Facultad de Ingeniería
División de Estudios de Posgrado

Modelos de regresión en pronósticos

Maestro en Ingeniería
(Investigación de Operaciones)

Nereo Elías Mata

1992

UNAM – Dirección General de Bibliotecas Tesis Digitales Restricciones de uso

CONTENIDO

PROLOGO

Entre las técnicas de pronóstico cuantitativo, las de mayor aplicación son las técnicas basadas en el análisis de regresión. Ello se debe a una serie de factores tales como su facilidad de aplicación, su respaldo estadístico y la confianza que genera el método, ya que los principales resultados son familiares para cualquier persona que cuente con conocimientos básicos de inferencia estadística.

Pero ante todo, ha influido que, por una parte, las técnicas de regresión permiten tratar muchos casos que están fuera del alcance de las técnicas de series de tiempo y, por otra parte, que los responsables sientan que sus pronósticos están mejor fundados, por la manera que están construidos los modelos.

En las series de tiempo se busca conocer el valor de la variable de interés hurgando en su pasado, para de esta manera identificar algún patrón de comportamiento y generar un modelo de la forma

$$Y_{t+h} = f(Y_t, Y_{t-1}, Y_{t-2}, \cdots, Y_{t-d})$$

donde f( ) es una función que permite calcular el valor futuro de Y a partir de un conjunto de datos históricos, tomando en cuenta factores de variación cíclica, tendencias, estacionalidad, autocorrelaciones, etc.

De manera implícita se asume que el conocimiento del futuro está en el pasado, pues la historia se repite o bien, sigue un curso regular en el tiempo (fig. 1). A esta clase de técnicas se les conoce como métodos de extrapolación.

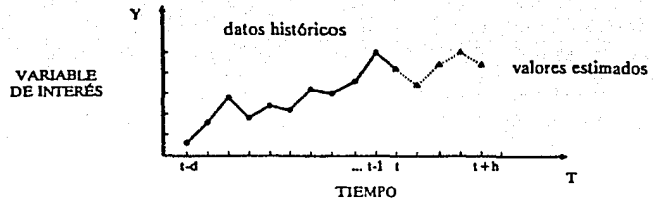


FIGURA 1 Pronóstico por medio de una serie de tiempo

En cambio, en el análisis de regresión se busca conocer el futuro de Y a partir de otras variables relacionadas X's, haciendo uso de un modelo que expresa dicha interdependencia funcional

$$Y = f(X_1, X_2, \ldots, X_k)$$

donde:

- Y es la variable por estimar, conocida como variable dependiente.
- X_i es la i-ésima variable relacionada, conocida como v. independiente, v. explicativa o regresor.
- f( ) es la función que indica la relación que guarda la Y con las X's.

Así, por ejemplo, podemos estimar la demanda futura de energía a partir de los elementos que configuran el consumo: (X_1) tamaño de la población; (X_2) producción industrial; (X_3) número de establecimientos comerciales, etc.

El papel de las técnicas de regresión consiste en identificar cuál es la función lineal que mejor representa a un conjunto de datos, así como estimar el valor de los parámetros asociados. De hecho es este el objetivo del presente trabajo: con la ayuda de estas técnicas y el uso de paquetes computacionales (básicamente lotus y statgraphics) realizar predicciones de producción de productos agrícolas denominados alimentos básicos (maíz, frijol, trigo y arroz).

El foco de atención lo tendremos en la explicación de la producción anual de los granos básicos en función de parámetros como: lluvias, créditos, fertilizantes, hectáreas cosechadas, precios de garantía, etc.

Para lograr esto y con la idea de introducir los conceptos básicos, iniciaremos (capítulo 1) con la explicación de, qué es un problema de regresión lineal simple, cómo se estiman los parámetros, y posteriormente medir la bondad de ajuste. En el capítulo 2 desarrollaremos el modelo de regresión múltiple con sus correspondientes conceptos generalizados en el capítulo 1. En el capítulo 3 veremos brevemente las instrucciones elementales para el manejo del paquete statgraphics. Finalmente, el capítulo 5 será destinado a la parte aplicativa del trabajo, que es la selección de variables y la interpretación de resultados.

CAPITULO 1

INTRODUCCION AL ANALISIS DE REGRESION

1.1 ASPECTOS GENERALES

El análisis de regresión es una técnica estadística para investigar y modelar la relación funcional entre variables, siendo numerosas sus aplicaciones en el pronóstico.

La idea básica consiste en identificar cuál es la curva que mejor se ajusta a un conjunto de N pares de datos $\{X_i, Y_i\}$, y con ello establecer una ecuación que permita estimar el valor de Y dado que se conoce el valor de la variable independiente (X_0), lo cual se ilustra en la figura 1.1.

