UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

VERIFICACION DEL SUPUESTO DE NORMALIDAD EN EL MODELO DE REGRESION LINEAL

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

P R E S E N T A N :

ARACELI AGUILAR CASTELLANOS

GERARDO RAFAEL KROEPFLY SAURY

1996

TESIS

COMPLETA

VNIVERSIDAD NACIONAL
AVFNOMA DE
MEXICO

M. en C. Virginia Abrín Batule
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis:

VERIFICACION DEL SUPUESTO DE NORMALIDAD EN EL MODELO DE REGRESION LINEAL

realizado por AGUILAR CASTELLANOS ARACELI
KROEPFLY SAURY GERARDO RAFAEL

con número de cuenta 9052038-9 , 9052130-0 pasante de la carrera de

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	DR. FEDERICO O'REILLY TOGNO	
Propietario	DR. JOSE MENDOZA BLANCO	José Rafa Mendoza Blanco
Propietario	MAT. MARGARITA CHAVEZ CANO	
Suplente	ACT. CLAUDIA LARA PEREZ SOTO	
Suplente	ACT. PILAR ALONSO REYES	

Claudia Larrillo G.
Consejo Departamental de Matemáticas

" A DIOS TODOPODEROSO "

ARACELI AGUILAR CASTELLANOS .

" Cuando se es exigente no hay mejor evaluación
que la de uno mismo ."

A MIS PADRES :

Rogelio Aguilar Espejel y **Socorro Castellanos Buenrostro** con mucho cariño y respeto, gracias por toda la confianza y el apoyo que me han brindado.

A MI NOVIO :

Gerardo Kroepfly Saury , que siempre ha estado conmigo apoyandome en todo , además de su gran ayuda y entrega en la realización de este trabajo , para tí con todo mi amor.

A MIS HERMANOS :

Que me dieron su apoyo , **Lizbeth Aguilar Castellanos**, **Erika Aguilar Castellanos**, **Rogelio Aguilar Castellanos** y **Jessica Aguilar Castellanos** , gracias por todo.

A MI HERMANA Y SU FAMILIA :

Marlem Aguilar Castellanos, Esteban Camacho del Monte y **Lisette Camacho Aguilar** , gracias por su apoyo .

A MI MEJOR AMIGO :

Gil Salgado González , gracias por tu gran ayuda .

A TODA MI FAMILIA :

Principalmente a aquellos que se preocuparon por mí y que me brindarón su apoyo .

A MIS MAESTROS :

Por su dedicación y tiempo requerido para mí educación especialmente a **Tomás Fernández Cruz**.y **Beatriz Rodríguez Fernández.**

Y A MIS ESCUELAS :

Próceres de la Revolución e Instituto Marillac , por la formación que me proporcionarón .

GERARDO KROEPFLY SAURY .

" La mayor satisfacción la da el saber que diste el mayor esfuerzo."

A MIS PADRES :

Marcelo Kroepfly Ortega y **Aurelia Esther Saury de Kroepfly** por su ejemplo de responsabilidad y honradez , por su consejo , comprensión y por sus esfuerzos por tratar de darme siempre lo mejor, con todo mi cariño , respeto y gratitud. .

A MI NOVIA :

Aracell Aguilar Castellanos , por su apoyo ,entrega y capacidad reflejada en la realización de este trabajo , con todo mi amor.

A MIS HERMANOS :

Por su apoyo, estímulo y ejemplo , **Act. Marcelo Kroepfly Saury** e **Ing. Diana Laura Kroepfly Saury.**

A TODOS MIS FAMILIARES , AMIGOS Y COMPAÑEROS :

Principalmente a aquellos que de alguna u otra forma me han ayudado y motivado a conseguir esta meta , con mucho cariño.

A MIS ABUELITAS :

Lidia Alvarez Tostado e **Isabel Ortega Aduna** , con mucho cariño.

ESPECIALMENTE A MIS MEJORES AMIGOS :

Mi primo **Antonino Saury Lomelí** y mis hermanos **Carlos Alberto García de León Rico** y **Carlos Manuel Gutiérrez González** .

A MIS MAESTROS :

Especialmente a los que contribuyerón en mi formación estadística y matemática , en particular a **Beatriz Rodríguez Fernández** , a **Enrique Andrade Solís** y a **Tomás Fernández Cruz**.

Y A MIS ESCUELAS :

Colegio del Distrito Federal e **Instituto Centro Unión** , con todo mi agradecimiento .

DEDICATORIA CONJUNTA :

A LA UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO.

A LA FACULTAD DE CIENCIAS .

AL DIRECTOR DE TESIS :

Dr. Federico O' Reilly Togno , un testimonio de gratitud por su amable orientación y su valioso tiempo en la elaboración de este trabajo .

A los sinodales por su intervención en la realización del presente trabajo :

Dr. José Mendoza Blanco , Mat. Margarita Chavez Cano, Act. Claudia Lara Perez Soto y Act. Pilar Alonso Reyes . Muchas Gracias.

I N D I C E

CAPITULO III

METODOS DE COMPROBACION DEL SUPUESTO DE NORMALIDAD

CAPITULO IV

EJEMPLOS

CAPITULO V

SIMULACION

INTRODUCCION

Debido a que actualmente en la literatura existe poca información acerca de pruebas para la verificación del supuesto de normalidad en las perturbaciones estocásticas del modelo de regresión lineal , se realizó este trabajo ; el cual proporciona la información que existe sobre las pruebas más utilizadas y principalmente da a conocer un método sencillo que ayuda a verificar dicho supuesto del modelo de regresión lineal a distintos niveles de significancia y con una base teórica más firme que otras pruebas ya existentes. Dicho método se basa en la Función de Distribución Empírica (FDE) de los residuos.

En el primer capítulo se realiza un resumen del modelo de regresión lineal, en el cual se proporciona en primer lugar la definición del modelo al que se refiere este trabajo. A continuación se efectúa una breve descripción de los elementos teóricos que se consideran en el modelo, tales como los supuestos en los que se basa, sus estimadores, las propiedades de los estimadores mencionados y los elementos necesarios para sustentar a posteriori que el modelo utilizado fué el adecuado para representar el fenómeno de estudio.

En el capítulo II se menciona la importancia de los residuos en el análisis de regresión y se describen los distintos tipos de residuos que existen ; así como también los diferentes tipos de observaciones que pueden ocasionar alteraciones en el modelo. Finalmente, se mencionan algunos tipos de gráficas que se realizan para analizar el comportamiento de los residuos y de tal manera determinar si el modelo que se está utilizando es adecuado.

En el tercer capítulo se menciona en primer lugar la importancia del supuesto de normalidad en las perturbaciones estocásticas dentro del modelo de regresión lineal . Después se presentan algunos métodos que existen para la verificación del supuesto antes mencionado ; a continuación se realiza una breve explicación del método de bondad de ajuste en general y del papel que desempeñan en este método la función de distribución empírica y las estadísticas basadas en dicha función. Al final de este capítulo se proporciona el desarrollo del método de bondad de ajuste para la comprobación del supuesto de normalidad en las perturbaciones estocásticas.

En el capítulo IV se presentan ejemplos extraídos de literatura especializada en regresión lineal para ilustrar la teoría descrita en los capítulos anteriores y principalmente para verificar el supuesto de normalidad en las perturbaciones estocásticas a través del método descrito en este trabajo, así como para realizar comparaciones con los otros métodos ya existentes mencionados en el capítulo III.

En el capítulo V se realiza una simulación con perturbaciones estocásticas que se distribuyen normalmente cuyo objetivo principal es hacer una comparación entre los resultados obtenidos por dicha simulación para la estadística $\bar{A}^2$, de este trabajo , y la distribución límite de la estadística Anderson-Darling A^2 para la verificación de normalidad, con localización y escala desconocidas. Además se genera por medio de esa simulación una tabla de los valores críticos de la estadística $\bar{A}^2$.

Finalmente se presentan las conclusiones obtenidas en este trabajo.

CAPITULO I

EL MODELO DE REGRESION LINEAL

I . 1 DEFINICION DEL MODELO.

En el análisis de regresión se tiene como principal objetivo representar el comportamiento de un fenómeno por medio de un modelo (Modelo de regresión) que involucre la relación entre las variables que a través de un análisis teórico profundo parezcan estar asociadas a tal fenómeno. El modelo tiene el propósito de adentrarse en la relación de las variables elegidas para explicarnos el comportamiento de una de ellas (efecto) y la identificaremos como variable dependiente (Y), en función de las otras (causas) que serán identificadas como variables independientes o explicativas (X_j).

Para ejemplificar se puede mencionar la relación existente entre las variables consumo e ingreso, ya que al respaldarse en la teoría económica, se puede suponer que el consumo o gasto realizado por una entidad se puede explicar aproximadamente al conocer el comportamiento de los ingresos de dicha entidad. En este caso, la variable dependiente sería el consumo realizado y se debería expresar funcionalmente en términos de la variable explicativa que representaría el ingreso.

Para tratar de explicar el comportamiento de la variable dependiente a partir de las variables explicativas es necesario contar con una serie de datos (muestra) de las variables que fueron seleccionadas para el análisis. Los datos obtenidos pueden ser utilizados para establecer las variables (o funciones de las mismas) que deben ser incluidas en el modelo de regresión y pudieran ser aquellas que al graficarse individualmente contra la variable dependiente tengan un efecto importante ; en el caso de considerar un modelo de regresión lineal el comportamiento de dichas variables debe ser lineal sobre la variable dependiente.

A continuación se puede proponer un modelo de regresión lineal que se espera en promedio nos describa la relación existente entre las variables :

$$Y = \beta_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_k X_k + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1),$$

donde : Y es la variable dependiente.

X_j son las variables explicativas. $(j = 2, 3, \ldots\ldots, k)$

β_1 es el valor esperado para Y cuando todas las variables explicativas son iguales a cero.

β_j son los coeficientes asociados a cada variable. $(j = 2, 3, \ldots\ldots, k)$

ε es el término de perturbación estocástica.

El modelo queda completo al considerar los supuestos que se mencionan en la sección I . 2 .

El modelo descrito cuenta con k-1 variables explicativas. El término de perturbación estocástica puede representar la influencia de todas aquellas variables que no fueron consideradas en dicho modelo.

La utilidad de contar con un modelo consiste principalmente en la posibilidad de realizar la predicción del valor de la variable dependiente a partir de las condiciones presentadas por las variables explicativas y/o en la posibilidad de realizar pruebas de hipótesis en relación a los valores de los coeficientes asociados a las variables explicativas para poder determinar la influencia individual de cada una de éstas con respecto a la variable dependiente.

I . 2 SUPUESTOS DEL MODELO DE REGRESION LINEAL

Los supuestos del modelo de regresión lineal son las condiciones deseables a priori sobre la manera como se generan tanto el término de perturbación estocástica (ε) así como las variables independientes (Xj); dichos supuestos son indispensables para realizar estimaciones con respecto a la regresión.

Antes de mencionar los supuestos es importante establecer la notación para representar a las variables independientes por medio de la matriz (X) y el término de perturbación estocástica por medio del vector ($\underline{\varepsilon}$) para considerar las n observaciones de las variables que se relacionan en el modelo de la siguiente manera :

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix} \quad , \qquad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} ,$$

donde : X_{ij} es la i - ésima observación de la j - ésima variable explicativa, con ($i = 1, 2, 3, \ldots, n$) y ($j = 1, 2, 3, \ldots, k$), siendo (x_{i1}) = 1 $\forall$ i.

ε_i es el término de la perturbación estocástica de la i - ésima observación, ($i = 1, 2, 3, \ldots, n$)

SUPUESTO 1) INDEPENDENCIA ENTRE PERTURBACIONES ESTOCASTICAS Y VARIABLES EXPLICATIVAS.

Las variables explicativas son fijas , o en su defecto, si son aleatorias se distribuyen independientemente del término de perturbación estocástica.

En el caso de que las variables explicativas sean fijas obviamente la distribución de las perturbaciones estocásticas (ε_i) no depende de los valores fijos de tales variables.

Si las variables explicativas fueran aleatorias :

$$\begin{aligned} \text{Cov}\,(\varepsilon_i, X_{ij}) &= E\left[\{\varepsilon_i - E(\varepsilon_i)\}\ \{X_{ij} - E(X_{ij})\}\right] \\ &= E\{\varepsilon_i X_{ij} - E(\varepsilon_i) X_{ij} - E(X_{ij})\,\varepsilon_i + E(\varepsilon_i)\,E(X_{ij})\} \\ &= E\{\varepsilon_i X_{ij}\} - E(X_{ij})\,E(\varepsilon_i) = E(\varepsilon_i)E(X_{ij}) - E(X_{ij})\,E(\varepsilon_i) = 0 \end{aligned}$$

$$i = 1, 2, \ldots, n \quad y \quad j = 2, \ldots, k \quad ,$$

donde : Cov representa la covarianza entre las variables que por hipótesis son independientes.

SUPUESTO 2) PERTURBACION ESPERADA IGUAL A CERO.

El valor esperado de las perturbaciones estocásticas es igual a cero.

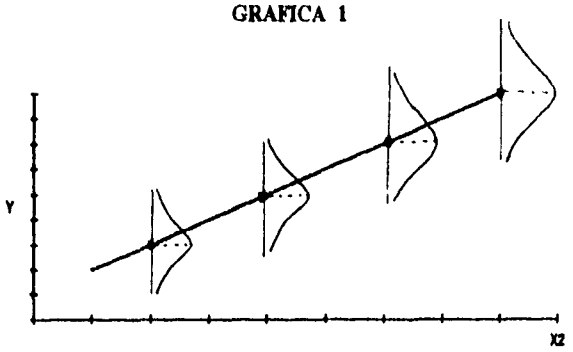
$$E(\varepsilon_i) = 0 \quad , \qquad i = 1, 2, \ldots\ldots, n$$

En notación vectorial :

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

donde : $\underline{\varepsilon}$ y $\underline{0}$ son vectores columna, siendo $\underline{0}$ el vector nulo.

GRAFICA 1



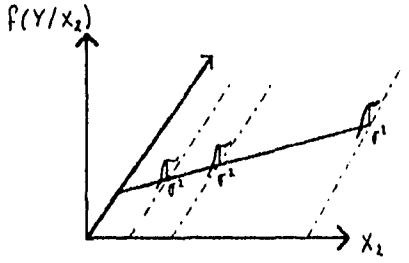
Función de regresión : —— $E(Y/X_2) = \beta_1 + \beta_2 X_2$ * $E(\varepsilon_i) = 0$
Supuesto 2 : La media de las perturbaciones estocásticas es cero.

SUPUESTO 3) HOMOSCEDASTICIDAD.

La varianza de las perturbaciones estocásticas es constante (σ^2).

$$Var(\varepsilon_i) = E(\varepsilon_i^2) - E^2(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2 \qquad i = 1, 2, \ldots\ldots, n$$

GRAFICA 2 (HOMOSCEDASTICIDAD)

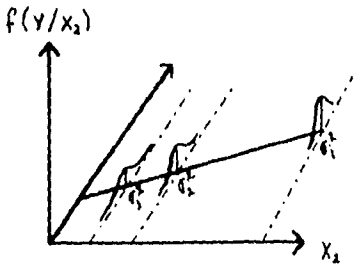


Función de regresión : —— $E(Y/X_2) = \beta_1 + \beta_2 X_2$
Supuesto 3 : Igual varianza para cada valor de la variable explicativa X_2.

Si este supuesto no se cumpliera tendríamos que la varianza de las perturbaciones estocásticas no es constante (σ_i^2).

$$\text{Var}(\varepsilon_i) = E(\varepsilon_i^2) - E^2(\varepsilon_i) = E(\varepsilon_i^2) = \sigma_i^2 \qquad i = 1, 2, \ldots\ldots, n$$

GRAFICA 3 (HETEROSCEDASTICIDAD)



Función de regresión : ——— $E(Y/X_2) = \beta_1 + \beta_2 X_2$

Violación al supuesto 3 : Distinta varianza para cada valor de la variable explicativa . En este caso la varianza decrece con los valores de X_2.

Es importante mencionar que las gráficas 1, 2 y 3 toman como referencia el modelo de regresión lineal con una sola variable explicativa (X_2).

SUPUESTO 4) INDEPENDENCIA ENTRE LAS PERTURBACIONES ESTOCASTICAS.

Las perturbaciones estocásticas no presentan dependencia alguna entre sí.

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = E[\{\varepsilon_i - E(\varepsilon_i)\}\{\varepsilon_j - E(\varepsilon_j)\}]$$
$$= E\{\varepsilon_i \varepsilon_j - E(\varepsilon_i)\varepsilon_j - E(\varepsilon_j)\varepsilon_i + E(\varepsilon_i)E(\varepsilon_j)\} = E\{\varepsilon_i \varepsilon_j\} = 0$$

siendo i, j dos índices distintos y COV representa la covarianza entre las variables.

Representación matricial de los supuestos 3 y 4.

$$E(\underline{\varepsilon}\underline{\varepsilon}') = E\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix}(\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_n) = E\begin{pmatrix}(\varepsilon_1)^2 & \varepsilon_1\varepsilon_2 & \cdots & \varepsilon_1\varepsilon_n\\ \varepsilon_2\varepsilon_1 & (\varepsilon_2)^2 & \cdots & \varepsilon_2\varepsilon_n\\ \vdots & \vdots & \ddots & \vdots\\ \varepsilon_n\varepsilon_1 & \varepsilon_n\varepsilon_2 & \cdots & (\varepsilon_n)^2\end{pmatrix}$$

$$= \begin{pmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n)\\ E(\varepsilon_1\varepsilon_2) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_n)\\ \vdots & \vdots & \ddots & \vdots\\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \cdots & E(\varepsilon_n^2)\end{pmatrix} = \begin{pmatrix}\sigma^2 & 0 & \cdots & 0\\ 0 & \sigma^2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sigma^2\end{pmatrix} = \sigma^2 I_n\ ,$$

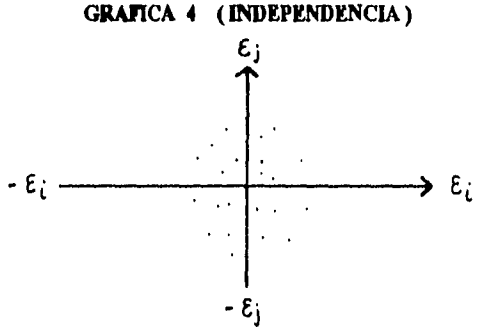
donde : $\underline{\varepsilon}$ es el vector columna que representa a las perturbaciones estocásticas.

$\underline{\varepsilon}'$ es el vector fila que representa a las perturbaciones estocásticas .

σ^2 es la varianza desconocida de cada perturbación estocástica.

I_n es la matriz identidad de (n x n).

GRAFICA 4 (INDEPENDENCIA)



Supuesto 4 : Dado un valor de cada una de las variables explicativas, las perturbaciones estocásticas de dos observaciones distintas cualesquiera de Y no guardan entre si relación alguna.

SUPUESTO 5) LA MATRIZ X ES DE RANGO COMPLETO.

Las columnas de la matriz X son linealmente independientes , por lo que el rango de la matriz X es igual a k.

SUPUESTO 6) NORMALIDAD EN LAS PERTURBACIONES ESTOCASTICAS .

Las perturbaciones estocásticas poseen una distribución normal . Su media y varianza han sido descritas en los supuestos 2 y 3 respectivamente , es decir , $E(\varepsilon_i) = 0$ y $VAR(\varepsilon_i) = \sigma^2$.

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

En notación vectorial :

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

SUPUESTO 7) ESPECIFICACION ADECUADA DEL MODELO .

El modelo de regresión esta correctamente especificado , es decir , no existe ningún sesgo en su representación, en cuanto a lo siguiente:

A) Las variables explicativas utilizadas para representar el modelo son correctas y suficientes.

B) Los parámetros βj utilizados para representar el modelo, son constantes.

C) El valor esperado de Y es lineal en los parámetros (βj).

A partir de dichos supuestos se estiman los parámetros (βj). La estimación de dichos parámetros por el método de Máxima verosimilitud coincide con aquel en el que se minimiza la suma de cuadrados de los errores en el ajuste.

I . 3 ESTIMADORES DEL MODELO DE REGRESION LINEAL

Al contar con el modelo de regresión lineal se busca estimar los parámetros desconocidos que lo componen de tal manera que las desviaciones entre dichos parámetros y su estimación sean mínimas en algún sentido.

El método de mínimos cuadrados ordinarios, obtiene los mejores estimadores lineales insesgados para los parámetros basándose en los supuestos antes mencionados (a excepción del supuesto 6) y utilizando los datos de las variables del modelo. Este método (en presencia del supuesto 6) coincide con el de máxima verosimilitud , obteniéndose la estimación del modelo de regresión lineal por medio de la siguiente función :

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_k x_{ik} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2),$$

donde : $\hat{y}_i$ es el estimador para la i-ésima observación (i = 1, 2, ..., n) de la variable dependiente del modelo.

x_{ij} son los valores de las variables explicativas (j = 2, 3, ..., k) , para la i-ésima observación .

$\hat{\beta}_1$ es el estimador de la ordenada al origen del modelo.

$\hat{\beta}_j$ son los estimadores de los parámetros (" pesos " de cada variable) (j = 2, 3, ..., k).

Los estimadores obtenidos por mínimos cuadrados ordinarios y que se representan mediante el vector ($\hat{\underline{\beta}}$) son :

$$\hat{\underline{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = (X'X)^{-1} X'Y \ ,$$

donde :

$$X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{pmatrix} \quad , \qquad \underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad ,$$

X contiene la información de las variables explicativas.
X' es la transpuesta de la matriz (X).
Y es el vector columna que contiene la información de la variable dependiente.

Al contar con la estimación de la variable dependiente para cada observación, se calcula el i - ésimo residuo (e_i) de la siguiente manera :

$$e_i = y_i - \hat{y}_i \quad ,$$

donde : e_i es el estimador de la i - ésima perturbación estocástica (ε_i) .

Para representar las varianzas y covarianzas de los coeficientes estimados se utiliza la siguiente matriz :

$$VCov \left(\underline{\hat{\beta}} \right) = \begin{pmatrix} Var(\hat{\beta}_1) & Cov(\hat{\beta}_1,\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_1,\hat{\beta}_k) \\ Cov(\hat{\beta}_2,\hat{\beta}_1) & Var(\hat{\beta}_2) & \cdots & Cov(\hat{\beta}_2,\hat{\beta}_k) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_k,\hat{\beta}_1) & Cov(\hat{\beta}_k,\hat{\beta}_2) & \cdots & Var(\hat{\beta}_k) \end{pmatrix}$$

$$= \sigma^2 (X'X)^{-1} \quad ,$$

donde : σ^2 es la varianza constante y desconocida de las perturbaciones estocásticas de acuerdo con el supuesto 3.

La varianza σ^2 que necesitamos para obtener la matriz de varianzas y covarianzas es desconocida por lo cual se utiliza un estimador insesgado ($\hat{\sigma}^2$) para estimar dicha matriz :

$$\hat{\sigma}^2 = \frac{\underline{e}'\underline{e}}{n-k} \ ,$$

donde : $\underline{e}$ es el vector columna de residuos.

$\underline{e}'$ es el vector transpuesto de $\underline{e}$.

n es el número de observaciones.

$k - 1$ es el número de variables explicativas utilizadas para representar el modelo. (Ya que se adiciona una variable $X_1 = 1$, para representar la ordenada al origen del modelo) .

I . 4 PROPIEDADES DE LOS ESTIMADORES

El estimador de $\underline{\beta}' = (\beta_1 , \beta_2 , ..., \beta_K)$ que se obtiene por el método de mínimos cuadrados ordinarios ($\underline{\hat{\beta}}$) tiene propiedades muy importantes que se mencionan a continuación :

1) Cada componente de $\underline{\hat{\beta}}$ es una combinación lineal de los componentes del vector Y respectivamente.

2) Cada componente de $\underline{\hat{\beta}}$ es un estimador insesgado del verdadero valor de la correspondiente componente de $\underline{\beta}$, es decir .

$$E(\hat{\beta}_j) = \beta_j \qquad j = 1,2, ... , k$$

En notación vectorial:

$$E(\hat{\underline{\beta}}) = \underline{\beta}$$

Demostración .

$$\begin{aligned} E[\hat{\underline{\beta}}] &= E[(X'X)^{-1}X'\underline{Y}] = E[(X'X)^{-1}X'(X\underline{\beta} + \underline{\varepsilon})] \\ &= E[(X'X)^{-1}X'X\underline{\beta}] + E[(X'X)^{-1}X'\underline{\varepsilon}] \\ &= E[\underline{\beta}] + (X'X)^{-1}X'E[\underline{\varepsilon}] = \underline{\beta} \end{aligned}$$

$$\therefore \qquad E[\hat{\underline{\beta}}] = \underline{\beta}$$

3) Los componentes de $\hat{\underline{\beta}}$ son los estimadores de varianza mínima y las varianzas y covarianzas de $\hat{\underline{\beta}}$ se estiman al estimar la matriz de varianza-covarianza por medio de la siguiente expresión:

$$\widehat{VCov}(\hat{\underline{\beta}}) = \hat{\sigma}^2(X'X)^{-1}$$

Por otra parte el estimador $\hat{\sigma}^2$ de la varianza constante pero desconocida σ^2 de las perturbaciones estocásticas, es un estimador insesgado, es decir ,

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

Demostración .

$$E[\,\hat{\sigma}^2\,] = E\left[\frac{\underline{e}'\underline{e}}{n-k}\right] = \frac{1}{(n-k)}E[\,\underline{e}'\underline{e}\,]$$

$$\begin{aligned}
\underline{e}'\underline{e} &= (\underline{Y} - X\hat{\underline{\beta}})'(\underline{Y} - X\hat{\underline{\beta}}) \\
&= \left(\underline{Y} - X(X'X)^{-1}X'\underline{Y}\right)'\left(\underline{Y} - X(X'X)^{-1}X'\underline{Y}\right) \\
&= \underline{Y}'\left(I_n - X(X'X)^{-1}X'\right)\left(I_n - X(X'X)^{-1}X'\right)\underline{Y} \\
&= \underline{Y}'\left(I_n - X(X'X)^{-1}X'\right)\underline{Y} \\
&= (X\underline{\beta} + \underline{\varepsilon})'\left(I_n - X(X'X)^{-1}X'\right)(X\underline{\beta} + \underline{\varepsilon}) \\
&= \underline{\varepsilon}'\left(I_n - X(X'X)^{-1}X'\right)\underline{\varepsilon} \qquad \textit{que es un escalar}
\end{aligned}$$

$$\Rightarrow \quad \underline{\varepsilon}'\left(I_n - X(X'X)^{-1}X'\right)\underline{\varepsilon} = tr\left\{\underline{\varepsilon}'\left(I_n - X(X'X)^{-1}X'\right)\underline{\varepsilon}\right\}$$

$$y \quad tr\left\{\underline{\varepsilon}'\left(I_n - X(X'X)^{-1}X'\right)\underline{\varepsilon}\right\} = tr\left\{\left(I_n - X(X'X)^{-1}X'\right)\underline{\varepsilon}\underline{\varepsilon}'\right\}$$

Como la traza y la esperanza son operadores lineales , por lo tanto :

$$\begin{aligned}
E\left[tr\left\{\left(I_n - X(X'X)^{-1}X'\right)\underline{\varepsilon}\underline{\varepsilon}'\right\}\right] &= tr\left\{E\left[\left(I_n - X(X'X)^{-1}X'\right)\underline{\varepsilon}\underline{\varepsilon}'\right]\right\} \\
&= tr\left\{\left(I_n - X(X'X)^{-1}X'\right)E[\underline{\varepsilon}\underline{\varepsilon}']\right\} \\
&= tr\left\{\left(I_n - X(X'X)^{-1}X'\right)\sigma^2 I_n\right\}
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \quad E[\,\hat{\sigma}^2\,] &= \frac{1}{(n-k)}tr\left\{\left(I_n - X(X'X)^{-1}X'\right)\sigma^2 I_n\right\} \\
&= \frac{\sigma^2}{(n-k)}\left(tr\{I_n\} - tr\left\{X(X'X)^{-1}X'\right\}\right) \\
&= \frac{\sigma^2}{(n-k)}\left(tr\{I_n\} - tr\left\{X'X(X'X)^{-1}\right\}\right) \\
&= \frac{\sigma^2}{(n-k)}\left(tr\{I_n\} - tr\{I_k\}\right) = \sigma^2
\end{aligned}$$

$$\therefore \qquad E[\,\hat{\sigma}^2\,] = \sigma^2$$

El estimador $\hat{\underline{\beta}}$ tiene la propiedad de ser el mejor estimador lineal insesgado de $\underline{\beta}$.

