



**ESPECIALIDAD EN MÉTODOS ESTADÍSTICOS**  
**FACULTAD DE ESTADÍSTICA E INFORMÁTICA**  
**UNIVERSIDAD VERACRUZANA**

**“ Pruebas de Normalidad para los Residuos  
de un Ajuste de Regresión”**

Trabajo Recepcional que como requisito  
parcial para obtener el diploma de esta  
Especialidad presenta:

***Francisco Javier Hernández Loeza***

Tutor Académico:

***Dr. Héctor Coronel Brizio***

---

**Xalapa, Ver., diciembre de 1995**

## **CON MUCHO CARIÑO:**

*A mis padres Hugo  
Hernández Castillo y Vianey  
Loeza de Hernández, a  
quienes admiro, respeto y  
amo. Gracias por brindarme  
su amor de padres, motivo  
para superarme cada día.*

*A ti Juliana, que supiste  
comprenderme, apoyarme  
y motivarme, para  
realizar este objetivo en  
mi vida. Tu cariño es un  
factor importante para  
seguir mejorando (t.q.m.)*

*A mis hermanos Jorge y  
Hugo, de quienes he  
recibido motivación para  
lograr metas en la vida.*

*A mi sobrino Huguito,  
para que recuerde  
siempre que en la vida no  
hay imposibles.*

## ***AGRADECIMIENTOS:***

*Al Dr. Héctor Coronel Brizio*, por su interés y apoyo incondicional en la realización de este trabajo. Gracias Héctor por tu amistad.

*Al Ing. Miguel Marengo Canales y Lic. J. Raúl López Gutiérrez* por su apoyo y comprensión para poder realizar este objetivo.

A mi amigo *Alberto Munguía Herrera*, por sus palabras de aliento cuando las necesité.

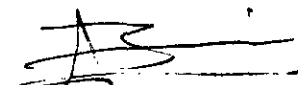
A mis compañeros de la especialidad, en especial a *Teresa, Cristina y Cecilia* por compartir sus conocimientos.

El comité académico de la Especialidad en Métodos Estadísticos y el respectivo Tutor Académico del trabajo recepcional “ *Pruebas de Normalidad para los Residuos de un Ajuste de Regresión*”, una vez cubiertos todos los requisitos académicos y administrativos establecidos, autorizan la impresión y la constitución del jurado para la defensa del mismo.

POR EL COMITÉ ACADÉMICO:

---

Lic. Claudio Rafael Castro López



---

Dr. Mario Miguel Ojeda Ramírez

---

Lic. Sergio Hernández González

TUTOR:



---

Dr. Héctor Coronel Brizio

# ÍNDICE

## INTRODUCCIÓN

### 1. FUNDAMENTOS DE ANÁLISIS DE REGRESIÓN LINEAL

1.1.	Introducción .....	1
1.2.	Significado de la Regresión .....	1
1.2.1.	Conceptos Básicos.....	2
1.3.	Estimadores para $\beta_i$ por el Método de Mínimos Cuadrados .....	7
1.4.	Supuestos Básicos al Estimar los $\beta_i$ .....	9
1.5.	El supuesto de Normalidad .....	11
1.6.	Modelo Estimado y Resultados Importantes.....	13
1.7.	Intervalos de Confianza y Pruebas de Hipótesis.....	15
1.8.	Comprobación de los Supuestos.....	18
1.9.	Comentarios Generales.....	22

### 2. PRUEBAS DE BONDAD DE AJUSTE PARA LA DISTRIBUCIÓN NORMAL

2.1.	Introducción .....	24
2.2.	Objetivo de una Prueba de Bondad de Ajuste .....	25
2.2.1.	Clasificación de las pruebas.....	26
2.3.	Prueba de Ajuste Basada en el Gráfico de Probabilidad .....	28
2.4.	Pruebas de Ajuste Basados en la Función de Distribución Empírica .....	30
2.4.1.	Introducción.....	30
2.4.2.	La función de Distribución Empírica.....	30
2.4.3.	Estadísticos de Prueba de la Función de Distribución Empírica .....	31
2.4.4.	Prueba de Anderson - Darling.....	34
2.4.5.	Prueba de Cramer - Von Mises .....	36
2.4.6.	Prueba de Watson .....	38
2.5.	Pruebas de Bondad de Ajuste Shapiro-Wilk y D'Agostino .....	40

2.5.1.	Introducción.....	40
2.5.2.	La prueba de Shapiro- Wilk.....	40
2.5.3.	Prueba de D' Agostino .....	42
2.6.	Pruebas de Normalidad para los Residuos de un ajuste de Regresión .....	44
2.6.1.	Introducción.....	44
2.6.2.	Investigación de Pierce y Kopecky.....	46
2.6.3.	Comentarios.....	48
3.	<b>PROBANDO NORMALIDAD EN EXCEL 5.0 (MACRO)</b>	
3.1.	Introducción .....	49
3.2.	Explicación del Macro .....	49
3.3.	Ejemplo de Aplicación .....	58
	<b>CONCLUSIONES.....</b>	<b>63</b>
	<b>REFERENCIAS .....</b>	<b>64</b>

# INTRODUCCIÓN

El análisis de regresión es una herramienta estadística que permite estudiar la relación que existe entre una variable dependiente con una o más variables explicatorias, por medio de un modelo estadístico.

Una vez que se ajusta el modelo de regresión, se prueban los supuestos que el método de mínimos cuadrados considera al estimar los coeficientes de regresión, por medio de técnicas gráficas (análisis de residuos) o por medio de algunas pruebas analíticas usuales, sin embargo, se han diseñado considerando que la muestra de los residuos es aleatoria e independiente.

Las técnicas empleadas para probar Normalidad requieren que las observaciones correspondan a una realización de una muestra aleatoria de tamaño  $n$ . Sin embargo, los residuos calculados a partir del ajuste de un modelo de regresión no cumplen con esta consideración, sino que están correlacionados y sus varianzas difieren. Por lo anterior, es necesario utilizar pruebas que consideren estos aspectos en los residuos y verificar que el supuesto de normalidad se cumpla.

En este trabajo se examinan algunas pruebas de bondad y ajuste diseñadas recientemente para probar el supuesto de normalidad en los errores después de un ajuste al modelo de regresión y se implementan estas pruebas por medio de una rutina (Macro) computacional en la hoja de cálculo EXCEL 5.0.

El trabajo se divide en tres temas de importancia. En el capítulo uno se dan los fundamentos del análisis de regresión, haciendo notar lo importante que es comprobar los supuestos al modelo y en especial el de normalidad dado que si se cumple el modelo

estimado es confiable. En el capítulo dos se presentan algunas pruebas de bondad de ajuste para una distribución normal. Aquí se llega a la conclusión por el estudio realizado por Pierce y Kopecky que las pruebas que son recomendables para probar normalidad en los residuos de un ajuste de regresión son la pruebas basadas en la Función de Distribución Empírica. Y por último en el capítulo tres se presenta la rutina (Macro) computacional en la hoja de cálculo EXCEL 5.0 que permite probar normalidad en los residuos de un ajuste de regresión, por medio de los estadísticos de prueba de Anderson-Darling, Cramer-Von Mises y Watson.



# CAPÍTULO 1

---

## FUNDAMENTOS DE ANÁLISIS DE REGRESIÓN LINEAL

# CAPÍTULO 1: FUNDAMENTOS DE ANÁLISIS DE REGRESIÓN LINEAL

## 1.1 INTRODUCCIÓN

Una herramienta estadística que permite estudiar la relación entre dos o más variables, de tal manera que pueda predecirse una de las variables a partir de otra, o de las otras, es el análisis de regresión.

Este capítulo presenta de manera sencilla, clara y resumida los conceptos en torno del análisis de regresión lineal.

## 1.2 SIGNIFICADO DE REGRESIÓN

En diversas áreas de investigación se encuentran con frecuencia problemas en los cuales el resultado de una variable depende de los valores de otras. Por ejemplo podríamos mencionar: (a) la demanda de cierto producto depende de los gastos de propaganda que se realizan ; (b) una calificación de estadística de alumnos de la universidad depende de sus calificaciones en matemáticas; (c) el producto de cosecha de trigo depende de la temperatura, lluvia, cantidad de sol y la fertilización; (d) el salario para profesores de cierta universidad se ve influenciada por los años de experiencia el rango académico y la disciplina, etc.

Por el comportamiento que tienen las variables en la mayoría de estas relaciones, se distingue una variable "*dependiente*<sup>1</sup>", cuyo valor es determinado por una o más variables

---

<sup>1</sup> También recibe el nombre de variable de respuesta, predecida o explicada.

"explicatorias"<sup>2</sup>. La variable dependiente es aleatoria, a diferencia de las variables explicatorias que no lo son. Como ejemplos de variables dependientes tenemos: (a) la demanda del producto, (b) la calificación de una prueba Estadística, (c) el producto de una cosecha y (d) el salario de los profesores, etc. Como variables explicatorias tenemos; (a) los gastos de propaganda, (b) las calificaciones de Matemáticas, (c) la temperatura, cantidad de lluvia, cantidad de sol y la fertilización y (d) años de experiencia, rango académico y disciplina. Una variable dependiente generalmente es representada por  $Y$ , y una variable explicatoria por  $X$ , así nos referimos a ellas.

Estudiar al mismo tiempo el efecto de las variables explicatorias sobre la variable dependiente no es fácil si se requiere de precisión para asociarlas, por lo que se recurre a un método estadístico llamado ANÁLISIS DE REGRESIÓN. El análisis de regresión es una herramienta estadística que permite estudiar este tipo de relaciones<sup>3</sup>, es decir, la relación que existe entre una variable dependiente con una o más variables explicatorias.

### 1.2.1 Conceptos Básicos

Relaciones como las anteriores se pueden representar por medio de un "*modelo estadístico*", el cual su construcción se fundamenta en tener una parte del modelo que es explicada con exactitud y otra parte que es aleatoria, es decir la presencia de un "*error*". El modelo estadístico tiene la siguiente forma:

$$(1.2.1.1) \quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i \quad \text{para } i = 1, 2, \dots, N$$

<sup>2</sup> Llamadas también variables independientes, predictoras o de regresión. Estas variables en esencia pueden ser aleatorias pero para fines de análisis de regresión se consideran no aleatorias.

<sup>3</sup> Las relaciones estadísticas por muy fuertes y sugerentes que sean, nunca establecen una relación casual, la casualidad queda fuera de la Estadística.

y recibe el nombre de "***modelo de regresión lineal general***", en donde  $Y_i$  representa un valor individual de la variable dependiente  $Y$ ,  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  son parámetros desconocidos llamados "***coeficientes de regresión***",  $X_{ik}$  representa el  $i$ -ésimo valor de la variable explicatoria  $k$ ,  $\varepsilon_i$  representa un "***error aleatorio***" y  $N$  es el tamaño de la población.

Es de importancia mencionar que un modelo de regresión será "lineal" en los parámetros o coeficientes de regresión y no en las variables explicatorias, por lo tanto, los  $\beta$ 's deberán aparecer con exponente uno y no se encuentran multiplicados ni divididos por otro coeficiente.

El ***objetivo del análisis de regresión*** es ajustar un modelo lineal con cierta forma<sup>4</sup>, que describa razonablemente el comportamiento de la variable dependiente ( $Y$ ) dadas las variables explicatorias ( $X_1, X_2, X_3, \dots, X_K$ ), para estimar o predecir el valor "***promedio poblacional***" de la variable dependiente, a partir de los valores conocidos, valores fijos, de las variables explicatorias. La expresión  $E(Y/X_1, X_2, X_3, \dots, X_K)$ , representa el valor esperado poblacional de  $Y$  dados los valores de  $X_1, X_2, X_3, \dots, X_K$ .

Si el modelo determina la relación de una variable dependiente  $Y$  con dos o más variables explicatorias ( $X_1, X_2, X_3, \dots, X_K$ ) se llamará "***Modelo de regresión lineal múltiple poblacional***" y la forma de representarlo es:

$$(1.2.1.2) \quad E(Y_i/X_1, X_2, X_3, \dots, X_K) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik}$$

para  $i = 1, 2, \dots, N$

---

<sup>4</sup> Su forma es un hecho empírico en ocasiones, dependiendo de la complejidad del problema que analizamos y como primera aproximación suponemos un modelo lineal en los parámetros de regresión.

Si el modelo sólo trata la relación de una variable dependiente  $Y$ , con una sola variable explicatoria  $X$ , se llamará *"Modelo de regresión lineal simple poblacional"* y se representa así:

$$(1.2.1.3) \quad E(Y/X_i) = \beta_0 + \beta_1 X_{i1} \quad i = 1, 2, \dots, N$$

De los ejemplos al inicio del capítulo (a) y (b) se tratan con un modelo lineal de regresión simple, (c) y (d) con un modelo lineal de regresión múltiple.