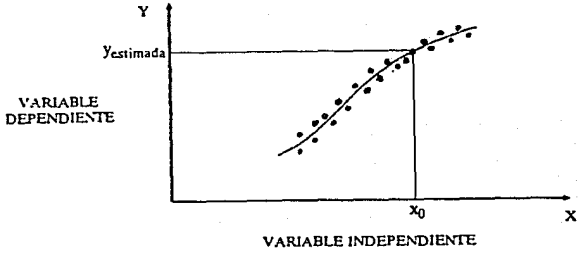


FIGURA 1.1 Estimación de Y dado que $X = X_0$

Por ejemplo, podemos estimar el consumo de gasolina dado un incremento en los precios, el consumo de agua per capita en función del tamaño de una población, los costos esperados según el nivel de producción, las ventas futuras de acuerdo al esfuerzo publicitario, fig (1.2), etc. Desde luego, suponiendo que contamos con la información necesaria para cada caso.

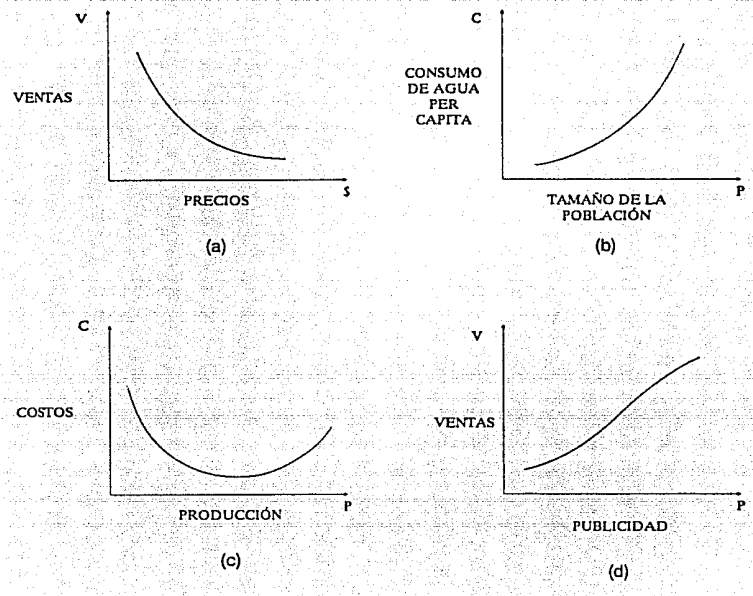


FIGURA 1.2 Ejemplos de aplicación del análisis de regresión

Los datos pueden provenir de observaciones de X y Y a lo largo del tiempo (anual, mensual, diario, etc.), sin que exista el requisito de que sean secuenciales; también pueden corresponder a observaciones sobre distintas unidades -equipos, poblaciones, regiones, etc.-, obtenidas o no en un mismo momento.

Cabe añadir que no solo buscamos una expresión matemática que diga de qué manera están relacionadas las variables, sino también definir con que precisión se puede hacer una predicción, lo que en buena medida depende del grado de dispersión de los datos y de la bondad de la curva seleccionada, correspondiendo al análisis de correlación establecer el grado de asociación y la calidad de ajuste.

1.2 DEPENDENCIA ESTADISTICA Y DEPENDENCIA FUNCIONAL

Se puede decir que el análisis de regresión se ocupa de lo que se conoce como dependencia estadística entre variables y no de la dependencia funcional o determinista. En la dependencia estadística entre variables se manejan esencialmente variables aleatorias o estocásticas, es decir, variables que tienen distribuciones de probabilidad; la dependencia funcional, en cambio, se ocupa de variables que no son aleatorias ni estocásticas.

La dependencia de una cosecha de la temperatura ambiente, las lluvias, el sol y los fertilizantes, por ejemplo, es de naturaleza estadístico en el sentido de que las variables explicatorias, aunque ciertamente son importantes, no permiten al decisor predecir el producto de la cosecha con seguridad por errores en la medición de tales variables o porque otros muchos factores (variables) afectan la cosecha, los cuales son difíciles de identificar individualmente. Por tanto, existe alguna variabilidad "intrínsica" o aleatoria en la variable dependiente, producto de la cosecha, que no puede ser completamente explicado, cualquiera que sea el número de variables explicatorias.

De otro lado, en fenómenos deterministas nos ocupamos de relaciones tales como las que se dan en la ley de gravitación de Newton, ley de los gases de Boyle, la ley de electriciadad de Kirchhoff, la ley del movimiento de Newton, etc.

1.3 REGRESION Y CAUSALIDAD

Aunque los modelos de regresión lineal se ocupa de la dependencia de una variable de otras, no implica necesariamente causalidad. En el ejemplo de la cosecha, citado previamente, no existe una razón para suponer que las lluvias no dependen estadísticamente del producto de la cosecha. El hecho de que tratemos a la cosecha como dependiente de las lluvias (entre otras causas) se debe a consideraciones no estadísticas. El sentido común nos sugiere que la relación no puede ser al contrario, pues no podemos controlar las lluvias mediante cambios en el producto de la cosecha.