Demostración .

$\hat{\underline{\beta}} = \left[(X'X)^{-1}X'\right]\underline{Y}$, *por lo cual* $\hat{\underline{\beta}}$ *es una función lineal de* $\underline{Y}$, *y como se demostró anteriormente* $\hat{\underline{\beta}}$ *es un estimador insesgado , por lo tanto ,* $\hat{\underline{\beta}}$ *pertenece a la clase de estimadores lineales e insesgados* .

Sea $\ddot{\underline{\beta}}$ *cualquier miembro de la clase de estimadores lineales e insesgados*

$\Rightarrow$ $\ddot{\underline{\beta}}$ *puede ser escrito como :* $\ddot{\underline{\beta}} = \left[(X'X)^{-1}X' + C\right]\underline{Y}$

$\Rightarrow$ $\hat{\underline{\beta}}$ *es el mejor estimador lineal insesgado de* $\underline{\beta}$ $\Leftrightarrow$ $\left[VCov(\ddot{\underline{\beta}}) - VCov(\hat{\underline{\beta}})\right]$ *es una matriz semidefinida positiva* $\forall$ $\ddot{\underline{\beta}}$ *perteneciente a la clase* .

$$E[\ddot{\underline{\beta}}] = E\left[\left((X'X)^{-1}X' + C\right)\underline{Y}\right] = E\left[\left((X'X)^{-1}X' + C\right)(X\underline{\beta} + \underline{\varepsilon})\right]$$

$$= E[\underline{\beta}] + (CX)E[\underline{\beta}] + (X'X)^{-1}X'E[\underline{\varepsilon}] + CE[\underline{\varepsilon}]$$

$$= \underline{\beta} + CX\underline{\beta}$$

y como $\ddot{\underline{\beta}}$ *pertenece a la familia de estimadores lineales e insesgados*

$\Rightarrow$ $CX\underline{\beta} = \underline{0}$ $\forall$ $\underline{\beta}$ $\Rightarrow$ $CX = 0$

$$(\ddot{\underline{\beta}} - \underline{\beta}) = \left\{\left((X'X)^{-1}X' + C\right)\underline{\varepsilon}\right\} \quad , \quad (\hat{\underline{\beta}} - \underline{\beta}) = \left[\left((X'X)^{-1}X'\right)\underline{\varepsilon}\right]$$

$$\Rightarrow VCov(\ddot{\underline{\beta}}) = E\left[(\ddot{\underline{\beta}} - \underline{\beta})(\ddot{\underline{\beta}} - \underline{\beta})'\right] = E\left[\left\{\left((X'X)^{-1}X' + C\right)\underline{\varepsilon}\right\}\left\{\left((X'X)^{-1}X' + C\right)\underline{\varepsilon}\right\}'\right]$$

$$= \left((X'X)^{-1}X' + C\right)E[\underline{\varepsilon}\underline{\varepsilon}']\left(X(X'X)^{-1} + C'\right)$$

$$= \sigma^2(X'X)^{-1} + \sigma^2 CC'$$

$$, \; VCov(\underline{\hat{\beta}}) = E\left[(\underline{\hat{\beta}} - \underline{\beta})(\underline{\hat{\beta}} - \underline{\beta})'\right] = E\left[\left\{\left((X'X)^{-1}X'\right)\underline{\varepsilon}\right\}\left\{\left((X'X)^{-1}X'\right)\underline{\varepsilon}\right\}'\right]$$
$$= \left((X'X)^{-1}X'\right) E[\underline{\varepsilon}\underline{\varepsilon}'] \left(X(X'X)^{-1}\right)$$
$$= \sigma^2 (X'X)^{-1}$$

$$\Rightarrow \left[VCov(\underline{\ddot{\beta}}) - VCov(\underline{\hat{\beta}})\right] = \sigma^2 CC'$$

σ^2 es una constante desconocida mayor que cero y CC' es una matriz semidefinida positiva .

$\Rightarrow \left[VCov(\underline{\ddot{\beta}}) - VCov(\underline{\hat{\beta}})\right]$ *es una matriz semidefinida positiva .*

$\therefore$ $\underline{\hat{\beta}}$ *es el mejor estimador lineal e insesgado de* $\underline{\beta}$.

I . 5 ELEMENTOS PARA SUSTENTAR A POSTERIORI EL MODELO

Los elementos que se utilizan principalmente para sustentar a posteriori el modelo son el coeficiente de determinación R^2 y las pruebas de hipótesis acerca de la significancia global de la regresión y del verdadero valor de los parámetros (β_i).

El coeficiente de determinación R^2 es un elemento utilizado para la " cantidad " de ajuste de la función de regresión y determina la proporción de variabilidad total en la variable dependiente que es explicada por las variables explicativas. El coeficiente de determinación tiene como límite inferior el valor de cero que indica que el comportamiento de la variable dependiente no es explicada en absoluto por las variables explicativas y como límite superior el valor de uno que indicaría que el comportamiento de la variable dependiente es totalmente explicado por las variables explicativas.

$$R^2 = \frac{SCR}{SCT} = \frac{\underline{\hat{Y}}'\underline{\hat{Y}} - n\bar{Y}^2}{\underline{Y}'\underline{Y} - n\bar{Y}^2} \qquad 0 \le R^2 \le 1 ,$$

donde : SCR es la suma de cuadrados debida a la regresión y representa a la suma total de cuadrados que es explicada por la regresión .

SCT es la suma total de cuadrados y es la medida de variación en la variable dependiente con respecto a su media .

Al calcular la diferencia entre la suma total de cuadrados (SCT) y la suma de cuadrados debida a la regresión (SCR) se obtiene la suma de cuadrados que no es explicada por la regresión (SCE) :

$$\begin{aligned} SCT - SCR &= (\underline{Y}'\underline{Y} - nY^2) - (\underline{\hat{Y}}'\underline{\hat{Y}} - nY^2) = \underline{Y}'\underline{Y} - \underline{\hat{Y}}'\underline{\hat{Y}} \\ &= \underline{Y}'\underline{Y} - (X\underline{\hat{\beta}})'(X\underline{\hat{\beta}}) \\ &= \underline{Y}'\underline{Y} - 2(X\underline{\hat{\beta}})'(X\underline{\hat{\beta}}) + (X\underline{\hat{\beta}})'(X\underline{\hat{\beta}}) \\ &= \underline{Y}'\underline{Y} - 2\underline{\hat{\beta}}'X'X(X'X)^{-1}X'\underline{Y} + \underline{\hat{\beta}}'X'X\underline{\hat{\beta}} \end{aligned}$$

Al factorizar se obtiene:

$$SCE = SCT - SCR = (\underline{Y} - X\underline{\hat{\beta}})'(\underline{Y} - X\underline{\hat{\beta}}) = \underline{e}'\underline{e}$$

$$\therefore \quad SCE = \underline{e}'\underline{e}$$

donde: SCE se conoce comúnmente como suma de cuadrados del error o suma de cuadrados de residuos.

Sin embargo, el coeficiente de determinación R^2 muestra cierta deficiencia al ser utilizado para comprobar la bondad de ajuste de modelos alternativos. La deficiencia consiste en que al incrementar las variables explicativas en un modelo el coeficiente de determinación nunca decrece; al contrario siempre aumenta.

Para la problemática señalada anteriormente es necesario utilizar un coeficiente de determinación alternativo conocido como R^2 ajustado ($\bar{R}^2$), ya que el coeficiente se ajusta por los grados de libertad asociados con las sumas de cuadrados en la siguiente ecuación:

$$\bar{R}^2 = 1 - \left[\frac{\underline{e}'\underline{e} \;/\; (n-k)}{\underline{Y}'\underline{Y} - n\bar{Y}^2/(n-1)} \right],$$

que es equivalente a,

$$\bar{R}^2 = 1 - \left[\frac{(n-1)}{(n-k)} (1 - R^2) \right]$$

Al usar $\bar{R}^2$ se elimina la deficiencia presentada por R^2, ya que al añadir variables explicativas en el modelo el coeficiente de determinación $\bar{R}^2$ produce una reducción en el término (1- R^2) que se compensa con el incremento que genera el término (n-1) / (n - k) .

Recordando el supuesto 6 que indica que las perturbaciones estocásticas se distribuyen normalmente con media cero y varianza constante (σ^2), tenemos en notación vectorial :

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

Con base en dicho supuesto, se puede afirmar que :

$$\hat{\underline{\beta}} \sim N(\underline{\beta}, \sigma^2 (X'X)^{-1})$$

Demostración .

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

La función generadora de momentos de $\hat{\underline{\beta}}$ *es :*

$$M_{\hat{\underline{\beta}}}(\underline{t}) = E\left[e^{\underline{t}'\hat{\underline{\beta}}} \right] = E\left[e^{\underline{t}'(\underline{\beta} + (X'X)^{-1}X'\underline{\varepsilon})} \right]$$
$$= E\left[e^{\underline{t}'\underline{\beta}}\, e^{\underline{t}'(X'X)^{-1}X'\underline{\varepsilon}} \right] = e^{\underline{t}'\underline{\beta}} E\left[e^{\underline{L}'\underline{\varepsilon}} \right]$$

donde : $\underline{L}' = \underline{t}'(X'X)^{-1}X'$

$$\Rightarrow M_{\hat{\underline{\beta}}}(\underline{t}) = e^{\underline{t}'\underline{\beta}}\, M_{\underline{\varepsilon}}(\underline{L})$$

Como $\underline{\varepsilon} \sim N(\underline{0},\ \sigma^2 I_n) \Rightarrow M_{\underline{\varepsilon}}(\underline{L}) = e^{-\frac{1}{2}\underline{L}'(\sigma^2 I_n)\underline{L}}$

por lo tanto , $M_{\hat{\underline{\beta}}}(\underline{t}) = e^{\underline{t}'\underline{\beta}}\, e^{-\frac{1}{2}\underline{L}'(\sigma^2 I_n)\underline{L}}$

$$= e^{\underline{t}'\underline{\beta} - \frac{1}{2}\underline{t}'(X'X)^{-1}X'(\sigma^2 I_n)X(X'X)^{-1}\underline{t}}$$

$$\Rightarrow M_{\hat{\underline{\beta}}}(\underline{t}) = e^{\underline{t}'\underline{\beta} - \frac{1}{2}\underline{t}'\sigma^2(X'X)^{-1}\underline{t}}$$

que es la función generadora de momentos de una distribución normal con media $\underline{\beta}$ *y matriz de varianza y covarianza* $\sigma^2(X'X)^{-1}$.

$$\therefore \qquad \hat{\underline{\beta}} \sim N\left(\underline{\beta},\ \sigma^2(X'X)^{-1}\right)$$

En la práctica como se desconoce el valor de σ^2, se estima por medio de $\hat{\sigma}^2$ y se puede obtener un valor (t) para cada componente del vector (β) que se define de la siguiente manera :

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{Var}(\hat{\beta}_j)}} \sim t_{(n-k)}$$

Por lo tanto la t se puede utilizar para evaluar la hipótesis acerca del verdadero valor de β_j a un nivel de significancia α ; siendo dicho nivel la probabilidad de rechazar la hipótesis dado que es verdadera.

Por otra parte si consideramos como hipótesis nula ($\beta_2 = \beta_3 = \ldots = \beta_k = 0$) , que indica que no existe ninguna relación entre las supuestas variables explicativas y la variable dependiente, se puede obtener el valor F que se define de la siguiente manera :

$$F = \frac{SCR/(k-1)}{SCE/(n-k)} \sim F_{[(k-1),(n-k)]}$$

donde : SCR es la suma de cuadrados debida a la regresión.
SCE es la suma de cuadrados del error.

Por lo tanto, por medio de la distribución F se puede realizar la prueba de significancia global de la regresión como un elemento para sustentar a posteriori el modelo.

CAPITULO II

RESIDUOS DEL MODELO DE REGRESION LINEAL

II . 1 EL RESIDUO COMO UTENSILIO PARA VERIFICAR SUPUESTOS DEL MODELO.

En el capítulo anterior se define el modelo de regresión lineal y se proporciona una estimación de sus parámetros, buscando que la diferencia entre las observaciones Y_i y su estimación (modelo $\hat{Y}_i$) sea mínima. Tal diferencia se expresa por medio del vector de residuos (e), que se define como el vector cuyos elementos están formados por la diferencia entre las observaciones de la variable dependiente y su estimación.

$$e_i = y_i - \hat{y}_i \qquad i = 1, 2, \ldots, n \ ,$$

por consiguiente :

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

donde :

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \qquad \text{y} \qquad \underline{\hat{Y}} = X\underline{\hat{\beta}}$$

Es importante realizar un análisis profundo de los elementos que constituyen el vector de residuos ya que se utilizan para detectar si existen violaciones en los supuestos del modelo de regresión lineal utilizando principalmente gráficas. Las violaciones que se pueden detectar por medio de los residuos son las siguientes :

1. Existencia de observaciones discrepantes "outliers" en los datos.
2. Varianza no constante de las perturbaciones estocásticas .
3. Presencia de autocorrelación entre los elementos del vector de perturbación estocástica.

4. Dependencia entre las observaciones de las variables explicativas y los elementos del vector de perturbación estocástica.

5. El vector de perturbación estocástica no sigue una distribución normal.

6. La especificación del modelo no es la adecuada :

   a) La forma funcional de las variables explicativas incluidas en el modelo no es la adecuada.
   b) Las variables explicativas incluidas en el modelo de regresión lineal no son suficientes.

II . 2 TIPOS DE RESIDUOS.

II . 2 . 1 RESIDUOS ORDINARIOS.

El vector de residuos definido en la sección anterior es el vector de residuos ordinarios cuyos elementos se calculan como la diferencia de las observaciones de la variable dependiente y su estimación.

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

donde :

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

Sustituyendo el valor del vector $(\underline{\hat{\beta}})$ se tiene :

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

Por lo tanto:

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

Por consiguiente :

$$\underline{e} = (I_n - H)\underline{Y} \quad \text{..... (3)}$$

Denotando a la matriz H como :

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

donde :

$$X = \begin{bmatrix} \underline{x}_1' \\ \vdots \\ \underline{x}_i' \\ \vdots \\ \underline{x}_n' \end{bmatrix} \text{ ó } , \; X = [\underline{c}_1, \cdots, \underline{c}_j, \cdots, \underline{c}_k]$$

y

$\underline{x}_i'$ es el vector fila que contiene la i-ésima observación de las variables explicativas.

$\underline{c}_j$ es el vector columna que contiene los valores de la j-ésima variable explicativa.

De tal forma , X es la matriz compuesta por los vectores $\underline{x}_i'$ o $\underline{c}_j$.

La matriz H es conocida como la matriz sombrero (" Hat - matrix ") ya que H aplicada a $\underline{Y}$ nos da $\hat{\underline{Y}}$, es decir, " le pone el sombrero ". Esta matriz tiene las características siguientes :

a) Es simétrica, es decir , $H = H'$

b) Es idempotente, es decir , $H = H^2$

Por otro lado la matriz $(I_n - H)$ hereda las características de la matriz sombrero. Estas matrices son transformaciones lineales que proyectan ortogonalmente cualquier vector en un espacio lineal; H en el espacio generado por las columnas de la matriz **X** y la matriz $(I_n - H)$ en el complemento ortogonal.

Para el análisis de residuos es importante conocer la relación que existe entre el vector de perturbación estocástico $(\underline{\varepsilon})$ y su estimador $(\underline{e})$, la cual se muestra partiendo de la ecuación (3).

$$\underline{e} = (I_n - H)\underline{Y}$$

Sustituyendo el vector $(\underline{Y})$ se tiene :

$$\begin{aligned}
\underline{e} &= (I_n - H)(X\underline{\beta} + \underline{\varepsilon}) \\
&= (X\underline{\beta} + \underline{\varepsilon} - HX\underline{\beta} - H\underline{\varepsilon}) \\
&= \left(X\underline{\beta} + \underline{\varepsilon} - X(X'X)^{-1}X'X\underline{\beta} - X(X'X)^{-1}X'\underline{\varepsilon}\right) \\
&= \left(X\underline{\beta} + \underline{\varepsilon} - X\underline{\beta} - X(X'X)^{-1}X'\underline{\varepsilon}\right) \\
&= \left(I_n - X(X'X)^{-1}X'\right)\underline{\varepsilon}
\end{aligned}$$

Por lo tanto :

$$\underline{e} = (I_n - H)\underline{\varepsilon} \quad \text{..... (4)}$$

Por medio de la ecuación (4) se puede observar que la relación entre el vector de perturbación estocástica y su estimador depende esencialmente de los elementos de la matriz H , es decir :

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

Considerando la relación anterior podemos determinar que si los elementos de la matriz sombrero (H) son lo suficientemente pequeños el vector estimador compuesto por los residuos ordinarios, sirve como sustituto del vector de perturbación estocástica , en otro caso, dicho estimador puede ser limitado como representante del vector de perturbación estocástica .

Tomando en cuenta la relación señalada en la ecuación (4) y los supuestos del modelo de regresión lineal referentes a las perturbaciones estocásticas podemos decir que el vector de residuos se distribuye normalmente con los parámetros siguientes :

$$E(\underline{e}) = \underline{0} \quad y \quad Var(\underline{e}) = \sigma^2(I_n - H) \quad ,$$

de tal forma :

$$\underline{e} \sim N(\underline{0}, \sigma^2(I_n - H))$$

Demostración .

$$Si \quad \underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_n) \Rightarrow M_{\underline{\varepsilon}}(\underline{t}) = e^{-\frac{1}{2}\underline{t}'(\sigma^2 I_n)\underline{t}}$$

$$\underline{e} = (I_n - H)\underline{\varepsilon}$$

$$\Rightarrow M_{\underline{e}}(\underline{t}) = E\left[e^{\underline{t}'\underline{e}} \right] = E\left[e^{\underline{t}'(I_n - H)\underline{\varepsilon}} \right]$$

$$= E\left[e^{\underline{t}^{*\prime}\underline{\varepsilon}} \right] = M_{\underline{\varepsilon}}(\underline{t}^*)$$

$$donde : \quad \underline{t}^* = \underline{t}'(I_n - H)$$

$$\Rightarrow M_{\underline{e}}(\underline{t}) = e^{-\frac{1}{2}\underline{t}'(I_n - H)(\sigma^2 I_n)(I_n - H)'\underline{t}}$$

$$= e^{-\frac{1}{2}\underline{t}'\sigma^2(I_n - H)\underline{t}}$$

$$\Rightarrow M_{\underline{e}}(\underline{t}) \quad es\ la\ función\ generadora\ de\ momentos$$

de una distribución normal con media $\underline{0}$ *y matriz de varianza y covarianza* $\sigma^2(I_n - H)$.

$$\therefore \quad \underline{e} \sim N(\underline{0}, \sigma^2(I_n - H))$$

Por lo tanto :

$$e_i \sim N(0\ ,\ \sigma^2\ (1-h_{ii}))\ \ ,$$

donde :

$$h_{ii} = \underline{x}_i'(X'X)^{-1}\underline{x}_i$$

y

$\underline{x}_i$ es el vector columna que contiene la i-ésima observación de las variables explicativas.

El objetivo del estudio de los residuos es examinar para determinar cualquier suposición incorrecta concerniente a los supuestos distribucionales del vector de perturbación estocástica a partir de un análisis realizado con su estimador (vector de residuos).

Debido a que la relación entre el vector de perturbación estocástico y su estimador no es perfecta en el sentido de igualdad, las alteraciones al modelo para el vector de perturbación estocástico generalmente no son transmitidas íteralmente a su estimador.

Es importante hacer notar que el vector de residuos se utiliza para verificar uno por uno los supuestos del modelo, ya que si fueran varios los supuestos que fallan simultáneamente, sería realmente complicado detectar las alteraciones ocasionadas al modelo.

II . 2 . 2 RESIDUOS ESTANDARIZADOS

Los residuos estandarizados son utilizados en técnicas de verificación de algunos supuestos del modelo de regresión lineal, principalmente para realizar gráficas con el propósito de detectar observaciones discrepantes " Outliers " y gráficas sobre papel normal, temas que serán abordados en secciones posteriores.

Estos residuos se definen de la siguiente manera :

$$w_i = \frac{e_i}{\hat{\sigma}} \qquad \text{..... (5)}$$

Tomando en cuenta la relación señalada en la ecuación (4) y los supuestos del modelo de regresión lineal referentes a las perturbaciones estocásticas podemos decir que los residuos estandarizados se distribuyen aproximadamente de la siguiente manera :

$$w_i \sim N(0\ ,\ (1-h_{ii}))$$

La idea de utilizar estos residuos se basa principalmente en que si se cumple el supuesto de normalidad en las perturbaciones estocásticas y los valores (h_{ii}) son pequeños, se asocia a los residuos estandarizados con una distribución normal con media cero y varianza aproximadamente unitaria .

II . 2 . 3 RESIDUOS ESTUDENTIZADOS

Los residuos ordinarios se presentan en unidades del problema original (unidades de las observaciones) tal como se puede observar en la ecuación (4), su distribución depende de σ^2 (parámetro desconocido) y de los elementos de H (conocidos), por tal razón los resultados obtenidos del análisis de estos residuos no se encuentran en una escala libre de parámetros desconocidos. Por otra parte cada residuo ordinario tiene una varianza que en general es diferente a la de los demás residuos .

Por esta razón se busca la manera de expresar a los residuos de tal forma que su distribución (en presencia del supuesto 6) no dependa de ningún parámetro desconocido, dicho de otra forma , que su escala sea única y que todos los componentes del vector de residuos tengan los mismos primeros dos momentos.

Los residuos estudentizados (Cook and Weisberg (1982 , capítulo 2)) son el resultado de haber dividido el residuo ordinario (cuya distribución depende de determinados parámetros desconocidos y de los elementos de la diagonal H) por estimadores de los parámetros desconocidos, de tal manera que el cociente tenga una distribución libre de dichos parámetros y todos los residuos tengan los mismos primeros dos momentos. De esta forma los residuos estudentizados cumplen con lo sugerido en el párrafo anterior .

Existen dos tipos de residuos estudentizados, los residuos estudentizados internos (r_i) y los residuos estudentizados externos (t_i).

a) ESTUDENTIZACION INTERNA

Es aquella en que la estadística y el estimador, en este caso, (e_i) y ($\hat{\sigma}$) respectivamente son dependientes y se calculan a partir de los mismos datos. Estos residuos se definen de la siguiente manera :

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \qquad i = 1, 2, \ldots, n \quad ,$$

donde : r_i = Residuo estudentizado interno.

e_i = Residuo ordinario.

$\hat{\sigma}^2$ = Cuadrado medio del error, también referido como cuadrado medio de los residuos.

Se puede observar que el estimador de la desviación estándar ($\hat{\sigma}$) y los residuos ordinarios (e_i) no son variables aleatorias independientes ya que el primer término se calcula utilizando los residuos ordinarios, por lo cual se dice que (r_i) no sigue una distribución t de Student.

b) ESTUDENTIZACION EXTERNA

Los residuos estudentizados externos requieren un estimador de (σ^2) que sea independiente de los residuos ordinarios , buscando de esta forma que los residuos estudentizados externos tengan una distribución t de Student.

Bajo el supuesto de normalidad de los errores, definamos a $\left(\hat{\sigma}^2_{(i)}\right)$ como el cuadrado medio de los residuos sin el i-ésimo caso, es decir ; originalmente

$$\hat{\sigma}^2 = \frac{e'e}{n-k} \qquad \text{..... (6)},$$

por lo tanto :

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-k} \qquad \Rightarrow \qquad \sum_{i=1}^{n} e_i^2 = \hat{\sigma}^2 (n-k)$$

Por consiguiente :

$$\hat{\sigma}^2_{(i)} = \frac{\hat{\sigma}^2(n-k) - \frac{e_i^2}{1-h_{ii}}}{n-k-1} = \frac{\hat{\sigma}^2\left((n-k) - \frac{e_i^2}{\hat{\sigma}^2(1-h_{ii})}\right)}{n-k-1} \quad ,$$

de tal forma :

$$\hat{\sigma}^2_{(i)} = \hat{\sigma}^2 \left(\frac{n-k-r_i^2}{n-k-1}\right)$$

Los residuos estudentizados externos se definen como :

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} \qquad i = 1, 2, \ldots, n \quad ,$$

donde : t_i = Residuo estudentizado externo.

e_i = Residuo ordinario.

$\left(\hat{\sigma}^2_{(i)}\right)$ = Cuadrado medio de los residuos exceptuando el i-ésimo caso.

En este caso la estadística y el estimador, (e_i) y $(\hat{\sigma}^2_{(i)})$ respectivamente son independientes, por lo cual podemos decir que los residuales estudentizados externos (t_i) se distribuyen como una t de Student con (n-k-1) grados de libertad.

Estos residuos estudentizados externos se relacionan con los residuos estudentizados internos (r_i) de la siguiente manera:

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = \frac{e_i}{\hat{\sigma}\sqrt{\frac{n-k-r_i^2}{n-k-1}}\sqrt{1-h_{ii}}}$$

$$= \frac{1}{\sqrt{\frac{n-k-r_i^2}{n-k-1}}} * \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}} = \frac{\sqrt{n-k-1}}{\sqrt{n-k-r_i^2}} * r_i$$

y por lo tanto:

$$t_i^2 = r_i^2\left(\frac{n-k-1}{n-k-r_i^2}\right)$$

Si se observa la ecuación anterior se puede afirmar que (t_i^2) es una transformación monótona de (r_i^2). Es importante señalar que los (t_i) no son independientes entre sí, pero sí idénticamente distribuidos.

II . 2 . 4 RESIDUOS PREDICTIVOS

Para obtener los residuos ordinarios y los residuos estudentizados se utiliza el vector estimado $(\hat{\underline{\beta}})$ que se calcula considerando todas las observaciones.

Con el propósito de contrastar dos cantidades independientes se utilizan los residuos predictivos, para los cuales se utiliza el vector estimado $(\hat{\underline{\beta}}_{(i)})$ que se calcula sin considerar la i-ésima observación de las variables.

Los residuos predictivos (Cook and Weisberg (1982 , capítulo 2)) se definen de la siguiente manera :

$$e_{i\,(i)} = Y_i - \hat{Y}_{i\,(i)} \qquad i = 1, 2, \ldots, n \quad ,$$

donde :

- $e_{i\,(i)}$ es el i-ésimo residuo ordinario que se calcula sin utilizar la i-ésima observación de las variables por medio del vector estimado $(\hat{\underline{\beta}}_{(i)})$.
- Y_i es la i-ésima observación de la variable dependiente .
- $\hat{Y}_{i\,(i)}$ es la estimación de la i-ésima observación que se calcula sin utilizar dicha observación de las variables por medio del vector estimado $(\hat{\underline{\beta}}_{(i)})$.

Se puede mostrar que el valor de $\hat{Y}_{i\,(i)}$ se puede calcular en términos de $(\hat{\underline{\beta}})$:

$$\hat{Y}_{i\,(i)} = \underline{x}_i' \hat{\underline{\beta}}_{(i)} = \underline{x}_i' \left[\hat{\underline{\beta}} - \frac{(X'X)^{-1} \underline{x}_i \, e_i}{1 - h_{ii}} \right] \quad ,$$

donde :

$\underline{x}_i'$ es el vector fila que contiene la i-ésima observación de las variables explicativas.

$\hat{\underline{\beta}}_{(i)}$ es el estimador de $\underline{\beta}$ calculado sin utilizar la i-ésima observación de las variables explicativas.

Los residuos predictivos se pueden calcular en términos de los residuos ordinarios de acuerdo con la relación siguiente :

$$e_{i(i)} = \frac{e_i}{1-h_{ii}} \qquad \text{..... (7)}$$

De tal relación y bajo el supuesto de normalidad en las perturbaciones estocásticas, podemos decir que el i-ésimo residuo predictivo se distribuye de la siguiente manera :

$$e_{i(i)} \sim N\left(0\,,\ \frac{\sigma^2}{1-h_{ii}}\right)$$

Es importante señalar que el i-ésimo residuo predictivo multiplicado por $\sqrt{1-h_{ii}}$ y dividido por la raíz cuadrada del cuadrado medio de los residuos exceptuando el i-ésimo caso coincide con el i-ésimo residuo estudentizado externo tal como se muestra a continuación :

$$\frac{e_{i(i)}\sqrt{1-h_{ii}}}{\hat{\sigma}_{(i)}} = \frac{e_{i(i)}\sqrt{1-h_{ii}}\sqrt{1-h_{ii}}}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

$$= \frac{e_{i(i)}(1-h_{ii})}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = t_i$$

Otra forma de escribir el i-ésimo residuo estudentizado externo en términos del i-ésimo residuo predictivo es la siguiente :

$$t_i = \frac{e_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1+h_{(i)ii}}} \quad ,$$

donde :

$$h_{(i)ii} = x_i' \left(X'_{(i)} X_{(i)} \right)^{-1} x_i$$

$X_{(i)}$ es la matriz $\mathbf{X}$ sin considerar el i-ésimo vector fila.