Las expresiones 1.2.1.1 y 1.2.1.2 son expresiones equivalentes ya que si obtenemos el valor esperado en ambos miembros de 1.2.1.1 llegamos a 1.2.1.2, considerando que  $E(\epsilon_i) = 0$ , el cual es un supuesto que se mencionará más adelante. Es decir,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i \quad \text{es equivalente a}$$

$$E(Y_i/X_i, X_2, X_3, \dots, X_K) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} \quad \text{para } i = 1, 2, \dots, N$$

para  $Y_i$  que representa un valor particular, lo expresamos también como:

$$(1.2.1.5) \quad Y_i = E(Y_i/X_{i1}, X_{i2}, X_{i3}, \dots, X_{iK}) + \epsilon_i$$

Lo mismo sucede con el modelo lineal simple poblacional, es decir:

$$(1.2.1.4) \quad Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i \quad \text{es equivalente con}$$

$$(1.2.1.3) \quad E(Y/X_i) = \beta_0 + \beta_1 X_{i1} \quad \text{para } i = 1, 2, \dots, N$$

por lo tanto  $Y_i$  es representado también como:

$$(1.2.1.6) \quad Y_i = E(Y/X_{i1}) + \varepsilon_i$$

Hasta el momento en las formas de los modelos lineales de regresión múltiple y simple, consideramos la información de una " población " pero en la realidad es muy difícil contar con esta, más bien, se cuenta con la información de una "*muestra*" de dicha población, por lo tanto, el modelo que se obtenga será una estimación del verdadero y se llama "*modelo lineal de regresión muestral*". Es decir,

$$(1.2.1.7) \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \dots + \hat{\beta}_k X_{ik} \quad i = 1, 2, 3, \dots, n$$

en donde  $\hat{y}_i$  estima el valor de  $E(Y_i/X_1, X_2, X_3, \dots, X_K)$ ,  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$  estiman el valor de los coeficientes de regresión  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ , y ,

$$(1.2.1.5) \quad Y_i = E(Y_i/X_1, X_2, X_3, \dots, X_K) + \varepsilon_i$$

es estimado por,

$$(1.2.1.8) \quad y_i = \hat{y}_i - e_i$$

en donde,  $y_i$  es una observación en la muestra,  $e_i$  son estimaciones de los  $\varepsilon_i$  y reciben el nombre de "*residuos o residuales*".

Las expresiones 1.2.1.7 y 1.2.1.8 son los "*modelos lineales de regresión múltiple muestral*". Para los modelos lineales de regresión simple " se tiene que:

$$(1.2.1.3) \quad E(Y/X_i) = \beta_0 + \beta_1 X_{i1} \quad \text{es estimado por}$$

$$(1.2.1.9) \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1}$$

y

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

una observación en particular,

$$(1.2.1.10) \quad y_i = \hat{y}_i - e_i$$

Los modelos 1.2.1.9 y 1.2.1.10 representan un “*modelo lineal de regresión simple muestral*”.

Como se mencionó anteriormente, un modelo de regresión lineal contiene una parte que es explicada con exactitud y otra que es aleatoria, la cual llamamos “*error aleatorio*”.

El *error aleatorio* es una variable estocástica no observable que puede tomar valores positivos o negativos y su presencia se debe a todas las variables explicatorias que no fueron incluidas en el modelo de regresión y por alguna razón afectan el comportamiento de  $Y_i$ . Cuando se tiene un modelo de regresión lineal, el estudio de estos errores aleatorios es de mucha importancia, por que a partir de ellos se encuentra información importante acerca de nuestro modelo propuesto.

En los modelos presentados se tiene que  $\varepsilon_i$  es estimado por  $e_i$ , en donde  $e_i$  es llamado  $i$ -ésimo residuo o residual. Los residuos representan las desviaciones de  $y_i$ , con respecto a su media:

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

Una manera efectiva de tratar el análisis de regresión es por medio del enfoque matricial, principalmente cuando se tiene más de una variable explicatoria.

3.- No existe relación lineal entre las variables explicatorias  $X_1, X_2, X_3, \dots, X_k$ , este supuesto se hace para el caso del análisis de regresión lineal múltiple y es conocido como "*No Multicolinealidad*".

El no cumplimiento de alguno de los supuestos o consideraciones presentados ocasiona errores graves en las inferencias que se obtienen a partir del modelo de regresión lineal. En muchas ocasiones el investigador sólo se limita a proporcionar el modelo considerándolo correcto y no considera el cumplimiento de los supuestos mencionados, por lo tanto, con la realización de algunas pruebas analíticas y métodos gráficos se puede comprobar la veracidad de estos y en consecuencia tener un estudio más confiable de nuestro análisis de regresión.

### **1.5 EL SUPUESTO DE NORMALIDAD**

Con los supuestos y consideraciones anteriores los estimadores de mínimos cuadrados ( $\hat{\beta}_i$ ) adquieren propiedades estadísticas relevantes como, ser insesgados y poseer varianza mínima sin embargo, el método no hace mención sobre la distribución de los  $\epsilon_i$ , es importante decir que para hacer inferencia por medio de una estimación puntual, no tiene mucha relevancia conocerla, pero si nuestro interés recae en inferir a través de la estimación por intervalo o prueba de hipótesis, si es importante, y la distribución que se supone es una distribución normal para los  $\epsilon_i$ , con media 0 y varianza  $\sigma^2$ , es decir,

$$\epsilon_i \sim N(0, \sigma^2) \text{ donde}$$

$$E(\epsilon_i) = 0 \quad \text{y} \quad \text{var}(\epsilon_i) = \sigma^2$$



***Propiedades de los Estimadores de Mínimos Cuadrados bajo el Supuesto de Normalidad***

El cumplimiento de este supuesto hace que los estimadores de mínimos cuadrados ( $\hat{\beta}_i, \hat{\sigma}^2$ ) obtengan las siguientes propiedades estadísticas:

Propiedades:

- 1.- Son insesgados
- 2.- Tienen varianza mínima. El ser insesgados y tener varianza mínima los hace ser estimadores eficientes.
- 3.- Son consistentes, es decir, conforme el tamaño de muestra crece indefinidamente, los estimadores se acercan al valor poblacional verdadero.
- 4.-  $\hat{\beta}_0$  está normalmente distribuido:

$$\hat{\beta}_0 \sim N(\beta_0, \sigma_{\hat{\beta}_0}^2) \text{ es decir}$$

$$E(\hat{\beta}_0) = \beta_0 \quad \text{y} \quad \text{var}(\hat{\beta}_0) = \sigma_{\hat{\beta}_0}^2 = \frac{\sum_{i=1}^n X_i^2}{N \sum_{i=1}^n X_i^2} \sigma^2$$

- 5.-  $\hat{\beta}_i$  ( $i=1,2,\dots,k$ ) está normalmente distribuido:

$$\hat{\beta}_i \sim N(\beta_i, \sigma_{\hat{\beta}_i}^2) \text{ es decir}$$

$$E(\hat{\beta}_i) = \beta_i \quad \text{y} \quad \text{var}(\hat{\beta}_i) = \sigma_{\hat{\beta}_i}^2 = \frac{\sigma^2}{\sum_{i=1}^n X_i^2}$$

6.-  $\frac{(N-2)\hat{\sigma}^2}{\sigma^2}$  está distribuida como una  $\chi^2$  con N-2 grados de libertad.

7.-  $(\hat{\beta}_0, \hat{\beta}_i)$  están distribuidos independientemente de  $\hat{\sigma}^2$ . Para  $i=1,2,\dots,m$ .

<sup>6</sup> 8.-  $(\hat{\beta}_0, \hat{\beta}_i)$  tienen varianza mínima para toda las clases de estimadores insesgados, lineales o no lineales

Como se puede apreciar el cumplimiento del supuesto de normalidad es fundamental, dado que permite obtener las distribuciones de probabilidad de  $\hat{\beta}_0$  (normal),  $\hat{\beta}_i$  (normal) y  $\hat{\sigma}^2$  (ji-cuadrada) importantes si nuestro interés radica en estimar intervalos de confianza o realizar pruebas de hipótesis.

Este trabajo tiene como esencia presentar algunas pruebas de bondad de ajuste, para validar el supuesto de normalidad de los residuos.

## **1.6 MODELO ESTIMADO Y RESULTADOS IMPORTANTES**

Una vez seleccionado la forma del modelo, y estimado los coeficientes de regresión ya se cuenta con un modelo de regresión,

$$(1.7) \quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \dots + \hat{\beta}_k X_{ik}$$

---

<sup>6</sup> Este resultado que se debe a Rao, es muy poderoso, porque a diferencia del Teorema de Gauss-Markov, no está restringido a la clases de estimadores lineales.

### ***Interpretación de los $\hat{\beta}_i$***

En el caso de un análisis de regresión lineal simple, el valor de  $\hat{\beta}_0$ <sup>7</sup> se considera sólo un factor de ajuste en nuestro modelo y  $\hat{\beta}_1$  que es el más importante, nos indica el número de unidades que cambia y por cada unidad de cambio de X. Para el caso del análisis de regresión múltiple, la interpretación para  $\hat{\beta}_0$  es la misma que en el caso simple, y para  $\hat{\beta}_i$ , ( $i=1,2,\dots,k$ ) es considerando el efecto parcial de cada  $\hat{\beta}_i$ , es decir,  $\hat{\beta}_1$  representa el número de unidades que cambia Y por cada unidad de cambio de  $X_1$  para  $X_2, X_3, \dots, X_k$  constantes;  $\hat{\beta}_2$  significa el número de unidades que cambia y por cada unidad de cambio de  $X_2$  para  $X_1, X_3, \dots, X_k$  constantes, la misma interpretación utilizamos para los restantes  $\hat{\beta}_3, \hat{\beta}_4, \dots, \hat{\beta}_k$ .

### ***Coefficiente de determinación y correlación***

El *coeficiente de determinación*<sup>8</sup> es una medida que indica que tanto la ecuación de regresión estimada se ajusta a los datos, es una medida de bondad del ajuste, representado generalmente por  $R^2$ . Las expresiones que permiten calcularlo son, para un análisis de regresión lineal simple,

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

<sup>7</sup> Tendrá interpretación si físicamente puede tomar el valor de 0.

<sup>8</sup> En regresión el coeficiente de determinación resulta de mayor importancia que el coeficiente de correlación.

en la que  $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$  se conoce como la suma de cuadrados de la regresión (SCR) y  $\sum_{i=1}^n (y_i - \bar{y})^2$  como la suma de cuadrados totales (SCT). Por lo que  $R^2 = \text{SCR}/\text{SCT}$ , mide que tanto de la variación en  $y$  la explica el modelo de regresión (ó las variables  $X_i$  en conjunto).

Para un análisis de regresión lineal múltiple  $R^2$  es calculado matricialmente por:

$$R^2 = \frac{\hat{B}' X' Y - n\bar{Y}^2}{Y' Y - n\bar{Y}^2}$$

y al igual que el caso simple  $\hat{B}' X' Y - n\bar{Y}^2$  representa la suma de cuadrados debida a la regresión (SCR) y  $Y' Y - n\bar{Y}^2$  es la suma de cuadrados totales (SCT).

El *coeficiente de correlación*, mide la fuerza y el grado de asociación lineal entre 2 variables, comúnmente es representado por  $(r)$  y se calcula a partir de:

$$r = \sqrt{R^2}$$

## 1.7 INTERVALOS DE CONFIANZA Y PRUEBAS DE HIPÓTESIS

Si el objetivo es estimar intervalos de confianza o el planteamiento de pruebas de hipótesis con respecto a los  $\hat{\beta}_i$  es necesario conocer la distribución de los  $\epsilon_i$ . En este sentido se sigue el supuesto de normalidad para los  $\epsilon_i$ , es decir  $\epsilon_i \sim N(0, \sigma^2)$ , de aquí se desprende (por el teorema del límite central) que los  $\hat{\beta}_i$  también siguen una distribución normal.

### ***Intervalos de confianza para los $\hat{\beta}_i$***

Las estimaciones que realizamos de los  $\beta_i$  pueden ser más precisas, es decir, si en lugar de realizar una estimación puntual se hace una estimación por intervalo.

La expresión que permite encontrar el intervalo de confianza para cualquier  $\hat{\beta}_i$ , es la siguiente;

$$(1.5) \quad \Pr [(\hat{\beta}_i - t_{(\alpha/2), (n-k)} \text{e}\hat{s}(\hat{\beta}_i)) < \beta_i < \hat{\beta}_i + t_{(\alpha/2), (n-k)} \text{e}\hat{s}(\hat{\beta}_i)] = 1 - \alpha$$

en donde,  $t_{(1-\alpha/2), (n-k)}$  sigue una distribución t-student,  $(n-k)$  son los grados de libertad, en donde  $k$  es el número de coeficientes de regresión;  $\text{e}\hat{s}(\hat{\beta}_i)$  es la desviación estándar estimada para  $\hat{\beta}_i$ ,  $\alpha$  recibe el nombre de nivel de significancia y  $1 - \alpha$  se llama *nivel de confianza*.

