos son no correlacionados se realiza una gráfica de los residuos en secuencia temporal para poder darnos una idea si los errores en un periodo se correlacionan con los de otros periodos.Si este supuesto no se cumple, los estimadores pierden la eficiencia, también existe la posibilidad de que se sobreestime el R^2 y de que las pruebas t y F dejen de validas, si se aplican, es probable que conduzcan a conclusiones erróneas. Se debe tener cuidado al realizar esté tipo de gráficas ya que cuando las observaciones tienen cierto orden en particular, por ejemplo, si los datos fuerón tomados en el tiempo. Si esto no ocurre, entonces se pueden obtener gráficas diferentes para diferentes ordenes.

  La correlación entre los errores del modelo en distintos periodos se llama autocorrelación. Una gráfica como la de la figura 3.3a indica una autocorrelación positiva, mientras que la figura 3.3b es característica de una autocorrelación negativa.

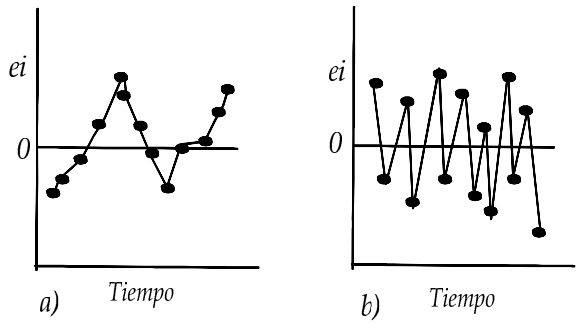


Figura 3.3: Gráficas prototipos que muestran autocorrelación en los errores.

La prueba Durbin-Watson ayuda a detectar problemas de autocorrelación o se pueden usar pruebas no paramétricas de aleatoriedad como la prueba de signos.

Si se escoge un modelo global, es necesario agregar dos supuestos:

- Todos los datos son expresados en las mismas unidades. Si diferentes conjuntos de datos están expresados en unidades diferentes darían a diferentes conjuntos de datos diferentes pesos.
- La dispersión es la misma para cada conjunto de datos. Para cada valor de X, para cada conjunto de valores la dispersión debería ser la misma (ajuste sin ponderar) o deberían variar de manera predecible con Y.

3.3. Los resultados de la regresión no lineal

3.3.1. Bandas de confianza y predicción

La gráfica de la mejor curva ajustada puede incluir la banda del 95 % de confianza de la mejor curva ajustada, o la banda del 95 % de predicción. Las dos son muy diferentes. La banda de confianza dice qué tan bien se conoce la curva. Se puede estar el 95 % seguro que la mejor curva ajustada está dentro de las bandas de confianza. La banda de predicción muestra la dispersión de los datos. Si se recolectan más puntos podrías esperar que el 95 % de ellos cayeron dentro de la banda de predicción.

Como la banda de predicción tiene que considerar la incertidumbre de la curva en sí, así como la dispersión alrededor de la curva, la banda de predicción es mucho más amplia que la banda de confianza. Cuando se incremente el número de puntos, las bandas de confianza se

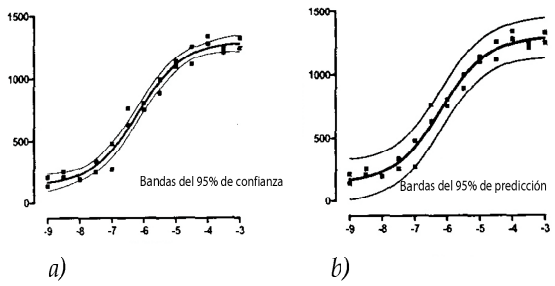


Figura 3.4: Gráficas de confianza y predicción.

van acercando más y más a la mejor curva ajustada, mientras que las bandas de predicción no cambian de manera predecible. En el ejemplo se muestra que las bandas de confianza (gráfica 3.4a) contienen una minoría de los datos (que es lo esperado). Las bandas de confianza tienen el 95 % de probabilidad de contener la mejor curva ajustada y aún con tantos datos, estás bandas contienen mucho menos de la mitad de los datos. En contraste, la banda de predicción (gráfica 3.4b) incluye el 95 % de los datos.

3.3.2. La Matriz de Correlación

Si los errores estándar son grandes (y por tanto los intervalos de confianza anchos) es necesario hacer una investigación más profunda. Una posibilidad es que el modelo sea redundante, entonces hay al menos dos parámetros que están relacionados. Algunos programas reportan una matriz de correlación para ayudar a diagnosticar este problema.

Para cualquier par de parámetros, la matriz de correlación reporta un valor que muestra qué tan vínculados están los parámetros. Este valor es como un coeficiente de correlación que toma valores entre -1 y +1.

El programa encuentra los mejores valores ajustados de cada parámetro. Esto significa que si se cambia el valor de cualquier parámetro (sin cambiar el resto), la suma de los cuadrados crecerá (el ajuste será peor). ¿Pero qué pasa si cambias el valor de un parámetro y lo fijas y luego le pides al programa encontrar un nuevo valor mejor ajustado? Un extremo es cuando los parámetros están completamente desligados. Cambiado un parámetro el ajuste es peor y no puedes compensar nada cambiando los otros, en este caso extremo el valor reportado por la matriz de correlación sería cero. El otro extremo es cuando los dos parámetros están completamente ligados. Cambiando un parámetro hace un ajuste peor, pero esto puede ser exactamente

compensado por el cambio de los otros. El valor reportado por la matriz de correlación sería +1 (si se compensa por un incremento en un parámetro incrementando el otro) o -1 (si se compensa por un incremento en un parámetro disminuyendo el otro).

Capítulo 4

Aplicación

4.1. Relación Plomo y Coeficiente Intelectual

La contaminación por plomo es un problema detectado hace décadas, primero en el ambiente laboral y posteriormente en el ambiente de sectores urbanos o rurales cercanos a fundiciones, mineras u otras fuentes de emisión. El plomo se encuentra presente en la corteza terrestre en forma natural y se produce primariamente por fundición del mineral. Se ha utilizado en la fabricación de baterías, pigmentos para pinturas, cerámica vidriada, recubrimiento de cables y como antidetonante de la gasolina.

El plomo es un elemento muy tóxico para el ser humano. Estudios realizados en población infantil han demostrado que los daños pueden ocurrir con la presencia de pequeñas cantidades en sangre debido a ciertas condiciones especiales: menor masa corporal, sistema nervioso en desarrollo, mayor tasa de absorción intestinal de plomo y menor tasa de eliminación, proximidad al suelo y tendencia de llevar objetos y tierra a la boca.

El conocimiento acerca de la toxicidad de la exposción crónica al plomo en dosis bajas en niños ha ido creciendo. Es evidente el daño en globulos rojos y sus precursores, causando anemia en riñones, en sistema nervioso central y periférico. También se ha visto una asociación entre el nivel de plomo en la sangre y el coeficiente intelectual y otros indicadores del desarrollo neuropsicológico de los niños expuestos.

A pesar de que el plomo en el ambiente ha ido disminuyendo en México, en el momento del embarazo se libera el plomo de los depósitos del útero de la madre y eso ocasiona que exista en el niño.

4.2. Análisis de Regresión Lineal Simple

Los datos provienen de un estudio real y se está tomando un ejemplo del modelo simulado al que ya le han sido ajustadas otras variables explicativas; sólo se busca la relación con el plomo en la sangre en el tercer trimestre del embarazo (variable explicativa) y el coeficiente intelectual (variable dependiente), que abreviamos IQ por sus siglas en inglés. Para ajustar el modelo se utiliza el programa Statistica versión 6 con una muestra de tamaño 99.

En la gráfica de dispersión, figura (4.1), se observa que los datos no tienen una relación lineal y al mismo tiempo muestra como la variable IQ decrece conforme el plomo va aumentando, por lo que la correlación es negativa.

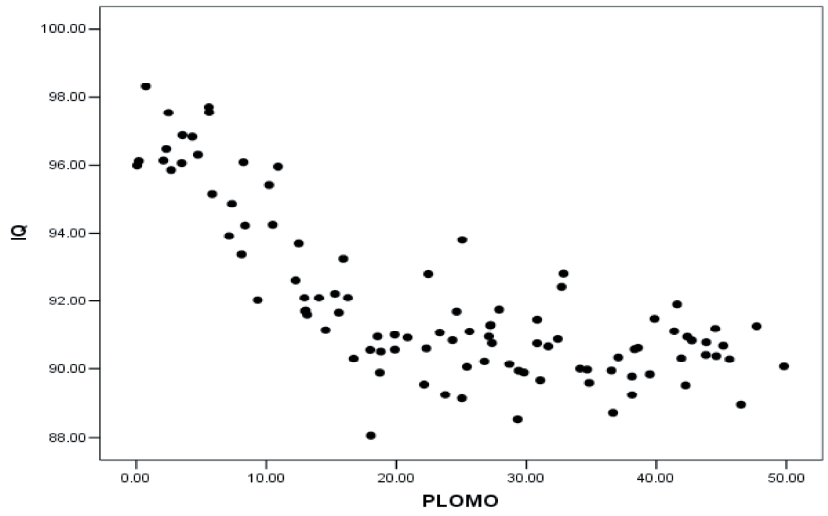


Figura 4.1: Gráfica de dispersión PLOMO vs IQ

Al realizar diversas transformaciones a la variable explicativa se escoge la transformación $\sqrt{plomo}$ que llamaremos $\sqrt{x}$, como aquella que ayuda a linealizar el modelo. La gráfica con la transformación se puede ver en la figura (4.2), en donde se conserva la relación negativa. En la figura (4.1) se presentan los resultados del ajuste de la regresión con mínimos cuadrados ordinarios. Es modelo ajustado queda como:

$$\hat{y} = 97{,}5031 - 1{,}2125 * \sqrt{x} \tag{4.1}$$

esto es que por cada unidad que cambia $\sqrt{x}$, se tendrá una disminución promedio en el IQ de 1.2125 unidades. Los valores de $\sqrt{x}$ se encuentran cercanos a cero, por lo que β_0 debe permanecer en el modelo, indicando que en ausencia de plomo el IQ es de 97,5031 unidades en promedio.