El punto esencial es entonces que, una relación estadística per se no puede implicar logicamente a causalidad. Para aducir causalidad se debe apelar a consideraciones teóricas o apriorísticas.

1.4 ASOCIACION Y CAUSALIDAD

Entre las ventajas de los modelos elaborados con las técnicas de regresión (modelos explicativos o causales) es que contribuyen a un mejor entendimiento del problema bajo consideración, lo que a su vez permite explorar distintas alternativas y diseñar políticas de intervención.

En otros términos, dado que Y está en función de X, podemos buscar cómo actuar sobre X para producir un efecto deseado en Y; a diferencia de los métodos de extrapolación en los que se limita a predecir el valor de Y, para luego ver cómo adaptarnos a la nueva circunstancia.

Sin embargo, el identificar una función que ajusta bien a un conjunto de datos no es razón suficiente para inferir que un cambio en una variable va a influir en que la otra también cambie.

Para ilustrar lo anterior, pensemos en que hemos obtenido una muestra de la altura de N mujeres y la de sus respectivos esposos, cuyos resultados se presentan en la figura 1.3

Al observar la gráfica, se nota la existencia de una relación funcional y una alta correlación entre ambas variables, de suerte que al conocer la altura de la esposa podemos estimar con cierta precisión la altura del esposo. No obstante, a nadie se le ocurriría pensar que el esposo va a reflejar algún efecto si la esposa ingiere la vitamina de crecimiento.

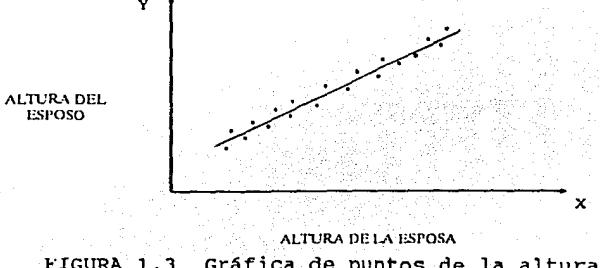


FIGURA 1.3 Gráfica de puntos de la altura de N parejas

En este caso, más que una relación de causa-efecto encontramos una relación de asociación, la cual es el resultado de la influencia de factores como las preferencias personales, patrones culturales, clases económicas, etc. (fig 1.4)

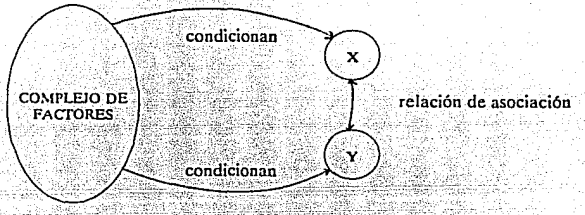


FIGURA 1.4 Relación de asociación como resultado de la influencia de un complejo de factores

Ejemplos como el anterior se encuentran con frecuencia en cualquier campo, basta que X (o un complejo de factores) sea causa de Y y Z, para que exista una relación de asociación entre las dos últimas variables (fig. 1.5)

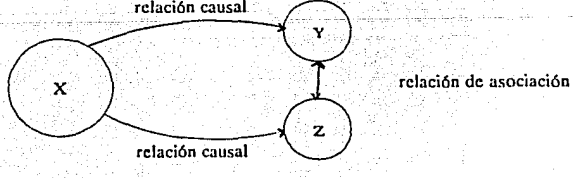


FIGURA 1.5 Relaciones de asociación y de causalidad

El porqué de una relación, está fuera del dominio de la estadística, el único medio que se dispone para hablar de relaciones causa-efecto es el conocimiento acerca del fenómeno y del mecanismo bajo el que opera, donde la regresión es sólo un instrumento de apoyo para el análisis de los datos.

Todo lo anterior no quiere decir que exista algún impedimento para utilizar las relaciones de asociación en el pronóstico, de hecho muchos trabajos son realizados sobre estas bases; generalmente porque se cuenta con mayor información de la variable asociada o porque es más fácil el acceso a ella, además de que se conocen sus tendencias.

CAPITULO 2

REGRESION LINEAL SIMPLE Y SUS EXTENSIONES

El término de regresión lineal simple es aplicable cuando tratamos con una sola variable independiente y la relación de ésta con la variable dependiente es lineal en los parámetros, por lo que queda un modelo de la forma $Y = a + b X$.

Este modelo es de interés no solo por los problemas que así pueden ser abordados, sino también porque, con ligeros cambios, es posible aplicar los mismos resultados en otros casos donde la relación no es lineal, pero si intrínsicamente lineal.