### ***Pruebas de Hipótesis***

En el análisis de regresión es de interés hacer pruebas de hipótesis acerca de los parámetros del modelo, esto utilizando los valores estimados. Una prueba de hipótesis de mucha importancia es la que plantea que los  $\hat{\beta}_i$  son iguales a cero en contra de que al menos uno no lo es. Cuando los  $\hat{\beta}_i$  son ceros entonces los cambios en las  $X_i$  no afectan a la variable  $y$ , por lo que la regresión no será significativa. A continuación se describe, en forma general, el procedimiento a seguir para el desarrollo de esta prueba de hipótesis

**Prueba de significancia para  $\hat{\beta}_1$** 

La hipótesis que se desea probar es:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

esta se debe probar a un nivel de significancia dado  $\alpha$ .

Pasos a seguir:

1.-Calcular el estadístico de prueba en base a la distribución del parámetro, por medio de:

$$t_c = \hat{\beta}_1 \frac{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}{\sigma^2} \quad \text{ó} \quad t_c = \frac{\hat{\beta}_1}{\text{es}(\hat{\beta}_1)}$$

2.- Se busca un valor de tablas de la distribución t-student con  $(n-2)$  grados de libertad y un nivel de significancia  $\alpha/2$ .

3.- Regla de decisión: si el  $|t_c| > t_{(1-\alpha/2), (n-k)}$  se rechaza  $H_0$ .

El no rechazar  $H_0$  implica que la variable dependiente (y) no tiene relación lineal con la variable explicatoria (X) y puede ser por causa de alguna deficiencia en el modelo, la violación para algún supuesto o bien que se encuentran en los datos algunos valores "extraños".

### ***Análisis de varianza***

Una forma alternativa para probar la hipótesis sobre si el valor de  $\hat{\beta}_1 = 0$ , es por medio del análisis de varianza. El análisis de varianza es básico si se realiza un análisis de regresión lineal múltiple y divide la variación total de las observaciones  $y_i$  para el modelo lineal en dos componentes, una debida a la regresión (SCR) y otra que se debe a un error aleatorio  $\epsilon_i$ . Es decir,

$$SCT = SCR + SCE$$

La hipótesis que se desea probar es:

$$H_0: \beta_1 = \beta_2 = \dots = \beta_k$$

$$H_1: \text{al menos algún } \beta_i \text{ es diferente de } 0$$

esta se debe probar a un nivel de significancia dado  $\alpha$ .

La estadística de prueba F que se calcula por medio de la tabla ANOVA (generada por la mayoría de los paquetes estadísticos) se compara contra una F de tablas con (1,n-2) grados de libertad en el caso simple y (k,n-k) en el caso múltiple. La regla de decisión que se toma es la siguiente: Si F calculada > F de tablas entonces Rechazamos  $H_0$ , si se rechaza  $H_0$  ratificaremos que la variable independiente y es influenciada por la(s) variable(s) X('s).

### ***1.8. COMPROBACIÓN DE LOS SUPUESTOS***

Comprobar los supuestos es fundamental en un análisis de regresión y se puede hacer por medio de un análisis gráfico conocido como "análisis de residuos"<sup>9</sup> o por medio de algunas

---

<sup>9</sup> Es importante mencionar que el análisis de residuos sólo descubre la violación a algún supuesto o anomalías en el modelo y nos marca el camino para corregir el problema que se presente.

pruebas analíticas. Como primera instancia se recomienda un estudio gráfico, puesto que existen patrones ya establecidos que nos permiten detectar la violación de algún supuesto o deficiencias del modelo, como podrían ser puntos extremos o simplemente un modelo mal especificado.

### ***Análisis de residuos***

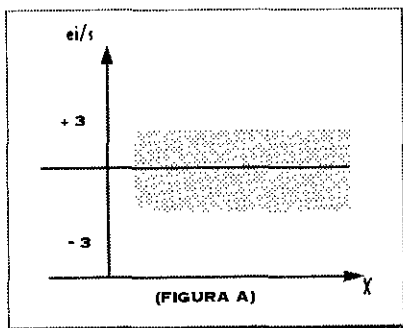
Para encontrar algún problema a través del análisis de residuos, se recomienda emplear los residuos estandarizados. Dado que la media de los residuos es igual a cero, define el  $i$ -ésimo residual estandarizado

$$e_i = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)}{s}$$

donde  $s$  es la desviación estándar residual (raíz de CME).

Las gráficas que usualmente se ven en una análisis de residuos se analizan a continuación.

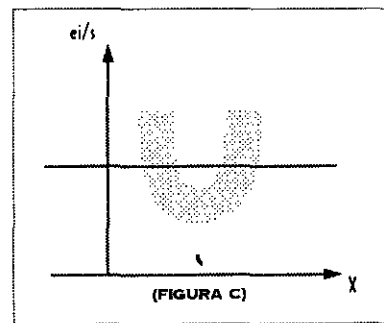
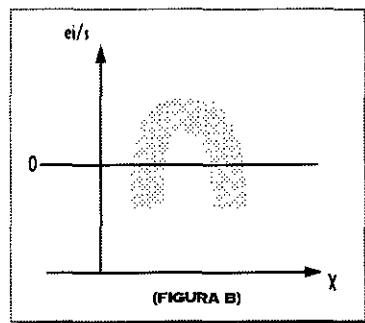
*Figura A:* Si el modelo de regresión estimado está prácticamente libre de cualquier deficiencia o violación de supuestos, entonces los residuos estandarizados tienden a no



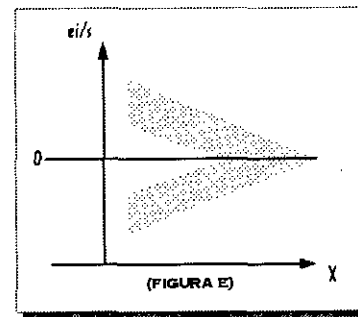
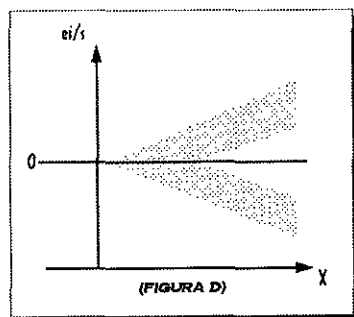


presentar un patrón sistemático positivo o negativo al rededor de 0, y muy raramente fuera del intervalo  $(+3)$ .

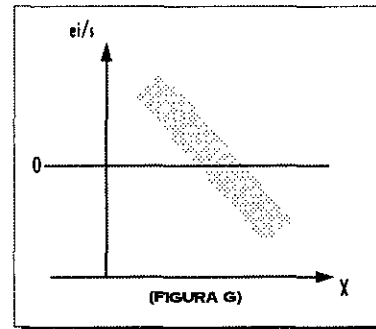
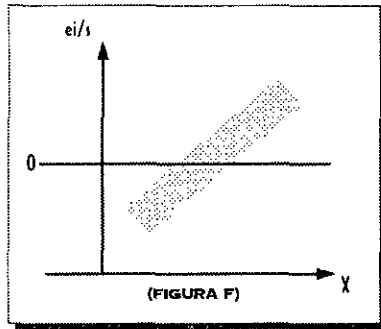
*Figuras B y C:* muestran un patrón sistemático, reflejando que es necesario un término cuadrático (causado por la variable  $X_{ij}$ ) en el modelo de regresión, una posible causa es que el modelo sea inadecuado.



*Figura D y E:* representan un patrón donde la varianza del error se incrementa directamente con  $X$ , es decir la varianza del error no es constante, violándose así el supuesto de homogeneidad de varianza. Se recomienda para eliminar este problema emplear el método de mínimos cuadrados con factores de peso.

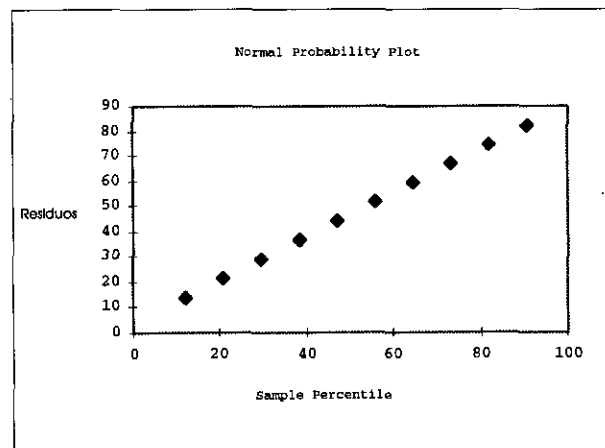


*Figuras F y G:* Por medio de esta gráfica se determina si una variable explicatoria  $X_{ip}$  que no está incluido en modelo, debe ser incluida, dado que existe una fuerte relación entre los residuos y esta variable  $X_{ip}$ .



Para el supuesto de normalidad, que es la parte esencial de nuestro trabajo, como primer estudio, el análisis de residuos presenta una gráfica con los residuos estandarizados y sus frecuencias acumuladas, para ello, se ordena los residuos estandarizados de menor a mayor y si la gráfica muestra una tendencia a *línea recta*, entonces el supuesto se cumple<sup>10</sup>.

Ver gráfica.



<sup>10</sup> Cuando se concluye gráficamente que el supuesto se cumple, queremos decir que gráficamente los datos no presentan alguna razón de que sea incorrecto.

característica de esta gráfica es que la frecuencia acumulada para una distribución normal se presenta como una línea recta con media 0 y varianza 1, y la podemos usar como medida para evaluar si los datos reflejan cualquier desviación de normalidad.

En muchas ocasiones una desviación de normalidad es causada por algunos factores, como el tener un modelo de regresión inadecuado, no existir homogeneidad de varianzas, o simplemente por que el número de residuos estandarizados es muy pequeño proporcionando un patrón de inestabilidad. Por lo tanto si se logra solucionar las causas, el supuesto de normalidad se cumplirá.

Para probar analíticamente los supuestos se encuentran varias pruebas. Para el supuesto de autocorrelación se puede emplear la pruebas de Durbin-Watson, en el caso del homogeneidad de varianzas se tiene la prueba de correlación de Sperman y para el supuesto de normalidad se tienen comúnmente las pruebas  $\chi^2$  y Kolgomorov. Las pruebas de normalidad utilizadas tienen un inconveniente muy grande, dado que consideran a los residuos una muestra aleatoria e independientemente distribuida, cuando en realidad no lo son, dado que resultan de un ajuste de regresión, lo cual son problemas distintos.

## ***1.9 COMENTARIOS GENERALES***

En este capítulo se presentó de manera resumida los aspectos más importantes de un análisis de regresión y el interés principal se enfocó en el supuesto de normalidad dado que si se cumple, los estimadores de los coeficientes de regresión adquieren propiedades relevantes como ser insesgados, tener varianza, como se analizó en la sección 1.5.

El aspecto computacional es indispensable para la realización de cálculos en el análisis de regresión. Existen diversidad de paquetes estadísticos que permiten realizar un estudio completo de regresión, tales como STATA, SYSTAT, SOLO, SAS, MINITAB, entre los más importantes. Sin embargo, este trabajo presenta la utilidad de la hoja de cálculo EXCEL 5.0, en la realización de un análisis de regresión y así como la presentación de una rutina (Macro) para probar el supuesto de normalidad en los residuos.

# **CAPÍTULO 2**

---

## **PRUEBAS DE BONDAD DE AJUSTE PARA LA DISTRIBUCIÓN NORMAL**

## **CAPÍTULO 2: PRUEBAS DE BONDAD DE AJUSTE PARA LA DISTRIBUCIÓN NORMAL**

### **2.1 INTRODUCCIÓN**

La distribución usada con mayor frecuencia en análisis estadísticos es la distribución normal. Su uso puede verse de dos maneras. La primera relacionada a la clase de estadística en la cual se toma a la distribución normal debida a un tamaño de muestra grande, tal como lo es el teorema del límite central. La segunda forma es la situación en la que la distribución normal es asumida como un supuesto del modelo matemático para el fenómeno bajo estudio. En el caso que la normalidad se considera un supuesto, si no se cumple ocasiona dar inferencias equivocadas por lo que es una necesidad probar que realmente se cumpla, para ello se requiere realizar una prueba de bondad de ajuste.

En este capítulo se presentan algunas pruebas de bondad de ajuste para verificar la normalidad de una muestra aleatoria, independiente e idénticamente distribuida. Es importante mencionar que cuando se requiere probar normalidad para los residuos de un ajuste de regresión ya se cuenta con una muestra aleatoria, por lo que se debe tener cuidado al aplicar alguna prueba. Por fortuna se encuentra un estudio realizado por Pierce y Kopecky que permite probar normalidad a los residuos, el cual se discutirá en la sección 2.6.2. A continuación se explica cual es la finalidad que una prueba de bondad de ajuste persigue.