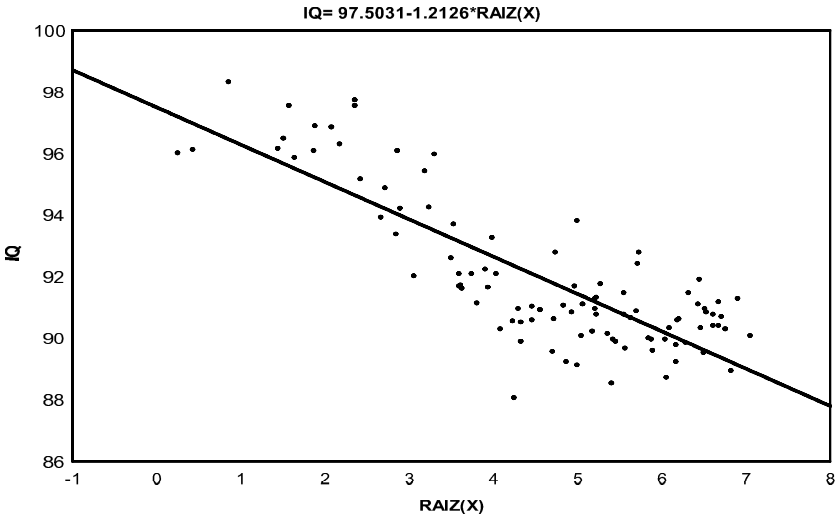


Figura 4.2: IQ VS $\sqrt{plomo}$

Effect	Parameter Estimates (transformafciones) Sigma-restricted parameterization			
	IQ Param.	IQ Std.Err	IQ t	IQ p
Intercept	97.50308	0.417459	233.5634	0.00
raiz(x)	-1.21265	0.085876	-14.1209	0.00

Cuadro 4.1: Estimación de los Coeficientes

En la tabla (4.2) se observa que el coeficiente de correlación es alto, 0,8202 que aparece en valor absoluto y con ayuda de la figura (4.2) se puede observar una relación negativa alta, es decir, que a medida que el plomo aumenta el IQ también aumenta. El coeficiente de determinación es de 0,6727 e indica que aproximadamente un 67,27 % de la variación del IQ es explicado por el plomo.

Dependnt Variable	Test of SS Whole Model vs. SS Resi		
	Multiple R	Multiple R^2	Adjusted R^2
IQ	0.820207	0.672739	0.669365

Cuadro 4.2: Coeficiente de Determinación

Se realiza la prueba F de significancia de la Regresión de reportada en la tabla ANOVA de la figura (4.3) como sigue:
Hipótesis

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_1 : \beta_1 \neq 0$$

Estadístico de Prueba:

	Univariate Results for Each DV (transformafciones Sigma-restricted parameterization Effective hypothesis decomposition				
GENERAL Effect	Degr. of Freedom	SS	MS	F	p
Intercept	1	107763	107763.168	54551.85	0.00
raiz(x)	1	394	393.899	199.40	0.00
Error	97	192	1.975		
Total	98	586			

Cuadro 4.3: Análisis de Varianza

$$F_0 = \frac{MS_R}{MS_{Res}} = \frac{393{,}899}{1{,}975} = 199{,}4425$$

Regla de Decisión:
Rechazar H_0 si

$$199{,}4425 = F_0 > F_{(\alpha \backslash 2,1,n-2)} = F_{(,025,1,97)} = 5{,}183$$

Por lo tanto rechazamos la hipótesis $H_0 : \beta_1 = 0$. Lo que indica que la variable $\sqrt{x}$ explica a la variable IQ.

El valor P-value para esta prueba es 0,00 lo que ayuda a explicar a la prueba F.

Para analizar los residuos comenzaremos con el supuesto de normalidad. En el histograma de los residuos estandarizados de la figura (4.3), se observa que la distribución está más cargada hacia la izquierda, posiblemente la barra dentro del círculosea un punto discrepante, sin embargo el histograma nos da la idea de que los residuales siguen una distribución normal.

En cuanto a la gráfica en papel normal, figura(4.4), los puntos se encuentran prácticamente sobre la diagonal, salvo por los extremos, en especial la observación 22, que en el histograma corresponde al dato en el rectángulo. La prueba Skapiro-Wilks rechaza la normalidad al tener un p-value mayor a 0.5. Por lo tanto se rechaza el supuesto de normalidad.

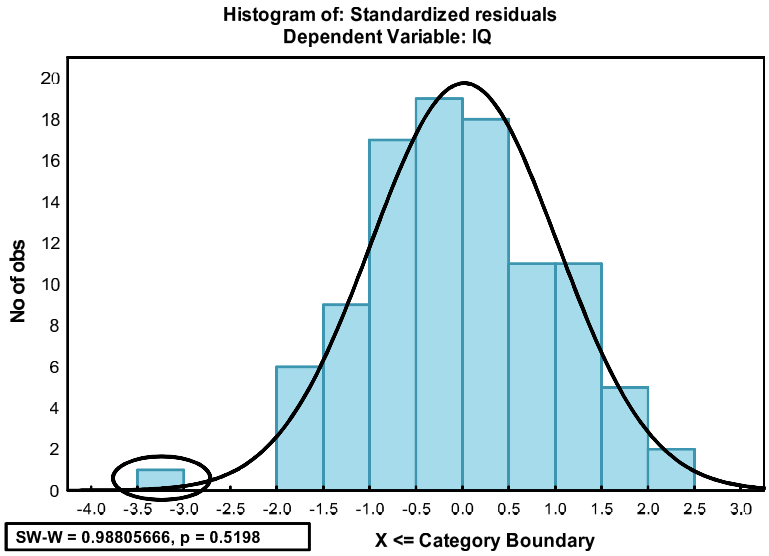


Figura 4.3: Gráfica de probabilidad Normal para residuos estandarizados

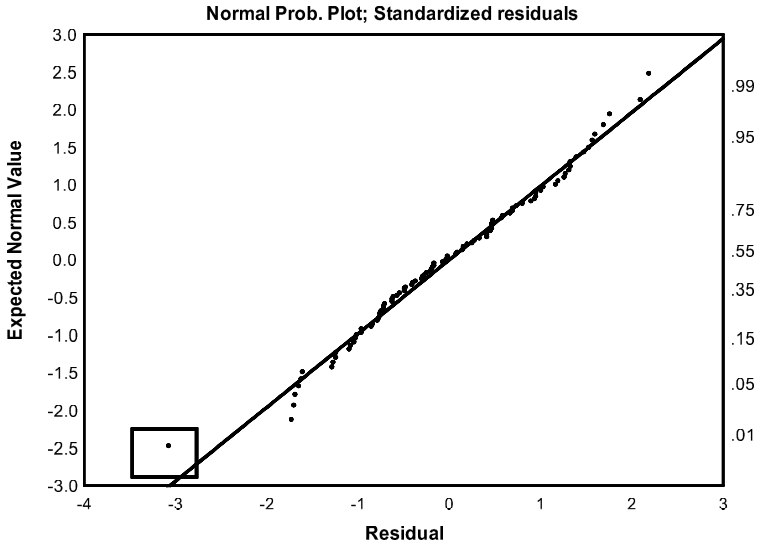


Figura 4.4: Residuales Estándarizados Papel Normal $\sqrt{x}$

Para la varianza observamos la figura (4.5), con respecto al eje y los datos se mantienen en un rango de -2 a 2, a pesar de que algunos datos quedan fuera la mayoría puede encerrarse en un rectángulo, entre los casos 60 y 70 notamos una aglomeración cerca de la recta horizontal, aparentemente el supuesto de análisis de varianza se cumple.

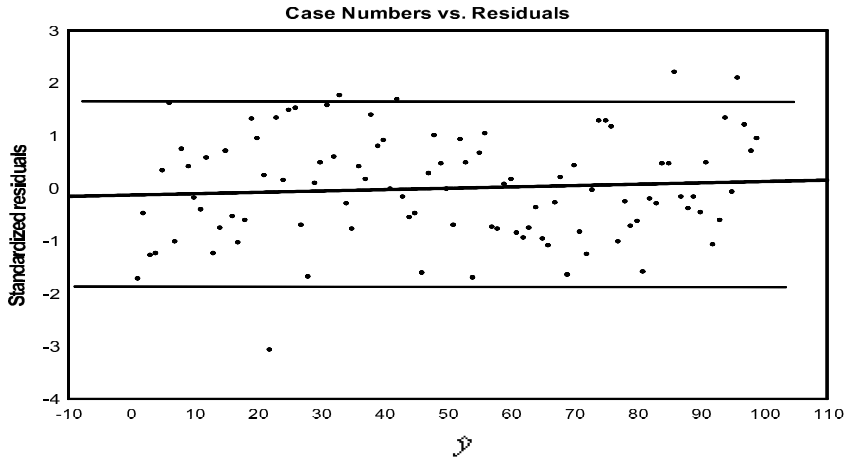


Figura 4.5: Gráfica de valores de y ajustados, $\hat{y}$, vs residuos estandarizados

Para el supuesto de residuos no correlaciónados, figura (4.6), los residuales se distribuyen de manera aleatoria a lo largo de los tiempos de observación, por lo que se puede decir que la no correlación es un supuesto que se acepta.