2.1 ESTIMACION DE LOS COEFICIENTES DE REGRESION

El punto de partida de la regresión es un conjunto de N pares de datos (X_i, Y_i), los cuales pueden provenir de registros históricos, observaciones de campo, experimentos, etc. Datos que representaremos en un diagrama de puntos, llamado también diagrama de dispersión (fig. 2.1)

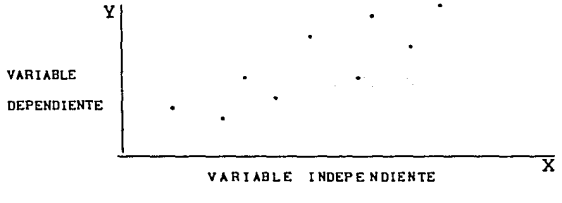


FIGURA 2.1 Diagrama de puntos o de dispersión

Con base en el conocimiento del fenómeno y con el apoyo del diagrama de puntos, se propondrá el tipo de función que mejor represente la

relación entre ambas variables, que por ahora supondremos de carácter lineal

$$Y = a + b X$$

con lo que queda una ecuación que representa una familia de rectas (fig 2.2); el paso siguiente consiste en fijar el valor de los coeficientes de regresión, la constante "a" y la pendiente "b", para identificar la recta más apropiada.

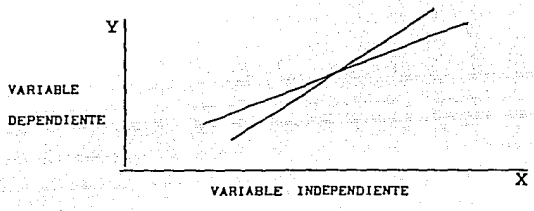


Figura 2.2 Diagrama de puntos y líneas de regresión posibles

El criterio comúnmente empleado para este fin es el de mínimos cuadrados, que busca minimizar la suma de los cuadrados de las desviaciones entre el valor real de Y y el valor estimado haciendo uso de la recta de regresión (fig. 2.3)

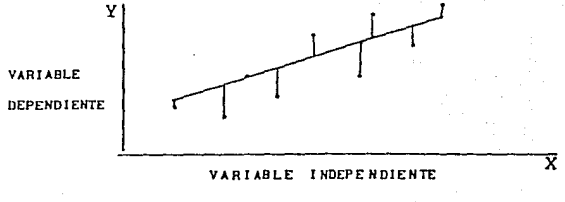


Figura 2.3 Desviaciones verticales respecto a la línea de regresión

Este criterio no es el único posible, también podría pensarse en minimizar la suma del valor absoluto de las desviaciones, calcular alguna desviación perpendicular a la recta, seleccionar la recta que pase por el mayor número de puntos, ajustar a simple ojo, etc.

Sin embargo, el método de los mínimos cuadrados ofrece ventajas como las siguientes: los cálculos son fáciles, intuitivamente los resultados son buenos y la técnica está muy difundida; aunque lo más importante es que, bajo ciertas condiciones, los coeficientes de regresión tienen tales propiedades estadísticas que permiten trazar intervalos de confianza y hacer distintas pruebas para ver la significancia de los resultados.

A continuación describiremos el procedimiento para calcular a y b :

Dado el conjunto de observaciones (X_1,Y_1), $(X_2,Y_2),\ldots,(X_N,Y_N)$, el valor real de Y para cada caso es Y_i, mientras que su valor estimado es $\hat{Y}_i$, que resulta de sustituir X_i en la ecuación de regresión

$$\hat{Y} = a + bX_i \qquad (2.1)$$

así, la Y real puede ser planteada como la suma de la Y estimada más la desviación vertical, llamada error o residual

$$\begin{aligned} Y_i &= \hat{Y} + \epsilon_i \\ &= a + bX_i + \epsilon_i \end{aligned} \qquad (2.2)$$

Conforme al criterio adoptado, debemos minimizar la suma de los cuadrados de los errores

$$\begin{aligned} \epsilon_i &= Y_i - \hat{Y}_i \\ &= Y_i - a - b\,X_i \end{aligned}$$

$$\min S = \sum \epsilon_i^2 = \sum (Y_i - a - bX_i)^2$$

Para ello obtenemos la derivada parcial de S con respecto a "a " y "b", e igualamos a cero

$$\frac{\delta S}{\delta a} = -2\sum (Y_i - a - bX_i) = 0$$

$$\frac{\delta S}{\delta b} = -2\sum X_i(Y_i - a - bX_i) = 0$$

al simplificar llegamos a un sistema de ecuaciones

$$Na + b \sum X_i = \sum Y_i$$

$$a \sum X_i + b \sum X_i^2 = \sum X_i Y_i$$

cuya solución para "a" y "b" es la siguiente

$$a = (\sum Y_i/N) - (b \sum X_i/N)$$
$$= \overline{Y} - b\overline{X} \qquad (2.3)$$

$$b = \frac{N \sum X_i Y_i - \sum X_i \sum Y_i}{N \sum X_i^2 - (\sum X_i)^2} \qquad (2.4)$$