## 2.2 OBJETIVO DE UNA PRUEBA DE BONDAD DE AJUSTE

Las hipótesis estadísticas son afirmaciones relacionadas a una característica que se desconoce de una población. Una *prueba de bondad de ajuste* es aquella que analiza las pruebas de hipótesis en las que la característica que se desconoce es alguna propiedad de la forma funcional de la distribución que se muestrea. El conocer la función de distribución  $F(x)$  de una variable  $x$  es de relevancia dado que especifica toda la información probabilística acerca de dicha variable. El conocimiento de  $F(x)$  es fundamental en muchos problemas.

Las pruebas de bondad de ajuste hacen comparaciones con los resultados de una muestra aleatoria y aquellos que se esperan observar si la hipótesis nula es correcta. Las comparaciones se hacen mediante la clasificación de los datos que se observan en cierto número de categorías y así se comparan las frecuencias observadas con las esperadas para cada categoría. Una prueba de bondad de ajuste dice que tan cercana se encuentra nuestra suposición a cerca de la función de distribución verdadera y no me asegura sólo dice si hay o no evidencia en contra.

Para fines de este trabajo se asume lo siguiente: Se tiene a  $X_1, X_2, \dots, X_n$  como una muestra aleatoria de tamaño  $n$ , con una función de densidad de probabilidad  $f(x)$  y una función de distribución acumulativa  $F(x)$ . La  $f(x)$  y  $F(x)$  de la distribución normal son  $\phi(x)$  y  $\Phi(x)$  respectivamente.

La hipótesis nula que se desea probar es:

$$H_0: f(x) = \phi(x) \quad \text{ó} \quad H_0: F(x) = \Phi(x)$$

La f.d. de la distribución normal está dada por:

$$(2.1.1) \quad \phi(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{1/2((x-\mu)/\sigma)^2} \quad \text{para} \quad -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0$$

Probando como inicio la hipótesis nula sobre la distribución normal  $H_0$ , se considera que la variable  $x$  está distribuida como una variable normal, o en otras palabras  $x$  tiene función de densidad dada por (2.1.1). Si alguno de los valores de  $\mu$  o  $\sigma$  no son especificados completamente, entonces la hipótesis nula será considerada como una hipótesis compuesta. Este trabajo considera solamente las hipótesis compuestas con  $\mu$  y  $\sigma$  desconocidos.

Para un valor específico de error tipo I, la hipótesis nula será rechazada si existe una diferencia suficiente entre las frecuencias observadas y las esperadas. Vale hacer notar que no existe una hipótesis alternativa dado que pueden ser muchas distribuciones diferentes las que podrían ser la adecuada. En consecuencia, una prueba de bondad de ajuste no debe usarse por sí misma para aceptar la afirmación de la hipótesis nula. La decisión es no rechazar  $H_0$  (más que aceptarla) si la diferencia que existe entre las frecuencias observadas y esperadas, es en forma relativamente, pequeña.

### ***2.2.1 Clasificación de las Pruebas***

Para probar normalidad se cuentan con diversidad de pruebas de bondad de ajuste, sin embargo, algunas de ellas han demostrado ser mejores al compararse con otras. Las pruebas de normalidad se agrupan en las siguientes categorías:

- 1.-  $J_i$  - Cuadrada
- 2.- Basadas en la Función de Distribución Empírica
  - Anderson - Darling



- Cramer - Von Mises
  - Watson
  - Kolgomorov
  - Kuiper V.
- 3.- Basadas en Momentos
- Ji - Cuadrada de D'Agostino - Pearson
  - Ji - Cuadrada de Bowman - Shenton
- 4.- Basadas en Regresión
- Shapiro - Wilk
  - D'Agostino
  - Shapiro - Francia
- 5.- Otras Pruebas
- Prueba de Locke-Spurrier's
  - Prueba de Gap

Cada una de las pruebas mencionadas adquieren importancia cuando se aplican bajo determinadas condiciones, no obstante existen algunas pruebas que han mantenido su eficiencia para diversas condiciones, entre ellas encontramos las de:

- Anderson - Darling
- Cramer - Von Mises
- Watson
- Shapiro - Wilk
- D'Agostino

las cuales se describe su aplicación en este capítulo. Es de importancia decir que estas pruebas permiten probar normalidad para una muestra aleatoria,

independiente e idénticamente distribuida. Antes de describirlas se presenta como probar normalidad por medio del un Gráfico de probabilidad.

### **2.3 PRUEBA DE AJUSTE BASADA EN EL GRÁFICO DE PROBABILIDAD**

Una herramienta de mucha utilidad para explorar normalidad, es *el gráfico de probabilidad normal*. Este gráfico se construye con las observaciones ordenadas (dando origen a las estadísticas de orden) y ciertas constantes que dependen de la distribución investigada, en nuestro caso es la distribución normal. El gráfico elige las constantes de tal manera que si la distribución hipotética es correcta se espera la forma de *una línea recta*.

La construcción del gráfico de probabilidad se hace de la siguiente forma:

1. Se tiene una muestra aleatoria  $X_1, X_2, \dots, X_n$  y sean  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  los valores ordenados de la muestra aleatoria (estadísticas de orden) de la característica de interés.

2. Considerar la secuencia de valores:

$$P_{(i)} = (i - 0.5)/n$$

que aproxima a la función de distribución empírica (ver. 2.4.2) sobre el eje de los valores ordenados  $X_{(i)}$ .

Para la distribución normal estándar, los cuantiles teóricos serían los valores  $q_{(i)}$  tal que:

$$P\{(Z \leq q(i))\} = \int_{-\infty}^{q(i)} 1/\sqrt{2\pi} e^{(-z^2/2)} dz = P(i)$$

3. Se grafican los valores  $\{X_{(i)}, q_{(i)}\}$  se obtiene una gráfica que indica que tanto concuerda la función de distribución acumulativa empírica  $F_n(x)$  con la teórica  $F(x)$ .
4. Si coordinan los valores, la gráfica corresponderá a puntos sobre una *línea recta* y se dice que no existe evidencia en contra de que la distribución de esos datos sea la distribución normal.

A partir de los datos que se producen el gráfico se puede realizar una prueba analítica basada en el coeficiente de correlación.

$$r_Q = \frac{\sum_{i=1}^n (x_{(i)} - \bar{x})(q_{(i)} - \bar{q})}{\sqrt{(x_{(i)} - \bar{x})(q_{(i)} - \bar{q})}}$$

Si  $r_Q$  es menor que  $r_\alpha$  de tablas se rechaza a este nivel la hipótesis de normalidad.

TABLA: Para la prueba de Normalidad basada en el coeficiente de correlación  $r_Q$

Tamaño de Muestra $n$	Nivel $\alpha = 0.10$	Nivel $\alpha = 0.05$	Nivel $\alpha = 0.01$
10	0.880	0.918	0.935
15	0.911	0.938	0.951
20	0.929	0.950	0.960
25	0.941	0.958	0.966
30	0.949	0.964	0.971
40	0.960	0.972	0.977
50	0.966	0.976	0.981
60	0.971	0.980	0.984
75	0.976	0.984	0.987
100	0.981	0.986	0.989
150	0.987	0.991	0.992
200	0.990	0.993	0.994

## 2.4 PRUEBAS DE NORMALIDAD BASADAS EN LA FUNCIÓN DE DISTRIBUCIÓN EMPÍRICA

(*Anderson-Darling, Cramer-Von Mises y Watson*)

### 2.4.1 Introducción

De las pruebas de bondad de ajuste consideradas de las más potentes para probar normalidad se encuentran las pruebas basadas en la Función de Distribución Empírica, sus estadísticos de prueba miden la discrepancia que existe entre la función de distribución empírica y la hipotética. En esta sección se describen las pruebas que se consideran más conocidas, *Anderson-Darling, Cramer-Von Mises y Watson*.

### 2.4.2 La Función de Distribución Empírica

La Función de Distribución Empírica se define como el número de observaciones  $X_i$  menores o iguales que  $x$ , en donde  $X_1, X_2, \dots, X_n$  se considera una muestra aleatoria, la cual proviene de una población cuya función de distribución es  $F(x)$ .

La Función de Distribución Empírica se representa como

$$F_n(x) = \#(X_i \leq x) / n \quad \text{para } -\infty < x < \infty$$

aquí se tiene que  $\#(X_i \leq x)$  representa el número de  $X_i$  que son menores o iguales que  $x$ . La  $F_n(x)$  se considera un estimador de la  $F(x)$  poblacional y es un estimador consistente, es decir, conforme  $n$  crece se acerca a la misma forma de  $F(x)$ . Las características que presenta son las siguientes:

1.  $F_n(x)=0$  si  $x < X_{(1)}$
2.  $F_n(x)=i/n$  si  $X_{(i)} \leq x \leq X_{(i+1)}$
3.  $F_n(x)=1$  si  $X_{(n)} \leq x$  para  $i=1,2,\dots,n-1$

### 2.4.3 Estadísticos de Prueba Basados en la Función de Distribución Empírica

En esencia las pruebas basadas en la Función de Distribución Empírica para probar normalidad, implican medir la discrepancia entre la f.d.a.

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} e^{-1/2((t-\mu)/\sigma)^2} dt$$

de la distribución normal y la función distribución empírica:

$$F_n(x) = (X \leq x)/n \quad \text{de la muestra.}$$

$\mu$  y  $\sigma$  frecuentemente no son especificados, por lo que se estiman por medio del método de máxima verosimilitud con  $\bar{x}$  y  $s$  donde:

$$\bar{X} = \frac{\sum_{i=1}^n \bar{X}_i}{n} \quad \text{y} \quad s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

La distribución de los estadísticos que se basan en la Función de Distribución Empírica, dependen del número de parámetros estimados y las distribución que se propone, sin embargo, cuando los parámetros desconocidos son los parámetros de localización y escala y estos se estiman por el método de máxima verosimilitud, la distribución de los estadísticos no dependen de los valores de los parámetros

desconocidos . Aquí es importante saber cuales son los parámetros de localización y escala.

Si se cuenta con una muestra aleatoria  $X_1, X_2, \dots, X_n$  con  $F_0(x)$  hipotética y  $X_{(1)}, X_{(2)}, \dots, X_{(n)}$  son las estadísticas de orden,  $F_0(x)$  frecuentemente es de la forma  $F(w)$  con  $w=(x-\alpha)/\beta$ , donde  $\alpha$  es el *parámetro de localización* y  $\beta$  es el *parámetro de escala*. Si  $u=F(w)$  entonces  $w=F^{-1}(u)$ , es decir,  $F^{-1}(\cdot)$  es la función inversa de  $F(\cdot)$  con  $f(w)$  su correspondiente función de densidad de  $F(w)$ .

Existen 2 clases de estadísticos de prueba basados en la Función de Distribución Empírica.

1. **Estadísticos Supremos.** Representados por  $D^+$  y  $D^-$ . En donde

$$D^+ = \sup_x (F_n(x) - F(x))$$

indica la diferencia vertical más larga cuando  $F_n(x)$  es mayor que  $F(x)$

$$\text{y } D^- = \sup_x (F(x) - F_n(x))$$

representa la diferencia vertical más larga cuando  $F_n(x)$  es menor que  $F(x)$

2.- **Clase Cuadrática:** Estos estimadores se basan en la familia de Cramer-Von Mises dado por :

$$(2.4.3.1) \quad Q = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \varphi(x) dF(x)$$

en donde  $\varphi(x)$  pondera las diferencias de las distribuciones empírica y hipotética al cuadrado. Cuando  $\varphi(x)=1$  resulta un estadístico llamado *Cramer- Von Mises* y cuando  $\varphi(x)=\{(F(x)(1-F(x)))\}^{-1}$  el estadístico que se origina es el de *Anderson-Darling*. Para resolver la fórmula anterior se aplica el Método de Transformación integral de Probabilidad, en cual se explica.

### ***Transformación integral de Probabilidad***

Es el fundamento de las pruebas basadas en la Función de Distribución Empírica y consiste en lo siguiente.

Si  $x$  es una variable aleatoria con distribución continua  $F(x)$  cualesquiera, entonces la variable aleatoria

$$Z=F(x)$$

*tiene una distribución uniforme (0,1)*. Por lo que el problema se reduce a probar que  $Z$  se distribuye como una uniforme  $(0,1)$  y se puede así concluir si hay o no evidencia de que la función propuesta sea correcta.

En las siguientes secciones se describen algunas de las pruebas más importantes para probar normalidad basadas en la Función de Distribución Empírica como lo son *Anderson-Darling*, *Cramer- Von Mises* y *Watson*.