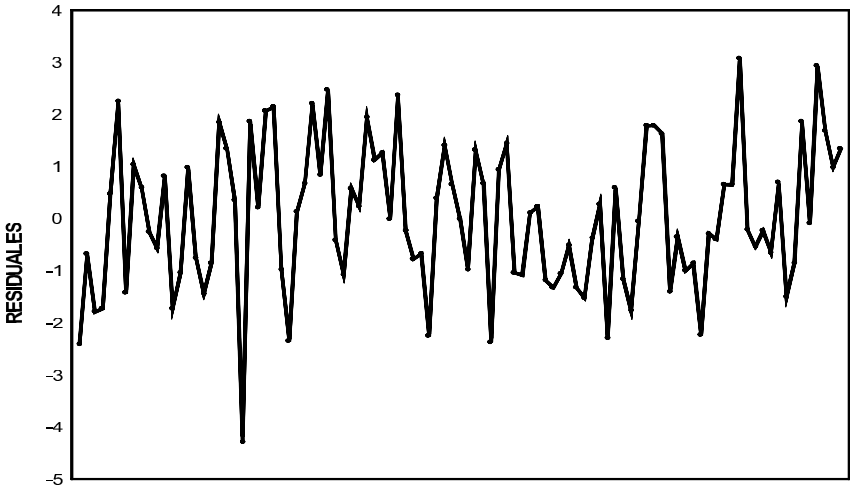


Figura 4.6: Gráfica de Independencia $\sqrt{x}$

Se han encontrado tres puntos discrepantes que se muestran en la figura (4.4), lo que procede

es determinar si estos puntos son influyentes o no, primero correremos la regresión sin cada uno de los valores.

Case	-5.	-4.	Standard Residual IQ (transformafciones) Outliers			
			Observed Value	Predicted Value	Residual	Standard Pred. v.
22	.	.	88.05991	92.35300	-4.29309	0.198059
86	.	.	97.71271	94.63760	3.07510	1.337603
96	.	.	97.55927	94.63239	2.92687	1.335007
Minimum	.	.	88.05991	92.35300	-4.29309	0.198059
Maximum	.	.	97.71271	94.63760	3.07510	1.337603
Mean	.	.	94.44396	93.87433	0.56963	0.956890
Median	.	.	97.55927	94.63239	2.92687	1.335007

Cuadro 4.4: Puntos discrepantes encontrados

Se corre la regresión sin el dato 22, que se sospecha es un Outlier, los resultados se muestran en la tabla (4.5)

Cuadro 4.5: Regresión sin dato discrepante 22

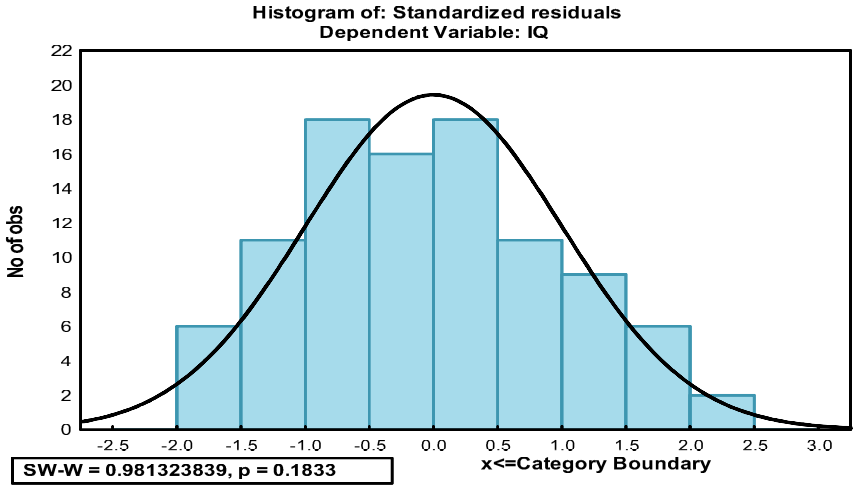
GENERAL Effect	Univariate Results for Each DV (transformafcio Sigma-restricted parameterization Effective hypothesis decomposition				
	Degr. of Freedom	IQ SS	IQ MS	IQ F	IQ p
Intercept	1	107610.1	107610.1	59717.71	0.00
raiz(x)	1	397.2	397.2	220.42	0.00
Error	96	173.0	1.8		
Total	97	570.2			

Effect	Parameter Estimates (transformafci Sigma-restricted parameterization			
	IQ Param.	IQ Std.Err	IQ t	IQ p
Intercept	97.57116	0.399273	244.3721	0.00
raiz(x)	-1.21795	0.082036	-14.8465	0.00

Dependnt Variable	Test of SS Whole Model vs. S		
	Multiple R	Multiple R^2	Adjusted R^2
IQ	0.834629	0.696605	0.693445

La recta de regresión ajustada es $\hat{y} = 97{,}57 - 1{,}21\sqrt{x}$, que es muy parecida al modelo con el dato. El coeficiente de determinación aumenta en 2 centésimas. Al analizar el histograma de los residuos, se observa que no está cargado hacia la izquierda, pero el histograma no se completa da la apariencia de estar truncado en la cola izquierda, también se observa que en la parte central de la normal, el histograma es picudo, así que al quitar el dato 22 se aparentemente la normalidad es aceptada.

Figura 4.7: Histograma sin el dato 22



Las siguientes tablas muestran los resultados del ajuste de la regresión sin la observación 86

Cuadro 4.6: Resultados de la regresión sin el dato 86

	Univariate Results for Each DV (transformafcio Sigma-restricted parameterization Effective hypothesis decomposition				
GENERAL Effect	Degr. of Freedom	IQ SS	IQ MS	IQ F	IQ p
Intercept	1	104631.8	104631.8	55225.59	0.00
raiz(x)	1	370.2	370.2	195.37	0.00
Error	96	181.9	1.9		
Total	97	552.0			

	Parameter Estimates (transformafci Sigma-restricted parameterization			
Effect	IQ Param.	IQ Std.Err	IQ t	IQ p
Intercept	97.35159	0.414260	235.0012	0.00
raiz(x)	-1.18652	0.084888	-13.9775	0.00

	Test of SS Whole Model vs. S		
Dependnt Variable	Multiple R	Multiple R^2	Adjusted R^2
IQ	0.818854	0.670521	0.667089

En este caso se tiene la recta ajustada $\hat{y} = 97{,}35 - 1{,}18\sqrt{x}$ con un R^2 de 67 %. Lo cual tampoco es un cambio significativo a la regresión con el dato incluido.

En la regresión del dato 96 la recta ajustada es $\hat{y} = 97{,}35 - 1{,}18\sqrt{x}$ con un R^2 de 67 %, arroja los mismos datos que para el dato 86.

Cuadro 4.7: Resultados de la regresión excluyendo la observación 96

GENERAL Effect	Univariate Results for Each DV (transformafcio Sigma-restricted parameterization Effective hypothesis decomposition) Degr. of Freedom	IQ SS	IQ MS	IQ F	IQ p
Intercept	1	104656.6	104656.6	54961.74	0.00
raiz(x)	1	371.0	371.0	194.83	0.00
Error	96	182.8	1.9		
Total	97	553.8			

Effect	Parameter Estimates (transformafci Sigma-restricted parameterization) IQ Param.	IQ Std.Err	IQ t	IQ p
Intercept	97.35912	0.415285	234.4392	0.00
raiz(x)	-1.18783	0.085098	-13.9583	0.00

Dependnt Variable	Test of SS Whole Model vs. S Multiple R	Multiple R^2	Adjusted R^2
IQ	0.818483	0.669914	0.666476

Las regresiones correspondientes sin los datos 22, 86 y 96 son muy parecidos al ajuste que los incluye, por lo que se decide conservarlos.

Se ajusta una nueva regresión considerando la transformación $ln(plomo) = ln(x)$, ya que en la literatura se encuentra que el plomo se distribuye de forma lognormal. Con mínimos cuadrados ordinarios se obtiene:

$$\hat{y} = 96{,}8389 - 1{,}7271 * \ln x \tag{4.2}$$

La gráfica (4.8) muestra que la relación es negativa y no precisamente lineal, teniendo una acumulación al final de la recta de regresión. En el rectángulo se muestran dos puntos que son atípicos, ya que están demasiado alejados de los demás datos.