Una expresión alternativa de la ecuación de regresión es

$$\hat{Y}_i = \overline{Y} - b\,(X_i - \overline{X}) \qquad (2.5)$$

que resulta de sustituir la ec. (2.3) en (2.2)

Una vez que tenemos las ecuaciones anteriores definiremos algunas equivalencias que nos serán de utilidad posteriormente:
Si definimos S_{XY} como

$$S_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y}) \qquad (2.6a)$$

se puede ver que las siguientes expresiones son equivalentes

$$S_{XY} = \sum (X_i - \overline{X}) Y_i \qquad (2.6b)$$
$$= \sum X_i (Y_i - \overline{Y}) \qquad (2.6c)$$
$$= \sum X_i Y_i - (\sum X_i)(\sum Y_i)/N \qquad (2.6d)$$
$$= \sum X_i Y_i - N\overline{X}\,\overline{Y} \qquad (2.6e)$$
$$= [\sum X_i Y_i - (\sum X_i)(\sum Y_i)]/N \qquad (2.6f)$$

$$S_{XX} = \sum (X_i - \bar{X})^2 \qquad (2.7a)$$
$$= \sum (X_i - \bar{X}) X_i \qquad (2.7b)$$
$$= \sum X_i^2 - (X_i)^2/N \qquad (2.7c)$$
$$= \sum X_i^2 - N \bar{X}^2 \qquad (2.7d)$$
$$= [N \sum X_i^2 - (\sum X_i)^2]/N \qquad (2.7e)$$

De manera similar que 2.6, tenemos

$$S_{YY} = \sum (Y_i - \bar{Y})^2 \qquad (2.8a)$$
$$= \sum (Y_i - \bar{Y}) Y_i \qquad (2.8b)$$
$$= \sum Y_i^2 - (Y_i)^2/N \qquad (2.8c)$$
$$= \sum Y_i^2 - N \bar{Y}^2 \qquad (2.8d)$$
$$= [N \sum Y_i^2 - (\sum Y_i)^2]/N \qquad (2.8e)$$

Al sustituir (2.6f) y (2.6e) en (2.4) obtenemos una fórmula más sencilla para b

$$b = \frac{S_{XY}}{S_{XX}} \qquad (2.9)$$

EJEMPLO 2.1 Consideremos un conjunto de N pares de datos (C_i, P_i) con $i = 1, 2, \ldots, 20$, donde C_i es el consumo de gas natural y P_i es el precio. En el apéndice C se muestran estos datos, que corresponden a 20 ciudades de Texas. Determinemos la relación entre demanda y precio para tomar decisiones sobre su impacto en el futuro.

En la figura 2.4 muestra que una línea recta no es el mejor ajuste de los datos. Sin embargo, para propósitos ilustrativos, usaremos el modelo lineal, aunque posteriormente propondremos una mejor aproximación.

Entonces, la ecuación de la regresión lineal (tabla I) es

$$\hat{C} = 138.561 - 1.10414\ P \qquad (2.10)$$

donde $\hat{C}$ es el consumo estimado

FIGURA 2.4 Diagrama de Consumo contra Precio

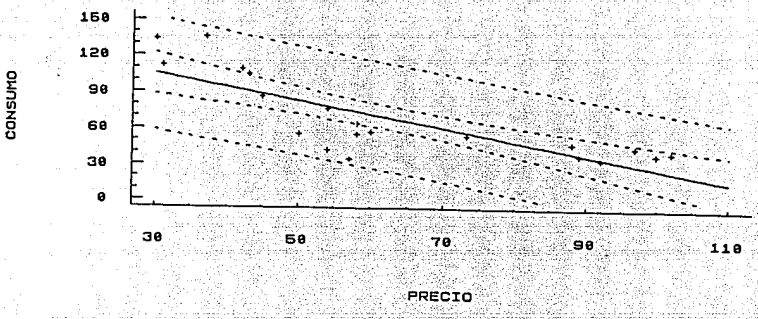


TABLA I Modelo Lineal: Y = a+bX

Variable Dependiente: CONSUMO Variable Independiente: PRECIO

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	138.561	13.5515	10.2247	.00000
Slope	-1.10414	0.201961	-5.46712	.00003

2.2 COEFICIENTE DE CORRELACION

2.2.1 DEFINICION Y CALCULO DEL COEFICIENTE DE CORRELACION

El coeficiente de correlación, r_{xy}, juega un papel muy importante, ya que nos brinda una medida del grado de asociación lineal de las variables estudiadas, esto es, qué tanto se alejan o se acercan los datos a la línea de regresión, lo que nos indica con qué precisión puede ser estimada Y dado un valor de X.