La forma alternativa para calcular el i-ésimo residuo estudentizado externo se desprende de la siguiente relación :

$$1 - h_{ii} = \frac{1}{1+h_{(i)ii}}$$

De la ecuación (7) se observa que cuando los elementos (a_{ii}) son pequeños la diferencia entre los residuos predictivos y los residuos ordinarios es mínima, sin embargo, si los elementos (a_{ii}) son grandes, existe una gran diferencia entre los residuos ordinarios y los residuos predictivos. Si los elementos (a_{ii}) son grandes puede resultar muy distinto verificar los supuestos con los residuos predictivos en lugar de utilizar los residuos ordinarios.

II.2.5 RESIDUOS RAO-BLACKWELL

El residuo que aquí se denomina Rao-Blackwell (O'Reilly and Quesenberry (1973, sección 4)) proviene del procedimiento que se describe a continuación :

1. Estimar insesgadamente la función de distribución normal estandarizada de la variable dependiente.

Bajo el supuesto de normalidad en las perturbaciones estocásticas la variable dependiente también se distribuye normal , es decir,

$$\varepsilon_i \sim N(0\ ,\ \sigma^2) \quad \Rightarrow \quad Y_i \sim N(\underline{x}_i'\underline{\beta}\ ,\ \sigma^2)$$

Demostración.

$$Y_i = \underline{x}_i'\underline{\beta} + \varepsilon_i$$

$$\Rightarrow M_{Y_i}(t) = E\left[e^{tY_i}\right] = E\left[e^{t(\underline{x}_i'\underline{\beta}+\varepsilon_i)}\right]$$
$$= e^{t\underline{x}_i'\underline{\beta}}\, E\left[e^{t\varepsilon_i}\right] = e^{t\underline{x}_i'\underline{\beta}}\, M_{\varepsilon_i}(t)$$

Como $\varepsilon_i \sim N(0\ ,\ \sigma^2) \quad \Rightarrow \quad M_{\varepsilon_i}(t) = e^{\frac{1}{2}t^2\sigma^2}$

y entonces $M_{Y_i}(t) = e^{t\underline{x}_i'\underline{\beta}-\frac{1}{2}t^2\sigma^2}$ *que es la función generadora de momentos de una distribución normal con media* $\underline{x}_i'\underline{\beta}$ *y varianza* σ^2.

$$\therefore \quad Y_i \sim N(\underline{x}_i'\underline{\beta}\ ,\ \sigma^2)$$

Entonces :

$$Z_i = \frac{Y_i - \underline{x}_i'\underline{\beta}}{\sigma} \sim N(0,1)$$

De esta manera la función de distribución de Z_i evaluada en z_i es :

$$F(z_i;\ \underline{x}_i'\underline{\beta}\ ,\ \sigma^2) = \Phi\left(\frac{y_i - \underline{x}_i'\underline{\beta}}{\sigma}\right)$$

Siendo :

$\Phi\left(\frac{y_t - \underline{x}_t'\underline{\beta}}{\sigma}\right)$ La función de distribución N(0,1) evaluada en z.

El estimador insesgado de varianza mínima de la función de distribución anterior resulta ser :

$$\hat{F}(z_i;\ \underline{x}_i'\underline{\beta},\ \sigma^2) = G\left\{\frac{(n-k-1)^{1/2}(y_i - \underline{x}_i'\hat{\underline{\beta}})}{\sqrt{(1-h_{ii})(n-k)\hat{\sigma}^2 - (y_i - \underline{x}_i'\hat{\underline{\beta}})^2}}\right\},$$

(Ver Ghurye and Olkin (1969 , página 1268))

donde la función G es la función de distribución t de Student con (n-k-1) grados de libertad y el residuo Rao-Blackwell (s_i) es precisamente el argumento de G pero sustituyendo el valor y_i por la variable aleatoria Y_i.

$$s_i = \frac{(n-k-1)^{1/2}\ (Y_i - \underline{x}_i'\hat{\beta})}{\sqrt{(1-h_{ii})(n-k)\hat{\sigma}^2 - (Y_i - \underline{x}_i'\hat{\beta})^2}}$$

2. Si se transforma una variable aleatoria con su propia función de distribución (continua) se sabe que la correspondiente variable aleatoria transformada sigue una distribución uniforme en el intervalo (0,1). Ahora bien, si se transforma a una variable aleatoria con su función de distribución estimadora Rao-Blackwell y ésta es continua, entonces la variable aleatoria transformada sigue una distribución uniforme en el intervalo (0,1) que se denota por U(0,1).

Por lo anterior $G(s_i)$ sigue una distribución U(0,1) y es obvio que s_i tiene como distribución a G ; es decir, s_i se distribuye como una t de Student con (n-k-1) grados de libertad.

Es importante mostrar que el residuo Rao-Blackwell coincide con el residuo estudentizado externo (t_i), para lo cual se realiza el desarrollo siguiente :

$$s_i = \frac{(n-k-1)^{\frac{1}{2}}\left(Y_i - x_i'\hat{\beta}\right)}{\sqrt{(1-h_{ii})(n-k)\hat{\sigma}^2 - \left(Y_i - x_i'\hat{\beta}\right)^2}}$$

$$= \frac{(n-k-1)^{\frac{1}{2}}\left(Y_i - x_i'\hat{\beta}\right)}{\sqrt{(1-h_{ii})\left[(n-k)\hat{\sigma}^2 - \frac{\left(Y_i - x_i'\hat{\beta}\right)^2}{(1-h_{ii})}\right]}}$$

$$= \frac{e_i}{\sqrt{\frac{(1-h_{ii})\left[(n-k)\hat{\sigma}^2 - \frac{e_i^2}{(1-h_{ii})}\right]}{(n-k-1)}}}$$

$$= \frac{e_i}{\sqrt{\frac{\left[(n-k)\hat{\sigma}^2 - \frac{e_i^2}{(1-h_{ii})}\right]}{(n-k-1)}}\sqrt{(1-h_{ii})}}$$

$$= \frac{e_i}{\sqrt{\hat{\sigma}^2_{(i)}}\sqrt{(1-h_{ii})}} = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{(1-h_{ii})}} = t_i$$

Considerando todos los residuos que se analizaron en este capítulo y la relación existente entre ellos, se puede señalar que el residuo de Rao-Blackwell que coincide con el residuo estudentizado externo es el que debe ser utilizado para realizar un análisis eficiente con respecto a la verificación de los supuestos del modelo de regresión lineal y particularmente en el supuesto de normalidad en las perturbaciones estocásticas ; que es el tema central de este trabajo. Lo anterior por las propiedades distribucionales ya mencionadas y por la base teórica que lo sustenta.

II . 3 PUNTOS DISCREPANTES " OUTLIERS " , PALANCA E INFLUYENTES.

Al considerar las observaciones de las variables que son utilizadas en un modelo de regresión lineal, es importante mencionar que pueden existir algunas de ellas que ocasionen severas alteraciones al modelo. Tales observaciones pueden ser de tres tipos y comúnmente se conocen como :

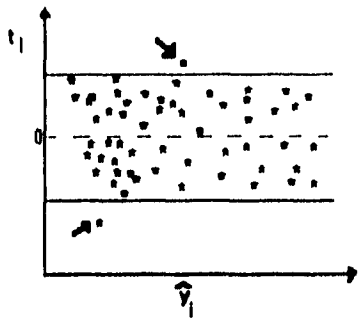
1 . *Puntos discrepantes "outliers".*
2 . *Puntos palanca.*
3 . *Puntos influyentes.*

II . 3 . 1 PUNTOS DISCREPANTES " OUTLIERS "

Los puntos discrepantes "outliers", son aquellas observaciones cuyos residuos salen de un rango de confianza previamente definido.

Estas observaciones discrepantes se detectan usualmente al graficar los residuos t_i , r_i o w_i contra cualquiera de las variables explicativas o contra la estimación de la variable dependiente.

Al realizar la gráfica los puntos discrepantes serán aquellas observaciones que abandonen la banda de confianza predeterminada alrededor del cero.



Para determinar los residuos que se deben utilizar para realizar las gráficas se puede considerar la relación que existe entre los residuos estudentizados internos y los residuos estandarizados :

$$r_i = \frac{w_i}{\sqrt{1-h_{ii}}}$$

Dado que h_{ii} es mayor que cero, se puede ver que el valor absoluto del residuo estudentizado interno es mayor que el valor absoluto del residuo estandarizado, por lo que r_i refleja mejor las desviaciones que w_i.

Al considerar la relación que existe entre los residuos estudentizados externos y los residuos estudentizados internos tenemos que :

$$t_i = r_i \sqrt{\frac{n-k-1}{n-k-r_i^2}} \quad ,$$

por lo que para un residuo estudentizado interno tal que $|r_i| > 1$, se tiene que el valor absoluto del residuo estudentizado externo es mayor que el valor absoluto del residuo estudentizado interno. Por lo tanto se puede ver que los residuos estudentizados externos reflejan más dramáticamente grandes desviaciones que los residuos estudentizados internos.

Bajo el supuesto de normalidad en las perturbaciones estocásticas se puede decir que los residuos estudentizados externos se distribuyen como una t de Student con n-k-1 grados de libertad , por lo cual si utilizamos dichos residuos para detectar observaciones discrepantes " outliers " se puede determinar la banda de confianza en base al valor t_α con el siguiente criterio :

> Si el valor absoluto de t_i excede el valor t_α, la observación i - ésima se considera una observación discrepante , donde :
>
> t_α es el valor de la variable t obtenida de la distribución t de Student con n-k-1 grados de libertad al nivel de significancia α .

Cuando detectamos observaciones discrepantes, se procede a detectar el origen de dichas observaciones. Si los puntos discrepantes provienen de errores en la recopilación de las observaciones, estas son anuladas y se procede a realizar el análisis sin considerarlas.

II.3.2 PUNTOS PALANCA

Los puntos palanca son (Belsley, Kuh and Welsch (1980 , capítulo 2)) son aquellas observaciones de las variables explicativas (vector columna z_j) que generan un elemento h_{ii} demasiado grande. Estas observaciones causan alteraciones en el modelo debido a que la relación que existe entre las perturbaciones estocásticas ε_i y su estimación e_i es la siguiente :

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

Considerando que e_i es una estimación de ε_i, es indispensable que los elementos h_{ii} de la matriz (H) sean pequeños, en caso contrario se ocasionan alteraciones en el modelo.

Para determinar si alguna observación de variables explicativas genera un punto palanca es importante considerar la siguiente relación :

$$\sum_{i=1}^{n} h_{ii} = k$$

De la relación anterior se puede establecer que el valor " promedio " de los elementos de la diagonal de la matriz (H) es $\left(k/n \right)$. De tal manera el valor $\left(2k/n \right)$ es utilizado como medida puntual siendo el límite superior y se dice que si un elemento h_{ii} rebasa tal límite se considera como un punto palanca .

En los párrafos anteriores se habla de puntos palanca generedos por una observación de variables explicativas, sin embargo, puede suceder que tales puntos palanca se originen por una variable explicativa en particular, a dicha ocurrencia se le danomina palanca parcial.

Para explicar el significado de palanca parcial es necesario introducir el concepto de matriz de predicción por lo cual se desarrolla a continuación :

La matriz (X^{*}) es aquella en la cual a la matriz (X) se le aumenta el vector (Y) por lo cual se obtienen una matriz de la siguiente forma :

$$(X^{*}) = (X : Y) \quad ,$$

donde :

(X^{*}) es la matriz de predicción de $[n \times (k+1)]$ de la matriz (X).

De tal forma, la correspondiente matriz de predicción (H^{*}) de (X^{*}) está dada por :

$$H^{*} = H + \frac{(I_n - H)YY'(I_n - H)}{Y'(I_n - H)Y}$$

Por lo tanto :

$$h_{ii}^{*} = h_{ii} + \frac{e_i^2}{\varepsilon'\varepsilon} \qquad i = 1,\ldots,n \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (8)$$

Ahora consideremos a la matriz de predicción construida de la siguiente forma :

$$X^{P} = \left(X_{[j]} : \varepsilon_j\right) \ ,$$

donde :

X^{P} es la matriz (X).

$X_{[j]}$ es la matriz (X) sin considerar el vector (ε_j) que contiene los datos de la j-ésima variable explicativa .

Por lo cual desarrollando un análisis similar tenemos :

$$H^{P} = H_{[j]} + \frac{\left(I_n - H_{[j]}\right)\varepsilon_j \varepsilon_j'\left(I_n - H_{[j]}\right)}{\varepsilon_j'\left(I_n - H_{[j]}\right)\varepsilon_j} \ ,$$

donde :

H^{P} es la matriz H .

Por lo tanto, tenemos que :

$$h_{ii}^{P} = h_{ii\,[j]} + \delta_{ij} \qquad i = 1,\ldots,n \quad \ldots\ldots\ldots\ldots\ldots\ldots \quad (9),$$

donde :

h_{ii}^{P} son los elementos de la diagonal de la matriz (H^{P}).

$h_{ii\,[j]}$ son los elementos de la diagonal de la matriz ($H_{[j]}$) que se calcula sin considerar los datos del vector (ε_j).

De tal forma :

$$\delta_{ij} = h^{*}_{ii} - h_{ii\,[j]} \quad ,$$

siendo :

$$\delta_j = (\delta_{1j}, \delta_{2j}, \ldots, \delta_{nj})'$$

El vector δ_j se conoce como el vector normalizado de residuos cuadrados obtenidos de la regresión de la variable explicativa (ξ_j) sobre todas las demás variables explicativas.

Para determinar si alguna observación de cualquier variable explicativa genera un punto palanca, es importante considerar la siguiente relación :

$$\sum_{i=1}^{n} \delta_{ij} = 1$$

De la anterior relación se puede establecer que el valor promedio de los elementos δ_{ij} es ($1/n$). Por lo tanto el valor ($2/n$) es utilizado como medida puntual siendo el límite superior y se dice que si un elemento δ_{ij} rebasa dicho límite se detecta un punto palanca en la i-ésima observación de la j-ésima variable explicativa.

Comparando las ecuaciones (8) y (9) se puede decir que un punto palanca es una observación discrepante en el espacio de las variables explicativas.

Observando la ecuación (8) se aprecia que el valor de h^{*}_{ii} es grande por alguna de las siguientes razones :

1. El valor de h_{ii} es grande, es decir, existe un punto palanca.

2. La observación i-ésima en las variables es una observación discrepante " outlier ".

Si se encuentra un punto palanca en la i-ésima observación de las variables explicativas, el problema se puede solucionar realizando una transformación sobre las variables explicativas. Si al realizar un análisis se llega a la conclusión de que la palanca es parcial, se aplica una transformación sobre la variable explicativa que ocasione el problema.

II . 3 . 3 PUNTOS INFLUYENTES

Los puntos influyentes (Belsley, Kuh and Welsch (1980 , capítulo 2)) son aquellas observaciones que individual o conjuntamente demuestran mayor impacto en los valores estimados que el resto de las observaciones.

A partir de la definición anterior se pueden describir dos tipos de influencia :

a) influencia parcial
b) influencia conjunta

a) INFLUENCIA PARCIAL

La influencia parcial se detecta en la observación que individualmente demuestra mayor impacto en los valores estimados, que el resto de las observaciones.

Para determinar si la i-ésima observación debe ser considerada como una observación influyente, se utiliza como medida la distancia de Welsh - Kuh denominada $|DFFITS_i|$. Esta es una distancia que refleja la diferencia que existe al realizar la estimación de la variable dependiente Y_i considerando y sin considerar la i-ésima observación ; de tal forma que mientras mayor es $|DFFITS_i|$, mayor es la influencia de la i-ésima observación.

$$DFFITS_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\hat{Var}(\hat{Y}_i)}} \quad ,$$

donde :

$\hat{Y}_i$ es la estimación de la i-ésima observación de la variable dependiente

$\hat{Y}_{i\,(i)}$ es la estimación de la i-ésima observación de la variable dependiente que se calcula sin utilizar tal observación de las variables por medio del vector estimado ($\hat{\underline{\beta}}_{(i)}$).

Por lo tanto :

$$DFFITS_i = \frac{\underline{x}'_i \left[\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)}\right]}{\sigma\sqrt{h_{ii}}} = \frac{\underline{x}'_i \left[\hat{\underline{\beta}} - \left(\hat{\underline{\beta}} - \frac{(X'X)^{-1}\underline{x}_i\, e_i}{1-h_{ii}}\right)\right]}{\sigma\sqrt{h_{ii}}}$$

Como σ es desconocida se utiliza $\hat{\sigma}_{(i)}$ como estimador, quedando la distancia de Welsh - Kuh de la siguiente manera :

$$DFFITS_i = \left[\frac{h_{ii}}{1-h_{ii}}\right]^{1/2} t_i$$

Suponiendo un balance perfecto en la matriz (X), el valor para $|DFFITS_i|$ debe ser $\sqrt{k/n}$. De tal manera el valor $2\sqrt{k/n}$ es utilizado como medida puntual para determinar si la i - ésima observación es influyente , y se dice que si $|DFFITS_i|$ es mayor que $2\sqrt{k/n}$ dicha observación es influyente.

Es importante considerar la influencia de una observación con respecto a un sólo coeficiente, para lo cual se utiliza la siguiente medida denominada $\left| DFBETAS_{ji} \right|$ que es una medida que refleja la diferencia que existe en la estimación del coeficiente β_j al considerar la i-ésima observación y al no considerarla ; de tal forma se obtiene la influencia de la i-ésima observación en el coeficiente ($\hat{\beta}_j$).

$$DFBETAS_{ji} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{\hat{Var}(\hat{\beta}_j)}} \quad ,$$

donde :

$\hat{\beta}_j$ es la estimación del coeficiente β_j.

$\hat{\beta}_{j(i)}$ es la estimación del coeficiente β_j sin considerar la i-ésima observación.

Se tiene que:

$$\hat{\beta} - \hat{\beta}_{(i)} = \frac{(X'X)^{-1} x_i' e_i}{(1 - h_{ii})} \quad ,$$

por lo tanto , para obtener $\hat{\beta}_j - \hat{\beta}_{j(i)}$ se considera el elemento (j , i) de la matriz:

$$C = (X'X)^{-1} X' \quad ,$$

y se le denomina C_{ji} , de tal forma :

$$\hat{\beta}_j - \hat{\beta}_{j(i)} = \frac{C_{ji}}{1 - h_{ii}} e_i$$

Por otro lado $\sqrt{Var\ (\hat{\beta}_j)}$ se estima a través del producto entre la raíz cuadrada del j-ésimo elemento en la diagonal de la matriz $(X'X)^{-1}$ al cual se le denomina a_{jj} y la raíz cuadrada de $\hat{\sigma}^2_{(i)}$ que es un estimador de σ^2.

Por lo tanto :

$$DFBETAS_{ji} = \frac{C_{ji}\, e_i}{(1-h_{ii})\sqrt{\hat{\sigma}^2_{(i)}}\, a_{jj}} = \frac{C_{ji}\, t_i}{\sqrt{(1-h_{ii})}\, a_{jj}}$$

El valor $2/\sqrt{n}$ es utilizado como medida puntual para determinar la influencia citada anteriormente, de tal manera si $|DFBETAS_{ji}|$ excede dicho valor, se dice que la i-ésima observación es influyente con respecto al coeficiente ($\hat{\beta}_j$).

b) INFLUENCIA CONJUNTA

La influencia conjunta se detecta en aquellas observaciones que conjuntamente demuestran mayor impacto en los valores estimados que el resto de las observaciones. Puede suceder que al analizar individualmente una observación no presente influencia parcial, pero si dicha observación es tomada en conjunto con otras, puede ser altamente influyente.

Para detectar influencia conjunta se puede utilizar la generalización para las medidas analizadas anteriormente que ayudan a detectar influencia parcial. Dicha generalización consiste en utilizar coeficientes que se calculan sin considerar en tal estimación un subconjunto de (m) observaciones.

II . 4 TIPOS DE GRAFICAS

Para poder verificar si se cumplen los supuestos del modelo se realizan distintos tipos de gráficas tomando como base los residuos. Debido a que los diferentes tipos de residuos se encuentran relacionados entre ellos, se pueden elaborar las gráficas utilizando los residuos estudentizados externos ya que estos como se mencionó anteriormente reflejan más dramáticamente las desviaciones y no dependen de las unidades del contexto del problema por lo cual ayudan a realizar un mejor análisis que los otros residuos.

A continuación se mencionan algunos tipos de gráfices que son las más utilizadas para verificar si se cumplen los supuestos del modelo.

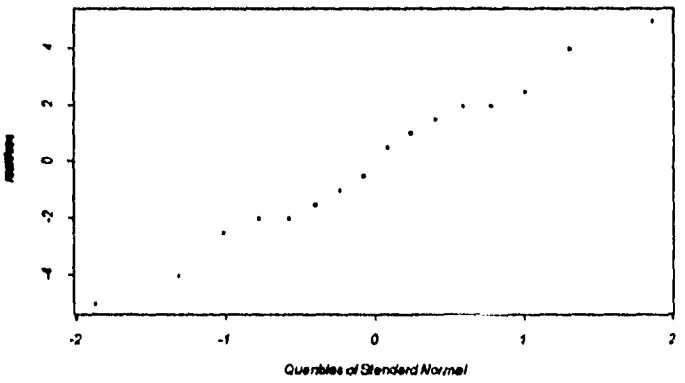
II . 4 . 1 GRAFICA SOBRE PAPEL NORMAL

En este tipo de gráfica se puede observar si el supuesto de normalidad en los errores es violado , para lo cual se grafican los residuos estandarizados en un papel especial "papel normal" que presente una escala relacionada con los porcentiles de una distribución normal estándar.

Para poder determinar si el supuesto de normalidad en los errores es violado se observa el comportamiento de los residuos al ser graficados . Si los residuos se presentan de manera semejante a una línea recta se dice que las perturbaciones estocásticas ε_i se distribuyen normalmente, en caso contrario, es decir, que los residuos presenten un comportamiento semejante a una línea curva se dice que el supuesto de normalidad no se cumple en las perturbaciones estocásticas. Para mayor información sobre el uso de "papel normal" , véase por ejemplo , el libro de D' Agostino y Stephens (1986, capítulo 2).

La violación de este supuesto se puede solucionar utilizando transformaciones en las variables.

GRAFICA SOBRE PAPEL NORMAL



No hay una clara evidencia para considerar que las perturbaciones estocásticas no se distribuyen normalmente.

II . 4 . 2 GRAFICA DE RESIDUOS CONTRA VALORES AJUSTADOS

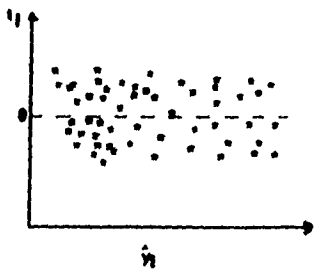
En este tipo de gráfica se pueden obtener conclusiones con respecto a dos supuestos del modelo de regresión lineal.

1) Varianza constante en las observaciones
2) Especificación correcta del modelo.

En este caso se grafican los residuos contra los valores ajustados de la variable dependiente. La razón por la cual se utilizan en la gráfica los valores ajustados de la variable dependiente y no la variable dependiente se debe a que los vectores $(\underline{e})$ y $(\underline{Y})$ están correlacionados y el análisis se puede ver alterado, sin embargo, los vectores $(\underline{e})$ y $(\underline{\hat{Y}})$ no presentan correlación por lo que se puede realizar un análisis correcto a través de ellos.

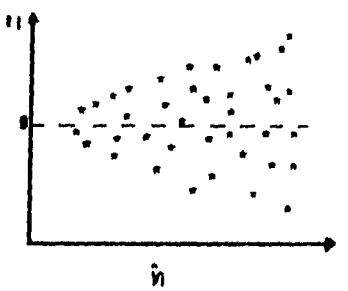
Al realizar la gráfica se pueden presentar diferentes comportamientos significativos que se resumen principalmente en tres y que se enlistan a continuación :

a) Si los residuos se presentan en la gráfica dentro de una banda horizontal sin comportamiento específico alguno, se dice que la varianza es constante en las observaciones y la especificación del modelo es adecuada por lo cual el análisis de mínimos cuadrados es válido.



Los residuos se presentan en una banda horizontal sin comportamiento específico alguno.

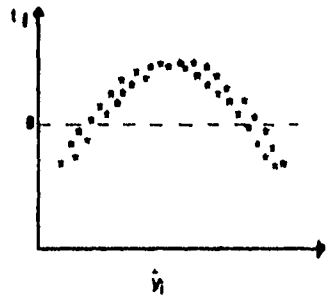
b) Si los residuos no se presentan en la gráfica en una banda horizontal se dice que la varianza no es constante.



Los residuos abandonan la banda horizontal, en este caso la varianza se incrementa con los valores de $\hat{y}_i$.

La violación de este supuesto se puede solucionar por medio de la aplicación de mínimos cuadrados generalizados (considerar varianza no constante) o por medio de la realización de una transformación en la variable dependiente buscando evitar que los residuos abandonen la banda horizontal.

c) Si los residuos siguen un comportamiento específico se dice que el modelo es inadecuado ya que se presume que los errores absorben la influencia de un error de especificación en el modelo.



Modelo inadecuado

Para solucionar la violación del supuesto presentado en la gráfica anterior (especificación incorrecta del modelo) se pueden incluir en el modelo términos cruzados y / o cuadráticos en las variables explicativas o también se puede aplicar una transformación en la variable dependiente.

II . 4 . 3 GRAFICA DE RESIDUOS CONTRA CADA VARIABLE EXPLICATIVA

Este tipo de gráfica es utilizado para obtener conclusiones con respecto a los supuestos de varianza constante y especificación correcta del modelo, de igual manera que el tipo de gráfica anterior, sólo que en este caso se grafican los residuos contra cada variable explicativa para realizar un análisis con respecto a cada una de estas variables. El objetivo de estas gráficas consiste en aislar la influencia individual de cada una de las variables explicativas.

Al realizar la gráfica se pueden presentar en los residuos los mismos comportamientos (a), (b) y (c) descritos en el tipo de gráfica anterior , pero ahora contra cada variable explicativa, obteniendo las mismas conclusiones y realizando los mismos ajustes para solucionar el problema en caso de que alguno de los supuestos sea violado.

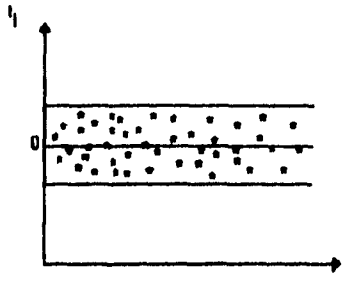
La diferencia al realizar este tipo de gráfica radica principalmente en que si el modelo no esta correctamente especificado y se requiere la aplicación de una transformación en las variables explicativas, podemos determinar cuales de estas requieren dicha transformación.

II . 4 . 4 GRAFICA DE RESIDUOS CONTRA EL TIEMPO

Las gráficas de este tipo se utilizan para verificar los supuestos de varianza constante y especificación correcta del modelo. Para realizar ésta gráfica es necesario conocer el orden en que fueron obtenidos los datos de las variables explicativas que se están utilizando en el modelo debido a que se grafican los residuos contra el tiempo.

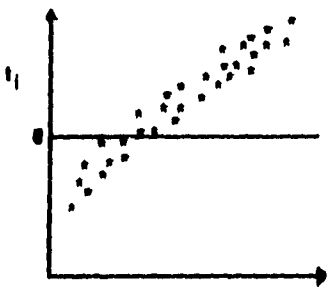
Al realizar ésta gráfica se pueden presentar diferentes comportamientos :

a) Si los residuos se presentan en la gráfica dentro de una banda horizontal sin comportamiento especifico alguno, se dice que la varianza es constante en las observaciones y la especificación del modelo es correcta.