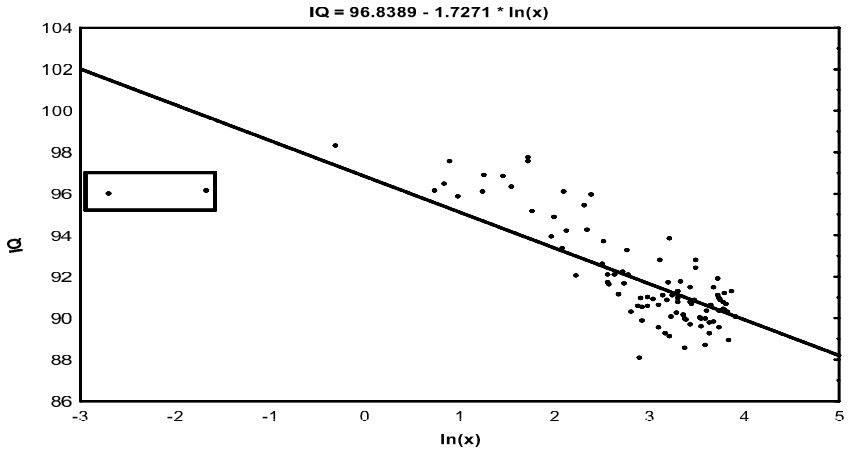


Figura 4.8: Gráfica de dispersión $\ln(x)$ IQ

Se realiza la prueba F de significancia de la Regresión:
Hipótesis

$$H_0 : \beta_1 = 0 \qquad vs \qquad H_1 : \beta_1 \neq 0$$

Cuadro 4.8: Resultados del ajuste de una Regresión Lineal para el modelo $ln(x)$

GENERA Effect	Univariate Results for Each DV (transformafci Sigma-restricted parameterization Effective hypothesis decomposition Degr. of Freedom	IQ SS	IQ MS	IQ F	IQ p
Intercept	1	122825	122825.0	52808.55	0.00
ln(x)	1	360	359.9	154.74	0.00
Error	97	226	2.3		
Total	98	586			

Effect	Parameter Estimates (transformafci Sigma-restricted parameterization IQ Param.	IQ Std.Err	IQ t	IQ p
Intercept	96.8389	0.42140	229.801	0.0
ln(x)	-1.7271	0.13884	-12.440	0.0

Dependnt Variable	Test of SS Whole Model vs. S Multiple R	Multiple R^2	Adjusted R^2
IQ	0.78402	0.61468	0.61071

Estadístico de Prueba:

$$F_0 = \frac{MS_R}{MS_{Res}} = \frac{359{,}9}{2{,}3} = 154{,}74$$

Regla de Decisión:

$$154{,}74 = F_0 > F_{(,025,1,97)} = 5{,}183$$

Por lo tanto rechazamos la hipótesis $H_0 : \beta_1 = 0$. Lo que nos indica que la variable $\ln x$ ayuda a explica a la variable IQ. La prueba t concluye que los parámetros β_0 y β_1 son distintos de cero, por lo que se deben de conservar en el análisis.

El coeficiente de determinación ; R^2, nos indica que el 61.46 % de los datos son explicados por el modelo. Y muy similar a obtenida usando la transformación $\sqrt{x}$ como variable independiente.

En el histograma de los residuos estándarizados, figura (4.9), se observa una distribución muy "picuda", la distribución esta inclinada a la izquierda como en la transformación anterior; sin embargo, gráfica (4.10) hay ciertos datos, encerrados en rectángulos, que no se ajusta a la recta normal. La prueba Shapiro-Wilks acepta la prueba de normalidad.

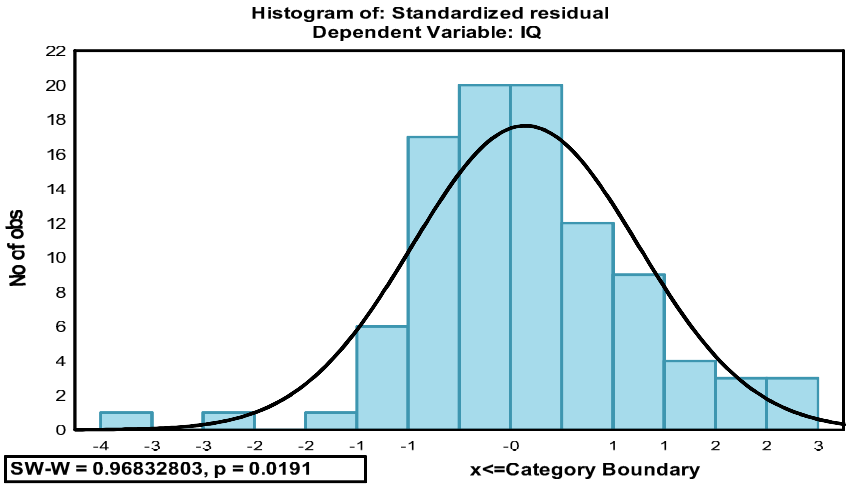


Figura 4.9: Histograma de los residuos estandarizados

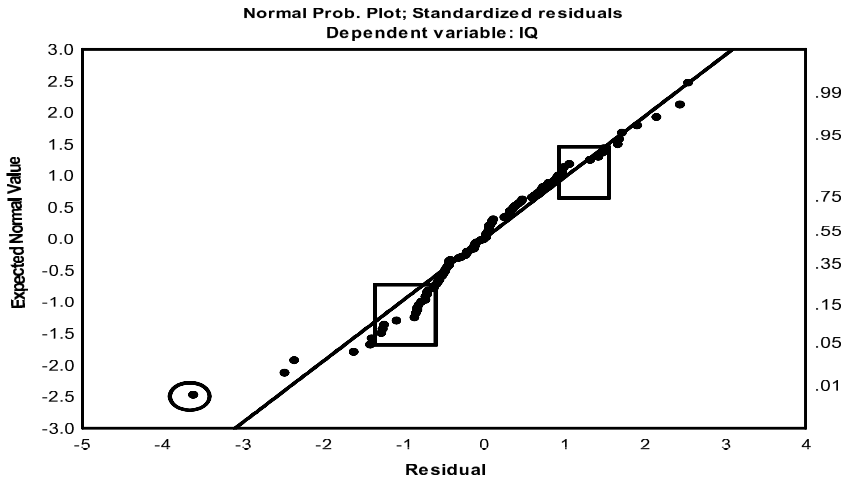


Figura 4.10: Gráfica de probabilidad Normal para residuos estandarizados

Con respecto a la varianza constante; gráfica (4.11), los residuos oscilan entre $-4{,}5$ y $2{,}5$; por debajo de la línea horizontal del cero la dispersión disminuye, en comparación a la gráfica de $\sqrt{x}$, se observan tres datos que no siguen la tendencia de la mayoría, que corresponden a los mismos datos atípicos que en la gráfica de dispersión. Sin embargo, la mayoría de las observaciones se pueden encerrar en un rectángulo, por lo que aparentemente el supuesto de varianza constante es aceptado.

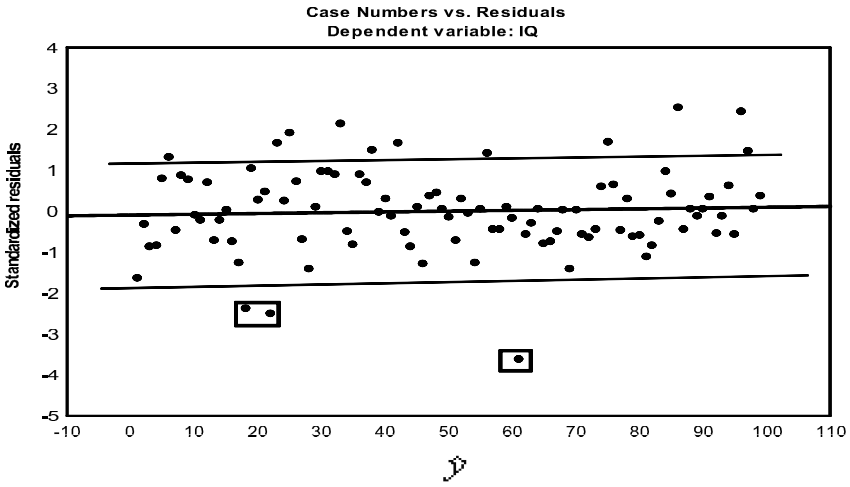


Figura 4.11: Gráfica de valores de y ajustados, $\hat{y}$, vs residuos estandarizados

Finalmente el supuesto de no correlación no es violado de acuerdo a la gráfica (4.12) que no muestra ningún patrón en los datos.

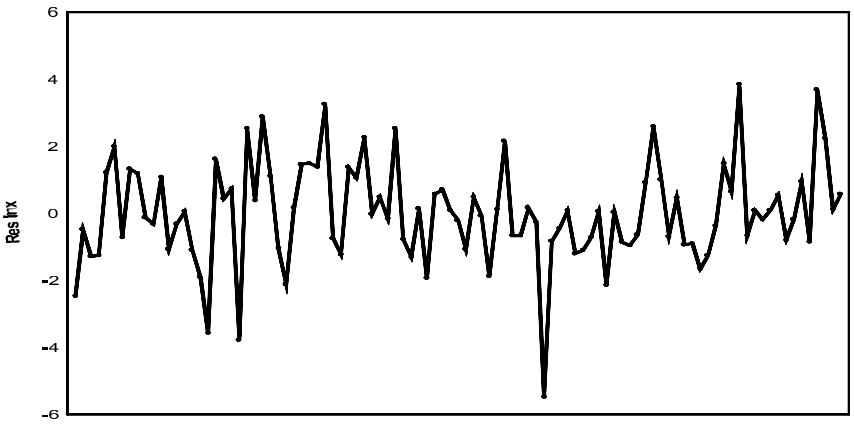


Figura 4.12: Gráfica de independencia $ln(x)$

Se corre la regresión sin los datos 18 y 61. La recta ajustada obtenida es $\hat{y} = 98{,}94275 - 2{,}41312 \ln x$ con un, coeficiente de determinación, R^2, de 72.71 %. Que muy parecido al ajuste que los concluye .