#### 2.4.4 Prueba de Anderson-Darling

Las pruebas basadas en la Función de Distribución Empírica son consideradas de las más potentes para probar normalidad en una muestra aleatoria, independiente e idénticamente distribuida, una de ellas es la prueba de *Anderson-Darling (1952)*. Esta prueba a diferencia de las Cramer-Von Mises y Watson, pone mayor peso en los extremos de la distribución. Su estadístico de prueba se representa por  $A^2$  y es el siguiente:

##### *Estadístico de Prueba:*

$$A^2 = -\sum_{i=1}^n [(2i-1)\{\log P_i + \log(1 - P_{n+1-i})\} / n] - n$$

Para valores grandes  $A^2$  es significativa. El uso de  $A^2$  es basado sobre los trabajos de Stephens (1974,1976) quien calculó una fórmula a un nivel de significancia.

A continuación se presenta un procedimiento para realizar la prueba de Anderson-Darling  $A^2$ .

##### *Procedimiento*

1. Se ordena la muestra en orden ascendente.  $X_{(1)} \leq \dots \leq X_{(n)}$
2. Se calculan los valores estandarizados  $Y_{(i)}$ , donde

$$Y_{(i)} = (X_{(i)} - \bar{X})/s \quad \text{para } i = 1, 2, \dots, n$$



3. Calcular  $P_i$  para  $i = 1, 2, \dots, n$  donde

$$P_i = \Phi(Y_{(i)}) = \int_{-\infty}^{Y_{(i)}} \frac{e^{-t^2}}{\sqrt{2\pi}} dt$$

$\Phi(v)$  representa la función de distribución acumulativa de la distribución normal estándar y  $P_i$  es la probabilidad acumulada correspondiente a los valores estandarizados de  $Y_{(i)}$ .  $P_i$  puede encontrarse en las tablas de la distribución normal estándar.

4.- Calcular la estadística de Anderson Darling:

$$A^2 = -\sum_{i=1}^n [(2i-1) \{ \log P_i + \log(1 - P_{n+1-i}) \} / n] - n$$

donde  $\log$  es el logaritmo en base  $e$  ( $\ln$ ).

5.- Rechazar  $H_0$  de que existe normalidad si:

$A^2 >$	0.341	para $\alpha = 0.50$
	0.470	para $\alpha = 0.25$
	0.561	para $\alpha = 0.15$
	0.631	para $\alpha = 0.10$
	0.752	para $\alpha = 0.05$
	0.873	para $\alpha = 0.025$
	1.035	para $\alpha = 0.01$
	1.159	para $\alpha = 0.005$

### 2.4.5 Prueba de Cramer - Von Mises

La prueba de *Cramer - Von Mises* se basa en la Función de Distribución Empírica y se representa por  $W^2$ . Como se mencionó en la sección 2.4.3 el estadístico de prueba pertenece a la clase cuadrática y se origina cuando en la fórmula

$$Q = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \varphi(x) dF(x)$$

donde  $\varphi(x)=1$ , el cual pondera las diferencias entre la Función de Distribución Empírica y teórica. La forma del estadístico de prueba es:

**Estadístico de Prueba:**

$$W^2 = - \sum_{i=1}^n \left( P_i - \frac{2i-1}{2n} \right) + \frac{1}{12n}$$

Para valores grandes  $W^2$  es significativa. A continuación se presenta un procedimiento para realizar la prueba de Cramer - Von Mises  $W^2$ .

**Procedimiento**

1. Se ordena la muestra en orden ascendente.  $X_{(1)} \leq \dots \leq X_{(n)}$
2. Se calculan los valores estandarizados  $Y_{(i)}$ , donde

$$Y_{(i)} = (X_{(i)} - \bar{X})/s \quad \text{para } i = 1, 2, \dots, n$$

3. Calcular  $P_i$  para  $i = 1, 2, \dots, n$  donde

$$P_i = \Phi(Y_{(i)}) = \int_{-\infty}^{Y_{(i)}} \frac{e^{-t^2}}{\sqrt{2\pi}} dt$$

$\Phi(v)$  representa la función de distribución acumulativa de la distribución normal estándar y  $P_i$  es la probabilidad acumulada correspondiente a los valores estandarizados de  $Y_{(i)}$ .  $P_i$  se encuentra en las tablas de la distribución normal estándar.

4.- Calcular la estadística de Cramer-Von Mises:

$$W^2 = - \sum_{i=1}^n \left( P_i - \frac{2i-1}{2n} \right)^2 + \frac{1}{12n}$$

5.- Rechazar  $H_0$  de que existe normalidad  $W^2$  rebasa el valor de tablas.

$W^2 >$	0.051	para $\alpha = 0.50$
	0.074	para $\alpha = 0.25$
	0.091	para $\alpha = 0.15$
	0.104	para $\alpha = 0.10$
	0.126	para $\alpha = 0.05$
	0.148	para $\alpha = 0.025$
	0.179	para $\alpha = 0.01$
	0.201	para $\alpha = 0.005$

### 2.4.6 Prueba de Watson

La prueba de *Watson* se representa por  $U^2$ . Su estadístico de prueba tiene la siguiente forma:

**Estadístico de Prueba:**

$$U^2 = W^2 - n(\bar{P} - 0.5)^2$$

en donde

$$\bar{P} = \sum_{i=1}^n \frac{\bar{P}_i}{n}$$

En seguida se presenta un mecanismo para realizar la prueba de Watson  $U^2$ .

#### **Procedimiento**

1. Se ordena la muestra en orden ascendente.  $X_{(1)} \leq \dots \leq X_{(n)}$
2. Se calculan los valores estandarizados  $Y_{(i)}$ , donde

$$Y_{(i)} = (X_{(i)} - \bar{X})/s \quad \text{para } i = 1, 2, \dots, n$$

3. Calcular  $P_i$  para  $i = 1, 2, \dots, n$  donde

$$P_i = \Phi(Y_{(i)}) = \int_{-\infty}^{Y_{(i)}} \frac{e^{-t^2}}{\sqrt{2\pi}} dt$$

$\Phi(V)$  representa la función de distribución acumulativa de la distribución normal estándar y  $P_i$  es la probabilidad acumulada correspondiente a los valores estandarizados de  $Y_{(i)}$ .  $P_i$  puede encontrar en las tablas de la distribución normal estándar.

4.- Calcular la estadística de prueba de Watson:

$$U^2 = W^2 - n(\bar{P} - 0.5)^2$$

5.- Rechazar  $H_0$  de que existe normalidad  $W^2$  rebasa el valor de tablas.

$U^2 >$	0.048	para $\alpha = 0.50$
	0.070	para $\alpha = 0.25$
	0.085	para $\alpha = 0.15$
	0.096	para $\alpha = 0.10$
	0.117	para $\alpha = 0.05$
	0.136	para $\alpha = 0.025$
	0.164	para $\alpha = 0.01$
	0.183	para $\alpha = 0.005$

Un ejemplo de la aplicación de estas estadísticas de prueba, se da en el capítulo tres.

## **2.5. PRUEBAS DE BONDAD DE AJUSTE SHAPIRO-WILK Y D'AGOSTINO**

### **2.5.1 Introducción**

Otras pruebas que permiten probar normalidad son las de Shapiro-Wilk y D'Agostino las cuales son llamadas pruebas de regresión. Estas pruebas consideran un modelo lineal:

$$(2.5.1.1) \quad X_{(i)} = \mu + \sigma E(Z_i) + \varepsilon_i$$

y estimando en particular el parámetro  $\mu$  y  $\sigma$  por medio de regresión. En (2.5.1)  $X_{(i)}$  es la  $i$ -ésima estadística de orden para una muestra de tamaño  $n$ ,  $EZ_{(i)}$  es el valor esperado de la  $i$ -ésima estadística de orden para una muestra de tamaño  $n$ , tomados de la distribución normal estándar (con  $\mu = 0$  y  $\sigma=1$ ) y  $\varepsilon_i$  es el error aleatorio. En las siguientes secciones se da un breve descripción de estas pruebas.

### **2.5.2 Prueba de Shapiro-Wilk**

La prueba de  $W$  de Shapiro-Wilk (1965), considerada de las más potentes para probar normalidad, fue realizada para intentar resumir formalmente algunos aspectos de los gráficos de probabilidad. En un gráfico de probabilidad las observaciones son ordenadas como primer paso, entonces se comparan con los valores esperados de las estadísticas de orden normal. Una desviación de linealidad indica no normalidad y el hecho de que tal desviación es independiente de las constantes de regresión hacen que la aproximación sea libre de escala y localización.

La estadística  $W$  de Shapiro-Wilk es proporcional a el cuadrado de la pendiente de regresión (quien se espera para normalidad sea proporcional a la varianza  $\sigma^2$ ) dividido por

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

excepto para alguna constante, esta es la razón de dos estimadores de  $\sigma^2$  (bajo la hipótesis de normalidad); y una desviación de normalidad trae como consecuencia disminuir el valor de  $W$ .

Sean  $Y_{(1)}, Y_{(2)}, \dots, Y_{(n)}$  un conjunto de observaciones escritas en orden ascendente y  $m_i$  denota el valor esperado de la  $i$ -ésima estadística de orden de  $n$  variables normal estándar, y  $V$  corresponde a la matriz de varianzas y covarianzas. Entonces la pendiente cuadrada  $B$  de la línea de regresión, es proporcional a

$$\mathbf{m}'\mathbf{V}^{-1} (Y_{(1)}, Y_{(2)}, \dots, Y_{(n)})'.$$

La estadística de prueba de  $W$  es definida de la siguiente forma:

$$W = \frac{\left\{ \sum_{i=1}^n a_i Y_{(i)} \right\}^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

donde  $\mathbf{a}' = (a_1, a_2, \dots, a_n)$  es proporcional a  $\mathbf{m}'\mathbf{V}^{-1}$  y normalizado por conveniencia. (que es,  $\mathbf{a}'\mathbf{a}=1$ ). Por lo tanto

$$(2) \quad \mathbf{a} = (\mathbf{m}'\mathbf{V}^{-1}) / (\mathbf{m}'\mathbf{V}^{-1}\mathbf{V}^{-1}\mathbf{m})^{1/2}$$

Aquí  $W^{1/2}$  es el coeficiente de correlación muestral entre  $a'$  y  $(Y_{(1)}, Y_{(2)}, \dots, Y_{(n)})$  también  $W$  es siempre  $\leq 1$ , cuando  $W = 1$  indica que  $(Y_{(1)}, Y_{(2)}, \dots, Y_{(n)})$  es un múltiplo de  $a'$ . En realidad,  $m'V^{-1}$  es aproximadamente una constante múltiplo de  $m'$ , por lo tanto esto nos conduce exactamente a una línea recta (en la gráfica de  $Y_{(i)}$  en contra de  $m_i$ ) de  $W$  muy cercano a 1 y viceversa. También si  $W$  es significativamente más pequeño que 1, la hipótesis de normalidad será rechazada.

Los  $a_i$  son los pesos óptimos para los estimadores de mínimos cuadrados de  $\sigma$ , dado que la población es normalmente distribuida.  $W$  puede verse también como  $R^2$  (coeficiente de determinación) obtenido desde una gráfica de probabilidad normal y de esta manera la noción de una prueba de correlación.

Los valores de  $a_i$  para  $n = 3$  a 50 fueron dados por Shapiro- Wilk (tabla 5.4 pag. 209, Ralph B. D'Agostino, Michael A. Stephens) Puesto que  $W$  es similar a una  $R^2$ , valores grandes (valores cercanos a 1) indican normalidad y valores pequeños (lejos de 1) indican no normalidad. Por lo que valores chicos de  $W$  harán rechazar  $H_0$ . Las tablas presentadas (tabla 5.5 pag. 212, 213, Ralph B. D'Agostino, Michael A. Stephens) muestran los valores críticos de  $W$  para  $n=3$  a 50.

### **2.5.3 Prueba de D'Agostino**

Una prueba alternativa para comprobar normalidad cuando  $n > 50$  fue la realizada por D'Agostino (1971). Esta se considera como una modificación de la prueba  $W$  de Shapiro-Wilk y la cual no requiere de la tabla de pesos  $a$ . Algunas de las razones por la que se recomienda aplicarla son:



1. es simple de calcular
2. es mas potente para simetría de extremos grandes

Esta prueba es particularmente de utilidad para muestras de tamaño grande, porque su cálculo es fácil de realizar.

La forma de su estadístico de prueba es la siguiente:

$$D = \frac{T}{n^2 \sqrt{m_2}}$$

$$= \frac{T}{n^{3/2} \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}}$$

donde  $T = \sum_{i=1}^n (i - \frac{1}{2}(n+1))X_{(i)}$

La estadística **D** es igual a una constante, a la razón del estimador lineal Downton's de la desviación estándar o desviación estándar muestral.