Los residuos se presentan en una banda horizontal y sin comportamiento especifico alguno a través del tiempo.

b) Si los residuos al ser graficados se van incrementando o decrementando a través del tiempo, se dice que la varianza de las perturbaciones estocásticas no es constante.

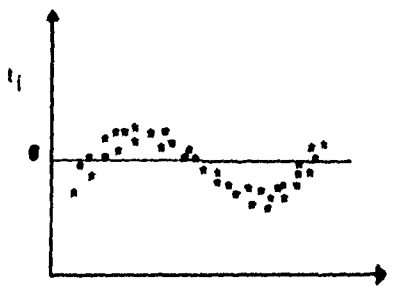


tiempo

Los residuos se van incrementando con respecto al tiempo.

La violación de este supuesto se puede solucionar por medio de la aplicación de mínimos cuadrados generalizados.

c) Si al graficar los residuos estos presentan un comportamiento similar a términos cruzados y / o cuadráticos en el tiempo, se dice que el supuesto de especificación correcta del modelo esta siendo violado.



tiempo

Los residuos presentan términos cruzados a través del tiempo.

La violación de este supuesto se puede solucionar modificando el modelo con términos cruzados y / o cuadráticos en las variables .

Para verificar el supuesto de normalidad en las perturbaciones estocásticas existen algunos métodos, entre los cuales se encuentra la gráfica sobre papel normal mencionada brevemente en este capítulo. En lo que resta de este trabajo, se describe un procedimiento más objetivo para determinar si las perturbaciones estocásticas se distribuyen normalmente en los modelos de regresión lineal múltiple, tal procedimiento se menciona en el siguiente capítulo.

CAPITULO III

METODOS DE COMPROBACION DEL SUPUESTO DE NORMALIDAD

III . 1 SUPUESTO DE NORMALIDAD EN LAS PERTURBACIONES ESTOCASTICAS.

El supuesto de normalidad en las perturbaciones estocásticas es una hipótesis muy importante en el modelo de regresión lineal múltiple, ya que en tal consideración se basan la prueba de significancia global del modelo y las pruebas de significancia para cada variable explicativa del mismo. Para poderlo verificar es necesario contar con métodos que nos ayuden a determinar si el comportamiento de las perturbaciones estocásticas se puede asociar a la distribución normal.

Tal como se mencionó en el capítulo anterior, uno de los métodos utilizados para comprobar el supuesto de normalidad es el conocido como " Gráfica sobre papel normal ", sin embargo, en este método la decisión depende de la apreciación del individuo que está haciendo el análisis, por lo que no es un método que establezca objetiva y estadísticamente, el resultado de lo que se desea probar.

III . 2 PRUEBA BERA-JARQUE.

Existe un método muy utilizado por economistas para comprobar el supuesto de normalidad en las perturbaciones estocásticas que se conoce como la prueba Bera-Jarque (Judge and Hill (1988 , capítulo 22)) y tiene la siguiente interpretación :

Considerando el modelo de regresión lineal :

$$\underline{Y} = X\underline{\beta} + \underline{\varepsilon} \quad ,$$

donde:

$$E(\underline{\varepsilon}) = 0 \quad , \qquad E(\underline{\varepsilon}\,\underline{\varepsilon}') = \sigma^2 I_n \quad ,$$

y bajo la hipótesis :

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

entonces el tercer y cuarto momentos para un elemento en $\underline{\varepsilon}$ dados respectivamente por :

$$\mu_3 = E(\varepsilon_i^3) = 0 \qquad y \qquad \mu_4 = E(\varepsilon_i^4) = 3\sigma^4.$$

Esta prueba de normalidad se basa en considerar qué tan lejos se encuentran las estimaciones del tercer y cuarto momentos, $\hat{\mu}_3$ y $\hat{\mu}_4$, de los correspondientes en una normal ; esto es , de 0 y $3\hat{\sigma}^4$ respectivamente , donde :

$$\hat{\sigma}^2 = \frac{1}{n}\sum_1^n e_i^2$$

Para realizar esta prueba se consideran las versiones a escala de $\hat{\mu}_3$ y $\hat{\mu}_4$ que se conocen como medidas de sesgo y curtosis respectivamente.

La medida de sesgo de una distribución se refiere a su grado de asimetría y está dada por :

$$\sqrt{b_1} = \frac{\mu_3}{\sigma^3} \quad ,$$

mientras que la curtosis de una distribución se refiere tanto a lo picudo de ésta como al peso de sus colas y está dada por :

$$b_2 = \frac{\mu_4}{\sigma^4}$$

Como la distribución normal tiene una curtosis igual a 3, se puede decir que una medida del exceso de curtosis es :

$$b_2 - 3 = \frac{\mu_4 - 3\sigma^4}{\sigma^4}$$

La prueba Bera-Jarque es un método conjunto que consiste en verificar si las estimaciones de $\sqrt{b_1}$ y/o $(b_2 - 3)$ son significativamente distintas de cero, en cuyo caso podemos afirmar que las perturbaciones estocásticas no se distribuyen normalmente .Dado que $\underline{\varepsilon}$ no es observable, las estimaciones de $\sqrt{b_1}$ y $(b_2 - 3)$ se realizan a través de los residuos, por lo cual se obtienen los estimadores siguientes:

$$\tilde{\mu}_3 = \frac{1}{n}\sum_1^n e_i^3 \qquad \tilde{\mu}_4 = \frac{1}{n}\sum_1^n e_i^4$$

$$\sqrt{b_1} = \frac{\tilde{\mu}_3}{\tilde{\sigma}^3} \qquad (b_2 - 3) = \frac{\tilde{\mu}_4 - 3\tilde{\sigma}^4}{\tilde{\sigma}^4}$$

La prueba Bera-Jarque queda estructurada para el problema de prueba de hipótesis :

Ho : Las perturbaciones estocásticas se distribuyen normalmente. vs.

Ha : Las perturbaciones estocásticas no se distribuyen normalmente.

La estadística Bera-Jarque que está dada por:

$$\lambda = n\left(\frac{\left[\sqrt{b_1}\right]^2}{6} + \frac{[b_2 - 3]^2}{24}\right) ,$$

y bajo la hipótesis nula de que las perturbaciones estocásticas se distribuyen normalmente, se distribuye asintóticamente como $\chi^2_{(2)}$ y tiende a ser grande si $\sqrt{b_1}$ y/o $(b_2 - 3)$ difieren significativamente de cero. Se aclara que para n finito pueden (y deben) utilizarse los puntos de rechazo obtenidos por Monte Carlo que aparecen en el libro de Jarque y Bera (1987, pág 169).

La prueba Bera-Jarque se base en el parecido del tercer y cuarto momentos con los de una distribución normal por lo que si la no-normalidad se debe a otres diferencias, esta prueba no las detectaría.

III . 3 METODO FDE DE BONDAD DE AJUSTE.

El problema de bondad de ajuste se considera formalmente como un problema de prueba de hipótesis en el cual la distribución (F) es miembro de una familia no paramétrica $\mathcal{F}$ como lo puede ser la familia de todas las distribuciones continuas, o en su caso, de todas las distribuciones discretas. Para introducir el problema de bondad de ajuste se utilizará el caso simple con (F) una función de distribución continua y tomando como base una muestra $X = (X_1, X_2, \ldots, X_n)$ de la mencionada función de distribución de la siguiente forma:

Probar :

$$H_0: F = F_0 \qquad vs. \qquad H_1: F \neq F_0 \quad ,$$

donde :

F_0: es un sólo elemento perfectamente identificado de $\mathcal{F}$.

H_0: hipótesis nula.

H_1: hipótesis alterna.

En los métodos de bondad de ajuste para probar (H_0), la idea es contrastar la hipótesis con la evidencia, representadas en este caso por (F_0) y la muestra $X = (X_1, X_2, \ldots, X_n)$ respectivamente.

Asociada a cualquier muestra siempre existe una función de distribución conocida como función de distribución empírica ($F_n(x)$) la cual permite representar la evidencia muestral y se defina de la siguiente manera :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[x_i \leq x]} \quad ,$$

donde :

$1_{[c]}$: es la indicadora del evento c .

Existen varios resultados relacionados con el comportamiento asintótico de la función de distribución empírica, siendo uno de estos el lema de Glivenko - Cantelli (O'Reilly Togno Federico (1990 , sección 3)) que afirma que bajo (H_0) :

$$D_n = Sup_x |F_n(x) - F_0(x)| \to 0 \quad \textit{cuando } n \to \infty ,$$
$$\textit{con probabilidad 1} .$$

Lo anterior sugiere que si se cumple (H_0), la función de distribución empírica ($F_n(x)$) y ($F_0(x)$) como funciones, deben estar muy cerca, es decir, al contrastar la evidencia contra la hipótesis la diferencia debe ser mínima. La estadística o métrica (D_n) es conocida como la estadística del supremo.

Para cualquier estadística o métrica de este tipo $\gamma_n(F_n(x), F_0(x))$ se busca obtener su distribución bajo (H_0) para poder evaluar la probabilidad de que (γ_n) tome un valor " grande ", caso en el que podríamos cuestionar si (H_0) es cierta. El motivo por el cual se desea conocer la probabilidad de que la estadística (γ_n) tome un valor " grande " siendo (H_0) cierta es para controlar la probabilidad de rechazar la hipótesis nula (H_0) equivocadamente.

El cómputo para obtener las distribuciones exactas de una métrica (γ_n)suele ser muy complicado, por ello el uso de las distribuciones asintóticas que son presentadas, usualmente, como tablas con valores críticos para niveles de significancia preespecificados.

Al quedar definido el problema de bondad de ajuste y establecer los elementos utilizados en tal problema, a continuación se desarrollan los pasos básicos que constituyen el procedimiento seguido para la bondad de ajuste en el caso simple:

1. Transformar cada (X_i) perteneciente a la muestra con (F_0), esto es, $\varepsilon_i = F_0(X_i)$, y obtener ($\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$).

El objetivo de este paso es usar la transformación de una variable aleatoria continua con su propia función de distribución (considerando (H_0) cierta), resaltando la idea de transformar al [0,1] y entonces probar uniformidad.

2. Calcular la estadística (γ_n) con la cual se decidió trabajar.

La estadística (γ_n) que se elija debe ser adecuada según algún criterio (potencia , factibilidad ,etc.) .

3. Comparar el resultado del valor de la estadística con los valores de las tablas correspondientes al nivel de significancia previamente seleccionado.

4. Tomar la decisión de aceptar o rechazar la hipótesis nula (H_0) con base en lo observado en el paso anterior.

El problema compuesto de bondad de ajuste es similar al problema simple pero en presencia de parámetros desconocidos.

Al pasar del procedimiento desarrollado para el problema de bondad de ajuste simple al compuesto, la extensión más utilizada sólo requiere de la modificación del primer paso en cuanto al cómputo de la estadística , y el uso en el tercer paso de las tablas apropiadas que utilizan la distribución asintótica que corresponde a este caso compuesto. Para el paso uno, se emplean estimadores apropiados sustituyendo a $F_0(\cdot,\theta)$ por $F_0(\cdot,\hat{\theta})$, con $\hat{\theta}$ el vector de parámetros estimados.

III . 4 ESTADISTICAS BASADAS EN LA FUNCION DE DISTRIBUCION EMPIRICA.

Para realizar pruebas de bondad de ajuste se pueden utilizar varias estadísticas o métricas del tipo $\gamma_n(F_n(x), F_0(x))$ que surgen a partir de la idea de contrastar la función de distribución empírica con la función de distribución dada por (H_0). Estas métricas como se observa , se basan en la función de distribución empírica (FDE) y se clasifican según su construcción en las siguientes dos clases :

1. Estadísticas del Supremo.
2. Estadísticas Cuadráticas.

Las estadísticas del supremo (D'Agostino and Stephens (1986 , capítulo 4)) resaltan puntos en los cuales la diferencia entre la función de distribución empírica y la función de distribución propuesta en (H_0) es mayor. Las mas utilizadas son D^+, D^-, D, V y se calculan de la siguiente forma:

Para calcular la estadística (D^+) se utilizan los valores en los puntos tales que ($F_n(x)$) es mayor que ($F_0(x)$), de tal forma :

$$D^+ = Sup_x\{F_n(x) - F_0(x)\}$$

Para calcular la estadística (D^-) se utilizan los valores en los puntos tales que ($F_0(x)$) es mayor que ($F_n(x)$), de tal forma :

$$D^- = Sup_x\{F_0(x) - F_n(x)\}$$

La estadística (D), fue introducida por Kolmogorov y es la métrica más utilizada de esta clase. Su cálculo requiere únicamente de considerar el resultado mayor entre las dos estadísticas anteriores, de tal manera :

$$D = Max(D^+, D^-) = Sup_x|F_n(x) - F_0(x)|$$

La estadística (V) es utilizada para puntos en un círculo, fue introducida por Kuiper y se calcula sumando (D^+) y (D^-) , de tal manera :

$$V = D^+ + D^-$$

Las estadísticas cuadráticas (D'Agostino and Stephens (1986 , capítulo 4)) están dadas por la familia de Crámer - Von - Mises que se define de la siguiente manera :

$$Q = n\int_{-\infty}^{+\infty}\{F_n(x) - F_0(x)\}^2\psi(x)dF_0(x) \quad ,$$

donde :

$\psi(x)$ es una función que se elige según el peso que se le desea dar a la diferencia $\{F_n(x) - F_0(x)\}^2$ para cada valor de X .

Cuando $\psi(x) = 1$, resulta la estadística de Crámer- Von - Mises, conocida como W^2 y cuando $\psi(x) = [\{F_0(x)\}\{1 - F_0(x)\}]^{-1}$ se obtiene la estadística Anderson - Darling comúnmente denotada por A^2; que otorga mucha importancia a las discrepancias que ocurran en valores muy grandes de X (en valor absoluto) ; esto es, en las colas.

La estadística Anderson - Darling es la más utilizada y adecuada para realizar pruebas de normalidad en bondad de ajuste como lo han mostrado numerosos estudios de simulación reportados en literatura.

La velocidad de convergencia de las distribuciones exactas (n finito) a la asintótica es extraordinaria, y basta mencionar el caso de la estadística Anderson - Darling para la cual se recomienda el uso de su distribución límite para $n \geq 5$ en el caso simple. La distribución límite tabulada extraída del libro de D'Agostino y Stephens (1986) para el caso de prueba de una distribución normal con ambos parámetros desconocidos es :

PRUEBA A^2

(CASO NORMAL CON AMBOS PARAMETROS DESCONOCIDOS)

α (NIVEL)							
0.50	0.25	0.15	0.10	0.05	0.025	0.01	0.005
VALORES CRITICOS							
0.341	0.470	0.561	0.631	0.752	0.873	1.035	1.159

(D'Agostino and Stephens (1986, pag. 123))

Esta misma distribución es la utilizada para el caso de la prueba de normalidad en regresión (ver O' Reilly, 1993).

III . 5 APLICACION DE BONDAD DE AJUSTE EN LA COMPROBACION DEL SUPUESTO DE NORMALIDAD.

Para comprobar el supuesto de normalidad en las perturbaciones estocásticas del modelo de regresión lineal múltiple se utilizará un método de bondad de ajuste que use como muestra a los residuos Rao-Blackwell por sus propiedades y como estadística a la Anderson - Darling (A^2) por ser la más adecuada.

El problema de bondad de ajuste para comprobar el supuesto de normalidad en las perturbaciones estocásticas queda planteado de la siguiente manera :

Probar :

H_0 : Las perturbaciones estocásticas $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ se distribuyen normalmente $N(0, \sigma^2)$.

vs.

H_1 : Las perturbaciones estocásticas $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ no se distribuyen normalmente $N(0, \sigma^2)$.

Al quedar planteado el problema de bondad de ajuste se procede de acuerdo a los siguientes pasos :

1. Se calculan los residuos Rao-Blackwell del modelo de regresión lineal múltiple.

$$s_i = \frac{(n-k-1)^{1/2}\left(Y_i - x_i'\hat{\beta}\right)}{\sqrt{(1-h_{ii})(n-k)\hat{\sigma}^2 - \left(Y_i - x_i'\hat{\beta}\right)^2}} \qquad i = 1, 2, \ldots, n-1$$

Los residuos Rao-Blackwell s_i tal como se mencionó en el capítulo anterior se distribuyen como una t de Student con ($n-k-1$) grados de libertad, pero no son independientes.

2. Se calculan las estadísticas de orden $s_{(i)}$; esto es , se ordenan de menor a mayor los referidos residuos :

$$s_{(i)} \leq s_{(i+1)} \qquad i = 1, 2, \ldots, n-1$$

3. Se calculan los valores $z_{(i)}$ transformando los residuos Rao-Blackwell ordenados con la función de distribución t de Student con ($n-k-1$) grados de libertad (G_{n-k-1}) de la siguiente manera :

$$z_{(i)} = G_{n-k-1}\left(s_{(i)}\right) \quad ,$$

donde :

$G_\gamma(*)$ es la función de distribución de una variable aleatoria t de Student con γ grados de libertad, evaluada en $*$.

El paso anterior con el objeto de utilizar la transformación de una variable aleatoria continua con su propia función de distribución (transformación con la integral de probabilidad) y al transformar al [0 , 1] probar uniformidad en los $z_{(i)}$ obtenidos.

4. Calcular el valor de la estadística Anderson - Darling A^2 ; aquí referida como $\overline{A}^2$, y utilizar la tabla ya mencionada.

$$\overline{A}^2 = -n - \left(\frac{1}{n}\right)\left(\sum_{i=1}^{n}(2i-1)\left[\ln\left(z_{(i)}\right)+\ln\left(1-z_{(n+1-i)}\right)\right]\right)$$

5. Elegir el nivel de significancia (α) que determina la probabilidad de rechazar la hipótesis nula (H_0) siendo verdadera.

6. Comparar el resultado de la estadística $\overline{A}^2$ con el valor crítico de la tabla de distribución límite presentada en la sección anterior.

CAPITULO IV

EJEMPLOS

Para ilustrar la teoría que se muestra en las secciones anteriores, se presentan a continuación algunos ejemplos que fueron extraídos de libros especializados en el análisis de regresión con el objeto de llevar a cabo algunas pruebas y en particular la prueba sobre la distribución normal, que fue detallada en el capítulo anterior y de tal forma, obtener conclusiones.

El primer ejemplo se extrajo de Lyman Ott (1984, p. 475) ; del libro "An Introduction to Statistical Methods and Data Analysis".

EJEMPLO 1.

Una compañía está interesada en desarrollar un modelo de regresión que le permita predecir adecuadamente las ventas mensuales de sus automóviles (estándar y de lujo) efectuadas en una ciudad determinada.

Tomando como base un análisis empírico, se utilizan los datos de muestreos realizados en la ciudad por esa compañía para los siguientes conceptos:

1) Número de Ventas mensuales (en miles).

2) Precio de la Gasolina por galón (dólares).

3) Tasa de interés aplicable al comprar una unidad a plazo (%).

4) Modelo del automóvil (1) estándar ó 0) de lujo).

Los datos obtenidos durante los dieciocho meses anteriores por estos conceptos para los automóviles estándar y de lujo se presentan a continuación:

MES	NUMERO DE VENTAS (V)	PRECIO DE LA GASOLINA (PG)	TASA DE INTERES (TI)	MODELO DEL AUTOMOVIL (M)
1	22.1	1.39	12.1	1
1	7.2	1.39	12.1	0
2	15.4	1.44	12.2	1
2	5.4	1.44	12.2	0
3	11.7	1.45	12.3	1
3	7.6	1.45	12.3	0
4	10.3	1.32	14.2	1
4	2.5	1.32	14.2	0
5	11.4	1.35	15.8	1
5	2.4	1.35	15.8	0
6	7.5	1.28	16.3	1
6	1.7	1.28	16.3	0
7	13.0	1.26	16.5	1
7	4.3	1.26	16.5	0
8	12.8	1.26	14.7	1
8	3.7	1.26	14.7	0
9	14.6	1.25	13.4	1
9	3.9	1.25	13.4	0
10	18.9	1.24	12.9	1
10	7.0	1.24	12.9	0
11	19.3	1.20	11.2	1
11	6.6	1.20	11.2	0
12	30.1	1.20	10.9	1
12	10.1	1.20	10.9	0
13	28.2	1.18	10.3	1
13	9.4	1.18	10.3	0
14	25.6	1.10	9.7	1
14	7.9	1.10	9.7	0
15	37.5	1.11	9.6	1
15	14.1	1.11	9.6	0
16	36.1	1.14	9.1	1
16	14.5	1.14	9.1	0
17	39.8	1.17	7.8	1
17	14.9	1.17	7.8	0
18	44.3	1.18	8.3	1
18	15.6	1.18	8.3	0

El modelo que se propone consiste en hacer una regresión de la variable dependiente, representada en este caso por el número de ventas mensuales (V), en las variables explicativas que se definen a continuación:

PG = Precio de la Gasolina por galón (dólares).

T I = Tasa de interés aplicable al comprar una unidad a plazo (%).

M = Modelo del automóvil $\begin{cases} 1 & \text{) estándar} \\ 0 & \text{) de lujo} \end{cases}$.

Modelo de regresión lineal múltiple:

$$V = \beta_1 + \beta_2(PG) + \beta_3(TI) + \beta_4(M) + \varepsilon$$

Al realizar la estimación de los parámetros con el método de mínimos cuadrados ordinarios se obtienen los resultados siguientes:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{pmatrix} = \begin{pmatrix} 56.0744 \\ -16.1436 \\ -2.3322 \\ 14.4222 \end{pmatrix} \qquad \hat{\sigma}^2 = 22.6642$$

$SCR = 3713.1462$	$n = 36$, $k = 4$
$SCE = 725.2538$	$R^2 = 0.8366$
$SCT = 4438.4000$	$\overline{R}^2 = 0.8213$

De tal forma la estimación del modelo de regresión lineal múltiple está dada por la siguiente función:

$$\hat{V}_t = 56.0744 - 16.1436(PG_t) - 2.3322(TI_t) + 14.4222(M_t) \qquad t = 1,2,\ldots,36.$$

La estimación de la matriz de varianzas y covarianzas de los coeficientes estimados es :

$$\widehat{VCov}(\hat{\beta}) = \begin{pmatrix} 100.1597 & -86.1623 & 0.7371 & -1.2591 \\ -86.1623 & 86.7844 & -1.8567 & 0 \\ 0.7371 & -1.8567 & 0.1314 & 0 \\ -1.2591 & 0 & 0 & 2.5182 \end{pmatrix}$$

PRUEBAS DE SIGNIFICANCIA DEL MODELO.

Antes de realizar las pruebas de significancia es importante aclarar que si las perturbaciones estocásticas no se distribuyen normalmente la validez de dichas pruebas podría cuestionarse.

La prueba de significancia global del modelo consiste en contrastar las siguientes hipótesis :

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0 \qquad vs. \qquad H_1: \beta_j \neq 0 \quad \textit{para algún} \quad j = 2,3,4.$$

$$F = \frac{(S.S.R.)/(k-1)}{(S.S.E.)/(n-k)} \sim F_{(k-1,\,n-k)}$$

Por lo que :

$$F = 54.6111$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$F^{\alpha}_{(3,32)} = 2.904$$

Se rechaza H_0 si $F > F^{\alpha}_{(3,32)}$

Por lo tanto se rechaza H_0 , es decir $\beta_j \neq 0$ para algún $j = 2,3,4$.

Para evaluar hipótesis acerca del verdadero valor de cualquier β_j se realiza la prueba t , por lo tanto para comprobar si los parámetros son significativos se realizarán estas pruebas bajo la hipótesis ($\beta_j = 0$)

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

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_1 = 0$

Por lo tanto:

$$t = 5.6030$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(32)}^{0.025} = 2.038$$

Rechazamos H_0 si $|t| > t_{(32)}^{0.025}$

Por lo tanto se rechaza la hipótesis nula , es decir , $\beta_1 \neq 0$, lo cual indica que el valor del modelo con todas las variables explicativas iguales a cero es distinto de cero.

$$H_0: \beta_2 = 0 \qquad vs. \qquad H_1: \beta_2 \neq 0$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{Var}(\hat{\beta}_2)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_2 = 0$

Por lo tanto:

$$t = -1.7329$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(32)}^{0.025} = 2.038$$

Por lo que se acepta H_0 , es decir , $\beta_2 = 0$, lo cual indica que la variable explicativa PG no es significativa y no debería incluirse en el modelo.

$$H_0 : \beta_3 = 0 \qquad vs. \qquad H_1 : \beta_3 \neq 0$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{\sqrt{\widehat{Var}(\hat{\beta}_3)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_3 = 0$

Por lo tanto:

$$t = -6.4346$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(32)}^{0.025} = 2.038$$

Por lo que se rechaza H_0 , es decir , $\beta_3 \neq 0$, lo cual indica que la variable explicativa TI es significativa en éste modelo.

$$H_0: \beta_4 = 0 \qquad vs. \qquad H_1: \beta_4 \neq 0$$

$$t = \frac{\hat{\beta}_4 - \beta_4}{\sqrt{\widehat{Var}(\hat{\beta}_4)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_4 = 0$

Por lo tanto:

$$t = 9.0883$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(32)}^{0.025} = 2.038$$

Por lo que se rechaza H_0 , es decir , $\beta_4 \neq 0$, lo cual indica que la variable explicativa M es significativa en éste modelo.

Al realizar las pruebas t se determina que la variable explicativa PG no debe incluirse en el modelo , por lo cual el modelo de regresión lineal múltiple que debe ser utilizado es el siguiente :

$$V = \beta_1 + \beta_2(TI) + \beta_3(M) + \varepsilon$$

Al realizar la estimación de los parámetros con el método de mínimos cuadrados ordinarios se obtienen los resultados siguientes:

$$\underline{\hat{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 40.0465 \\ -2.6776 \\ 14.4222 \end{pmatrix} \qquad \hat{\sigma}^2 = 24.0399$$

$$SCR = 3645.0847 \qquad n = 36, \quad k = 3$$
$$SCE = 793.3153 \qquad R^2 = 0.8213$$
$$SCT = 4438.4000 \qquad \bar{R}^2 = 0.8104$$

De tal forma la estimación del modelo de regresión lineal múltiple está dada por la siguiente función:

$$\hat{V}_t = 40.0465 - 2.6776(TI_t) + 14.4222(M_t) \qquad t = 1, 2, \ldots, 36.$$

La estimación de la matriz de varianzas y covarianzas de los coeficientes estimados es :

$$\widehat{VCov}(\hat{\beta}) = \begin{pmatrix} 15.5020 & -1.1735 & -1.3355 \\ -1.1735 & 0.0972 & 0 \\ -1.3355 & 0 & 2.6711 \end{pmatrix}$$

PRUEBAS DE SIGNIFICANCIA DEL NUEVO MODELO.

La prueba de significancia global del modelo es la siguiente :

$$H_0: \beta_2 = \beta_3 = 0 \qquad vs. \qquad H_1: \beta_j \neq 0 \quad para\ algún\ \ j = 2, 3.$$

$$F = \frac{(S.S.R.)/(k-1)}{(S.S.E.)/(n-k)} \sim F_{(k-1,\, n-k)}$$

Por lo que :

$$F = 75.8134$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$F^{\alpha}_{(2,33)} = 3.293$$

Por lo tanto se rechaza H_0 , es decir $\beta_j \neq 0$ para algún $j = 2,3.$

Pruebas t bajo la hipótesis ($\beta_j = 0$)

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

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_1 = 0$

Por lo tanto:

$$t = 10.1712$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t^{0.025}_{(33)} = 2.036$$

Por lo que se rechaza H_0 , es decir , $\beta_1 \neq 0$, lo cual indica que el valor del modelo con todas las variables explicativas iguales a cero es distinto de cero.

$$H_0 : \beta_2 = 0 \qquad vs. \qquad H_1 : \beta_2 \neq 0$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{Var}(\hat{\beta}_2)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_2 = 0$

Por lo tanto:

$$t = -8.5881$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(33)}^{0.025} = 2.036$$

Por lo que se rechaza H_0 , es decir , $\beta_2 \neq 0$, lo cual indica que la variable explicativa TI es significativa en éste modelo.

$$H_0 : \beta_3 = 0 \qquad vs. \qquad H_1 : \beta_3 \neq 0$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{\sqrt{\widehat{var}(\hat{\beta}_3)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_3 = 0$

Por lo tanto:

$$t = 8.8244$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(33)}^{0.025} = 2.036$$

Por lo que se rechaza H_0 , es decir , $\beta_3 \neq 0$, lo cual indica que la variable explicativa M es significativa en éste modelo.