El Análisis de la varianza, tabla (4.9), concluye que la variable independiente explica al plomo.

GENERA Effect	Univariate Results for Each DV (transformafc Sigma-restricted parameterization Effective hypothesis decomposition Degr. of Freedom	IQ SS	IQ MS	IQ F	IQ p
Intercept	1	72656	72656.44	46243.13	0.00
ln(x)	1	402	401.87	255.78	0.00
Error	95	149	1.57		
Total	96	551			

Effect	Parameter Estimates (transformafci Sigma-restricted parameterization IQ Param.	IQ Std.Err	IQ t	IQ p
Intercept	98.94275	0.460109	215.0422	0.00
ln(x)	-2.41312	0.150886	-15.9931	0.00

Dependnt Variable	Test of SS Whole Model vs. S Multiple R	Multiple R^2	Adjusted R^2
IQ	0.853917	0.729173	0.726323

Cuadro 4.9: Análisis de los Residuos excluyendo las observaciones 18 y 61

Se analizan los residuos comenzando con la normalidad

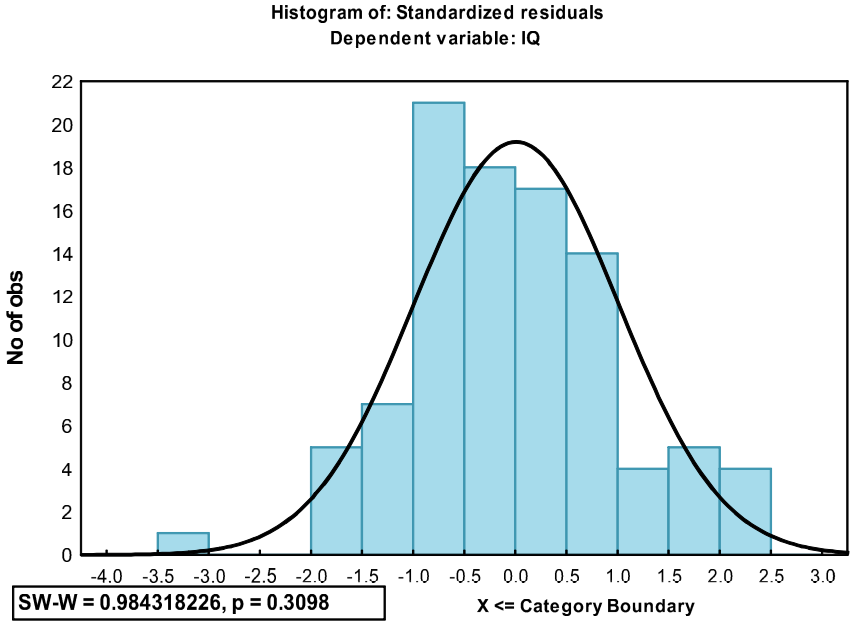


Figura 4.13: Histograma de los Residuos excluyendo las observaciones 18 y 61

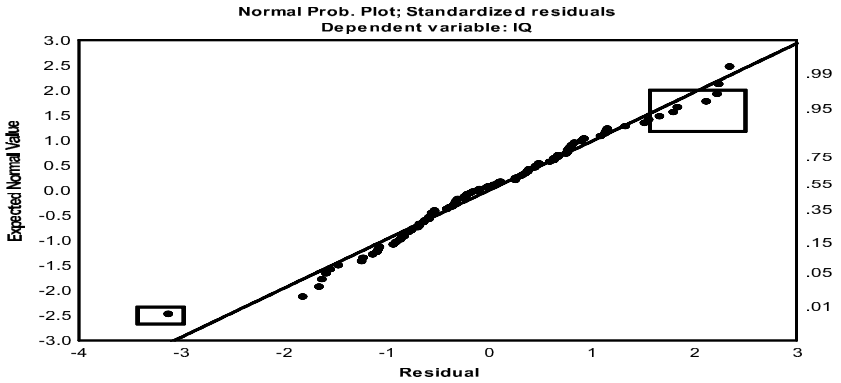


Figura 4.14: Gráfica de probabilidad normal de los Residuos excluyendo las observaciones 18 y 61

se observa que el histograma sigue cargado hacia la izquierda, la gráfica en papel normal mantiene las separaciones originales, pero en esta ocasión que siguen siendo muy pronunciadas. La prueba no parámetrica Shapiro-Wilks rechaza la normalidad. Por lo tanto el supuesto es rechazado.

Ahora se ve el comportamiento de la varianza:

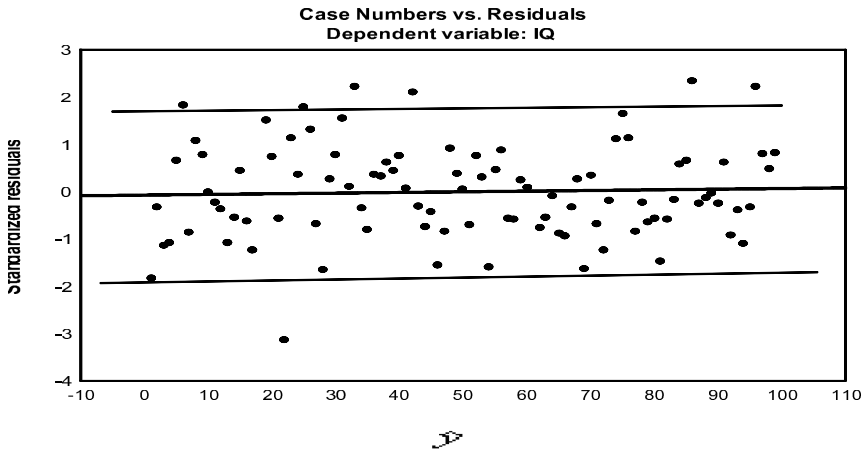


Figura 4.15: Grafica de Varianza de los Residuales excluyendo las observaciones 18 y 61

La aglomeración de los datos entre entre 60 y 70 continúa, aunque aún se puede encerrar en su mayoría los datos en un rectángulo, escapandose pocos puntos.

Por último observamos la independencia

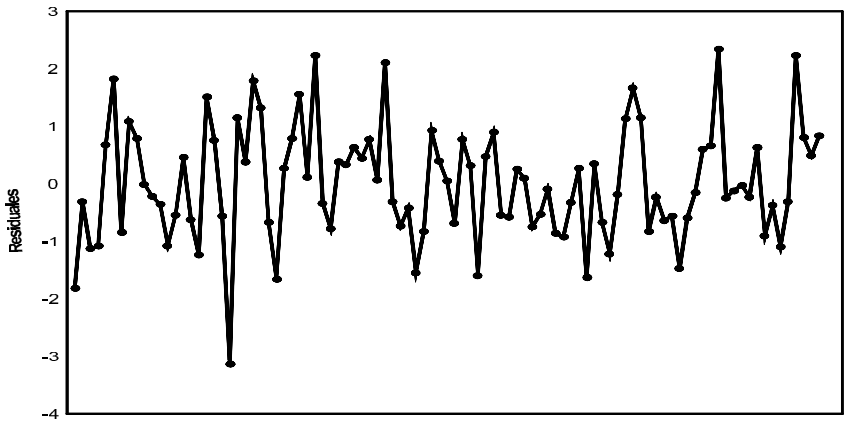


Figura 4.16: Independencia de los Residuales excluyendo las observaciones 18 y 61

No se observa un patrón en los datos, por lo que hay evidencia para pensar que los residuos no están correlacionados.

Como se hemos observado los supuestos de residuales mejorarón y el R^2 mejoro en 1 centécima. Por lo que obtamos por la transformación sin los datos 18 y 61, quedandonos la recta ajustada

$$\hat{y} = 98{,}94275 - 2{,}41312 * \ln x$$

4.3. Análisis de Regresión no Lineal

Se utiliza el método de estimación Gauss-Newton. Los datos son los mismos que en el caso anterior.

El modelo que se utiliza es logístico y esta dado por la siguiente expresión:

$$IQ = \theta_0 + \theta_1 * \frac{1}{1 + exp(-\theta_2 * (plomo - \theta_3))} \tag{4.3}$$

En la gráfica (4.17) se puede observar una relación negativa con respecto a las dos variables, y una relación no lineal ya que los datos no siguen la relación de su tendencia.

Los valores iniciales que se toman son : $\theta_0 = 70$, $\theta_1 = 80$, $\theta_2 = 25$ y $\theta_3 = 13$. El algoritmo Gauss-Newton converge hacia $\hat{\theta} = [97{,}01379, -6{,}48832, 0{,}35141, 10{,}43034]$. Por consiguiente, el modelo ajustado obtenido por linealización es:

$$IQ = 97{,}01379 - 6{,}48832 * \frac{1}{1 + exp(-0{,}35141 * (x - 10{,}43034))} \tag{4.4}$$

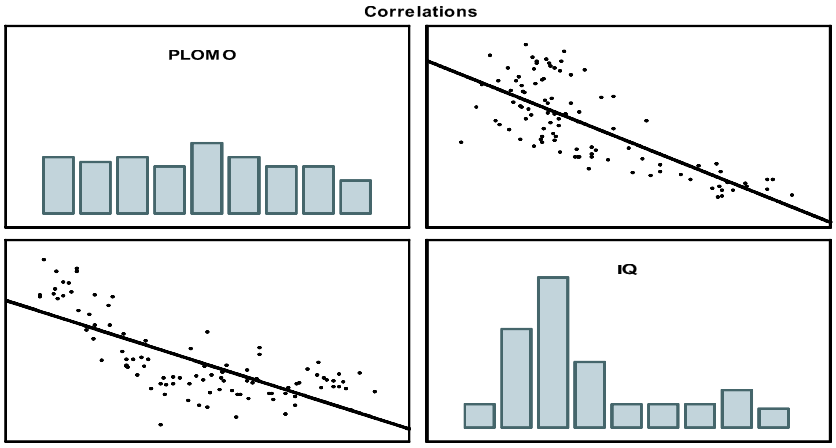


Figura 4.17: Gáfica de correlación entre las variables IQ vs Plomo

La tabla (4.18), muestra las 28 iteraciones que se realizaron para minimizar la función $S(\theta)$. Los estimadores cumplen con el criterio de convergencia que se mencionó en la sección (2.4.2).