El valor esperado de D es aproximadamente  $\frac{1}{2\sqrt{\pi}} = 0.28209479$  y la desviación estándar es asintóticamente

$$\left[ \frac{12\sqrt{3} - 27 + 2\pi}{24\pi n} \right]^{1/2} = \frac{0.02998598}{\sqrt{n}}$$

una variable estandarizada aproximada está dada por:

$$Y = \frac{\sqrt{n}(D - 0.28209479)}{0.02998598}$$

Si la hipótesis nula de normalidad es falsa  $Y$  tenderá a ser diferente de 0. Estudios de simulación realizados por D'Agostino indican que para una distribución alternativa con Kurtosis menor que la normal ( $\beta_2 < 3$ )  $Y$  tiende a ser mayor que 0. Para distribuciones alternativas con ( $\beta_2 > 3$ )  $Y$  tiende a ser menor que 0.

La tabla que permite decidir rechazar o no la hipótesis de normalidad fue realizada por D'Agostino por medio de simulación para  $n=50$  a  $n=2000$ , basándose en los estudios de Cornish-Fisher y Pearson. (tabla 9.7, pag. 396, 397, Ralph B. D'Agostino, Michael A. Stephens)

## **2.6 PRUEBAS DE BONDAD DE AJUSTE PARA LOS RESIDUOS DE UN AJUSTE DE REGRESIÓN**

### **2.6.1 Introducción**

El probar normalidad en los residuos de un ajuste de regresión lineal es un problema totalmente distinto al de probar normalidad en una muestra aleatoria. Los residuos no se consideran una muestra aleatoria independiente, si no lo contrario, es decir, los residuos se encuentran correlacionados entre si. Por lo anterior, para comprobar el supuesto de normalidad de los residuos no es posible aplicar cualquier prueba de bondad de ajuste, si no que la prueba que se realice debe considerar este aspecto. Por fortuna existe una investigación realizada por *Donald A. Pierce y Kenneth J. Kopeccky en el año 1979* el cual afirma que entre las diversas pruebas de bondad de ajuste, las basadas en la Función de Distribución

Empírica son válidas para probar normalidad en los residuos de un ajuste de regresión. Este capítulo presenta de manera muy general los aspectos más importantes de esta investigación. Antes se define lo que es una distribución asintótica y parámetros de localización y escala.

### ***Distribución Asintótica***

Sea  $x_n$  una sucesión de variables aleatorias y sea  $X$  una variable aleatoria. Los resultados denominados asintóticos se refieren al comportamiento de la distribución de probabilidad de  $x_n$ , a medida que  $n$  tiende al infinito.

Se dice que la distribución asintótica  $x_n$  es aquella de la variable aleatoria  $X$  si, para un valor “grande de  $n$ ” la distribución de  $x_n$  y la  $X$  son prácticamente indistinguibles, es decir, a medida que  $n$  tiende al infinito ( $n \rightarrow \infty$ ).

$$P[X_n \leq x] \rightarrow P[X \leq x]$$

para cualquier  $x$  finito.

### ***Modelo de Localización y Escala***

Sea  $Y$  con función de distribución  $F^*$ , se define  $F_{\alpha,\beta}$  como la función de distribución de  $\alpha + \beta y$ . Sea la familia  $F = \{F_{\alpha,\beta} : -\infty < \alpha < \infty, \beta > 0\}$ . La familia  $F$  se denomina una *familia paramétrica de localización y escala*, en donde  $\alpha$  se conoce como el parámetro de localización y  $\beta$  parámetro de escala.

estadística de bondad de ajuste, la cual depende solamente de la distribución empírica de los  $\hat{e}_i$ ; es decir, son funciones simétricas permutables de los residuos.

Las estadísticas de bondad de ajuste son amplias y clases naturales, aunque otras son dignas de consideración después de que los  $\hat{e}_i$  no son idénticamente distribuidos.

Una atención considerable se debe de dar a este problema, pero muy en especial al caso de que la distribución de las observaciones es idéntica para una familia de localización y escala. *El estudio realizado por Pierce y Kopecky, que es un estudio sorprendente, es el hecho de que a pesar de la dimensión fijada de  $\beta$ , cuando  $n$  tiende al infinito ( $n \rightarrow \infty$ ), la distribución bajo la hipótesis nula de cualquier estadística de bondad de ajuste que depende únicamente de la Función de Distribución Empírica es precisamente la misma para el modelo de regresión como para el modelo ordinario de localización y escala.*

Se insiste que la estimación de los parámetros de localización y escala tienen un efecto substancial sobre la distribución límite, semejante a la de las estadísticas de bondad de ajuste, en comparación con algunos cálculos para los verdadero residuos,

$$\hat{e}_i = (y_i - X_i' \beta) / \sigma$$

Aquí se conoce que la estimación mayor de los parámetros de regresión tiene un efecto no incluido sobre la distribución límite.

Para probar normalidad cuando  $\sigma$  es conocido, el efecto asintótico de la estimación de  $\beta$  depende solamente de:

$$\sum X_i' (\hat{\beta} - \beta)$$

La distribución de esta cantidad dependerá de la forma del modelo de regresión, excepto cuando el modelo contiene un término constante, en tal caso

$$\sum X_i'(\hat{\beta} - \beta) = \sum (y_i - X_i' \hat{\beta})$$

Muchos resultados asintóticos están disponibles para los problemas de localización y escala. Para probar normalidad, con ambos parámetros desconocidos, la distribución límite de la mayor parte de los tipos generales comunes de estadísticas son ahora disponibles.

### 2.6.3 Comentarios

La importancia de la investigación realizada por Pierce y Kopecky, radica en que las pruebas de bondad de ajuste basadas en la Función de Distribución Empírica son las recomendables para probar el supuesto de normalidad de los residuos de un ajuste de regresión sin depender del número de parámetros en el modelo que se estima. Es importante decir que la investigación fue realizada en el año 1979 y no se encontró algún otra investigación posterior a esta fecha.

La demostración de la investigación se encuentra disponible en el artículo "*Testing goodness of fit for the Distribution of errors in Regression Models*" *Biometrika* (1979), 66, 1, pp.1-5, printed in Great Britain. En la sección anterior sólo se menciona la conclusión de la investigación, dado que la demostración va más allá de los objetivos de este trabajo.

# **CAPÍTULO 3**

---

## **PROBANDO NORMALIDAD EN EXCEL 5.0 (MACRO)**

## **CAPÍTULO 3: PROBANDO NORMALIDAD EN EXCEL 5.0 (MACRO)**

### **3.1 INTRODUCCIÓN**

En el capítulo anterior se describieron algunas pruebas de bondad de ajuste para validar el supuesto de normalidad en una muestra aleatoria, independiente e idénticamente distribuida, sin embargo, de acuerdo con el estudio realizado por Pierce y Kopecky en 1979, para probar normalidad en los residuos de un ajuste de regresión las pruebas que son adecuadas son las basadas en la Función de Distribución Empírica, tales como Anderson-Darling, Cramer-Von Mises, Watson, etc.

En este capítulo se presenta una rutina computacional en la hoja de cálculo EXCEL 5.0, que permite realizar un análisis de regresión, obtener algunas estadísticas descriptivas y calcular los estadísticos de prueba de Anderson-Darling, Cramer-Von Mises y Watson para probar el supuesto de normalidad de los residuos de un ajuste de regresión. Con esta rutina (Macro) se cumple uno de los objetivos en la realización de este trabajo.

### **3.2 EXPLICACIÓN DEL MACRO**

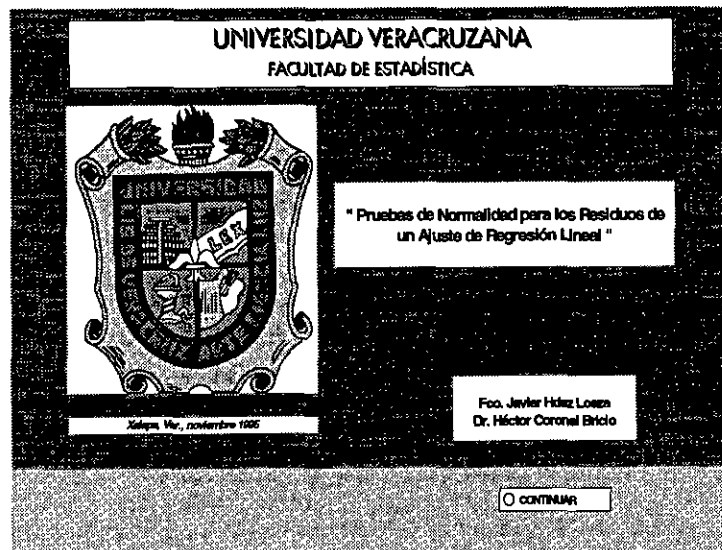
La hoja de cálculo EXCEL es una herramienta poderosa para resolver problemas numéricos y cuenta con un programa que permite realizar operaciones estadísticas. Un Macro en Excel es un conjunto de instrucciones ordenadas que permiten resolver problemas determinados. Al realizar un macro se graban las instrucciones indicadas y se genera un programa en *Visual Basic* que contiene dichas instrucciones. El Macro realizado en este trabajo recibe el nombre de REG-NOR.XLS y el programa generado se llama PERSONAL.XLS. Para ejecutar REG-NOR.XLS se requiere direccionar

adecuadamente al programa PERSONAL. Estos archivos están disponibles en disco con un pequeño manual que indica como instalarse y su uso.

A continuación se presentan las partes que integran el Macro:

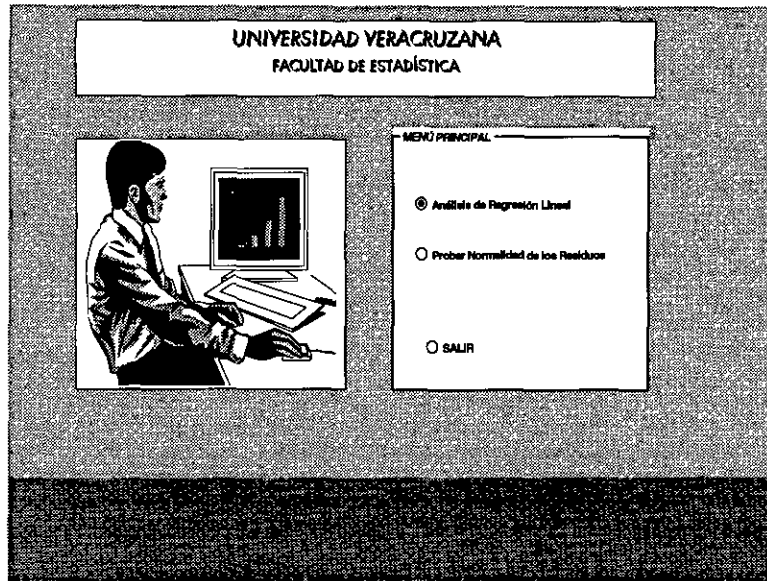
1. Presentación
2. Menú Principal
3. Hoja de Captura y análisis de regresión
4. Salidas
5. *Pruebas de Normalidad*

*La Pantalla de Presentación es la siguiente:*



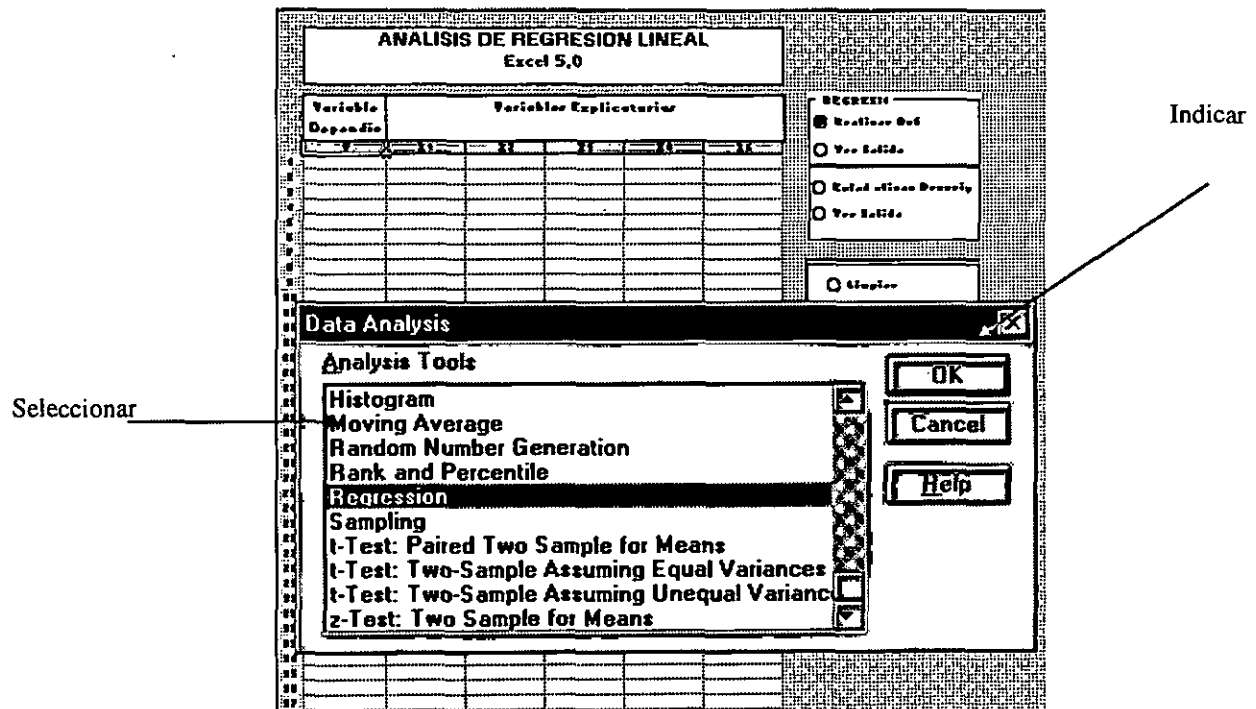
La siguiente imagen que aparece al dar *continuar* es el *Menú Principal*.