Al contar con un modelo adecuado en base a las pruebas de significancia se presenta a continuación la tabla que contiene a todos los residuos mencionados en el segundo capítulo.

Residuos ordinarios ei	Residuos estandarizados wi	Residuos est. internos ri	Residuos predictivos ei(i)	Residuos Rao-Blackwell eimi
0.029933	0.006106	0.006282	0.031693	0.006186
-0.447845	-0.091340	-0.093988	-0.474191	-0.092566
-6.402310	-1.305782	-1.343686	-6.779390	-1.360923
-1.980088	-0.403849	-0.415571	-2.096710	-0.410301
-9.834553	-2.005806	-2.064185	-10.415389	-2.178145
0.487670	0.099483	0.102358	0.516471	0.100811
-6.147163	-1.253744	-1.302780	-6.637417	-1.317209
0.475059	0.096891	0.100680	0.512947	0.099158
-0.763045	-0.155627	-0.165126	-0.859038	-0.162672
4.659177	0.950262	1.008264	5.245314	1.008526
-3.324258	-0.677999	-0.725986	-3.811476	-0.720680
5.297984	1.080545	1.157024	6.074455	1.163196
2.711256	0.552974	0.594502	3.133780	0.588586
8.433478	1.720049	1.848225	9.747756	1.923371
-2.308376	-0.470805	-0.491778	-2.519622	-0.486054
3.013846	0.614689	0.642071	3.288346	0.636255
-3.969222	-0.813621	-0.840387	-4.256006	-0.836556
-0.267000	-0.054456	-0.056247	-0.284855	-0.055391
-1.029009	-0.209867	-0.216063	-1.091682	-0.212915
1.494214	0.304752	0.314048	1.586764	0.309717
-5.179884	-1.056462	-1.088866	-5.502505	-1.092038
-3.257661	-0.664416	-0.684794	-3.460560	-0.679182
4.816844	0.982419	1.013887	5.130370	1.014331
-0.760933	-0.155196	-0.160167	-0.810462	-0.157783
1.310300	0.267242	0.276859	1.406287	0.272948
-3.067478	-0.625627	-0.648137	-3.292186	-0.642343
-2.896244	-0.590703	-0.615266	-3.142319	-0.609397
-6.174022	-1.259222	-1.311625	-6.696589	-1.326642
8.735999	1.781749	1.857874	9.498426	1.933414
-0.241779	-0.049312	-0.051419	-0.262880	-0.050636
5.997212	1.223181	1.283121	6.599597	1.296290
-1.190566	-0.240782	-0.252595	-1.299147	-0.248970
6.216366	1.267858	1.358784	7.139968	1.377116
-4.261411	-0.869136	-0.931467	-4.894554	-0.929546
12.065153	2.458707	2.610765	13.592350	2.886194
-2.222625	-0.453315	-0.481350	-2.506040	-0.476674

A continuación se presenta una tabla con los residuos estudentizados externos(t_i), los elementos de la diagonal de la matriz (H) y el vector normalizado de residuos cuadrados (δ_j); Sus cotas superiores e inferiores respectivas para detectar puntos discrepantes "outliers" y puntos palanca se presentan en la última fila de la tabla y dichos puntos aparecen sombreados en la tabla :

ti	hii	di1	di2	di3
0.006186	0.055559	0.000003	0.000003	0.027778
-0.092566	0.055558	0.004556	0.000003	0.027778
-1.360923	0.055622	0.000060	0.000066	0.027778
-0.410301	0.055622	0.003772	0.000066	0.027778
-2.178146	0.055765	0.000192	0.000210	0.027778
0.100811	0.055765	0.003062	0.000210	0.027778
-1.317209	0.073862	0.016729	0.018307	0.027778
0.099158	0.073862	0.003619	0.018307	0.027778
-0.162672	0.111745	0.051349	0.056189	0.027778
1.008526	0.111745	0.024781	0.056189	0.027778
-0.720680	0.127829	0.068047	0.072273	0.027778
1.163196	0.127829	0.035274	0.072273	0.027778
0.588586	0.134829	0.072448	0.076278	0.027778
1.923871	0.134829	0.039988	0.076278	0.027778
-0.486054	0.083477	0.025516	0.027921	0.027778
0.636255	0.083477	0.008200	0.027921	0.027778
-0.836556	0.062684	0.006514	0.007129	0.027778
-0.055391	0.062684	0.000133	0.007129	0.027778
-0.212915	0.058326	0.002532	0.002771	0.027778
0.309717	0.058326	0.000356	0.002771	0.027778
-1.092038	0.058632	0.002811	0.003076	0.027778
-0.679182	0.058632	0.014934	0.003076	0.027778
1.014331	0.061112	0.005077	0.005556	0.027778
-0.157783	0.061112	0.019723	0.005556	0.027778
0.272948	0.068255	0.011606	0.012700	0.027778
-0.642343	0.068255	0.031298	0.012700	0.027778
-0.609397	0.078310	0.020794	0.022754	0.027778
-1.326642	0.078310	0.045533	0.022754	0.027778
1.669414	0.080269	0.022584	0.024713	0.027778
-0.050636	0.080269	0.048164	0.024713	0.027778
1.296260	0.091276	0.032643	0.035720	0.027778
-0.248970	0.091276	0.058539	0.035720	0.027778
1.377116	0.129357	0.087448	0.076901	0.027778
-0.929546	0.129357	0.108162	0.076901	0.027778
2.699194	0.113093	0.052580	0.067537	0.027778
-0.476574	0.113093	0.089084	0.067537	0.027778
LS=1.6644 y Li=-1.6644	LS=0.166666	LÍMITE SUPERIOR = 0.055556		

Los límites para identificar " outliers " se obtuvieron a un nivel de significancia $\alpha = 0.05$

En la tabla que se presenta a continuación se muestran la distancia de Welsh-Kuh (DFFITS) que se utiliza para determinar si la i-ésima observación es influyente y el valor absoluto de la medida DFBETAS que es utilizada para determinar si la i-ésima observación es influyente con respecto al j-ésimo coeficiente ($\hat{\beta}_j$). Sus respectivas cotas superiores para detectar observaciones influyentes se presentan en la última fila de la tabla y los puntos influyentes aparecen sombreados en la tabla.

/ DFFITS i /	/ DFBETAS i1 /	/ DFBETAS i2 /	/ DFBETAS i3 /
0.001500	0.000011	0.000011	0.001061
0.022451	0.006429	0.000168	0.015876
0.330290	0.010878	0.011379	0.233405
0.099575	0.025930	0.003431	0.070369
0.529333	0.031036	0.032466	[illegible]
0.024499	0.006741	0.001603	0.017291
0.371967	0.177034	0.185191	0.228121
0.028003	0.006199	0.013941	0.017173
0.057698	0.039112	0.040914	0.028767
0.357711	0.168452	0.253656	0.178347
0.275903	0.198320	0.207458	0.128615
0.445314	0.233926	[illegible]	0.207587
0.232354	0.170317	0.178165	0.105465
[illegible]	[illegible]	[illegible]	[illegible]
0.146[illegible]	0.081099	0.084836	0.084619
0.192018	0.080181	0.111052	0.110766
0.216337	0.069742	0.072955	0.144013
0.014324	0.000680	0.004831	0.009536
0.052999	0.011040	0.011549	0.036568
0.077081	0.006021	0.016800	0.053194
0.272536	0.059676	0.062426	0.187589
0.169501	0.095544	0.038825	0.116869
0.258762	0.074593	0.078030	0.174470
0.040255	0.022989	0.012138	0.027140
0.073876	0.030462	0.031866	0.047129
0.173865	0.117727	0.074992	0.110909
0.177630	0.091533	0.095760	0.105793
0.386696	0.294866	0.208446	0.230309
0.571173	0.302967	0.316927	[illegible]
0.014959	0.011587	0.008300	0.008800
0.410829	0.245685	0.257005	0.226837
0.079906	0.065256	0.049962	0.043529
0.530817	[illegible]	[illegible]	0.246980
0.358299	0.327634	0.270634	0.166036
[illegible]	[illegible]	[illegible]	[illegible]
0.1[illegible]	0.150769	0.121158	0.084182
LS= 0.577[illegible]	LIMITE SUPERIOR = 0.[illegible]		

Los resultados obtenidos en las tablas anteriores se pueden resumir en :

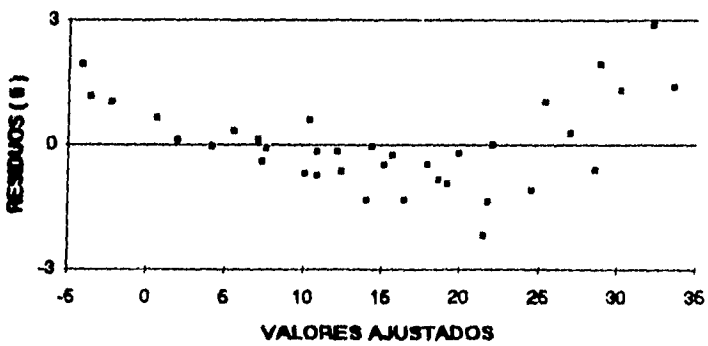
1. Las observaciones 5, 14, 29 y 35 son observaciones discrepantes "outliers" , es decir, su residuo sale del rango de confianza.

2. Aunque pueda haber sospecha de puntos palanca debido a varios valores δij que rebasan su cota superior generando palanca parcial de observaciones para alguna variable explicativa , se puede ver que todos los valores de la diagonal de la matriz (H) no rebasan su cota superior, por lo que no existen puntos palanca.

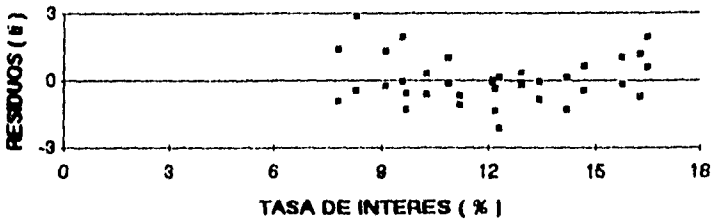
3. Considerando a la medida **DFFITS** las observaciones 14 y 35 son influyentes y en base a la medida **DFBETAS** las observaciones 14 y 35 influyen en la estimación de los parámetros $\beta 1$, $\beta 2$ y $\beta 3$.

GRAFICAS PARA LOS RESIDUOS.

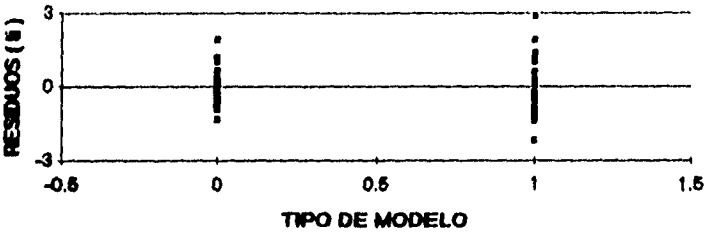
RESIDUOS (ti) CONTRA VALORES AJUSTADOS



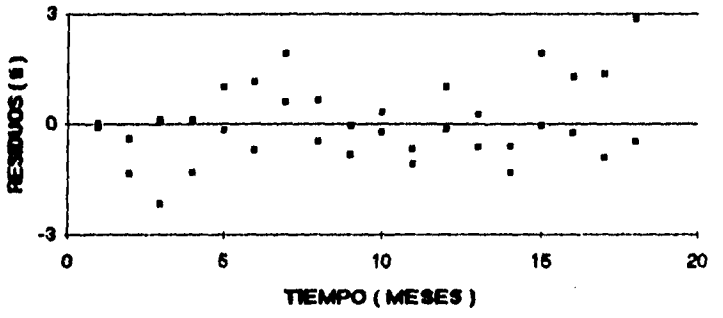
RESIDUOS (ti) CONTRA VARIABLE EXPLICATIVA (TI)



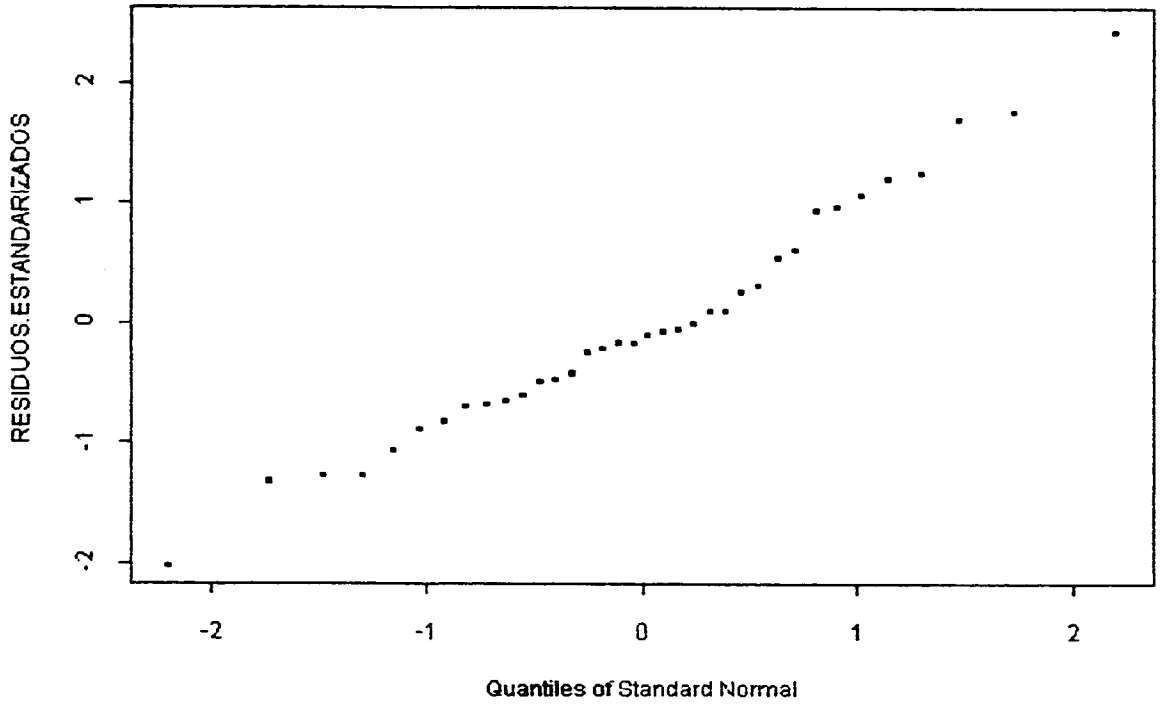
RESIDUOS (ti) CONTRA VARIABLE EXPLICATIVA (M)



RESIDUOS (ti) CONTRA EL TIEMPO



GRAFICA SOBRE PAPEL NORMAL



VERIFICACION DEL SUPUESTO DE NORMALIDAD.

Para verificar el supuesto de normalidad se utilizará la estadística Anderson-Darling modificada ($\overline{A}^2$), que se describe a continuación :

1. Se cuente con los residuos Rao-Blackwell (s_i)

$$s_i = \frac{(n-k-1)^{1/2}\left(Y_i - x_i'\hat{\beta}\right)}{\sqrt{(1-h_{ii})(n-k)\hat{\sigma}^2 - \left(Y_i - x_i'\hat{\beta}\right)^2}}$$

2. Se consideran los estadísticos de orden $s_{(i)}$, con :

$$s_{(i)} < s_{(i+1)} \qquad \forall \quad i = 1,\ldots,n-1$$

3. Sea $z_{(i)} = G_{n-k-1}(s_{(i)})$ en dónde $G_{n-k-1}(s_{(i)})$ es la función de distribución de una t de Student con n-k-1 grados de libertad evaluada en $(s_{(i)})$

4. Se calcula el valor $\overline{A}^2$

$$\overline{A}^2 = -n - \left(1/n\right)\sum_{i=1}^{n}(2i-1)\left[\ln z_{(i)} + \ln\left\{1 - z_{(n-i+1)}\right\}\right]$$

Los resultados obtenidos de los puntos anteriores se presentan en la siguiente tabla:

S_i	$S_{(i)}$	$Z_{(i)}$
0.006186	-2.178145	0.018438
-0.092568	-1.360923	0.091523
-1.360923	-1.326642	0.097009
-0.410301	-1.317209	0.098562
-2.178145	-1.092038	0.141486
0.100811	-0.929546	0.179783
-1.317209	-0.836556	0.204523
0.099158	-0.720680	0.238169
-0.162672	-0.679182	0.250953
1.006526	-0.642343	0.262613
-0.720680	-0.609397	0.273282
1.163196	-0.486054	0.316121
0.589586	-0.475674	0.318769
1.923371	-0.410301	0.342160
-0.486054	-0.248970	0.402488
0.636255	-0.212915	0.416372
-0.836556	-0.162672	0.435900
-0.055391	-0.157783	0.437810
-0.212915	-0.092586	0.463413
0.309717	-0.055391	0.479066
-1.092038	-0.050636	0.479965
-0.679182	0.006186	0.502449
1.014331	0.099158	0.539164
-0.157783	0.100811	0.539835
0.272948	0.272948	0.606676
-0.642343	0.309717	0.620608
-0.609397	0.589586	0.719864
-1.326642	0.636255	0.735432
1.933414	1.006526	0.839616
-0.050636	1.014331	0.840983
1.296280	1.163196	0.873325
-0.248970	1.296280	0.897925
1.377116	1.377116	0.910983
-0.929546	1.923371	0.968315
2.886194	1.933414	0.968963
-0.475674	2.886194	0.996536

De tal forma:

$$A^2 = 0.44225816$$

y se rechaza el supuesto de normalidad si el valor de la estadística excede al valor de tablas (según α).

PRUEBA A^2

(CASO NORMAL CON AMBOS PARAMETROS DESCONOCIDOS)

α (NIVEL)

0.50	0.25	0.15	0.10	0.05	0.025	0.01	0.005
VALORES CRITICOS							
0.341	0.470	0.561	0.631	0.752	0.873	1.035	1.159

Por lo tanto , con α = .25 o cualquier valor menor, se acepta la hipótesis nula y se concluye que las perturbaciones estocásticas se distribuyen normalmente.

Por otra parte si se verifica el supuesto de normalidad en las perturbaciones estocásticas a través de la prueba **Bera-Jarque** se obtiene el resultado siguiente:

$$\lambda = 1.3534996$$

La decisión en la prueba es rechazar H_0 si λ es mayor que el valor crítico de la tabla.

DISTRIBUCION *de* λ

α (NIVEL)

0.50	0.25	0.20	0.10	0.05	0.025	0.01	0.005
VALORES CRITICOS (asintótica)							
1.3863	2.7726	3.2190	4.6052	5.9915	7.3777	9.2103	10.5966
VALORES CRITICOS (N finito = 36)							
			2.616	3.878			

Por lo tanto , con α = .10 o cualquier valor menor, se acepta la hipótesis nula y se concluye que las perturbaciones estocásticas se distribuyen normalmente.

Se puede notar que aunque se detectaron cuatro puntos discrepantes en el análisis se aceptó el supuesto de normalidad en las perturbaciones estocásticas con las dos pruebas anteriores.

El segundo ejemplo se extrajo de Neter, Wasserman y Kutner (1990, p 439) ; del libro **"Applied Linear Statistical Models"**.

EJEMPLO 2.

La unidad quirúrgica de un hospital está interesada en obtener un modelo de regresión que le permita predecir adecuadamente el tiempo de supervivencia en pacientes a los cuales se les practicará una operación específica del hígado.

Se seleccionaron aleatoriamente 54 pacientes a los cuales se les realizó una evaluación preoperatoria que proporcionó información para los siguientes conceptos:

1) Función del Tiempo de supervivencia de los pacientes .

2) Nivel de Coagulación en la sangre .

3) Edad del paciente.

4) Funcionamiento de la Enzima en el paciente.

5) Funcionamiento del Hígado del paciente.

Los datos obtenidos para los cincuenta y cuatro pacientes por estos conceptos se presentan a continuación:

PACIENTE	LOG DEL TIEMPO DE SUPERVIVENCIA	NIVEL DE COAGULACION	EDAD	FUNCIONAMIENTO DE LA ENZIMA	FUNCIONAMIENTO DEL HIGADO
1	2 3010	6 7	62	81	2 59
2	2 0043	5 1	59	66	1 70
3	2 3096	7 4	57	83	2 16
4	2 0043	6 5	73	41	2 01
5	2 7067	7 8	65	115	4 30
6	1 9031	5 8	38	72	1 42
7	1 9031	5 7	46	63	1 91
8	2 1038	3 7	68	81	2 57
9	2 3054	6 0	67	93	2 50
10	2 3075	3 7	76	94	2 40
11	2 6172	6 3	84	83	4 13
12	1 8129	6 7	51	43	1 86
13	2 9191	5 8	96	114	3 95
14	2 5185	5 8	83	88	3 95
15	2 2253	7 7	62	67	2 40
16	2 3365	7 4	74	68	2 40
17	1 9395	6 0	85	28	2 96
18	1 5315	3 7	51	41	1 65
19	2 3324	7 3	68	74	2 66
20	2 2355	5 6	57	87	3 02
21	2 0374	5 2	52	75	2 85
22	2 1335	3 4	83	53	1 12
23	1 8451	6 7	26	68	2 10
24	2 3424	5 8	67	86	3 40
25	2 4409	6 3	59	100	2 95
26	2 1584	5 8	61	73	3 50
27	2 2577	5 2	52	86	2 45
28	2 7569	11.2	76	90	5 59
29	1 8573	5 2	54	56	2.71
30	2 2504	5 8	76	59	2 68
31	1 8513	3 2	64	65	0 74
32	1 7634	8 7	45	23	2 52
33	2 0546	5 0	59	73	3 50
34	2 4698	5 8	72	93	3 30
35	2 0607	5 4	58	70	2 64
36	2 2648	5 3	51	99	2 60
37	2 0719	2 6	74	86	2 05
38	2 0792	4 3	8	119	2 85
39	2 1790	4 8	61	76	2 45
40	2.1703	5 4	52	88	1 81
41	1 9777	5 2	49	72	1 84
42	1 8751	3 6	28	99	1 30
43	2 6840	8 8	86	88	5 40
44	2.1847	6 5	56	77	2 65
45	2 2810	3 4	77	93	1 46
46	2 0899	6 5	40	84	3 00
47	2 4926	4 5	73	106	3 05
48	2.5999	4 8	86	101	4 10
49	2 1987	5.1	57	77	2.86
50	2 4914	3 9	82	103	4 55
51	2 0934	6 6	77	46	1 95
52	2 0969	5 4	85	40	1 21
53	2 2967	6 4	59	85	2 33
54	2 4955	8 8	78	72	3 20

El modelo que se propone consiste en hacer una regresión de la variable dependiente, representada en este caso por el logaritmo del tiempo de supervivencia de los pacientes (T), en las variables explicativas que se definen a continuación:

C = Nivel de **C**oagulación en la sangre.

E = **E**dad del paciente.

FE = **F**uncionamiento de la **E**nzima en el paciente.

FH = **F**uncionamiento del **H**ígado del paciente.

Modelo de regresión lineal múltiple:

$$T = \beta_1 + \beta_2(C) + \beta_3(E) + \beta_4(FE) + \beta_5(FH) + \varepsilon$$

Al realizar la estimación de los parámetros con el método de mínimos cuadrados ordinarios se obtienen los resultados siguientes:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \end{pmatrix} = \begin{pmatrix} 0.4888 \\ 0.0685 \\ 0.0093 \\ 0.0095 \\ 0.0019 \end{pmatrix} \qquad \hat{\sigma}^2 = 0.0022$$

$$SCR = 3.8630 \qquad n = 54 \,,\; k = 5$$
$$SCE = 0.1098 \qquad R^2 = 0.9724$$
$$SCT = 3.9728 \qquad \overline{R}^2 = 0.9701$$

De tal forma la estimación del modelo de regresión lineal múltiple está dada por la siguiente función:

$$\hat{Y}_t = 0.4888 + 0.0685(C_t) + 0.0093(E_t) + 0.0095(FE_t) + 0.0019(FH_t) \qquad t = 1,2,\ldots,54.$$

La estimación de la matriz de varianzas y covarianzas de los coeficientes estimados es :

$$\widehat{Cov}(\hat{\beta}) = \begin{pmatrix} 0.0025 & -0.0002 & -1.40E-05 & -1.50E-05 & 0.0003 \\ -0.0002 & 2.96E-05 & 5.85E-07 & 1.07E-06 & -3.45E-05 \\ -1.40E-05 & 5.85E-07 & 1.91E-07 & 5.17E-08 & -1.98E-06 \\ -1.50E-05 & 1.07E-06 & 5.17E-08 & 1.57E-07 & -2.40E-06 \\ 0.0003 & -3.45E-05 & -1.98E-06 & -2.40E-06 & 9.43E-05 \end{pmatrix}$$

PRUEBAS DE SIGNIFICANCIA DEL MODELO.

Antes de realizar las pruebas de significancia es importante aclarar que si las perturbaciones estocásticas no se distribuyen normalmente la validez de dichas pruebas podría cuestionarse.

La prueba de significancia global del modelo consiste en contrastar las siguientes hipótesis :

$$H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \qquad vs. \qquad H_1: \beta_j \neq 0 \quad \text{para algún} \quad j = 2,3,4,5.$$

$$F = \frac{(S.S.R.)/(k-1)}{(S.S.E.)/(n-k)} \sim F_{(k-1,\, n-k)}$$

Por lo que :

$$F = 431.0972$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$F^{\alpha}_{(4,\infty)} = 2.574$$

Se rechaza H_0 si:

$$F > F^{\alpha}_{(4,\infty)}$$

Por lo tanto se rechaza la hipótesis nula , es decir , $\beta_j \neq 0$ para algún $j = 2,3,4,5.$

Para evaluar hipótesis acerca del verdadero valor de cualquier β_j se realiza la prueba t , por lo tanto para comprobar si los parámetros son significativos se realizarán estas pruebas bajo la hipótesis ($\beta_j = 0$)

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

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_1 = 0$

Por lo tanto:

$$t = 9.7297$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(40)}^{0.025} = 2.012$$

Se rechaza H_0 si $|t| > t_{(40)}^{0.025}$

Por lo tanto , se rechaza la hipótesis nula , es decir , $\beta_1 \neq 0$, lo cual indica que el valor del modelo con todas las variables explicativas iguales a cero es distinto de cero.

$$H_0: \beta_2 = 0 \qquad vs. \qquad H_1: \beta_2 \neq 0$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{Var}(\hat{\beta}_2)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_2 = 0$

Por lo tanto :

$$t = 12.5961$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(40)}^{0.025} = 2.012$$

Por lo que se rechaza H_0 , es decir , $\beta_2 \neq 0$, lo cual indica que la variable explicativa **C es significativa en éste modelo.**

$$H_0: \beta_3 = 0 \qquad vs. \qquad H_1: \beta_3 \neq 0$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{\sqrt{\hat{Var}(\hat{\beta}_3)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_3 = 0$

Por lo tanto :

$$t = 21.1893$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(40)}^{0.025} = 2.012$$

Por lo que se rechaza H_0 , es decir , $\beta_3 \neq 0$, lo cual indica que la variable explicativa **E es significativa en éste modelo.**

$$H_0: \beta_4 = 0 \qquad vs. \qquad H_1: \beta_4 \neq 0$$

$$t = \frac{\hat{\beta}_4 - \beta_4}{\sqrt{\widehat{Var}(\hat{\beta}_4)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_4 = 0$

Por lo tanto :

$$t = 23.9105$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(40)}^{0.025} = 2.012$$

Por lo que se rechaza H_0 , es decir , $\beta_4 \neq 0$, lo cual indica que la variable explicativa **FE es significativa en éste modelo.**

$$H_0: \beta_5 = 0 \qquad vs. \qquad H_1: \beta_5 \neq 0$$

$$t = \frac{\hat{\beta}_5 - \beta_5}{\sqrt{\widehat{Var}(\hat{\beta}_5)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_5 = 0$

Por lo tanto :

$$t = 0.1983$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(49)}^{0.025} = 2.012$$

Por lo que se acepta H_0 , es decir , $\beta_5 = 0$, lo cual indica que la variable explicativa **FH** no es significativa y no debería incluirse en el modelo.