Este coeficiente se define como sigue:

$$r_{XY} = \frac{\hat{cov}\ XY}{S_X\ S_Y} \qquad (2.11a)$$

donde:

cov XY es el estimador de la covarianza de X y Y, dado por

$$\hat{cov}\ XY = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{N}$$

Sx, Sy son los estimadores de la desviación estándar de X y Y, respectivamente

$$S_X = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N}} \quad ; \quad S_Y = \sqrt{\frac{\sum (X_i - \bar{X})^2}{N}}$$

Por tanto

$$r_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{N\ S_X\ S_Y} \qquad (2.11b)$$

$$= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{[\sum(X_i-\bar{X})^2]^{1/2}[\ \sum(Y_i-\bar{Y})^2]^{1/2}} \qquad (2.11c)$$

$$= \frac{N\sum X_iY_i - \sum X_i \sum Y_i}{[N\sum X_i^2-(\sum X_i)^2]^{1/2}[\ N\sum Y_i^2-(\sum Y_i)^2]^{1/2}} \qquad (2.11d)$$

siendo esta última fórmula la más apropiada para los cálculos.

Uno de los aspectos que debemos tener en cuenta es que el coeficiente de correlación es un indicador adimensional, por lo cual su valor no cambia al hacer una transformación lineal en X y/o en Y, esto es, si

$$U_i = cX_i + f,$$
$$V_i = dY_i + g$$

donde c y d son constantes mayores que cero, y f y g son cualquier constante, entonces

$$r_{UV} = r_{XY} = r_{UY} = r_{XV}$$

Por tal motivo, en general no afecta la unidad de medida que se adopte y podemos trabajar por igual con toneladas o kilogramos, meses o días, pesos o dólares, etc., así como modificar el punto en que iniciamos el conteo. Para constatar lo antes dicho, basta con hacer algunas operaciones:

Sabemos que

$$r_{UV} = \frac{\sum(U_i-\overline{U})(V_i-\overline{V})}{((\sum(U_i-\overline{U})^2)^{1/2}((\sum(V_i-\overline{V})^2)^{1/2}}$$

ahora bien, como

$$U_i = cX_i + f$$
$$V_i = dY_i + g$$

entonces

$$\overline{U} = \sum(cX_i + f)/N = c\overline{X} + f$$
$$\overline{V} = \sum(d\overline{Y}_i + g)/N = d\overline{Y} + g$$
$$(U_i - \overline{U}) = c(X_i - \overline{X})$$
$$(V_i - \overline{V}) = d(Y_i - \overline{Y})$$

sustituyendo estos resultados en r_{UV} obtenemos

$$r_{UV} = \frac{\sum c(U_i-U)\ d(V_i-V)}{(\sum c^2(X_i-\overline{X})^2)^{1/2}(\sum d^2(Y_i-\overline{Y})^2)^{1/2}}$$

$$= \frac{\sum (X_i-\overline{X})\quad (Y_i-\overline{Y})}{(\sum (X_i-\overline{X})^2)^{1/2}(\sum (Y_i-\overline{Y})^2)^{1/2}} = r_{XY}$$

que es lo que queríamos probar.

Lo anterior es de interés porque nos permite trabajar por igual, ya que no afecta la unidad de medida que se adopte.

Hay que mencionar, también, que los valores del coeficiente de correlación van de -1 a +1 pasando por el cero. Adquiere el valor de ± 1 cuando todos los puntos caen sobre una línea recta (figs. 2.5a y 2.5b) y su valor absoluto decrece conforme tal liga es menos estrecha (figs. 2.5c

y 2.5d), hasta llegar a cero cuando no existe ninguna relación, es decir, cuando las variables son independientes, (fig. 2.5e)

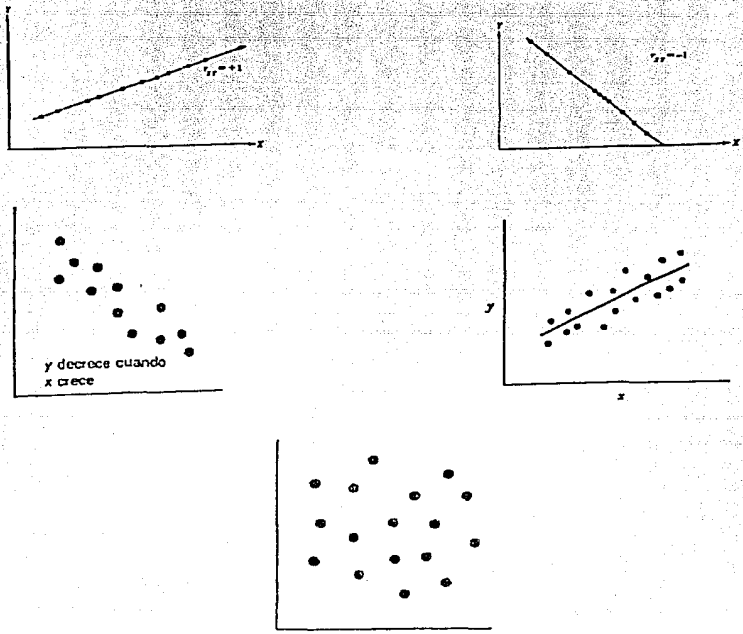


FIGURA 2.5 Posibles Gráficas de Datos Ordenados en Parejas

De esta manera, cuando r_{XY} vale -1 ó +1, basta conocer X para calcular Y con toda precisión, en tanto que si su valor es de cero, de nada servirá el conocimiento que se tenga sobre X para estimar Y.