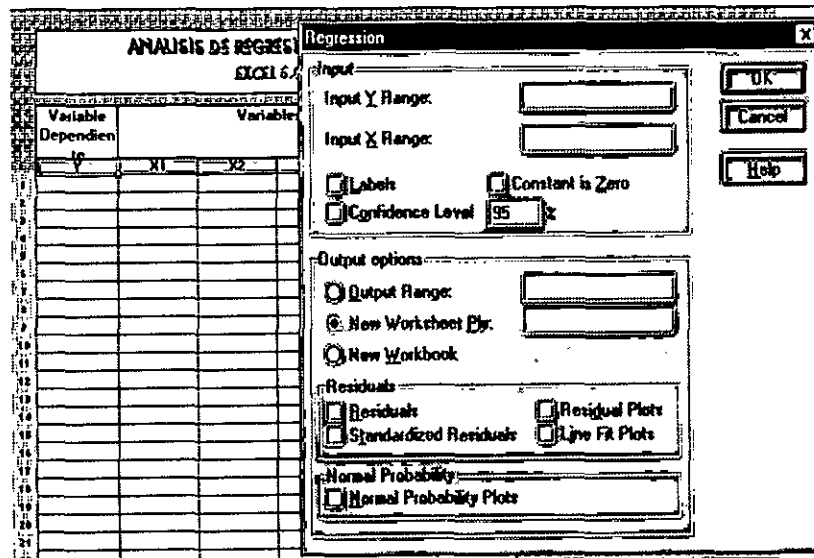


La pantalla de menú principal cuenta con 3 opciones. La primera opción envía a la hoja de edición para realizar un *análisis de regresión*, la segunda al menú de *pruebas de normalidad* y la tercera sale de menú principal. La opción se elige marcando el círculo correspondiente.

Cuando se elige la opción *análisis de regresión* aparece la siguiente hoja de edición en la cual se deben capturar los datos y posteriormente realizar la regresión o pedir algunas estadísticas descriptivas. Cuando se ejecuta la regresión aparece una caja de diálogo en la cual se especifica *regression* dando R y posteriormente *OK*.



Al elegir *regression* se presenta la siguiente caja en donde se realiza el análisis de regresión.



Los elementos de la caja son:

### *Valores de entrada (input)*

- I. El primer paso es marcar el rango de la variable dependiente (*input Y range*) y las variables explicatorias (*input X Range*).
- II. Se indica que se cuenta con etiqueta (*Labels*)
- III. Se especifica el intervalo de confianza para los  $\beta_i$  (*confidence level*)

### *Opciones de Salida (output options)*

- I. Definir una salida marcando el rango (*Output range*)
- II. Salida en una hoja en este programa (*New Worksheet ply*) si se elige esta opción se debe dar el nombre
- III. Salida en un archivo nuevo (*New Workbook*)

### *Análisis de Residuos (Residuals)*

- I. Generar residuos
- II. Generar residuos estandarizados
- III. Gráficas de Residuos
- IV. Línea de ajuste

### *Probabilidad Normal*

- I. Gráfico de Probabilidad Normal (*Normal Probability plots*)

Este gráfico permite probar si hay desviación de normalidad si no existe tendencia a una línea recta y se construye.

Realizado el análisis, independiente de la salida elegida se cuenta en el programa con una hoja de salida que permite presentar un reporte de los resultados obtenidos, para ello se marcan los resultados, se copian (CTRL-C), pasar a la hoja de análisis (REG-NOR) con ALT-W y al regresar al programa dando (CTRL-V) donde se indica *ver salida*:

# REPORTE DEL ANALISIS DE REGRESION Book#0

☐ Datos
 ☐ Imprime
 ☐ Limpia
 ☐ Regresión
 ☒ Anderson Darling
 ☐ Cramer-Von Mises
 ☐ Watson

## SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.92240498
R Square	0.85071056
Adjusted R Square	0.828776888
Standard Error	136.098993
Observations	35

## ANOVA

	df	SS	MS	F	Significance F
Regression	1	507.8291185	507.8291185	0.018870249	0.87145856
Residual	33	901745.1189	19254.64		
Total	34	902553.0037			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	81.9150	28.2344	1.7988	0.0882	-7.7466	115.1888
X1	0.1502	1.0022	0.1289	0.8974	-1.8897	2.1891

## RESIDUAL OUTPUT

## PROBABILITY OUTPUT


Observation	Predicted Y	Residuals	Standard Residuals	Probability	Y
1	86.6721	-49.8721	-0.8471		1.0000
2	88.48180249	-86.48180249	-0.410886188		4.266719288
3	81.4468979	-52.4468979	-0.3920989		7.142857143
4	84.67862888	-51.67862888	-0.361848076		10
5	88.48180249	-67.48180249	-0.428877083		12.88714286
6	88.47710016	-68.47710016	-0.447108956		15.71428571
7	82.84806548	-29.84806548	-0.217452188		18.67142857
8	82.84806548	-19.84806548	-0.143397979		21.42857143
9	82.84806548	-19.84806548	-0.143397979		24.28571429
10	82.84806548	-19.84806548	-0.143397979		27.14285714
11	82.84806548	-19.84806548	-0.143397979		30
12	82.84806548	-19.84806548	-0.143397979		32.85714286
13	82.84806548	-19.84806548	-0.143397979		35.71428571
14	82.84806548	-19.84806548	-0.143397979		38.57142857
15	88.47710016	-68.47710016	-0.447108956		41.42857143
16	82.84806548	-19.84806548	-0.143397979		44.28571429
17	81.87307779	-47.87307779	-0.386248881		47.14285714
18	82.84806548	-19.84806548	-0.143397979		50
19	82.84806548	-19.84806548	-0.143397979		52.85714286
20	82.84806548	-19.84806548	-0.143397979		55.71428571
21	82.84806548	-19.84806548	-0.143397979		58.57142857
22	82.84806548	-19.84806548	-0.143397979		61.42857143
23	82.84806548	-19.84806548	-0.143397979		64.28571429
24	82.84806548	-19.84806548	-0.143397979		67.14285714
25	82.84806548	-19.84806548	-0.143397979		70
26	82.84806548	-19.84806548	-0.143397979		72.85714286
27	82.84806548	-19.84806548	-0.143397979		75.71428571
28	88.47710016	-68.47710016	-0.447108956		78.57142857
29	82.84806548	-19.84806548	-0.143397979		81.42857143
30	82.84806548	-19.84806548	-0.143397979		84.28571429
31	82.84806548	-19.84806548	-0.143397979		87.14285714
32	82.84806548	-19.84806548	-0.143397979		90
33	82.84806548	-19.84806548	-0.143397979		92.85714286
34	88.47710016	-68.47710016	-0.447108956		95.71428571
35	82.84806548	-19.84806548	-0.143397979		98.57142857

Si existe interés en los resultados del análisis de residuos checar la nueva hoja que se generó con ALT-W y eligiendo la hoja de salida generada (book#).

Cuando se elige la opción *estadísticas descriptivas* la pantalla de salida es la siguiente:

<b>REPORTE DE ESTADÍSTICAS DESCRIPTIVAS</b> <small>Excel 5.0</small>						
<input type="radio"/> Datos <input type="radio"/> Imprime <input type="radio"/> Limpia <input checked="" type="radio"/> Regresión						
Estadísticas	Y	X1	X2	X3	X4	X5
Mean						
Standard Error						
Median						
Mode						
Standard Deviation						
Sample Variance						
Kurtosis						
Skewness						
Range						
Minimum						
Maximum						
Sum						
Count						
Confidence Level(95.000%)						

Para realizar las pruebas de normalidad de los residuos del ajuste realizados se selecciona en el menú principal la instrucción *probar normalidad de los residuos* y aparecerá la siguiente pantalla:

<b>Pruebas de Normalidad Basadas en la Función de Distribución</b>	
	<b>PREFERIR</b> <input type="radio"/> A <sup>2</sup> - Anderson-Darling <input type="text"/>
	<input type="radio"/> W - Cramer-Von Mises <input type="text"/>
	<input checked="" type="radio"/> U - Watson <input type="text"/>
	<input type="text"/>

Al seleccionar alguna de las pruebas presentadas aparece la siguiente pantalla, en la cual se realiza la prueba con los residuos de un ajuste o para una muestra aleatoria la cual debe capturarse en la hoja de edición.

PRUEBA A DE ANDERSON DARLING

1

2

3

Para realizar la prueba se siguen los siguientes pasos:

1. Si se realizó un análisis de regresión el reporte cuenta con los residuos estandarizados al dar *pega residuos*, localiza donde se encuentran.
2. Se marcan (SHIF-↓) y copian con CRTL-C.
3. Después CRTL-HOME para ir al inicio en la pantalla de reporte y marcar la prueba que se está realizando.
4. En la hoja de la prueba dar ALT-E-S-V para pegar los residuos estandarizados.
5. Se borra la parte baja de la hoja marcando con SHIF-→-END- y DEL.
6. Pasar al indicador 2 y dar tamaño de muestra.
7. Pasar al indicador 3 y realizar la prueba. Los resultados los pone en el menú de pruebas de normalidad.

El mecanismo es el mismo para cualquier prueba. Para hacer la comparación con tablas se marca *comparar con tablas* y aparece

Valores Calculados	
A <sup>2</sup> Anderson - Darling	0,537826
W <sup>2</sup> Cramer - Von Mises	0,0657942
U <sup>2</sup> Watson	0,0619014

VALORES DE TABLAS								
Nivel de Significancia								
	,50	,25	,15	,10	,05	,025	,01	,005
Upper tail								
A <sup>2</sup> Anderson - Darling	0,341	0,470	0,561	0,631	0,752	0,873	1,035	1,159
W <sup>2</sup> Cramer - Von Mises	0,051	0,074	0,081	0,104	0,126	0,148	0,179	0,201
U <sup>2</sup> Watson	0,048	0,070	0,085	0,096	0,117	0,136	0,164	0,183

Interpretación	
A <sup>2</sup> Anderson - Darling	Rechazar
W <sup>2</sup> Cramer - Von Mises	No Rechazar
U <sup>2</sup> Watson	No Rechazar

aquí se decide rechazar o no la hipótesis de normalidad. Si la estadística de prueba es mayor que el valor de tablas rechazar  $H_0$ , es decir no se cumple el supuesto de normalidad.

Este programa permite realizar pruebas hasta para 500 datos, regresión hasta con 5 variables explicativas, y es fácil de manejar usando comandos para Windows.

En la siguiente sección se presenta una aplicación del Macro en un problema real que se origina de una investigación realizada durante la especialidad en métodos estadísticos.

### 3.3 EJEMPLO DE APLICACIÓN

Con la finalidad de mostrar el uso del Macro para probar normalidad en los residuos de un ajuste de regresión, se presenta un ejercicio realizado en la especialidad de métodos estadísticos.

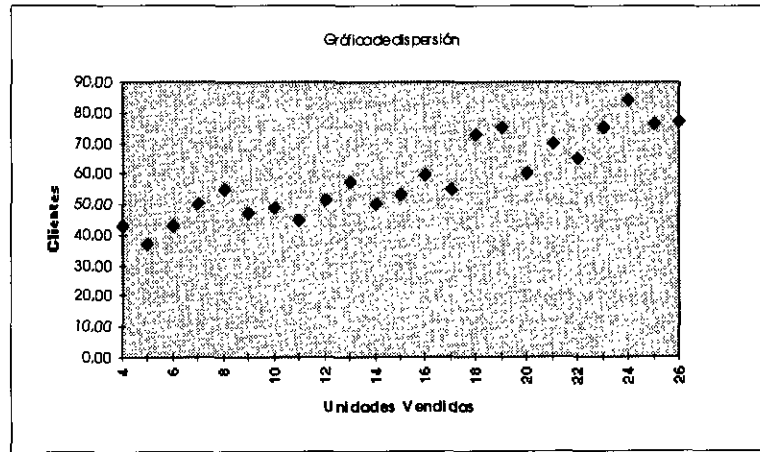
*Problema:*

Una compañía desea encontrar un modelo que permita describir la relación que existe entre las unidades vendidas (X) y el número de clientes (Y) que atienden. Para ello se obtuvo una muestra aleatoria de 23 datos y la información es la siguiente:

	Unidades_X	Clientes_Y
1	4	43
2	5	37
3	6	43
4	7	50
5	8	55
6	9	47
7	10	49
8	11	45
9	12	51
10	13	57
11	14	50
12	15	53
13	16	59,5
14	17	55
15	18	73
16	19	75
17	20	60
18	21	70
19	22	65
20	23	75
21	24	84
22	25	76
23	26	77



Al realizar un gráfico de dispersión se observa que la relación entre X y Y sigue el comportamiento parecido a una línea recta.