Al realizar las pruebas t se determina que la variable explicativa **FH** no debe incluirse en el modelo , por lo cual el modelo de regresión lineal múltiple que debe ser utilizado es el siguiente :

$$T = \beta_1 + \beta_2(C) + \beta_3(E) + \beta_4(FE) + \varepsilon$$

Al realizar la estimación de los parámetros con el método de mínimos cuadrados ordinarios se obtienen los resultados siguientes:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{pmatrix} = \begin{pmatrix} 0.4836 \\ 0.0692 \\ 0.0093 \\ 0.0095 \end{pmatrix} \qquad \hat{\sigma}^2 = 0.0022$$

$SCR = 3.8629$	$n = 54$, $k = 4$
$SCE = 0.1099$	$R^2 = 0.9723$
$SCT = 3.9728$	$\bar{R}^2 = 0.9707$

De tal forma la estimación del modelo de regresión lineal múltiple está dada por la siguiente función:

$$\hat{T}_t = 0.4836 + 0.0692(C_t) + 0.0093(E_t) + 0.0095(FE_t) \qquad t = 1,2,\ldots,54.$$

La estimación de la matriz de varianzas y covarianzas de los coeficientes estimados es :

$$\widehat{Var}(\hat{\beta}) = \begin{pmatrix} 0.0018 & -0.0001 & -8.56E-06 & -8.39E-06 \\ -0.0001 & 1.66E-05 & -1.37E-07 & 1.85E-07 \\ -8.56E-06 & -1.37E-07 & 1.46E-07 & 1.20E-09 \\ -8.39E-06 & 1.85E-07 & 1.20E-09 & 9.39E-08 \end{pmatrix}$$

PRUEBAS DE SIGNIFICANCIA DEL NUEVO MODELO PROPUESTO.

La prueba de significancia global del modelo es la siguiente :

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0 \qquad vs. \qquad H_1: \beta_j \neq 0 \quad para\ algún \quad j = 2,3,4.$$

$$F = \frac{(S.S.R.)/(k-1)}{(S.S.E.)/(n-k)} \sim F_{(k-1,\ n-k)}$$

Por lo que :

$$F = 586.0431$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$F^{\alpha}_{(3,\ \infty)} = 2.80$$

Por lo tanto se rechaza H_0 , es decir $\beta_j \neq 0$ para algún $j = 2,3,4.$

Pruebas t bajo la hipótesis ($\beta_1 = 0$)

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

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{var}(\hat{\beta}_1)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_1 = 0$

Por lo tanto:

$$t = 11.3450$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(50)}^{0.025} = 2.011$$

Por lo que se rechaza H_0 , es decir , $\beta_1 \neq 0$, lo cual indica que el valor del modelo con todas las variables explicativas iguales a cero es distinto de cero.

$$H_0 : \beta_2 = 0 \qquad vs. \qquad H_1 : \beta_2 \neq 0$$

$$t = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\widehat{var}(\hat{\beta}_2)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_2 = 0$

Por lo tanto :

$$t = 16.9755$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(50)}^{0.025} = 2.011$$

Por lo que se rechaza H_0 , es decir , $\beta_2 \neq 0$, lo cual indica que la variable explicativa C es significativa en éste modelo.

$$H_0: \beta_3 = 0 \qquad vs. \qquad H_1: \beta_3 \neq 0$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{\sqrt{\widehat{var}(\hat{\beta}_3)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_3 = 0$

Por lo tanto :

$$t = 24.2992$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(50)}^{0.025} = 2.011$$

Por lo que se rechaza H_0 , es decir , $\beta_3 \neq 0$, lo cual indica que la variable explicativa **E** es significativa en éste modelo.

$$H_0: \beta_4 = 0 \qquad vs. \qquad H_1: \beta_4 \neq 0$$

$$t = \frac{\hat{\beta}_4 - \beta_4}{\sqrt{\widehat{Var}(\hat{\beta}_4)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_4 = 0$

Por lo tanto :

$$t = 31.0817$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(50)}^{0.025} = 2.011$$

Por lo qua se rechaza H_0 , es decir , $\beta_4 \neq 0$, lo cual indica que la variable explicativa **FE es significativa en éste modelo.**

Al contar con un modelo adecuado en base a las pruebas de significancia se presenta a continuación la tabla que contiene a todos los residuos mencionados en el segundo capítulo.

Residuos ordinarios ei	Residuos estandarizados wi	Residuos est. internos ri	Residuos predictivos ei(i)	Residuos Piao-Blackwell si*ei
0.005895	0.125762	0.127453	0.006055	0.126193
-0.009307	-0.198547	-0.201542	-0.009590	-0.199597
-0.006537	-0.139463	-0.142725	-0.006847	-0.141319
0.001746	0.037241	0.038802	0.001895	0.038413
-0.016240	-0.346460	-0.370034	-0.010515	-0.366817
-0.020921	-0.446320	-0.460893	-0.022309	-0.457233
-0.002642	-0.056360	-0.057738	-0.002773	-0.057160
-0.039397	-0.840488	-0.863875	-0.041620	-0.861647
-0.102004	-2.176128	-2.210030	-0.105283	-2.304126
-0.033861	-0.722376	-0.749880	-0.036408	-0.746553
0.026258	0.560180	0.574742	0.027641	0.570854
-0.018067	-0.385433	-0.402096	-0.019663	-0.398699
0.056003	1.194759	1.295509	0.065847	1.304571
0.023847	0.508742	0.521906	0.025097	0.518074
-0.005699	-0.121585	-0.124601	-0.005985	-0.123368
0.005210	0.111155	0.113732	0.005455	0.112603
-0.016169	-0.344945	-0.374126	-0.019020	-0.370885
-0.072744	-1.551911	-1.661016	-0.083332	-1.691656
0.006658	0.142044	0.144683	0.006908	0.143258
0.005873	0.125302	0.126900	0.006024	0.125645
-0.013304	-0.283821	-0.288009	-0.013699	-0.285351
0.138314	2.950768	3.158781	0.158502	3.495164
0.008406	0.179324	0.191696	0.009595	0.189739
0.015507	0.330816	0.334668	0.016870	0.331676
0.020419	0.435624	0.446069	0.021410	0.442467
0.011081	0.236404	0.238753	0.011303	0.236488
0.111760	2.384259	2.422204	0.115345	2.552274
-0.063654	-1.358851	-1.578143	-0.086102	-1.602707
-0.021520	-0.459106	-0.470400	-0.022592	-0.466706
0.096894	2.069251	2.115279	0.101357	2.194512
-0.067728	-1.444697	-1.507098	-0.073685	-1.527038
0.040223	0.858110	0.966232	0.050998	0.965578
-0.008850	-0.188797	-0.191225	-0.009079	-0.189372
0.029768	0.635074	0.646354	0.030835	0.642547
-0.002474	-0.052785	-0.053426	-0.002536	-0.052091
-0.002576	-0.054945	-0.056311	-0.002705	-0.055747
-0.098535	-2.102122	-2.223127	-0.118205	-2.318344
0.090242	1.925201	2.285038	0.127128	2.390321
0.072335	1.543187	1.563758	0.074277	1.567344
-0.008533	-0.182031	-0.184990	-0.008812	-0.183194
-0.007026	-0.149884	-0.152599	-0.007282	-0.151100
-0.060819	-1.297489	-1.398550	-0.870662	-1.412395
-0.046212	-0.985878	-1.053480	-0.052767	-1.054663
-0.002698	-0.057563	-0.058343	-0.002772	-0.057758
-0.039364	-0.839784	-0.876890	-0.042920	-0.874830
-0.015451	-0.329630	-0.340555	-0.016492	-0.337524
0.009659	0.206068	0.213523	0.010371	0.211473
0.022781	0.486003	0.507836	0.024874	0.504034
0.005977	0.127511	0.129025	0.006120	0.127749
-0.006286	-0.112763	-0.116478	-0.005835	-0.117304
-0.000873	-0.018629	-0.019332	-0.000940	-0.019137
-0.000743	-0.015844	-0.016751	-0.000830	-0.016582
0.012152	0.259239	0.262725	0.012481	0.260264
-0.007978	-0.170191	-0.178881	-0.008813	-0.177140

A continuación se presenta una tabla con los residuos estudentizados externos(t_i), los elementos de la diagonal de la matriz (H) y el vector normalizado de residuos cuadrados (δ_j); Sus cotas superiores e inferiores respectivas para detectar puntos discrepantes "outliers" y puntos palanca se presentan en la última fila de la tabla y dichos puntos aparecen sombreados en la tabla:

ti	hii	di1	di2	di3	di4
0.126191	0.025264	0.001295	0.007124	0.000264	0.001379
-0.199597	0.029496	0.014393	0.004513	0.000909	0.000690
-0.141319	0.045194	0.003830	0.022744	0.003950	0.003459
0.039413	0.070951	0.008765	0.000416	0.005148	0.001078
0.366017	0.123355	0.081871	0.044291	0.000002	0.074975
0.457233	0.062240	0.022212	0.000211	0.043618	0.001247
-0.057160	0.047159	0.024854	0.000074	0.015882	0.008900
-0.861647	0.053411	0.000066	0.032716	0.003022	0.000001
2.364[illegible]	0.031147	0.095396	0.000995	0.000306	0.011443
-0.746553	0.072011	0.000001	0.030280	0.014673	0.007160
0.570954	0.050030	0.014295	0.001203	0.027509	0.002199
-0.398699	0.081160	0.028639	0.003076	0.011917	0.[illegible]
1.304571	0.[illegible]	0.[illegible]	0.000109	0.07775[illegible]	0.[illegible]
0.518074	0.049611	0.012974	0.000095	0.026194	0.009336
-0.123868	0.047636	0.000873	0.024914	0.000646	0.001722
0.112603	0.044803	0.004877	0.015409	0.005605	0.001430
-0.370895	0.[illegible]	0.014952	0.001959	0.029794	0.[illegible]
-1.[illegible]	0.127088	0.109215	0.[illegible]	0.007471	0.[illegible]
0.143288	0.036141	0.004126	0.015750	0.000733	0.000000
0.125645	0.025029	0.000223	0.000004	0.002388	0.000814
-0.285351	0.026860	0.010583	0.001918	0.007631	0.000247
2.[illegible]	0.127368	0.025007	0.[illegible]	0.031607	0.034046
0.197738	0.123988	0.029360	0.009515	0.[illegible]	0.002587
0.331676	0.022892	0.001148	0.000354	0.000870	0.003439
0.442467	0.046282	0.007034	0.004921	0.001369	0.024312
0.236488	0.019980	0.002149	0.000001	0.000349	0.000721
2.[illegible]	0.031065	0.003709	0.001163	0.007514	0.002465
-1.602707	0.261[illegible]	0.132744[illegible]	0.[illegible]	0.004056	0.024067
-0.466706	0.047443	0.031774	0.004172	0.005239	0.021402
2.1[illegible]	0.043047	0.001672	0.000636	0.010563	0.013713
-1.527[illegible]	0.080841	0.[illegible]	0.[illegible]	0.000629	0.012614
0.965576	0.[illegible]	0.031315	0.[illegible]	0.030524	0.[illegible]
-0.189372	0.026231	0.009156	0.004776	0.000836	0.001381
0.642547	0.034596	0.007170	0.000112	0.005244	0.010986
-0.052991	0.023871	0.008497	0.001331	0.001626	0.002690
-0.098747	0.047925	0.000031	0.000146	0.000974	0.018449
0.[illegible]	0.105998	0.009037	0.[illegible]	0.012691	0.000323
2.[illegible]	0.[illegible]	0.024539	0.002400	0.[illegible]	0.[illegible]
1.597344	0.026137	0.007178	0.007229	0.000118	0.000405
-0.183194	0.031738	0.001791	0.000218	0.007757	0.004261
-0.151100	0.035264	0.017558	0.002072	0.012568	0.001774
-1.412395	0.139301	0.036337	0.020609	0.[illegible]	0.012541
-1.054663	0.124222	0.[illegible]	0.[illegible]	0.026720	0.012537
-0.057758	0.026510	0.000235	0.004545	0.004167	0.000062
-0.874830	0.082842	0.000262	0.[illegible]	0.017295	0.005520
-0.337524	0.063179	0.012984	0.007330	0.[illegible]	0.002737
0.211473	0.058607	0.005973	0.008217	0.000748	0.029970
0.504034	0.084136	0.016191	0.006191	0.[illegible]	0.021139
0.127745	0.023326	0.001564	0.003874	0.001287	0.000095
-0.117304	0.094148	0.005279	0.023160	0.029619	0.021471
0.019137	0.071383	0.002542	0.000966	0.010809	0.026744
-0.016582	0.105363	0.002656	0.000005	0.029029	0.[illegible]
0.260264	0.026362	0.000675	0.004137	0.001504	0.003600
0.177140	0.034799	0.020653	0.[illegible]	0.009430	0.000045
LS=1.978 y Li=-1.978	LS=0.149146	LIMITE SUPERIOR = 0.097097			

Los límites para identificar " outliers " se obtuvieron a un nivel de significancia $\alpha = 0.05$

En la tabla que se presenta a continuación se muestran la distancia de Welsh-Kuh (DFFITS) que se utiliza para determinar si la i-ésima observación es influyente y el valor absoluto de la medida DFBETAS que es utilizada para determinar si la i-ésima observación es influyente con respecto al j-ésimo coeficiente ($\hat{\beta}_j$). Sus respectivas cotas superiores para detectar observaciones influyentes se presentan en la última fila de la tabla y los puntos influyentes aparecen sombreados en la tabla.

/ DFFITS /	/ DFBETAS 1 /	/ DFBETAS 2 /	/ DFBETAS 3 /	/ DFBETAS 4 /
0.020766	0.004777	0.010794	0.002155	0.004749
0.034797	0.024307	0.013611	0.006107	0.016572
0.030746	0.008678	0.021811	0.009090	0.008505
0.011239	0.003747	0.000817	0.002872	0.009045
0.137599	0.097528	0.082618	0.000601	0.107273
0.117795	0.070369	0.006867	0.097474	0.016673
0.012716	0.009232	0.000503	0.000257	0.005549
0.204675	0.079538	0.160193	0.048687	0.000914
0.413126	0.171955	0.073854	0.070443	0.259404
0.207964	0.000773	0.134854	0.093873	0.065577
0.131004	0.070025	0.020975	0.097143	0.027466
0.118494	0.070390	0.023069	0.045406	0.088257
[illegible]	[illegible]	0.019461	[illegible]	[illegible]
0.118618	0.058882	0.001132	0.086017	0.030825
0.027652	0.003735	0.019956	0.003213	0.005246
0.024387	0.008046	0.014302	0.008626	0.004357
0.155749	0.049188	0.017806	0.069434	0.121280
[illegible]	[illegible]	[illegible]	0.156501	[illegible]
0.027741	0.006373	0.018316	0.003950	0.000057
0.020131	0.001899	0.000243	0.006218	0.007858
0.049198	0.029789	0.012680	0.025295	0.004553
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
0.071362	0.034736	0.019775	0.063146	0.010312
0.050767	0.011371	0.002473	0.010451	0.019674
0.097471	0.037999	0.031782	0.016766	0.070645
0.033421	0.011073	0.000222	0.004463	0.006412
0.457153	0.157921	0.008439	0.224764	0.128701
[illegible]	[illegible]	[illegible]	0.116840	[illegible]
0.104156	0.085230	0.030885	0.034611	0.069955
0.465439	0.091743	0.056554	0.230558	0.262703
0.452665	[illegible]	[illegible]	0.039930	0.179029
0.499753	0.192399	0.233083	0.189954	[illegible]
0.030467	0.018353	0.013254	0.005546	0.007159
0.121637	0.055374	0.006916	0.047359	0.068543
0.008271	0.004935	0.001953	0.002159	0.003776
0.012507	0.000316	0.000691	0.005412	0.007760
[illegible]	0.243174	[illegible]	[illegible]	0.044078
[illegible]	[illegible]	0.138998	[illegible]	[illegible]
0.260045	0.136275	0.136759	0.017478	0.032371
0.033167	0.007879	0.002760	0.016397	0.012156
0.028888	0.020385	0.007003	0.017246	0.006480
[illegible]	[illegible]	0.210552	[illegible]	0.170490
0.397205	[illegible]	[illegible]	0.184220	0.126187
0.009535	0.000897	0.003947	0.003779	0.002463
0.262922	0.014796	0.184335	0.120132	0.067870
0.087615	0.016985	0.029861	0.067876	0.018244
0.057395	0.016934	0.019863	0.020019	0.037934
0.152769	0.067010	0.041440	0.100607	0.076575
0.019743	0.005112	0.008040	0.004038	0.001132
0.037817	0.008955	0.018757	0.020850	0.018960
0.005306	0.001001	0.000617	0.002005	0.003807
0.005691	0.000903	0.000036	0.002987	0.004082
0.042826	0.006860	0.016965	0.010227	0.015604
0.052365	0.032597	0.045977	0.018080	0.201249
LS=0.544981	LIMITE SUPERIOR = 0.272166			

Los resultados obtenidos en las tablas anteriores se pueden resumir en :

I. Las observaciones 9, 18, 22, 27, 30, 37 y 38 son observaciones discrepantes "outliers", es decir , sus residuos salen del rango de confianza.

II. Las observaciones de las variables explicativas que generan puntos palanca son:

1) La observación 13 en la primera variable explicativa utilizada para representar la constante del modelo y en las variables explicativas **E** y **FE**.

2) La observación 17 en la variable explicativa **FE**.

3) La observación 28 en la primera variable explicativa utilizada para representar la constante del modelo y en la variable explicativa **C**.

4) La observación 32 en las variables explicativas **C** y **FE**.

5) La observación 38 en las variables explicativas **E** y **FE**.

III. Considerando a la medida **DFFITS** las observaciones 13 ,18, 22, 28, 37, 38 y 42 son influyentes y en base a la medida **DFBETAS** podemos afirmar lo siguiente:

1) Las observaciones 13 y 38 influyen en la estimación de los parámetros $\beta 1$, $\beta 3$ y $\beta 4$.

2) Las observaciones 18 y 28 influyen en la estimación de los parámetros $\beta 1$, $\beta 2$ y $\beta 4$.

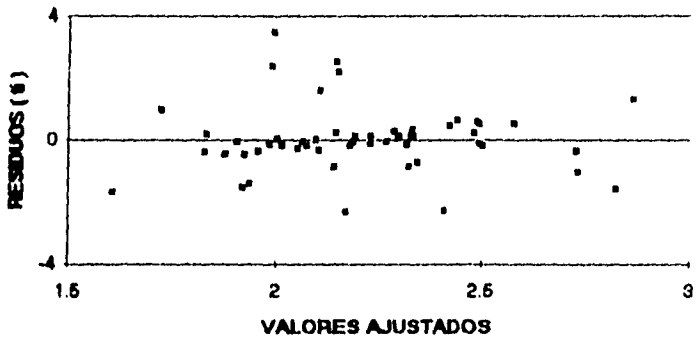
3) La observación 22 influye en la estimación de **todos** los parámetros .

4) La observación 37 influye en la estimación de los parámetros $\beta 2$ y $\beta 3$.

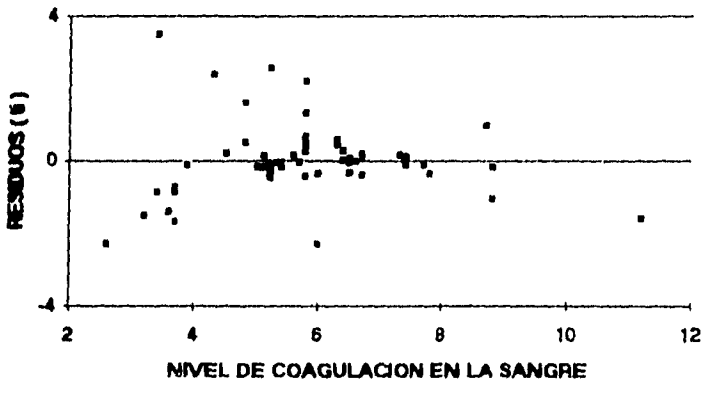
5) La observación 42 influye en la estimación de los parámetros $\beta 1$ y $\beta 3$.

GRAFICAS PARA LOS RESIDUOS

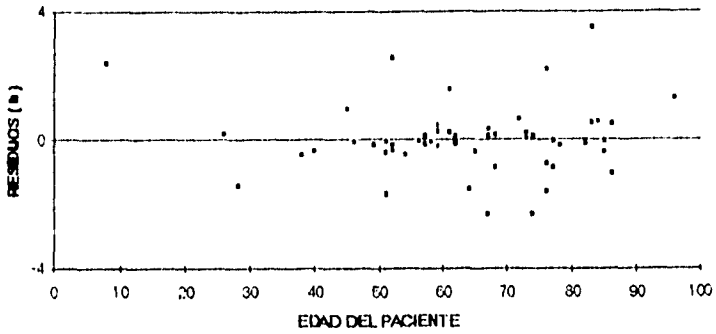
RESIDUOS (ti) CONTRA VALORES AJUSTADOS



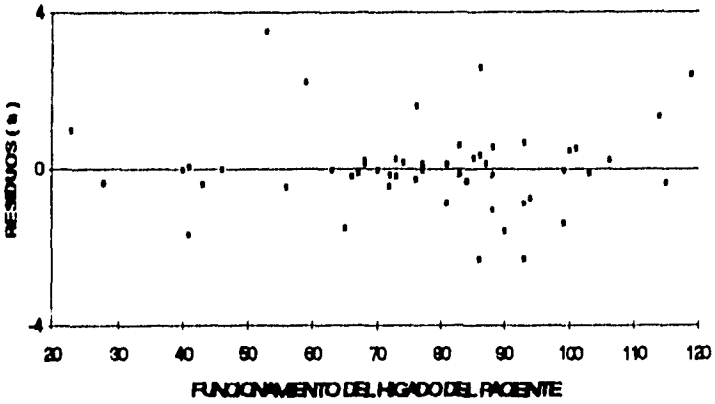
RESIDUOS (ti) CONTRA VARIABLE EXPLICATIVA (C)

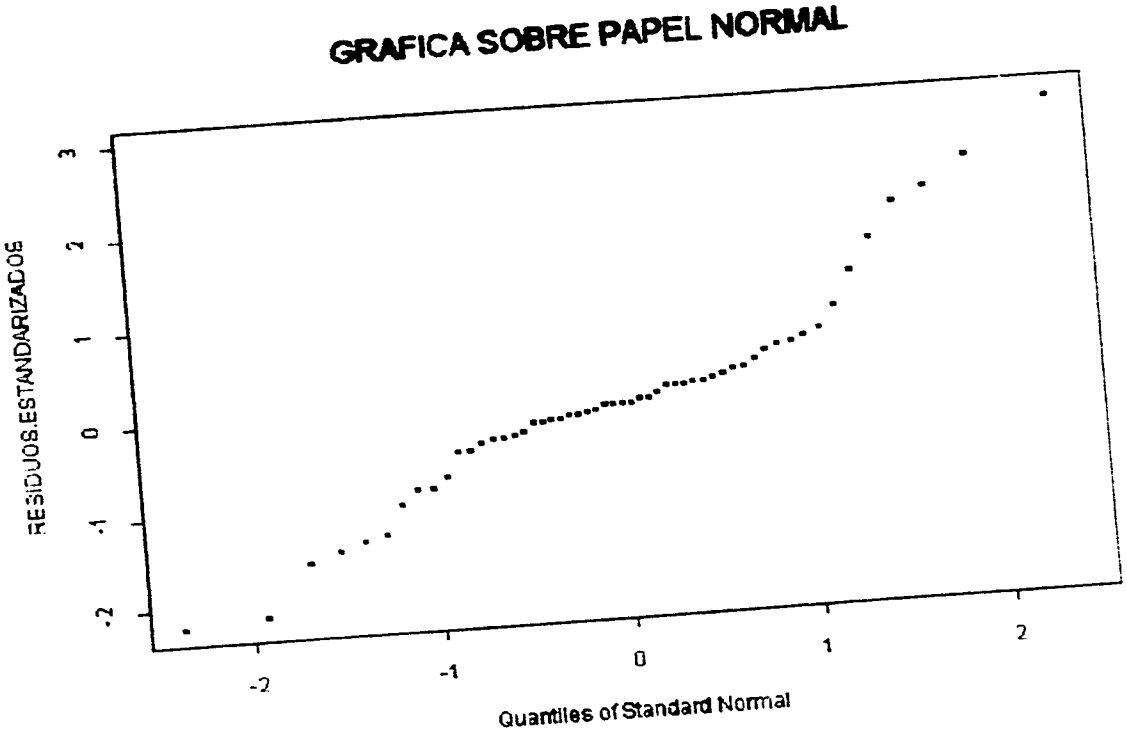


RESIDUOS (ti) CONTRA VARIABLE EXPLICATIVA (E)



RESIDUOS (ti) CONTRA VARIABLE EXPLICATIVA (FE)




GRAFICA SOBRE PAPEL NORMAL
RESIDUOS.ESTANDARIZADOS
3
2
1
0
-1
-2
-2
-1
0
1
2
Quantiles of Standard Normal

VERIFICACION DEL SUPUESTO DE NORMALIDAD.

Para verificar el supuesto de normalidad se utilizará la estadística Anderson-Darling modificada ($\overline{A}^2$), que se describe a continuación :

1. Se cuenta con los residuos Rao-Blackwell (s_i)

$$s_i = \frac{(n-k-1)^{1/2}\left(Y_i - \underline{x}_i'\hat{\underline{\beta}}\right)}{\sqrt{(1-h_{ii})(n-k)\hat{\sigma}^2 - \left(Y_i - \underline{x}_i'\hat{\underline{\beta}}\right)^2}}$$

2. Se consideran los estadísticos de orden $s_{(i)}$, con :

$$s_{(i)} < s_{(i+1)} \qquad \forall \quad i = 1,\ldots,n-1$$

3. Sea $z_{(i)} = G_{n-k-1}\left(s_{(i)}\right)$ en dónde $G_{n-k-1}\left(s_{(i)}\right)$ es la función de distribución de una t de Student con n-k-1 grados de libertad evaluada en $\left(s_{(i)}\right)$

4. Se calcula el valor $\overline{A}^2$

$$\overline{A}^2 = -n - \left(1/n\right)\sum_{i=1}^{n}(2i-1)\left[\ln z_{(i)} + \ln\left\{1 - z_{(n-i+1)}\right\}\right]$$

Los resultados obtenidos de los puntos anteriores se presentan en la siguiente tabla:

S_i	$S_{(i)}$	$z_{(i)}$
0.126193	-2.318344	0.012323
-0.199697	-2.304126	0.012748
-0.141319	-1.691658	0.048531
0.038413	-1.602707	0.057713
-0.366817	-1.527038	0.066591
-0.457231	-1.412395	0.082077
-0.057160	-1.054663	0.148376
-0.861647	-0.874830	0.192967
-2.304126	-0.861647	0.196540
-0.746553	-0.746553	0.229450
0.570854	-0.466706	0.321389
-0.398699	-0.457233	0.324762
1.304571	-0.398699	0.345923
0.518073	-0.370885	0.356181
-0.123368	-0.366817	0.357667
0.112603	-0.337524	0.368582
-0.370885	-0.285351	0.388289
-1.691656	-0.199597	0.421311
0.143258	-0.189372	0.425292
0.126645	-0.183194	0.427701
-0.285351	-0.177140	0.430064
3.495164	-0.151100	0.440259
0.189739	-0.141319	0.444099
0.331576	-0.123368	0.451160
0.442457	-0.117304	0.453550
0.236488	-0.057758	0.477088
2.552274	-0.057160	0.477325
-1.602707	-0.055747	0.477885
-0.458706	-0.052891	0.479017
2.194512	-0.019137	0.492405
-1.527038	-0.016582	0.493419
0.965578	0.038413	0.515243
-0.189372	0.112603	0.544597
0.642547	0.125645	0.549736
-0.052891	0.126193	0.549952
-0.055747	0.127749	0.550565
-2.318343	0.143258	0.556663
2.390321	0.189739	0.574851
1.587343	0.211473	0.583302
-0.183193	0.236488	0.59298
-0.151100	0.260263	0.602124
-1.412393	0.331576	0.629225
-1.054662	0.442467	0.669950
-0.057758	0.504033	0.691751
-0.874830	0.518073	0.696830
-0.337523	0.570854	0.714646
0.211473	0.642547	0.738243
0.504033	0.965578	0.830501
0.127749	1.304571	0.900933
-0.117303	1.587343	0.940566
-0.019137	2.194512	0.983515
-0.016582	2.390321	0.989638
0.260263	2.552273	0.993064
-0.177139	3.495164	0.999492

De tal forma:

$$A^2 = 1.84840111$$

PRUEBA A^2

(CASO NORMAL CON AMBOS PARAMETROS DESCONOCIDOS)

α (NIVEL)

0.50	0.25	0.15	0.10	0.05	0.025	0.01	0.005
VALORES CRITICOS							
0.341	0.470	0.561	0.631	0.752	0.873	1.035	1.159

De la tabla de valores críticos, con cualquier valor de α mayor o igual a .005 , se rechaza la hipótesis nula y se concluye que las perturbaciones estocásticas no se distribuyen normalmente. Como las tablas terminan con un valor de $\alpha = .005$, no se puede exhibir el nivel de significancia observado, en este caso (p - value).