	Loss Function	b0	b1	b2	b3
1	517.0184	70.00000	80.00000	25.00000	13.00000
2	12.4675	95.57263	-4.89666	27.27975	13.00328
3	12.3527	95.46678	-4.80350	23.20235	12.99724
4	12.2196	95.47490	-4.81063	16.25761	12.98769
5	12.1111	95.50421	-4.84721	8.67921	12.97447
6	12.0140	95.52336	-4.87632	5.00047	12.95766
7	11.5611	95.56132	-4.92939	1.18775	12.95473
8	11.2316	95.58497	-4.98728	0.58334	12.95470
9	11.1408	95.67953	-5.14321	0.50318	12.95467
10	11.0711	95.81537	-5.32067	0.47604	12.95460
11	11.0364	95.94063	-5.47951	0.44999	12.95445
12	11.0279	96.01552	-5.57446	0.43720	12.95414
13	11.0268	96.04178	-5.60766	0.43322	12.95351
14	11.0261	96.04684	-5.61385	0.43265	12.95223
15	11.0248	96.04799	-5.61496	0.43258	12.94969
16	11.0221	96.04951	-5.61623	0.43247	12.94461
17	11.0168	96.05252	-5.61872	0.43227	12.93450
18	11.0064	96.05848	-5.62365	0.43186	12.91449
19	10.9862	96.07022	-5.63339	0.43104	12.87523
20	10.9484	96.09300	-5.65235	0.42942	12.79972
21	10.8816	96.13593	-5.68834	0.42625	12.65968
22	10.7769	96.21262	-5.75345	0.42026	12.41708
23	10.6424	96.33735	-5.86157	0.40977	12.04319
24	10.5163	96.51213	-6.01750	0.39414	11.56180
25	10.4111	96.91398	-6.39071	0.34716	10.57447
26	10.4049	97.00630	-6.47976	0.35280	10.43668
27	10.4049	97.01452	-6.48923	0.35120	10.42995
28	10.4049	97.01379	-6.48832	0.35141	10.43034

Figura 4.18: Iteraciones del Método Gauss-Newton

La figura (4.19) muestra cómo la curva ajustada sigue la tendencia de los datos.

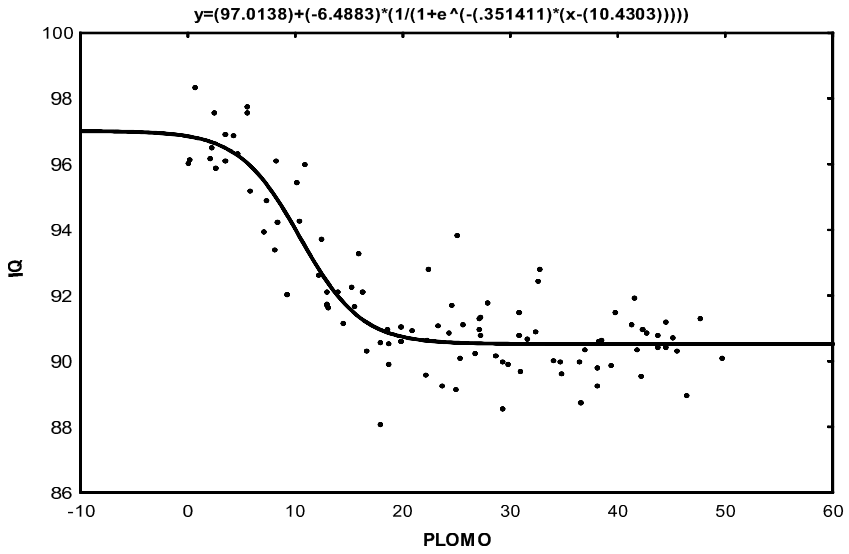


Figura 4.19: Análisis de Regresión No Lineal

Se realiza la prueba t para cada θ de acuerdo con la sección 2.8.1
Hipótesis

$$H_0 : \theta_i = 0 \qquad vs \qquad H_1 : \theta_i \neq 0 \qquad i = 1, 2, 3, 4$$

	Model is: v2=b0+b1*(1/(1+Euler^(-b2*(v1-b3)))) (trans Dep. Var.: **IQ** Level of confidence: 95.0% (alpha=0.050)					
	Estimate	Standard error	t-value df = 95	p-leve	Lo. Conf Limit	Up. Conf Limit
b0	97.01	0.548	177.0	0.000	95.93	98.10
b1	-6.49	0.593	-10.9	0.000	-7.67	-5.31
b2	0.35	0.081	4.4	0.000	0.19	0.51
b3	10.43	0.854	12.2	0.000	8.74	12.13

Cuadro 4.10: Prueba de Hipótesis para una sola θ

Regla de Decisión: Se rechaza H_0 si $|T_i| > t_{(1-\alpha/2;n-p)} = t_{(0,975,95)} = 1{,}985251$ y comparamos

$$10.9 > 1.985251$$
$$4.4 > 1.985251$$

$$12.2 > 1.985251$$

Por lo tanto, rechazamos la hipótesis de que los parámetros son igual a cero, lo que nos dice que todos los parámetros son significativos.

Oservamos que ningún intervalo tiene al cero incluido, que es otra forma de corraborar que los parámetros difieren de cero.

También se realiza la prueba F de la ANOVA, para varias θ de acuerdo a la sección 2.8.2

$$H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_1 = 0 \qquad vs \qquad H_1 : \ al \ menos \ una \ \theta_i \neq 0$$

Regla de Decisión: Se rechaza H_0 si

Effect	Model is: v2=b0+b1*(1/(1+Euler^(-b2*(v1-b3)))) (t Dep. Var. : **IQ** 1 Sum of Sqare	2 DF	3 Mean Squares	4 F-value	5 p-value
Regression	837610.6	4	209402.6	183752.5	0.00
Residual	108.3	95	1.1		
Total	837718.8	99			

Cuadro 4.11: Análisis de Regresión No Lineal

$$183752{,}5 = \frac{209402{,}6}{1{,}1} = F_0 > F_{(\alpha \backslash 2, k, n-k)} = F_{(,025,4,95)} = 2{,}923656$$

Por lo tanto se rechaza H_0. En la figura (4.20), se observa que los parámetros están fuertemente correlacionados esto podrá afectar los supestos de los residuos.

	Model is: v2=b0+b1*(1/(1+Euler^(-b2*(v1-b3) Dep. Var. : **IQ** b0	b1	b2	b3
b0	1.000000	-0.970326	-0.738102	-0.796890
b1	-0.970326	1.000000	0.770038	0.720500
b2	-0.738102	0.770038	1.000000	0.561768
b3	-0.796890	0.720500	0.561768	1.000000

Figura 4.20: Matriz de correlación de los parámetros

De acuerdo con la tabla ANOVA el error total que se puede tener entre los datos de la variable IQ es de 837718.8 unidades y la regresión explica 837610.6 unidades, dejando no explicado una cantidad pequeña de 108.3 unidades. Se calcula el $R^2 = 0{,}999870767$ indicando que el porcentaje de variabilidad explicada por el modelo es de 99,98 % mostrando que la curva estimada encontrada es un buen modelo.

Al revisar los supuestos de normalidad se observa una pequeña separación que se desprende de la recta normal;figura (4.21), pero en general los puntos no se desprenden de la recta por lo que se considera que el supuesto no es violado.

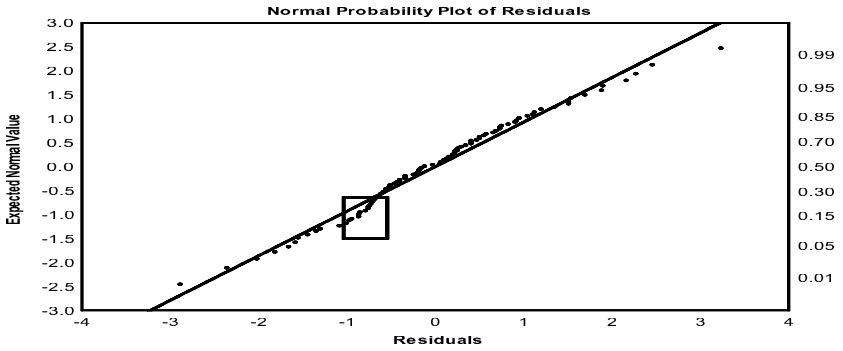


Figura 4.21: Residuales en Papel normal

Para probar el supuesto de varianza constante ; figura (4.22), no se puede encerrar a todos los puntos en un recrángulo, es probable que la varianza constante sea violada ya que aparenta disminuir conforme x aumenta.