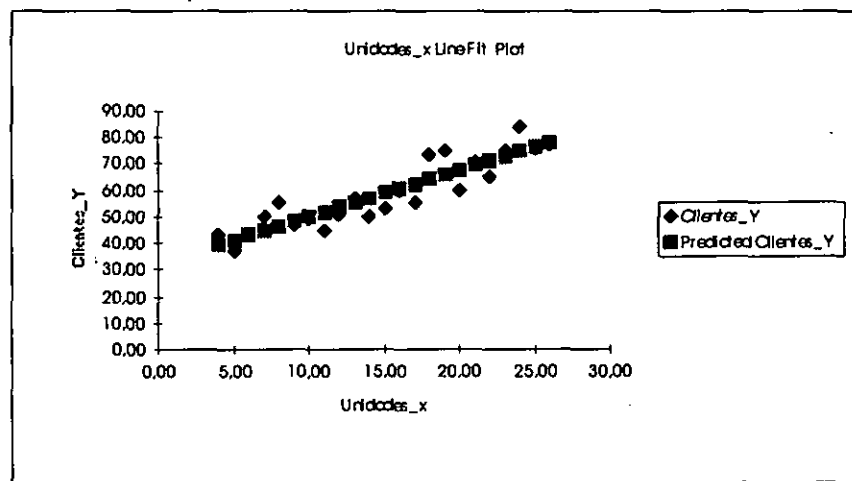
La salida generada por el macro al realizar un análisis de regresión lineal es la siguiente:

REPORTE DEL ANALISIS DE REGRESION						
Excel 5.0						
SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.90726124					
R Square	0.823104613					
Adjusted R Square	0.814691233					
Standard Error	5.68429468					
Observations	23					
ANOVA						
	df	SS	MS	F	Significance F	
Regression	1	3157.269022	3157.269022	97.71438638	2.37125E-09	
Residual	21	678.5353261	32.31206			
Total	22	3835.804348				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	32.1793	2.8306	10.9803	0.0000	26.0847	38.2739
Unidades_x	1.7663	0.1787	9.8851	0.0000	1.3947	2.1378
RESIDUAL OUTPUT						
Observation	Predicted Clientes	Residuals	Standard Residuals	Percentile	Clientes_Y	
1	39.2446	3.7554	0.6607	2.1739	37.0000	
2	41.01066957	-4.010669565	-0.705605496	6.52173913	43	
3	42.77717391	0.222826087	0.039200305	10.86956522	43	
4	44.54347826	5.456521738	0.959929428	15.2173913	45	
5	46.30978261	8.690217391	1.528811907	19.56521739	47	
6	48.07608696	-1.076086957	-0.189306792	23.91304348	49	
7	49.8423913	-0.842391304	-0.148196276	28.26086957	50	
8	51.60869565	-6.608695652	-1.162623689	32.60869565	50	
9	53.375	-2.375	-0.417817888	36.95652174	51	
10	55.14130435	1.856695652	0.326987913	41.30434783	53	
11	56.9076087	-6.907608696	-1.215209485	45.65217391	55	
12	58.67391304	-5.673913043	-0.998173628	50	55	
13	60.44021739	-0.940217391	-0.165406168	54.34782609	57	
14	62.20652174	-7.206521739	-1.26779524	58.69565217	59.5	
15	63.97282609	9.027173913	1.588090418	63.04347826	60	
16	65.73913043	8.260869565	1.629202933	67.39130435	65	
17	67.50543478	-7.505434783	-1.320381016	71.73913043	70	
18	69.27173913	0.72826087	0.128118071	76.08695652	73	
19	71.03804348	-5.038043478	-1.062232664	80.43478261	75	
20	72.80434783	2.19652174	0.386266423	84.7826087	76	
21	74.57065217	9.429347826	1.658642188	89.13043478	78	
22	76.33695652	-0.336956522	-0.05927851	93.47826087	77	
23	78.10326087	-1.10326087	-0.194089317	97.82608696	84	

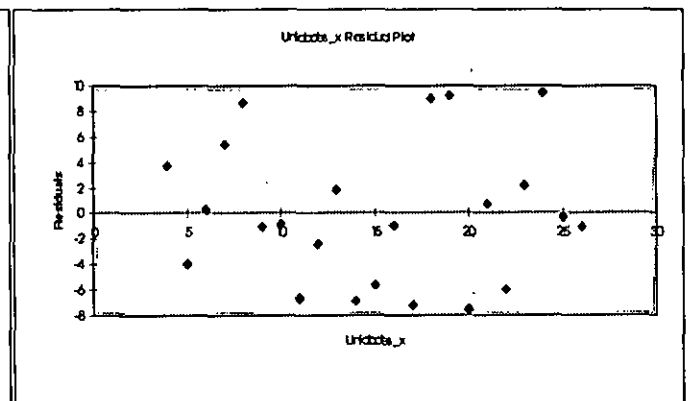
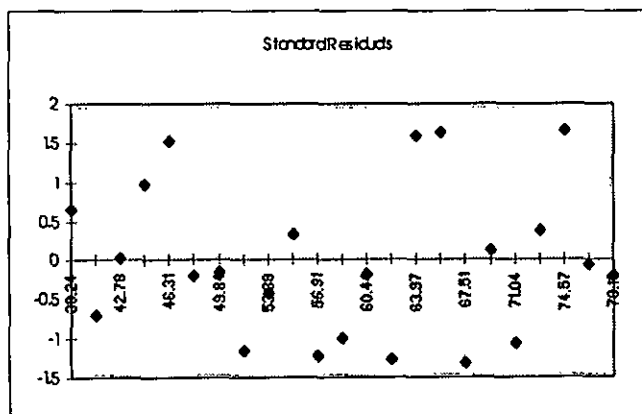
El modelo ajustado es:

$$\hat{y}_i = 32.17935 + 1.766304X_i$$

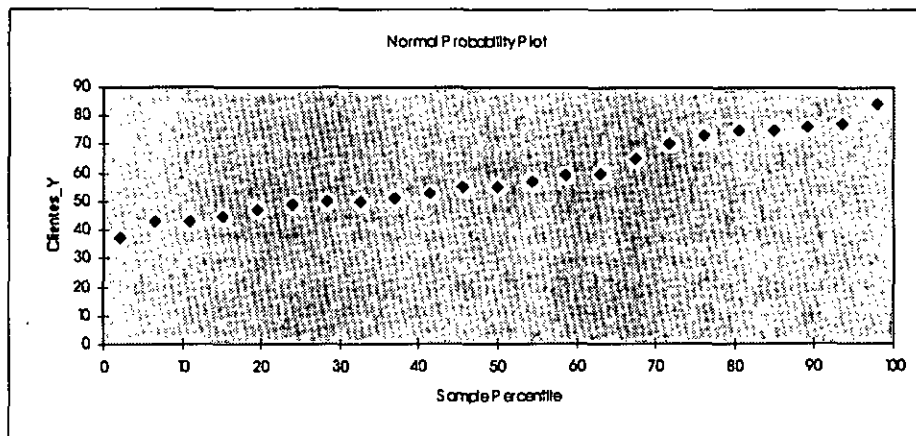
con un coeficiente de determinación de 0.83, un error estándar de 5.6842 y además los coeficientes de regresión son significativos. El gráfico de ajuste entre los valores predichos y los observados es el siguiente:



Como se puede apreciar gráficamente el ajuste no puede considerarse malo. Los siguientes gráficos de residuos, no presenta ningún patrón de comportamiento.




Con respecto al gráfico de probabilidad generado se observa que el supuesto de normalidad no está lejos de cumplirse.



Para comprobar el supuesto de normalidad analíticamente se ejecuta el macro realizado en este trabajo y los resultados son los siguientes:

**Pruebas de Normalidad Basadas en la Función de Distribución Empírica**



**PRUEBAS:**

- ☐ A<sup>2</sup> - Anderson-Darling **0.5378**
- ☐ W - Cramer-Von Mises **0.0658**
- ☒ U - Watson **0.0619**

Valores Calculados	
A <sup>2</sup> Anderson - Darling	0,537826
W <sup>2</sup> Cramer - Von Mises	0,0657842
U <sup>2</sup> Watson	0,0618014

VALORES DE TABLAS								
Nivel de Significancia								
	,50	,25	,15	,10	,05	,025	,01	,005
Upper tail								
A <sup>2</sup> Anderson - Darling	0,341	0,470	0,561	0,631	0,752	0,873	1,035	1,159
W <sup>2</sup> Cramer - Von Mises	0,051	0,074	0,091	0,104	0,126	0,148	0,179	0,201
U <sup>2</sup> Watson	0,048	0,070	0,085	0,096	0,117	0,136	0,164	0,183

Interpretación								
	,50	,25	,15	,10	,05	,025	,01	,005
A <sup>2</sup> Anderson - Darling	Rechazar	Rechazar	No Rechaza	No Rechaza	No Rechaza	No Rechaza	No Rechaza	No Rechaza
W <sup>2</sup> Cramer - Von Mises	Rechazar	No Rechazar	No Rechazar	No Rechazar	No Rechazar	No Rechazar	No Rechazar	No Rechazar
U <sup>2</sup> Watson	Rechazar	No Rechazar	No Rechazar	No Rechazar	No Rechazar	No Rechazar	No Rechazar	No Rechazar

Como lo muestra la tabla se rechaza la hipótesis de normalidad de los residuos a un nivel de significancia del 0.50 para los tres estadísticos de prueba, sin embargo cuando el nivel de significancia es del 0.25 Anderson-Darling rechaza la hipótesis y Cramer-Von Mises y Watson no lo hacen. Para los niveles de significancia menores o iguales que el 0.15 la hipótesis de normalidad no se rechaza. Si se prueba normalidad con un nivel de significancia de 0.05, no se rechaza la hipótesis nula y se dice que los datos no aportan evidencia significativa de que los residuos del ajuste de regresión se distribuyan normalmente.

Con los resultados anteriores el modelo ajustado  $\hat{y}_i = 32.17935 + 1.766304X_i$  puede ser considerado para resolver el problema planteado.

# CONCLUSIONES

El análisis de regresión siendo una de las herramientas estadísticas utilizada con frecuencia en diversidad de problemas, requiere ser aplicada eficientemente. En muchas ocasiones, quien realiza un análisis de regresión se limita a encontrar el modelo adecuado a su problema y estimar los coeficientes de regresión, considerándolo un buen modelo tomando en consideración que se tiene un error estándar pequeño, un coeficiente de determinación alto y coeficientes de regresión significativos. Sin embargo, un análisis de regresión para que sea completo debe comprobar los supuestos que el método de mínimos cuadrados hace al estimar los coeficientes de regresión, en especial para este trabajo el supuesto de normalidad de los errores, importante cuando nuestro interés recae en hacer estimaciones por intervalo o pruebas de hipótesis con respecto al verdadero valor de los coeficientes de regresión.

Este trabajo enfocó su atención al supuesto de normalidad, y tomando como fundamento la investigación realizada por Pierce y Kopecky en el año de 1979, que fue la investigación más reciente que se encontró, se llega a la conclusión de que las pruebas que deben ser utilizadas para probar el supuesto de normalidad en los residuos de un ajuste de regresión son las basadas en la Función de Distribución Empírica, entre las más importantes están: Anderson-Darling, Cramer-Von Mises y Watson, las cuales pueden ser calculadas por medio de la Macro presentada en este trabajo.

# REFERENCIAS

D' Agostino, R.B. y Stephens M.A. (1986) : "*Goodnes of fit techniques*"  
Marcel Dekker, New York.

Donald A. Pierce y Kenneth J. Kopecky : "*Testing goodnes of fit for distribution of errors in regression models*". Biometrika, 66, 1, pp 1-5. Printed in Great Britain.

G. Barrie Wetherill, P. Duncombe, M. Kenward, S. R. Paul, B.J. Vowden :  
"*Regression Analysis With Applicattions*". Chapman and Hall. London New York.

Menden Hall, Sheaffer y Wackerly : "*Estadística aplicada*". Grupo Editorial Iberoamérica, México.

Damodar Gujarati : "*Econometría Básica*" Mc Graw Hill. Colombia.

George C. Canavos : "*Probabilidad y Estadística, Aplicaciones y Métodos*".  
Mc Graw Hill, México.