Por otra parte si se verifica el supuesto de normalidad en las perturbaciones estocásticas a través de la prueba **Bera-Jarque** se obtiene el resultado siguiente:

$$\lambda = 8.3264948$$

DISTRIBUCION *de* λ

α (NIVEL)

0.50	0.25	0.20	0.10	0.05	0.025	0.01	0.005
VALORES CRITICOS (asintótica)							
1.3863	2.7726	3.2190	4.6052	5.9915	7.3777	9.2103	10.5966
VALORES CRITICOS (N finito = 54)							
			2.9304	4.2616			

Por lo tanto , con $\alpha = .05$ o cualquier valor mayor, se rechazaría la hipótesis nula .

A continuación se procede a realizar la verificación del supuesto de normalidad eliminando las observaciones discrepantes " outliers ", con los resultados siguientes:

$$A^2 = 1.01193646$$

De la tabla de valores críticos, con cualquier valor de α mayor o igual a .025 , se rechaza la hipótesis nula y se concluye que las perturbaciones estocásticas no se distribuyen normalmente. Se puede concluir que aún sin considerar las observaciones discrepantes, el supuesto de normalidad en las perturbaciones estocásticas sigue siendo rechazado inclusive con niveles de significancia muy pequeños.

Por otra parte si se verifica el supuesto de normalidad en las perturbaciones estocásticas sin considerar las observaciones discrepantes a través de la prueba **Bera-Jarque** se obtiene el resultado siguiente:

$$\lambda = 2.774676$$

DISTRIBUCION *de* λ

α (NIVEL)

0.50	0.25	0.20	0.10	0.05	0.025	0.01	0.005
VALORES CRITICOS (asintótica)							
1.3863	2.7726	3.2190	4.6052	5.9915	7.3777	9.2103	10.5966
VALORES CRITICOS (N finito = 47)							
			2.84	4.179			

Por lo tanto , con α = .10 o cualquier valor menor, se aceptaría la hipótesis nula . Se puede ver que al eliminar las observaciones discrepantes la decisión con respecto a la hipótesis de normalidad en las perturbaciones estocásticas es muy distinta .

El tercer ejemplo se extrajo de William H. Greene (1993, p. 206, 207) del libro "Ecomometric Analysis".

EJEMPLO 3.

La industria SIC 33 productora de metales está interesada en desarrollar un modelo de regresión que le permita predecir adecuadamente el valor agregado anual a declarar.

Tomando como base un análisis empírico, se utilizan los datos declarados en años anteriores basados en los siguientes conceptos:

1) Valor Agregado (costo final).

2) Labor (costo de mano de obra).

3) Capital (costo de las instalaciones y del equipo).

Los datos observados de los veintisiete años anteriores por estos conceptos se presentan a continuación:

OBSERVACIONES	LOGARITMO DEL VALOR AGREGADO (VA*)	LOGARITMO DE LABOR (L*)	LOGARITMO DE CAPITAL (C*)
1	6.488126	5.089508	5.634754
2	6.841541	5.367983	6.296188
3	7.012701	5.228109	8.581346
4	7.090818	5.504640	7.062774
5	6.959095	5.363752	6.699217
6	8.133300	6.537575	8.424644
7	7.794778	6.115428	8.029404
8	8.356428	6.571163	8.627842
9	7.393380	5.770007	7.389410
10	7.148385	5.534061	7.353774
11	6.911996	5.466694	6.496326
12	6.395045	4.946843	6.774647
13	6.748877	4.977010	7.436605
14	7.061017	5.481763	6.983595
15	7.558804	5.385679	7.664130
16	9.195142	7.355532	9.546066
17	6.992345	5.468969	6.784729
18	8.999080	6.987583	9.118192
19	8.063186	6.257169	8.645936
20	7.410577	5.719820	7.439007
21	8.548559	6.728258	8.557636
22	8.125158	5.648974	8.120983
23	6.384941	5.015755	5.979632
24	7.378996	5.560335	7.616741
25	7.633297	6.209797	7.821234
26	7.727996	5.421750	7.404527
27	6.614283	4.919981	8.644558

El modelo que se propone consiste en hacer una regresión de la variable dependiente, representada en este caso por el logaritmo del valor agregado anual (VA*), en las variables explicativas que se definen a continuación:

L* = Logaritmo de Labor o costo de mano de obra.

C* = Logaritmo de Capital o costo de las instalaciones y del equipo.

Modelo de regresión lineal múltiple:

$$VA^* = \beta_1 + \beta_2(L^*) + \beta_3(C^*) + \varepsilon$$

Al realizar la estimación de los parámetros con el método de mínimos cuadrados ordinarios se obtienen los resultados siguientes:

$$\hat{\underline{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 1.2100 \\ 0.5570 \\ 0.4088 \end{pmatrix} \qquad \hat{\sigma}^2 = 0.0386$$

$$SCR = 14.1316 \qquad n = 27, \quad k = 3$$
$$SCE = 0.9261 \qquad R^2 = 0.9385$$
$$SCT = 15.0576 \qquad \overline{R}^2 = 0.9334$$

De tal forma la estimación del modelo de regresión lineal múltiple está dada por la siguiente función:

$$\hat{VA}^*_t = 1.2100 + 0.5570(L^*_t) + 0.4088(C^*_t) \qquad t = 1, 2, \ldots, 27.$$

La estimación de la matriz de varianzas y covarianzas de los coeficientes estimados es:

$$\widehat{VCov}(\hat{\underline{\beta}}) = \begin{pmatrix} 0.1159 & -0.0208 & 0.0006 \\ -0.0208 & 0.0161 & -0.0096 \\ 0.0006 & -0.0096 & 0.0073 \end{pmatrix}$$

PRUEBAS DE SIGNIFICANCIA DEL MODELO.

La prueba de significancia global del modelo es la siguiente :

$$H_0: \beta_2 = \beta_3 = 0 \qquad vs. \qquad H_1: \beta_j \neq 0 \quad para\ algún \quad j = 2,3.$$

$$F = \frac{(S.S.R.)/(k-1)}{(S.S.E.)/(n-k)} \sim F_{(k-1,\, n-k)}$$

Por lo que :

$$F = 183.1188$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$F^{\alpha}_{(2,24)} = 3.40$$

Por lo tanto se rechaza H_0 , es decir $\beta_j \neq 0$ para algún $j = 2,3.$

Pruebas t bajo la hipótesis ($\beta_j = 0$)

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

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_1 = 0$

Por lo tanto:

$$t = 3.5539$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(24)}^{0.025} = 2.064$$

Por lo que se rechaza H_0 , es decir , $\beta_1 \neq 0$, lo cual indica que el valor del modelo con todas las variables explicativas iguales a cero es distinto de cero.

$$H_0 : \beta_2 = 0 \qquad vs. \qquad H_1 : \beta_2 \neq 0$$

$$t = \frac{\tilde{\beta}_2 - \beta_2}{\sqrt{\widehat{var}(\tilde{\beta}_2)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_2 = 0$

Por lo tanto :

$$t = 4.3932$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(24)}^{0.025} = 2.064$$

Por lo que se rechaza H_0 , es decir , $\beta_2 \neq 0$, lo cual indica que la variable explicativa TI es significativa en éste modelo.

$$H_0: \beta_3 = 0 \qquad vs. \qquad H_1: \beta_3 \neq 0$$

$$t = \frac{\hat{\beta}_3 - \beta_3}{\sqrt{\widehat{Var}(\hat{\beta}_3)}} \sim t_{(n-k)}$$

Bajo H_0 se tiene que $\beta_3 = 0$

Por lo tanto :

$$t = 4.7909$$

Utilizando un nivel de significancia ($\alpha = 0.05$) se obtiene un valor en tablas de :

$$t_{(24)}^{0.025} = 2.064$$

Por lo que se rechaza H_0 , es decir , $\beta_3 \neq 0$, lo cual indica que la variable explicativa M es significativa en éste modelo.

Al contar con un modelo adecuado en base a las pruebas de significancia se presenta a continuación la tabla que contiene a todos los residuos mencionados en el segundo capítulo.

Residuos ordinarios ei	Residuos estandarizados wi	Residuos est. internos ri	Residuos predictivos ei(i)	Residuos Rao-Blackwell ei=ti
0.139902	0.712216	0.823607	0.187086	0.817907
0.067835	0.345334	0.371352	0.078441	0.364582
0.200345	1.019920	1.056305	0.214895	1.058973
-0.072357	-0.368357	-0.376550	-0.075612	-0.369716
0.028574	0.145463	0.150188	0.030460	0.147095
-0.161915	-0.824275	-0.867855	-0.179488	-0.863235
-0.103736	-0.528096	-0.542305	-0.109392	-0.534169
-0.040554	-0.206454	-0.217925	-0.045186	-0.213548
-0.051125	-0.260265	-0.265577	-0.053233	-0.260368
-0.150127	-0.764270	-0.782209	-0.157258	-0.775691
0.002464	0.012545	0.013317	0.002777	0.013037
-0.339654	-1.729115	-1.838878	-0.384145	-1.942173
-0.273208	-1.390848	-1.625681	-0.373254	-1.687049
-0.057051	-0.290434	-0.297375	-0.059810	-0.291651
0.220168	1.120835	1.201032	0.252802	1.212755
-0.014079	-0.071675	-0.084310	-0.019481	-0.082547
-0.037252	-0.189643	-0.195866	-0.039737	-0.191896
0.169709	0.863958	0.954572	0.207176	0.952735
-0.166294	-0.846569	-0.897493	-0.186901	-0.893722
-0.026248	-0.133613	-0.136159	-0.027256	-0.133343
0.092769	0.472268	0.507970	0.107325	0.499970
0.449032	2.285936	2.482052	0.529384	2.818282
-0.021888	-0.111426	-0.121434	-0.025996	-0.118914
-0.041643	-0.211995	-0.219627	-0.044695	-0.215219
-0.232690	-1.184578	-1.228883	-0.250421	-1.242742
0.471297	2.399279	2.487869	0.506743	2.827179
-0.052278	-0.266137	-0.281895	-0.058652	-0.276418

A continuación se presenta una tabla con los residuos estudentizados externos(t_i), los elementos de la diagonal de la matriz (H) y el vector normalizado de residuos cuadrados (δ_i); Sus cotas superiores e inferiores respectivas para detectar puntos discrepantes "outliers" y puntos palanca se presentan en la última fila de la tabla y dichos puntos aparecen sombreados en la tabla :

ti	hii	di1	di2	di3
0.817907	0.252208	0.041125	0.000829	0.179498
0.364582	0.135217	0.014961	0.044109	0.006596
1.058973	0.067704	0.029374	0.000103	0.008178
-0.369716	0.043041	0.007553	0.000013	0.001541
0.147095	0.061941	0.017021	0.002111	0.012319
-0.863235	0.097909	0.049151	0.021591	0.001419
-0.534169	0.051706	0.008852	0.000696	0.001008
-0.213548	0.102504	0.053023	0.008195	0.000986
-0.260368	0.039802	0.000055	0.002437	0.002396
-0.776891	0.045342	0.008472	0.007963	0.004948
0.013037	0.112626	0.009793	0.039635	0.089424
-1.042178	0.115817	0.066407	0.080363	0.023767
-1.687049	0.200089	0.064737	0.250888	0.180182
-0.291651	0.046137	0.009703	0.000372	0.003671
1.212755	0.129089	0.016736	0.000288	0.001580
-0.092547	0.277288	0.218990	0.059493	0.000315
-0.191896	0.062540	0.009113	0.007636	0.019487
0.952735	0.190842	0.128182	0.029192	0.000043
-0.893722	0.110260	0.017546	0.014185	0.047782
-0.133343	0.037043	0.000551	0.000004	0.000002
0.499970	0.135630	0.078598	0.047919	0.007888
2.819282	0.151784	0.002684	0.000048	0.114200
-0.118914	0.158037	0.051975	0.020491	0.078279
-0.215219	0.089289	0.005582	0.030046	0.028745
-1.242742	0.070806	0.015714	0.027980	0.012669
2.527178	0.069950	0.013420	0.032845	0.024498
-0.276418	0.109675	0.070143	0.045383	0.012770
LS=1.714 y Li=-1.714	LS=0.225333	LIMITE SUPERIOR = 0.074074		

Los límites para identificar " outliers " se obtuvieron a un nivel de significancia $\alpha = 0.05$

En la tabla que se presenta a continuación se muestran la distancia de Welsh-Kuh (DFFITS) que se utiliza para determinar si la i-ésima observación es influyente y el valor absoluto de la medida DFBETAS que es utilizada para determinar si la i-ésima observación es influyente con respecto al j-ésimo coeficiente ($\hat{\beta}_j$). Sus respectivas cotas superiores para detectar observaciones influyentes se presentan en la última fila de la tabla y los puntos influyentes aparecen sombreados en la tabla.

/ DFFITS / /	/ DFBETAS j1 /	/ DFBETAS j2 /	/ DFBETAS j3 /
0.474993	0.191806	0.269068	0.[illegible]
0.144164	0.047953	0.082339	0.115331
0.285375	0.184743	0.011128	0.099181
0.078408	0.032846	0.001366	0.014838
0.037798	0.019814	0.006978	0.016857
0.284391	0.201498	0.133548	0.034241
0.124732	0.051609	0.014458	0.017418
0.072169	0.051905	0.020405	0.007078
0.052871	0.001968	0.013115	0.013004
0.169051	0.063869	0.070843	0.055838
0.004644	0.001298	0.002720	0.003646
0.702916	0.582259	0.507469	0.318423
1.020991	0.501799	0.947767	0.936978
0.084142	0.027958	0.005757	0.018092
0.466906	0.168117	0.[illegible]	0.371060
0.051126	0.045217	0.023683	0.001723
0.049564	0.018919	0.017319	0.027667
0.447649	0.373928	0.179853	0.006900
0.314815	0.125503	0.112845	0.207110
0.026153	0.003190	0.000284	0.000171
0.198049	0.150449	0.117720	0.047760
1.182187	0.158543	0.[illegible]	1.084102
0.051519	0.029545	0.019551	0.035557
0.058265	0.016658	0.038649	0.037903
0.343055	0.161610	0.215662	0.145113
0.775343	0.339601	0.[illegible]	0.468647
0.096519	0.077543	0.062373	0.033090
LS= 0.[illegible]	LIMITE SUPERIOR = 0.[illegible]		

Los resultados obtenidos en las tablas anteriores se pueden resumir en :

I. Las observaciones 12 , 22 y 26 son observaciones discrepantes "outliers" , es decir, su residuo sale del rango de confianza.

II. Las observaciones 1,13 y 16 son puntos palanca ya que su residuo sale del rango de confianza y existen varios valores de las δij que rebasan su cota superior generando palanca parcial de observaciones para alguna variable explicativa las cuales son :

1) Las observaciones 1 , 13 ,15 y 22 en las variables explicativas **L*** y **C***.

2) Las observaciones 2 y 23 en la variable explicativa **C***.

3) La observación 16, 18 y 21 en la primera variable explicativa utilizada para representar la constante del modelo .

III. Considerando a la medida **DFFITS** las observaciones 12, 13 y 22 son influyentes y en base a la medida **DFBETAS** se puede afirmar lo siguiente :

1) La observación 1 influye en la estimación del parámetro $\beta 3$.

2) La observación 12 influye en la estimación de los parámetros $\beta 1$ y.$\beta 2$.

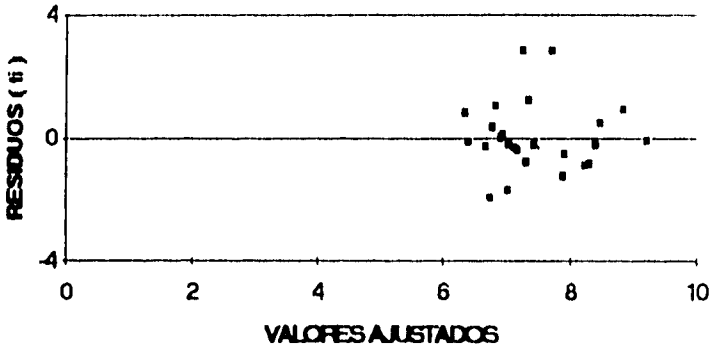
3) La observación 13 influye en la estimación de los parámetros $\beta 1$, $\beta 2$ y $\beta 3$.

4) La observación 15 influye en la estimación del parámetro $\beta 2$.

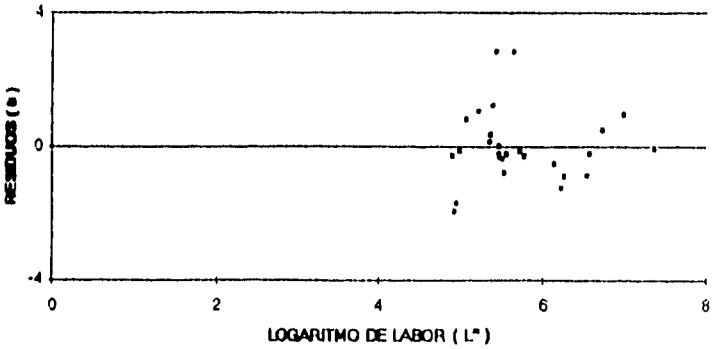
5) La observación 22 y 26 influye en la estimación de los parámetros $\beta 2$ y.$\beta 3$.

GRAFICAS PARA LOS RESIDUOS.

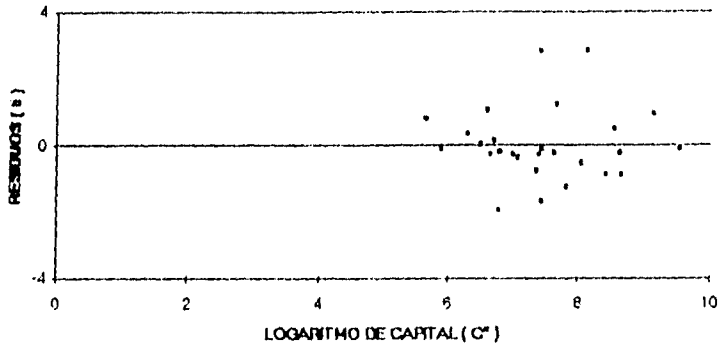
RESIDUOS (ti) CONTRA VALORES AJUSTADOS



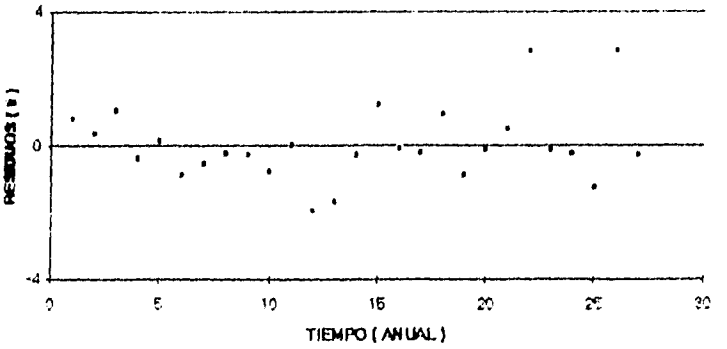
RESIDUOS (ti) CONTRA LA VARIABLE EXPLICATIVA (L*)



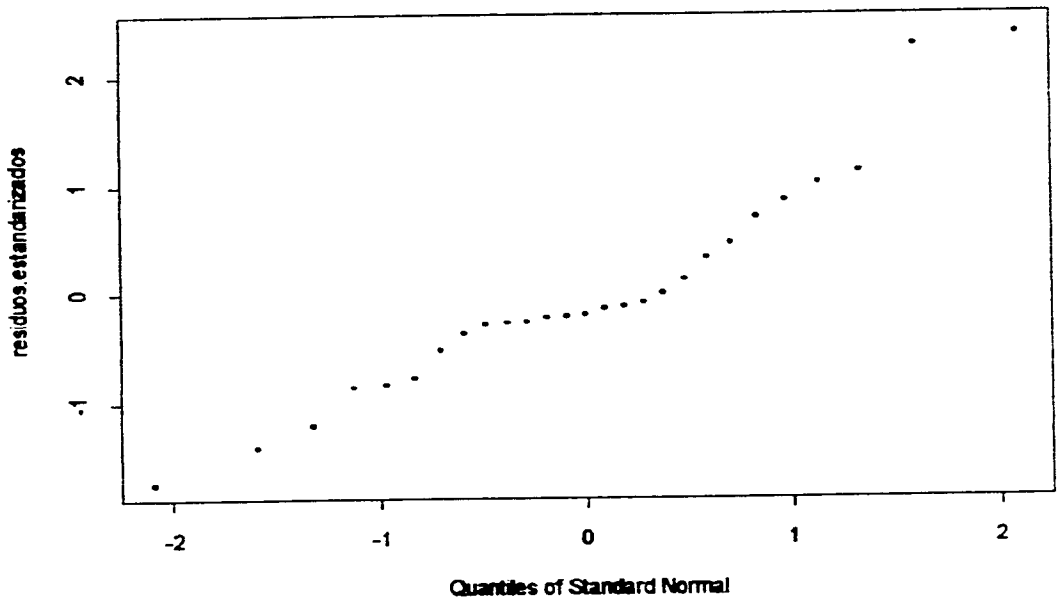
RESIDUOS (ûi) CONTRA VARIABLE EXPLICATIVA (C*)



RESIDUOS (ûi) CONTRA EL TIEMPO



GRAFICA SOBRE PAPEL NORMAL



VERIFICACION DEL SUPUESTO DE NORMALIDAD.

Para verificar el supuesto de normalidad se utilizará la estadística Anderson-Darling modificada ($\overline{A}^2$), que se describe a continuación :

1. Se cuenta con los residuos Rao-Blackwell (s_i)

$$s_i = \frac{(n-k-1)^{1/2}\left(Y_i - x_i'\tilde{\beta}\right)}{\sqrt{(1-h_{ii})(n-k)\tilde{\sigma}^2 - \left(Y_i - x_i'\tilde{\beta}\right)^2}}$$

2. Se consideran los estadísticos de orden $s_{(i)}$, con :

$$s_{(i)} < s_{(i+1)} \qquad \forall \quad i = 1,\ldots,n-1$$

3. Sea $z_{(i)} = G_{n-k-1}(s_{(i)})$ en dónde $G_{n-k-1}(s_i)$ es la función de distribución de una t de Student con n-k-1 grados de libertad evaluada en $(s_{(i)})$

4. Se calcula el valor $\overline{A}^2$

$$\overline{A}^2 = -n - \left(\frac{1}{n}\right)\sum_{i=1}^{n}(2i-1)\left[\ln z_{(i)} + \ln\left\{1 - z_{(n-i+1)}\right\}\right]$$

Los resultados obtenidos de los puntos anteriores se presentan en la siguiente tabla:

S_i	$S_{(i)}$	$z_{(i)}$
0.81790684	-1.94217319	0.03223017
0.36458217	-1.68704856	0.05255597
1.05897348	-1.24274235	0.11324192
-0.36971564	-0.89372150	0.19036388
0.14709535	-0.86323535	0.19846037
-0.86323535	-0.77569104	0.22291732
-0.53416939	-0.53416939	0.29917508
-0.21354808	-0.36971564	0.35748838
-0.26036800	-0.29165147	0.38658371
-0.77569104	-0.27641785	0.39234794
0.01303670	-0.26036800	0.39844838
-1.94217319	-0.21521870	0.41574691
-1.68704856	-0.21354809	0.41639058
-0.29165147	-0.19189577	0.42476399
1.21276452	-0.13334339	0.44754083
-0.08254661	-0.11891388	0.45318805
-0.19189577	-0.08254661	0.46746313
0.95273487	0.01303670	0.50514453
-0.89372150	0.14709535	0.55783072
-0.13334338	0.36458217	0.64062299
0.49996995	0.49996995	0.68907695
2.81828187	0.81790684	0.78909834
-0.11891388	0.95273487	0.82468474
-0.21521870	1.05897348	0.84969641
-1.24274235	1.21276452	0.88123324
2.82717938	2.81828187	0.99512436
-0.27641785	2.82717938	0.99522328

De tal forma:

$$\overline{A}^2 = 0.773866$$

PRUEBA A^2

(CASO NORMAL CON AMBOS PARAMETROS DESCONOCIDOS)

α **(NIVEL)**

0.50	0.25	0.15	0.10	0.05	0.025	0.01	0.005
			VALORES CRITICOS				
0.341	0.470	0.561	0.631	0.752	0.873	1.035	1.159

De la tabla de valores críticos, con cualquier valor de α mayor o igual a .05 , se rechaza la hipótesis nula y se concluye que a dichos niveles de significancia las perturbaciones estocásticas no se distribuyen normalmente. Por lo tanto el nivel de significancia observado (p - value) es menor de .05 .

Por otra parte si se verifica el supuesto de normalidad en las perturbaciones estocásticas a través de la prueba **Bera-Jarque** se obtiene el resultado siguiente:

$$\lambda = 3.472672$$

DISTRIBUCION *de* λ

α **(NIVEL)**

0.50	0.25	0.20	0.10	0.05	0.025	0.01	0.005
			VALORES CRITICOS (asintótica)				
1.3863	2.7726	3.2190	4.6052	5.9915	7.3777	9.2103	10.5966
			VALORES CRITICOS (N finito = 27)				
			2.382	3.575			

Por lo tanto , con α = .10 o cualquier valor mayor, se rechazaría la hipótesis nula . pero si se utiliza α = .05 , entonces ésta prueba no rechaza la hipótesis de normalidad.

A continuación se procede a realizar la verificación del supuesto de normalidad eliminando las observaciones discrepantes " outliers ", con los resultados siguientes:

$$\overline{A}^2 = 0.687921$$

De la tabla de valores críticos, con cualquier valor de α mayor o igual a .10 , se rechaza la hipótesis nula y se concluye que a dichos niveles de significancia las perturbaciones estocásticas no se distribuyen normalmente. Por lo tanto el nivel de significancia observado (p - value) es menor de .10 .

Se puede concluir que al no considerar las observaciones discrepantes y al utilizar un nivel de significancia $\alpha = 0.05$, la decisión de la prueba cambia, es decir , ahora se obtiene que las perturbaciones estocásticas se distribuyen normalmente.

Por otra parte si se verifica el supuesto de normalidad en las perturbaciones estocásticas a través de la prueba **Bera-Jarque** se obtiene el resultado siguiente:

$$\lambda = 3.243361$$

DISTRIBUCION *de* λ

α (NIVEL)

0.50	0.25	0.20	0.10	0.05	0.025	0.01	0.005
VALORES CRITICOS (asintótica)							
1.3863	2.7726	3.2190	4.6052	5.9915	7.3777	9.2103	10.5966
VALORES CRITICOS (N finito = 24)							
			2.274	3.44			

Por lo tanto , con α = .10 o cualquier valor mayor, se rechaza la hipótesis nula y se concluye que las perturbaciones estocásticas no se distribuyen normalmente. Se observa bajo esta prueba que al eliminar los puntos discrepantes del análisis, la decisión no cambia.

CAPITULO V

SIMULACION

V . 1 OBJETIVO Y PROCEDIMIENTO DE LA SIMULACION.

En este capítulo se presenta una simulación que consistió en generar perturbaciones estocásticas de tal forma que se distribuyan normalmente para poder evaluar el valor α (probabilidad de rechazar el supuesto de normalidad dado que se cumple dicho supuesto) con dos propósitos principales:

- Comparar los niveles de significancia presentados en la tabla de la distribución límite del estadístico Anderson-Darling (A^2) con los niveles de significancia obtenidos por la simulación para la estadística Anderson-Darling modificada ($\overline{A}^2$) .

- Obtener una tabla para la verificación del supuesto de normalidad a través de la estadística $\overline{A}^2$.