Analíticamente podemos ver lo anterior en los siguientes términos:

Si la relación es perfecta, el valor real de Y y su valor estimado coinciden

$$Y_i = \hat{Y}_i$$

así de acuerdo a la ec. (2.5) tenemos que

$$Y_i = \bar{Y} - \beta_1 (X_i - \bar{X})$$

de aquí que

$$Y_i - \bar{Y} = \beta_1 (X_i - \bar{X})$$

sustituyendo este resultado en la ec.(2.8c), obtenemos

$$r_{XY} = \frac{\sum (X_i - \bar{X}) b (X_i - \bar{X})}{\left[\sum (X_i - \bar{X})^2 \right]^{1/2} \left[b^2 \sum (Y_i - \bar{Y})^2 \right]^{1/2}}$$

$$= \frac{b}{|b|}$$

$$= +1 \text{ cuando la pendiente es positiva}$$

$$= -1 \text{ cuando la pendiente es negativa}$$

Por el contrario, si no hay relación entre ambas variables la covarianza es nula

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i - \bar{X}) \sum (Y_i - \bar{Y}) = 0$$

y, por tanto

$$r_{XY} = 0$$

Conviene hacer énfasis en que el coeficiente de correlación sólo es aplicable cuando la relación entre la variables estudiadas es lineal, en cualquier otro caso no tiene ningún sentido. Por ejemplo en la figura 2.6 representamos dos casos en los que la relación funcional es perfecta, no obstante, como no es lineal, r_{XY} toma valores no cercanos a la unidad.

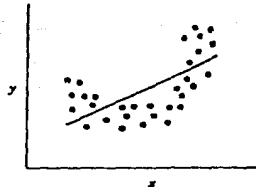


FIGURA 2.6 Relación no Lineal

En consecuencia, un valor bajo de r_{XY} (cercano a 0) no es razón suficiente para decir que no existe algún tipo de relación entre las variables estudiadas, como tampoco un valor alto (cercano a ±1) garantiza que este sea el mejor ajuste.

Lo anterior deja claro qué significa y cuáles son las implicaciones de que r_{XY} adopte un valor extremo, pero queda duda de cómo debemos interpretar un valor intermedio. Así, para el ejemplo 2.1, en la figura 2.4 podemos constatar que existe una relación negativa y que esta no es perfecta, de ahí que $r_{XY} = -0.79$; sin embargo, carecemos de los elementos de juicio necesarios para decidir si esto es bueno o malo, por lo que podemos caer en argumentos vacíos de contenido como el siguiente:

¡ la correlación es buena porque está arriba de la mitad !

sobre lo que cabe preguntar ¿la mitad de qué?, ¿en qué contribuye X para estimar Y?, ¿a partir de qué valor o en qué rango es bueno el resultado?...

Aún más, nuestras conclusiones difícilmente cambiarían si en lugar de -0.79 obtenemos 0.6, 0.8 ó 0.5, como tampoco sabríamos qué decidir si esos valores fueran más bajos.

Esta situación es fruto de la manera en que está construido este indicador, donde la división de una covarianza entre dos desviaciones estándar no se presta para una interpretación intuitiva; pese a ello, en la práctica es común que se haga uso de él para hablar de buenos o malos ajustes, apoyando o desapoyando hipótesis de trabajo.

Por último, haciendo uso de la notación S_{XY}, S_{XX}, S_{YY}, tenemos otras formas para expresar r_{XY}, que son las siguientes:

$$r_{XY} = \frac{S_{XY}}{N\, S_X\, S_Y} \qquad (2.13a)$$

$$= \frac{S_{XY}}{S_{XX}^{1/2} S_{YY}^{1/2}} \qquad (2.13b)$$

la primera la obtenemos al sustituir la ec. (2.6a) en (2.11b), y la

segunda al sustituir (2.6a), (2.7e) y (2.8e) en (2.11c).

Además, al comparar la ec (2.12b) con la ec (2.9), vemos que existe una fuerte similitud entre b y el coeficiente de correlación, resultando las siguientes equivalencias:

$$r_{XY} = b \frac{Sx^{1/2}}{Sy^{1/2}}$$

$$= b \frac{Sx}{Sy}$$

aunque cabe advertir que su significado es distinto, pues mientras r_{XY} mide el grado de asociación lineal entre ambas variables, b establece cuanto cambia Y dado un incremento de X en una unidad.