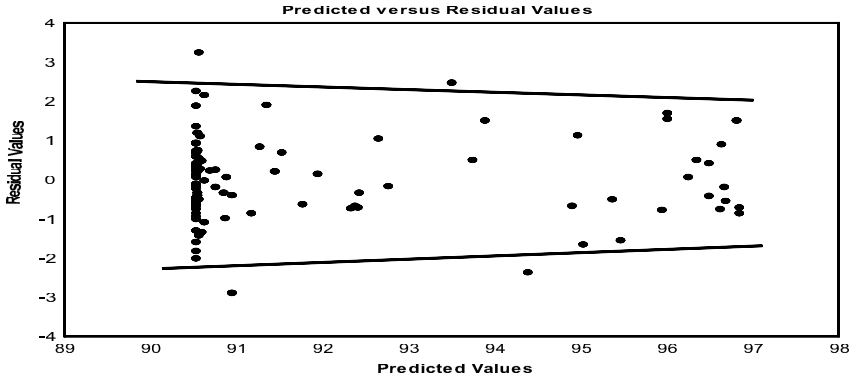


Figura 4.22: Dispersión de la desviación stándar de los residuales

En la gráfica (4.23), se observa que los residuos se distribuyen de manera aleatoria a lo largo de los tiempos de observación, por lo que se puede decir que la independencia es un supuesto cumplido.

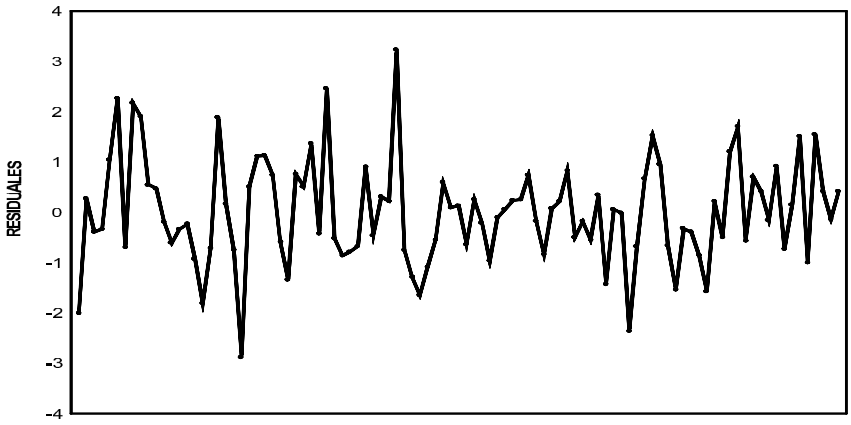


Figura 4.23: Independencia de los Residuales

El programa no encuentra valores discrepantes en este modelo, y no se observan en la gráfica de dispersión.

4.4. Conclusiones

De las dos transformaciones lineales que ser realizarón, se encuentran los siguientes puntos:

1. La gráfica de dispersión de la transformación $\sqrt{plomo}$ se apega más a la recta ajustada.
2. El coeficiente de determinación es menor en la transformación $\sqrt{plomo}$ que en $ln(plomo)$
3. Todos los supuestos de los residuales mejoran con la transformación $ln(plomo)$

En la literatura comúnmente se encuentra la transformación de $ln(plomo)$ por lo que sería útil comparar resultados. Por lo tanto se escoge a la transformación $ln(plomo)$.

De los tres modelos presentados se prefiere el modelo no lineal ya que sin ninguna transformación a las variables la curva ajustada sigue la tendencia de los datos, los valores encontrados de los parámetros no muestran ninguna irregularidad que pueda afectar al modelo por lo que se consideran plausibles, además no se tienen valores extremos. La curva se mantiene en niveles

de credibilidad, por ejemplo, en ausencia de plomo se tiene que el IQ está por las 98 unidades, y nunca se toma en cuenta una ausencia total de IQ, como en las transformaciones lineales, aquí se vuelve asintóta la curva en las 90 unidades. Por lo que no se puede concluir que con cantidades grandes de plomo el IQ es cero como en las transformaciones lineales. Lo cual es una incongruencia.

En cuanto a la correlación de los parámetros no se considera importante, ya que no afectó a ninguno de los supestos de los errores. El R^2 es el más alto de los tres modelos, prácticamente explica todo el modelo. Y se conservan las unidades iniciales sin perder información.

Se concluye que tiene más ventajas un modelo de regresión no lineal ya que ajustan mejor los datos por lo que una predicción futura tendrá mayor credibilidad, que al final es a lo que se pretende llegar. Tambien se observa la ventaja que los resultados se dan en las unidades originales ya que se puede ajustar los datos sin llegar muchas veces a una transformación.

Anexo A

Algunos resultados de álgebra matricial

A.1 $Rango(\mathbf{A})$=$R(\mathbf{A}')$=$R(\mathbf{A}'\mathbf{A})$=$R(\mathbf{A}\mathbf{A}')$

A.2 Si $\mathbf{X}$ y $\mathbf{Y}$ son $m \times 1$ y $n \times 1$ vectores de variables aleatorias y $\mathbf{A}$ y $\mathbf{B}$ son $l \times m$ y $p \times n$ matrices constantes, respectivamente, entonces:

$$Cov[\mathbf{AX}, \mathbf{BY}] = \mathbf{A}Cov[\mathbf{X}, \mathbf{Y}]\mathbf{B}'$$

De donde

$$\begin{aligned} Var[\mathbf{AX}] &= Cov[\mathbf{AX}, \mathbf{AX}] \\ &= \mathbf{A}Cov[\mathbf{X}, \mathbf{X}]\mathbf{A}' \\ &= \mathbf{A}Var[\mathbf{X}]\mathbf{A}'. \end{aligned}$$

A.3 Si $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ y E($\boldsymbol{\epsilon}$)=$\mathbf{0}$ entonces:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ E[\hat{\boldsymbol{\beta}}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{A.1}$$

Por lo que $\hat{\boldsymbol{\beta}}$ es un estimador insesgado para $\boldsymbol{\beta}$. Si además asumimos que los ϵ_i no están correlacionados y tienen la misma varianza σ^2, tenemos:

$$Var[\mathbf{Y}] = Var[\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}]$$

Y utilizando [A.2]

$$\begin{aligned} Var[\hat{\boldsymbol{\beta}}] &= Var[(\mathbf{X'X})^{-1}\mathbf{X'Y}] \\ &= (\mathbf{X'X})^{-1}\mathbf{X'}Var[\mathbf{Y}]\mathbf{X}(\mathbf{X'X})^{-1} \\ &= \sigma^2(\mathbf{X'X})^{-1}(\mathbf{X'X})(\mathbf{X'X})^{-1} \\ &= \sigma^2(\mathbf{X'X})^{-1} \end{aligned} \tag{A.2}$$

A.4 Si $\mathbf{Y} \sim N(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ y $\mathbf{C}$ es una matriz $p \times n$ de rango p, entonces $\mathbf{CY} \sim N(\boldsymbol{C\theta}, \boldsymbol{C\Sigma C'})$.

A.5 Si $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ es un vector de variables alatorias con función de densidad:

$$f(y_1, y_2, \ldots, y_n) = k^{-1} exp\Big[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\theta}) \Big]. \tag{A.3}$$

entonces:

(i) $k = (2\pi)^{(1/2)n}|\boldsymbol{\Sigma}|^{-1/2}$.

(ii) E[$\mathbf{Y}$|=$\boldsymbol{\theta}$ y Var[$\mathbf{Y}$|=$\boldsymbol{\Sigma}$.

(iii) $Q = (\mathbf{y} - \boldsymbol{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \boldsymbol{\theta}) \sim \chi^2_{(n)}$

A.6 Tenemos $\hat{\boldsymbol{\theta}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'Y} = \mathbf{PY}$ está transformación lineal representa la proyección ortogonal en el espacio Euclidiano n-dimensional,E_n, sobre Ω. Similarmente $\mathbf{I_n - P}$ representa la proyección ortogonal de E_n, sobre el complemento ortogonal de $\Omega^{\perp}$, de Ω. Esto es $\mathbf{Y} = \mathbf{PY} + (\mathbf{I_n - P})\mathbf{Y}$ representa una única descomposición de $\mathbf{Y}$ en dos componentes, una en Ω y la otra en $\Omega^{\perp}$. Algunas propiedades basicas de $\mathbf{P}$ y $(\mathbf{I_n - P})$ son:

(i) $\mathbf{P}$ y son simétricas e idempotentes.

(ii) $R(\mathbf{I_n - P})$=$tr[\mathbf{I_n - P}]$=$n - p$.

(iii) $(\mathbf{I_n - P})\mathbf{X} = \mathbf{0}$

A.7 Suponemos que $\mathbf{Y} \sim N_n(\boldsymbol{\theta}, \sigma^2\mathbf{I_n})$ y sea $\mathbf{U}$=$\mathbf{AY}$ y $\mathbf{V}$=$\mathbf{BY}$. Sea $\mathbf{A_1}$ el representante de las filas linealmente independientes de $\mathbf{A}$ y sea $\mathbf{U_1} = \mathbf{A_1Y}$. Entonces si $Cov[\mathbf{U}, \mathbf{V}] = \mathbf{0}$, tenemos:

(i) $\mathbf{U_1}$ es independiente de $\mathbf{V}$'$\mathbf{V}$.

(ii) $\mathbf{U}'\mathbf{U}$ y $\mathbf{V}'\mathbf{V}$ son independientes.