El procedimiento para evaluar el valor α para la estadística $\overline{A}^2$ toma como base el ejemplo 2 y es el que se presenta a continuación :

1) Se genera un vector de perturbaciones estocásticas ($\underline{\varepsilon}$) de tamaño 54 distribuido normalmente con media cero y varianza $\hat{\sigma}^2$ obtenida del ejemplo 2.

2) Por medio de dicho vector de perturbaciones estocásticas se obtiene el vector de residuos a través de la siguiente relación :

$$\underline{e} = (I - H)\,\underline{\varepsilon} \quad ,$$

donde:

H es la matriz sombrero ("Hat-Matrix") obtenida por las variables explicativas del ejemplo 2.

3) Utilizando el vector de residuos obtenido se calculan los residuos RAO - BLACKWELL :

$$s_i = \frac{(n-k-1)^{1/2} e_i}{\sqrt{(1-h_{ii})(n-k)\hat{\sigma}^2 - e_i^2}} \quad ,$$

donde:

n= 54 , k = 4 y $\hat{\sigma}^2 = 0.0022$ por ser tomados del ejemplo 2 .

h_{ii} son los elementos de la diagonal de la matriz H mencionada en el punto 2.

4) Se calculan las estadísticas de orden $s_{(i)}$ y se obtienen los valores de $z_{(i)}$ transformando los residuos RAO - BLACKWELL :

$$z_{(i)} = G_{n-k-1}\left(s_{(i)}\right) \quad ,$$

donde:

$G_\gamma(*)$ es la distribución de una variable aleatoria t de Student con γ grados de libertad evaluada en $*$.

5) Calcular el valor de la estadística $\bar{A}^2$ por medio de la estadística Anderson-Darling (A^2) .

6) Repetir el procedimiento anterior 1000 veces generando vectores de perturbaciones estocásticas ($\underline{\varepsilon}$) distintos de tamaño 54 .

V . 2 COMPARACION CON LA TABLA DE DISTRIBUCION LIMITE DEL ESTADISTICO A^2.

Los valores de la estadística $\bar{A}^2$ calculados con el procedimiento de la sección anterior son comparados con el valor de la tabla de distribución límite del estadístico A^2 y se obtiene el número de veces que se rechaza el supuesto de normalidad en las perturbaciones estocásticas (dado que dichas perturbaciones provienen de la distribución normal) para los distintos niveles de significancia.

Los resultados obtenidos en las mil veces que se realizó el procedimiento se presenta para los distintos niveles de significancia α en la tabla siguiente:

COMPARACION DEL NIVEL EMPIRICO α' Y TEORICO α

α Teórico	0.50	0.25	0.15	0.10	0.05	0.025	0.01	0.005
Valores críticos	0.341	0.470	0.561	0.631	0.752	0.873	1.035	1.159
* Nivel empírico α'	0.504	0.252	0.143	0.096	0.043	0.028	0.011	0.005

* Proporción de veces (de 1000) en que $\bar{A}^2$ excedió el valor crítico.

Se puede ver que la proporción de veces que se rechaza el supuesto de normalidad en la simulación (α') se aproxima al nivel teórico de significancia α de la tabla de distribución límite de la estadística A^2 .

V.3 GENERACION DE UNA TABLA PARA VERIFICACION DEL SUPUESTO DE NORMALIDAD A TRAVES DE LA ESTADISTICA $\bar{A}^2$.

En esta sección se construye la tabla para la verificación del supuesto de normalidad en las perturbaciones estocásticas por medio de la simulación realizada al inicio del capítulo.

Para la construcción de la tabla se busca el mínimo valor al cual el porcentaje de veces que se rechaza el supuesto de normalidad en la simulación (α') coincide con los niveles de significancia presentados en la tabla de distribución límite de la estadística Anderson-Darling (A^2). La tabla obtenida con una aproximación de cuatro decimales para el valor mínimo se presenta a continuación :

α'	0.50	0.25	0.15	0.10	0.05	0.025	0.01	0.005
V.M.T	0.3422	0.4715	0.5559	0.6262	0.7198	0.8969	1.0762	1.1292

Donde:

V.M.T. es el valor mínimo de tablas al cual el porcentaje de rechazo coincide con α'.

CONCLUSIONES

El método presentado para la verificación del supuesto de normalidad en las perturbaciones estocásticas tiene bases teóricas muy firmes que lo respaldan y su aplicación es sencilla, por lo cual se propone para ser utilizado en los modelos de regresión lineal y de tal forma poder apoyarse en esta prueba para justificar (en el caso de aceptar normalidad en las perturbaciones estocásticas) el uso en el análisis de regresión de dicho supuesto.

Al realizar la comparación de este método de verificación del supuesto de normalidad con otros ya existentes se puede concluir lo siguiente:

1) La Gráfica Sobre Papel Normal es un método muy sencillo para verificar el supuesto de normalidad , sin embargo , no es un método que establezca estadísticamente el resultado de lo que se desea probar ya que la decisión de aceptar o rechazar la normalidad depende de la apreciación.

   En el método ilustrado en este trabajo , la decisión no depende de la apreciación . Además dicho método determina con que probabilidad se cometería el error de afirmar que las perturbaciones estocásticas no se distribuyen normalmente siendo que sí lo hacen.

2) La prueba Bera-Jarque presenta algunas limitaciones de tipo teórico en su desarrollo ya que se basa en el tercer y cuarto momentos , situación que no ocurre con el método propuesto en este trabajo que cuenta con una base teórica muy firme tal como se estableció anteriormente. Adicionalmente la simulación realizada en el capítulo V , muestra en sus resultados un valor α' muy aproximado al valor α de la tabla de distribución límite del estadístico Anderson-Darling.

   Por otra parte , al considerar los resultados obtenidos por ambos métodos de verificación en los tres ejemplos del capítulo IV , aparece la prueba Bera-Jarque como un método en ocasiones distinto en cuanto a resultados que el método desarrollado en este trabajo . Se subraya que su distribución asintótica es una mala aproximación para n chico , por lo que deben usarse los puntos críticos obtenidos en Jarque y Bera (1987, pág 169).

Sin embargo , el problema de la normalidad en las perturbaciones estocásticas no está resuelto completamente , ya que la presencia de observaciones discrepantes " outliers " puede ocasionar que se rechace el supuesto de normalidad en las perturbaciones estocásticas y además para detectar observaciones discrepantes " outliers " se utiliza el supuesto de normalidad en las perturbaciones estocásticas , lo cual implica la ocurrencia de alguna de las siguientes tres situaciones:

1) Se detectan puntos discrepantes y se acepta el supuesto de normalidad en las perturbaciones estocásticas (Ejemplo 1).

   Lo anterior significa que es muy posible que al eliminar las observaciones discrepantes en los datos , la normalidad seguirá aceptándose en las perturbaciones estocásticas. En esta situación no hay ningún problema.

2) Se detectan puntos discrepantes y se rechaza el supuesto de normalidad en las perturbaciones estocásticas , pero al eliminar las observaciones discrepantes en el análisis , ahora se acepta el supuesto de normalidad. (Ejemplo 3).

   Obsérvese que la detección de puntos discrepantes presupone normalidad ; y si no hay normalidad , es posible que esos puntos en realidad no sean discrepantes.

   Lo anterior significa que las observaciones discrepantes eran la aparente causa de la no normalidad en las perturbaciones estocásticas. En este caso no se puede saber si la normalidad existía y también los " outliers " o si la normalidad no existía pero sí había " outliers ".

3) Se detectan puntos discrepantes y se rechaza el supuesto de normalidad en las perturbaciones estocásticas , sin embargo , al eliminar las observaciones discrepantes del análisis el supuesto de normalidad se sigue rechazando. (Ejemplo 2).

   Esto significa que las perturbaciones estocásticas definitivamente no se distribuyen normalmente , por lo cual se plantea el siguiente problema en relación a " outliers " :

   " Las observaciones que fueron consideradas como discrepantes en el análisis , ¿ realmente eran " outliers " ?.

APENDICE

TABLA DE VALORES OBTENIDOS DE LA ESTADISTICA MODIFICADA ($\bar{A}^2$)
CON LOS VECTORES DE PERTURBACIONES ESTOCASTICAS SIMULADOS.

VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)
1	0.134641	51	0.296855	101	0.439134	151	0.160157
2	0.472994	52	0.267337	102	0.211062	152	0.218471
3	0.729787	53	0.708684	103	0.209231	153	0.586449
4	0.233563	54	0.573830	104	0.419142	154	0.332639
5	0.419065	55	0.299211	105	0.482402	155	0.224828
6	0.334485	56	0.197077	106	0.169129	156	0.643233
7	0.177112	57	0.568349	107	0.453829	157	0.409966
8	0.341439	58	0.428381	108	0.930772	158	0.478582
9	0.586446	59	0.339470	109	0.328426	159	0.377762
10	0.340837	60	0.926884	110	0.372683	160	0.413455
11	0.422588	61	0.213997	111	0.415555	161	0.240716
12	0.216569	62	0.532435	112	0.406575	162	0.656081
13	0.315241	63	0.187958	113	0.276152	163	0.223674
14	0.355599	64	0.337378	114	0.365290	164	0.386025
15	0.347357	65	0.468788	115	0.114107	165	0.310528
16	0.207022	66	0.567371	116	0.328147	166	0.190735
17	0.319850	67	0.224568	117	0.323403	167	0.281405
18	0.350448	68	0.239924	118	0.483271	168	0.359615
19	0.342999	69	0.249602	119	0.578389	169	0.274659
20	0.116868	70	0.341457	120	0.273632	170	0.684936
21	0.192198	71	0.378264	121	0.308469	171	0.327995
22	0.378360	72	0.355816	122	0.662949	172	0.352599
23	0.387885	73	0.205499	123	0.240926	173	0.255463
24	0.896829	74	0.548828	124	0.616311	174	0.557244
25	0.450473	75	0.442523	125	0.182530	175	0.403348
26	0.254140	76	0.312236	126	0.220967	176	0.367429
27	0.434541	77	0.468485	127	0.155725	177	0.357857
28	0.292298	78	0.088302	128	0.314906	178	0.092459
29	0.449764	79	1.030790	129	0.587354	179	0.241624
30	0.687205	80	0.415880	130	0.417855	180	0.237319
31	0.254218	81	0.331060	131	0.599979	181	0.435158
32	0.390240	82	0.510400	132	1.686126	182	0.209065
33	0.637735	83	0.485735	133	0.401943	183	0.382320
34	0.290582	84	0.314992	134	0.314762	184	0.388411
35	0.400178	85	0.160050	135	0.348962	185	0.344358
36	0.600123	86	0.735045	136	0.236262	186	0.709470
37	0.115212	87	0.800078	137	0.392935	187	0.404116
38	0.672586	88	0.294345	138	0.306092	188	0.326057
39	0.211201	89	0.342151	139	0.254456	189	0.265009
40	0.472415	90	0.294543	140	0.433821	190	0.510865
41	0.243465	91	0.272949	141	0.269730	191	0.564396
42	0.253131	92	0.261792	142	0.248243	192	0.189575
43	0.626626	93	0.587770	143	0.373145	193	0.325913
44	0.288250	94	0.347720	144	0.592571	194	0.382115
45	0.406125	95	0.235811	145	0.210284	195	0.701051
46	0.451216	96	0.471174	146	0.218471	196	0.241622
47	0.359178	97	0.349312	147	0.409904	197	0.371599
48	0.413418	98	0.225932	148	0.251594	198	0.604890
49	0.186393	99	0.650387	149	0.843560	199	0.185217
50	0.955245	100	0.893951	150	0.201272	200	0.632495

TABLA DE VALORES OBTENIDOS DE LA ESTADISTICA MODIFICADA ($\bar{A}^2$)
CON LOS VECTORES DE PERTURBACIONES ESTOCASTICAS SIMULADOS.

VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)
201	0.588236	251	0.229100	301	0.208989	351	0.213007
202	0.296388	252	0.234143	302	0.272361	352	0.600545
203	0.263797	253	0.243680	303	0.308455	353	0.761498
204	0.615653	254	0.324027	304	0.422425	354	0.372623
205	0.249757	255	0.641254	305	0.355308	355	0.257448
206	0.311908	256	0.396383	306	0.297680	356	0.565879
207	0.383097	257	0.180711	307	0.350630	357	0.151078
208	0.368880	258	0.287164	308	0.426581	358	0.336592
209	0.269155	259	0.549607	309	0.423804	359	0.588395
210	0.527847	260	0.508182	310	0.444555	360	0.310012
211	0.223237	261	0.439645	311	0.309173	361	0.326351
212	0.289137	262	0.302137	312	0.438292	362	0.317799
213	0.399728	263	0.124591	313	0.759339	363	0.245174
214	0.189998	264	0.321096	314	0.375648	364	0.195421
215	0.541980	265	0.756811	315	0.496221	365	0.168547
216	0.422424	266	0.409686	316	0.812514	366	0.302256
217	0.253999	267	0.475999	317	0.340093	367	0.568222
218	0.559364	268	0.431656	318	0.497488	368	0.139017
219	0.382374	269	0.721889	319	0.289208	369	0.444838
220	0.209217	270	1.097459	320	0.484305	370	0.295770
221	0.307646	271	0.695182	321	0.656068	371	0.356775
222	0.252855	272	0.563261	322	0.695047	372	0.229927
223	0.312094	273	0.646847	323	0.217122	373	0.167241
224	0.188099	274	0.204591	324	0.228817	374	0.641609
225	0.299964	275	0.207025	325	0.527371	375	0.573469
226	0.641847	276	0.446516	326	0.718016	376	0.186517
227	0.406544	277	0.213649	327	0.192396	377	0.486159
228	0.175987	278	0.158833	328	0.156339	378	0.237166
229	0.337230	279	0.137532	329	0.337546	379	0.665348
230	0.317261	280	0.219487	330	0.762813	380	0.561258
231	0.576750	281	0.431359	331	0.355916	381	0.453741
232	0.561203	282	0.363905	332	0.255354	382	0.324953
233	0.387795	283	0.422603	333	0.248418	383	0.314257
234	0.596740	284	0.434350	334	0.202730	384	0.190236
235	0.351750	285	0.483803	335	0.259423	385	0.312006
236	0.192916	286	0.558056	336	0.507082	386	0.291910
237	0.287259	287	0.345700	337	0.327855	387	0.166007
238	0.636647	288	0.437390	338	0.556647	388	0.363491
239	0.272923	289	0.249693	339	0.329584	389	0.443787
240	0.212536	290	0.168994	340	0.113409	390	0.406654
241	0.273493	291	0.293342	341	0.224176	391	0.409642
242	0.387729	292	0.113858	342	0.208106	392	0.422750
243	0.739771	293	0.287786	343	0.361695	393	0.269501
244	0.262345	294	0.372844	344	0.283740	394	0.409697
245	0.522789	295	0.309584	345	0.403101	395	0.609540
246	0.679771	296	0.456273	346	0.186901	396	0.287438
247	0.469364	297	0.462495	347	0.169432	397	0.328320
248	0.200936	298	0.316969	348	0.255336	398	0.328753
249	0.257939	299	0.535760	349	0.515170	399	0.490918
250	0.202784	300	0.916240	350	0.236086	400	0.328269

TABLA DE VALORES OBTENIDOS DE LA ESTADISTICA MODIFICADA ($\bar{A}^2$) CON LOS VECTORES DE PERTURBACIONES ESTOCASTICAS SIMULADOS.							
VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)
401	0.482844	451	0.495193	501	0.313578	551	0.290284
402	0.545860	452	0.227670	502	0.450634	552	0.568933
403	0.441506	453	0.207518	503	0.961489	553	0.339408
404	0.473202	454	0.143217	504	1.033176	554	0.291260
405	0.638299	455	0.306862	505	0.323153	555	0.479554
406	0.389207	456	0.202696	506	0.523616	556	0.437595
407	0.512480	457	0.309816	507	0.229738	557	0.323327
408	0.476023	458	0.215547	508	0.256368	558	0.374014
409	0.609792	459	0.525565	509	0.412985	559	0.341938
410	0.384835	460	0.466635	510	0.370847	560	0.628884
411	0.309838	461	0.529041	511	0.337070	561	0.525269
412	0.441278	462	0.249742	512	0.235045	562	0.391753
413	0.269946	463	0.480934	513	0.426546	563	0.232635
414	0.207741	464	0.284796	514	0.316166	564	0.483834
415	0.656432	465	0.165733	515	0.610255	565	0.248663
416	0.275847	466	0.443909	516	0.169195	566	0.322114
417	0.187630	467	0.400819	517	0.497496	567	0.329725
418	0.169300	468	0.599511	518	0.262245	568	0.280146
419	0.153503	469	0.517846	519	0.187483	569	1.114463
420	0.227606	470	0.500399	520	0.449680	570	0.348973
421	0.535873	471	0.493697	521	0.238292	571	0.324448
422	0.273849	472	0.360479	522	0.677493	572	0.304961
423	0.156595	473	0.485979	523	0.302986	573	0.194831
424	0.333713	474	0.496780	524	0.360461	574	0.186055
425	0.262463	475	0.210876	525	0.175367	575	0.386072
426	0.177248	476	0.490850	526	0.256520	576	0.167876
427	0.340763	477	0.238736	527	0.536111	577	0.148842
428	0.179861	478	0.394039	528	0.327530	578	0.172973
429	0.674658	479	0.296390	529	0.199841	579	0.479445
430	0.244905	480	0.455772	530	0.700847	580	0.162965
431	0.116307	481	0.395322	531	0.351016	581	0.322972
432	0.202951	482	0.292258	532	0.556466	582	0.490110
433	0.269931	483	0.560743	533	0.308332	583	0.271333
434	0.271932	484	0.153097	534	0.350499	584	0.304626
435	0.392812	485	0.361543	535	0.199017	585	0.609609
436	1.129141	486	0.466015	536	0.166941	586	0.403472
437	0.235918	487	0.234968	537	0.386729	587	0.195265
438	0.449043	488	0.200149	538	0.296984	588	0.508492
439	0.300892	489	0.217291	539	0.188457	589	0.391019
440	0.660506	490	0.553997	540	0.424871	590	0.335708
441	0.632406	491	0.353649	541	0.610640	591	0.311956
442	0.364353	492	0.891522	542	0.437013	592	0.480417
443	0.171423	493	0.133910	543	0.436913	593	0.385409
444	0.181572	494	0.476094	544	0.539394	594	0.980464
445	0.306836	495	0.356456	545	0.187260	595	0.503548
446	0.411544	496	0.563896	546	0.324563	596	0.415345
447	0.586623	497	0.232866	547	0.325057	597	0.691228
448	0.221532	498	0.499555	548	0.462369	598	0.326301
449	0.291152	499	0.314009	549	0.667957	599	0.303421
450	0.247961	500	0.283778	550	0.207415	600	0.320715

TABLA DE VALORES OBTENIDOS DE LA ESTADISTICA MODIFICADA ($\bar{A}^2$) CON LOS VECTORES DE PERTURBACIONES ESTOCASTICAS SIMULADOS.							
VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)
601	0.626989	651	0.376812	701	0.346805	751	0.432090
602	0.699540	652	0.457242	702	0.218440	752	0.326066
603	0.251931	653	0.586594	703	0.240842	753	0.934229
604	0.379039	654	0.504004	704	0.581483	754	0.263658
605	0.240503	655	0.509179	705	0.807188	755	0.233737
606	0.250708	656	0.286845	706	0.275438	756	0.232674
607	0.173872	657	0.274775	707	0.956253	757	0.404196
608	0.407197	658	0.271191	708	0.468792	758	0.204189
609	0.227314	659	0.310693	709	0.423683	759	0.306544
610	0.375433	660	0.362298	710	0.202470	760	1.081974
611	0.186485	661	0.426536	711	0.228603	761	0.222866
612	0.265342	662	0.491275	712	0.271567	762	0.227301
613	0.286577	663	0.312013	713	0.744142	763	0.355047
614	0.953507	664	0.438882	714	0.248196	764	0.440857
615	0.478667	665	0.507316	715	0.525520	765	0.271245
616	0.698551	666	0.212112	716	0.391397	766	0.430181
617	0.209875	667	0.208724	717	0.400780	767	0.629634
618	0.292921	668	0.519737	718	0.346581	768	0.372201
618	0.255905	669	0.252336	719	0.281006	769	0.299237
620	0.706915	670	0.260181	720	0.363665	770	0.405206
621	0.441150	671	0.446613	721	0.439659	771	0.719722
622	0.638295	672	0.397225	722	1.076162	772	0.381993
623	0.342924	673	0.410906	723	0.478991	773	0.429854
624	0.215315	674	0.502619	724	0.687660	774	0.296853
625	0.329595	675	0.214896	725	0.521341	775	0.682298
626	0.376868	676	0.256603	726	0.459213	776	0.461973
627	0.248421	677	0.372403	727	1.110241	777	0.410674
628	0.364396	678	0.205295	728	0.302195	778	0.247567
629	0.335873	679	0.174135	728	0.522824	779	0.308787
630	0.229488	680	0.607913	730	0.645777	780	0.579642
631	0.756705	681	0.353678	731	0.247959	781	0.271132
632	0.147782	682	0.241235	732	0.469794	782	0.642618
633	0.267025	683	0.354021	733	0.266634	783	0.488370
634	0.196990	684	0.230993	734	0.342929	784	0.396778
635	0.459410	685	0.217317	735	0.492058	785	0.351531
636	0.429858	686	0.269396	736	0.297766	786	0.606805
637	0.249879	687	0.302085	737	0.319858	787	0.213023
638	0.325694	688	0.220065	738	0.254350	788	0.407913
639	0.474317	689	0.192617	739	0.365744	789	0.962876
640	0.126537	690	0.329922	740	0.548422	790	0.383188
641	0.333330	691	0.494034	741	0.220468	791	0.177016
642	0.134231	692	0.269396	742	0.287141	792	0.371423
643	0.212109	693	0.308712	743	0.289064	793	0.342404
644	0.420189	694	0.323214	744	0.363612	794	0.349309
645	0.182199	695	0.521094	745	0.301259	795	0.177827
646	0.559415	696	0.256123	746	0.236749	796	0.447821
647	0.245615	697	0.952857	747	0.385501	797	0.268742
648	0.506660	698	0.210493	748	0.520264	798	0.296927
649	0.230913	699	0.376613	749	0.340797	799	0.300826
650	0.275360	700	0.373321	750	0.218991	800	0.459101

TABLA DE VALORES OBTENIDOS DE LA ESTADISTICA MODIFICADA ($\bar{A}^2$) CON LOS VECTORES DE PERTURBACIONES ESTOCASTICAS SIMULADOS.

VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)	VECTOR	($\bar{A}^2$)
801	0.387932	851	0.508069	901	0.203617	951	0.370277
802	0.667622	852	0.471724	902	0.374640	952	0.506848
803	0.346351	853	1.645923	903	0.101722	953	0.503446
804	0.626183	854	0.421624	904	0.314532	954	0.276171
805	0.226804	855	1.225356	905	0.924386	955	0.264264
806	0.274590	856	0.310366	906	0.373495	956	0.110792
807	0.567061	857	0.506343	907	0.178683	957	0.143647
808	0.351558	858	0.356672	908	0.263098	958	0.244395
809	0.542516	859	0.347450	909	0.270323	959	0.376351
810	0.744163	860	0.155758	910	0.167791	960	0.479840
811	0.304776	861	0.596127	911	0.277838	961	0.440749
812	0.569867	862	0.188110	912	0.692109	962	0.221595
813	0.189799	863	0.211187	913	0.157027	963	0.316802
814	0.226751	864	0.740806	914	0.325111	964	0.237448
815	0.296769	865	0.524875	915	0.213214	965	0.350679
816	0.180826	866	0.375781	916	0.446586	966	0.277511
817	0.213540	867	0.337791	917	0.153488	967	0.305291
818	0.316553	868	0.295581	918	0.178238	968	0.315316
819	0.255849	869	0.280679	919	0.193501	969	0.257320
820	0.182969	870	1.244810	920	0.455384	870	0.897850
821	0.869280	871	0.180115	921	0.371969	971	0.508290
822	0.440812	872	0.385143	922	0.211580	972	0.527245
823	0.328696	873	0.311324	923	0.390656	973	0.201228
824	0.509734	874	0.141445	924	0.305043	974	0.482624
825	0.400721	875	0.395095	925	0.314982	975	0.403895
826	0.237498	876	0.202447	926	0.350492	976	0.342540
827	0.414567	877	0.624933	927	0.167338	977	0.200908
828	0.498691	878	0.471458	928	0.251878	978	0.253816
829	0.476156	879	0.348393	929	0.440458	979	0.417776
830	0.525779	880	0.372658	930	0.587108	980	0.189322
831	0.291907	881	0.293191	931	0.410407	981	0.229686
832	0.679720	882	0.626978	932	0.870618	982	0.249601
833	0.416301	883	0.194043	933	0.176186	983	0.403246
834	0.305286	884	0.349347	934	0.266153	984	0.215977
835	0.245289	885	0.390755	935	0.638650	985	0.424700
836	0.287778	886	0.316536	936	0.442746	986	0.244361
837	0.221189	887	0.353801	937	0.493818	987	0.170831
838	0.383586	888	0.337505	838	0.170307	988	0.340031
839	0.219761	889	0.171872	939	0.307490	989	0.514990
840	0.302964	890	0.181089	840	0.302519	990	0.184297
841	0.375937	891	0.133666	941	0.432264	991	0.450650
842	0.227169	892	0.180515	942	0.276587	992	0.298739
843	0.493360	893	0.501275	943	0.213061	993	0.492005
844	0.219401	894	0.216991	944	0.194094	994	0.314189
845	0.125130	895	0.644152	945	0.186602	995	0.212381
846	0.394245	896	0.190429	946	0.286111	996	0.481648
847	0.215640	897	0.188959	947	0.186166	997	0.260209
848	0.349156	898	0.316567	948	0.241448	998	0.495079
849	0.403670	899	0.666833	949	1.556061	999	0.377874
850	0.416135	900	0.476401	950	0.449348	1000	0.625361

BIBLIOGRAFIA

- Allen and Cady (1982) " Analyzing Experimental Data by Regression ". Van Nostrand Reinhold .
- Atkinson (1985) " Plots , Transformations and Regression". Clarendon Press .
- Barnett and Lewis (1984) " Outliers in Statistical Data ". Wiley .
- Belsley , Kuh and Welsch (1980) " Regression Diagnostics: Identifying Influential Data and Sources of Collinearity ". Wiley .
- Chatterjee and Price (1977) " Regression Analysis by Example ". Wiley .
- Cook and Weisberg (1982) " Residuals and Influence in Regression ". Chapman and Hall.
- D' Agostino and Stephens (1986) " Goodness of fit technics ". Marcel Dekker .
- Draper and Smith (1981) " Applied Regression Analysis ". Wiley .
- Ghurye, and I. Olkin (1969) .
- Greene, William H. (1993) " Econometric Analysis ". MacMillan .
- Gunst and Mason (1980) " Regression Analysis and its Application ". Marcel Dekker .
- Hoaglin and Welsch (1978) " The Hat Matrix in Regression and ANOVA ". The American Statistician 32 .
- Jarque and Bera (1987) " A Test for Normality of Observations and Regression Residuals ". International Statistical Review . .
- Judge and Hill (1988) " Introduction to the Theory and Practice of Econometrics ". Wiley .
- Kleinbaum, Kupper and Muller (1988) " Applied Regression Analysis and Other Multivariate Methods ". PWS-Kent .
- Lyman Ott (1984) " An Introduction to Statistical Methods an Data Analysis ". Duxbury .
- Montgomery and Peck (1982) " Introduction to Linear Regression Analysis ". Wiley .
- Mostellar and Tukey (1977) " Data Analysis and Regression ". Addison-Wesley .

- Neter, Wasserman and Kutner (1990) "Applied Linear Statistical Models". Irwin.

- O'Reilly Togno, Federico (1993) "Testing the fit of a Multivariate Distribution" Proceedings of the Conference in Statistical Inference and Biostatistics, CIMAT, Gto. Ed. D. A. Sprott, 179 - 200.

- O'Reilly Togno, Federico (1990) "Algunas consideraciones sobre la inferencia estadistica".Ciencia.

- O' Reilly and Quesenberry (1973) "The Conditional Probability Integral Transformation and Aplications to Obtain Composite Chi-Square Goodness of Fit Tests". The Annals of Statistics.

- Roussaeuw and Leroy (1987) "Robust Regression and Outlier Detection". Wiley.

- Seber (1977) "Linear Regression Analysis". Wiley.

- Weisberg (1985) "Applied Linear Regression". Wiley.