A.8 Las suposiciones normales sobre ϵ_i son que $E[\epsilon] = 0$ y $Var[\epsilon] = \sigma^2\mathbf{I}$, si además agregamos que se distribuyen normal, tenemos $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$ y de aquí que $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, donde $\mathbf{X}$ es una matriz n×p de rango p, tenemos los siguentes resultados:

(i) $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.

(ii) $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_p$.

(iii) $\boldsymbol{\beta}$ es independiente de S^2.

(iv) $(n-p)S^2/\sigma^2 \sim \chi^2_{(n-p)}$

Dem: (i) $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{C}\mathbf{Y}$, donde C es una matriz $p \times n$ tal que el *Rango*(**C**)=*Rango*(**X'**)=*Rango*(**X**)=p (por [A.1]), utilizando [A.4] $\hat{\boldsymbol{\beta}}$ tiene una distribución normal multivariada y de acuerdo a las ecuaciones (A.1) y (A.2) tenemos que $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$.
(ii) $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(Var[\hat{\boldsymbol{\beta}}])^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ por (i) y por [A.5 (iii)], se distribuye $\chi^2_{(n-p)}$.
(iii) Utilizando [A.6]

$$\begin{aligned} Cov[\hat{\boldsymbol{\beta}}, \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}] &= Cov[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, (\mathbf{I_n} - \mathbf{P})\mathbf{Y}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Var[\mathbf{Y}](\mathbf{I_n} - \mathbf{P})' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I_n} - \mathbf{P})' \\ &= \mathbf{0} \end{aligned}$$

Si $\mathbf{U_1} = \hat{\boldsymbol{\beta}}$ y $\mathbf{V} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ por [A.7], $\hat{\boldsymbol{\beta}}$ es independiente de $(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, por consiguiente de S^2.
(iv)

$$\begin{aligned} Q_1 &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})) \\ &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + 2(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= Q + Q_2 \end{aligned}$$

Observemos que:

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}) = 0$$

Ahora $Q_1/\sigma^2 = (\Sigma_i \epsilon_i/\sigma^2)$ es χ^2_n, y $Q_1/\sigma^2 \sim \chi^2_p$, Q es independiente de Q_2 se sigue que $Q/\sigma^2 \sim \chi^2_{n-p}$.

A.9 La suma de cuadrados de la regresión $\mathbf{SS_R}$
Por definición,

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

Se observa que $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, y que

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\mathbf{1}'\mathbf{y}$$

donde $\mathbf{1}$ es un vector de $n \times 1$, con todos sus elementos iguales a uno. Además, $n = \mathbf{1}'\mathbf{1}$, y en consecuencia $\bar{y} = (\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y}$. Por conseguiente, se puede escribir SS_R en la forma

$$\begin{aligned} SS_R &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = [\hat{\mathbf{y}} - \mathbf{1}\bar{y}]'[\hat{\mathbf{y}} - \mathbf{1}\bar{y}] \\ &= [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y}|'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'\mathbf{y}| \\ &= \mathbf{y}'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y} \end{aligned}$$

Nótese que $\mathbf{X} = [\mathbf{1}\mathbf{X}_\mathbf{R}|$, siendo $\mathbf{X_R}$ la matriz formada por los valores reales de los regresores, por lo que SS_R es un caso especial de una matriz dividida. Por lo anterior, se puede usar la identidad especial para matrices divididas, para demostrar que

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

Por consiguiente, se puede demostrar que $[X(X'X)^{-1}X' - 1(1'1)^{-1}1']$ es idempotente. De acuerdo con la premisa que $Var(\epsilon) = \sigma^2\mathbf{I}$.

$$\frac{SS_R}{\sigma^2} = \frac{1}{\sigma^2}\mathbf{y}'[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']\mathbf{y}$$

Sigue una distribución χ^2 no central, con parámetro λ de no centralidad, y con grados de libertad iguales al rango de $[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}']$. Como esta matriz es idempotente, su rango es igual a su traza. Se observa que

$$\begin{aligned} Traza[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'] &= traza[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] - traza[\mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'] \\ &= traza[\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] - traza[\mathbf{1}'(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}] \\ &= traza(\mathbf{I}_p) - traza(1) \\ &= p - 1 = k \end{aligned}$$

Suponiendo que el modelo es correcto

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} = [\mathbf{1}\boldsymbol{\beta}_R]\begin{bmatrix} \beta_0 \\ \beta_R \end{bmatrix} = \beta_0\mathbf{1} + \mathbf{X}_R\boldsymbol{\beta}_R$$

Si se define como sigue la matriz

$$\mathbf{X}_c = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1k} - \bar{x}_k \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2k} - \bar{x}_k \\ \vdots & \vdots & \vdots & \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{nk} - \bar{x}_k \end{pmatrix}$$

siendo $\bar{x}_1$ el valor promedio del primer regresor, $\bar{x}_2$ el valor promedio del segundo regresor, y así sucesivamente, se puede demostrar con facilidad que el parámetro de no centralidad se expresa como sigue:

$$\lambda = \frac{1}{\sigma^2}\boldsymbol{\beta}_R[\mathbf{X'}_C\mathbf{X}_C]\boldsymbol{\beta}_R$$

A.10 La suma de cuadrados de los residuales $\mathbf{SS_Res}$

Por definición,

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

se ve que SS_res se puede escribir en la forma siguiente:

$$\begin{aligned} SS_Res &= [\mathbf{y} - \hat{\mathbf{y}}]'[\mathbf{y} - \hat{\mathbf{y}}] \\ &= [\mathbf{y} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'y}]'[\mathbf{y} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'y}] \\ &= \mathbf{y'}[\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}]\mathbf{y} \end{aligned}$$

Se puede demostrar que $[\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}]$ es simétrica e idempotente. En consecuencia

$$\frac{SS_Res}{\sigma^2} = \frac{1}{\sigma^2}\mathbf{y'}[\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}]\mathbf{y}$$

Sigue una distribución χ^2 no central. Los grados de libertad se deben al rango de $[\mathbf{I} - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}]$, que a su vez es su traza, si se hacen los cálculos se obtiene que la traza es $n - p$. Bajo la premisa que el modelo es correcto,

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

Y como resultado

$$\frac{SS_Res}{\sigma^2} \sim \chi^2_{n-p}$$

A.11 La prueba F global o general

Un estadístico F es la razón de dos variables independientes χ^2, cada una dividida entre sus respectivos grados de libertad. Aquí se ha demostrado que tanto $\frac{SS_R}{\sigma^2}$ como $\frac{SS_Res}{\sigma^2}$ siguen una distribución χ^2; lo fundamental es demostrar que son independientes, suponiendo que $Var(\epsilon) = \sigma^2\mathbf{I}$, si

$$[\mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'} - \mathbf{1}(\mathbf{1'1})'\mathbf{1}]\sigma^2\mathbf{I}[\mathbf{I} - \mathbf{X}(\mathbf{X'X})'\mathbf{X}] = \mathbf{0}$$

Se observa que

$$\begin{aligned}
& [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})'\mathbf{1}]\sigma^2\mathbf{I}[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})'\mathbf{X}] \\
= \; & \sigma^2[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})'\mathbf{1}][\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})'\mathbf{X}] \\
= \; & \sigma^2[\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})'\mathbf{1} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' + \mathbf{1}(\mathbf{1}'\mathbf{1})'\mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'] \\
= \; & \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' - \mathbf{1}(\mathbf{1}'\mathbf{1})^{-\mathbf{1}}\mathbf{1}' + \mathbf{1}(\mathbf{1}'\mathbf{1})^{-\mathbf{1}}\mathbf{1}' \\
= \; & \mathbf{0}
\end{aligned}$$

Por consiguiente, SS_R y SS_Res son independientes. A continuación se observa que

$$\frac{SS_R}{k\sigma^2} = \frac{MS_R}{\sigma^2} \qquad y \qquad \frac{SS_Res}{(n-p)\sigma^2} = \frac{MS_Res}{\sigma^2}$$

son variables aleatorias χ^2, cada una dividida entre sus grados de libertad respectivos. Entonces

$$\frac{MS_R}{MS_Res} \sim F'_{k,n-p,\lambda}$$

siendo

$$\lambda = \frac{1}{\sigma^2}\boldsymbol{\beta}'_R\mathbf{X}'_C\mathbf{X}_C\boldsymbol{\beta}_R$$

Bibliografía

Bates, D.M. and D. G. Watts. 1988. Nonlinear regression analysis and its applications *John Wiley*, New York.

Gallant A. Ronald. 1987. Nonlinear Statistical Models. *John Wiley*, New York.

Montomery, D. C and Peck Elizabeth A. and Vining G. Geoffrey. 2001, Introduction to linear regression analysis, *John Wiley*, New York.

Motulsky, Harvey and Christopoulos Arthur. 2004, Fitting models to biological data using linear and nolinear regression *Oxford University Press*, New York.

Ross, H. William. 1987. The geometry of case deletion and the assessment of influence in nonlinear regression . *The Canadian Journal of Statistics*,Vol. 15 No. 2. Page 91.103.

Roy, T. St. Laurent R. Dennis Cook. 1993. Leverage, local influence and curvature in nonlinear regression . *Biometrika*,80, 1, Page 99-106.

Seber, G.A.F and C.J Wild. 1989. Nonlinear regression, *John Wiley*, New York.

Seber, G.A.F. 1977. Linear Regression Analysis, *John Wiley*, New York.