2.3 COEFICIENTE DE DETERMINACION

Como hemos comentado, el coeficiente de correlación nos brinda una medida acerca de que tan bien se ajusta una línea de regresión a los datos; el defecto es que ésta medida no resulta del todo clara, por ello vale la pena explorar más este tema a fin de lograr un mejor entendimiento.

En ese sentido, se ocurre que debe existir una estrecha vinculación entre la bondad de ajuste y algún indicador relacionado con los errores de estimación. Para estudiar este punto recordemos que la Y real es planteada como la suma de la Y estimada más el error o residual

$$Y_i = \hat{Y}_i + \epsilon_i$$

fórmula que representaremos como sigue (ver fig 2.7)

$Y_i - \overline{Y}$	=	$(Y_i - \hat{Y})$	+	$(\hat{Y}_i - \overline{Y})$
desviación total		desviación explicada por la regresión		desviación no explicada o error

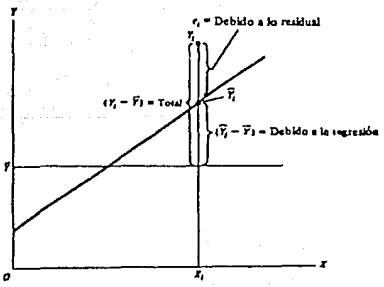


FIGURA 2.7 Partición de los Componentes de Variación de Y_i

Con base en lo anterior, intuitivamente podemos construir un indicador que hable más claro de la bondad de ajuste, al cual llamaremos coeficiente de determinación, denotado por R^2_{XY}, y que plantearemos como sigue:

$$R^2_{XY} = \frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2} = \frac{SCR}{SCT} \qquad (2.14a)$$

o alternativamente como

$$R^2_{XY} = 1 - \frac{\sum(Y_i-\hat{Y})^2}{\sum(Y_i-\bar{Y})^2} = 1 - \frac{SCE}{SCT} \qquad (2.14b)$$

donde :

SCT es la suma de cuadrados total, que es una medida de la dispersión de los datos respecto a su media

SCR llamada la suma de cuadrados debido a la regresión, que indica la parte de la variación total que es explicada con el apoyo de la línea de regresión;

SCE es la suma del cuadrado de los errores.

Así, el coeficiente de determinación nos da una medida acerca de la contribución de X para estimar Y, basada en la porción de la variación total que es explicada mediante la regresión. Sus propiedades más importantes son:

1) Es una cantidad no negativa.

2) Sus límites son $0 \leq R^2_{XY} \leq 1$. Un R^2_{XY} de 1 quiere decir ajuste perfecto, mientras que un R^2_{XY} de 0 indica que no hay relación entre la variable dependiente y las variables explicatorias.

De la tabla I tenemos que R^2_{XY} = 62.4, . Recordemos (ver sección 2.1) que $r_{CC\hat{}} = r_{PC} = -0.79$. De aquí que $r^2_{CC\hat{}} = r^2_{PC} = 0.624$ (que es equivalente a 62.4 %). Por tanto, $R^2_{XY} = r^2$, es decir, los dos nos dan una medida de ajuste lineal. . En general, el coeficiente de determinación, para un modelo de regresión lineal, es númericamente igual al coeficiente de correlación al cuadrado.

2.4 PRUEBA DE HIPOTESIS

La motivación principal de la inferencia estadística es ¿ qué tanto podemos decir acerca de la población, sí tan solo tenemos evidencias de muestras?

El problema que nos ocupa es, una vez estimada "a" y "b" de la regresión lineal Y = a + bX, ¿ cómo asegurar que ésta es la mejor representación?, es decir, ¿ésta estimación no depende de la muestra que hemos tomado? ¿ b es significativamente distinto de cero?. Como a y b son solo estimaciones basados en datos muestrales, implica la existencia de valores verdaderos correspondientes, denotados por α y β. Por tanto, la línea de regresión verdadero es $Y = \alpha + \beta X + \epsilon$. Ahora la pregunta es entonces ¿ qué tan buenos son a y b respecto a α y β?

Para responder a estas preguntas se requiere realizar un proceso que se denomina prueba de hipótesis. Sin embargo, para tal tarea es necesario el desarrollo de los siguientes puntos: los supuestos del análisis de regresión, error estándar de estimación y nivel de significancia.

2.4.1 SUPUESTOS DE LA REGRESION LINEAL

En el análisis de regresión lineal se supone que las X son constantes, no valores de variables aleatorias, y que para cada valor de X la variable que se va a predecir, Y, tiene distribución normal, que además cumple con lo siguiente

* SUPUESTO 1

$E(Y_i|X_i) = \alpha + \beta X$

Geométricamente este supuesto queda representado en la fig (2.8),

donde para algunos valores de X se presenta la población Y asociada con cada una de éstas X's. Puede verse que la población correspondiente a un X dado, está distribuido conforme a una normal.

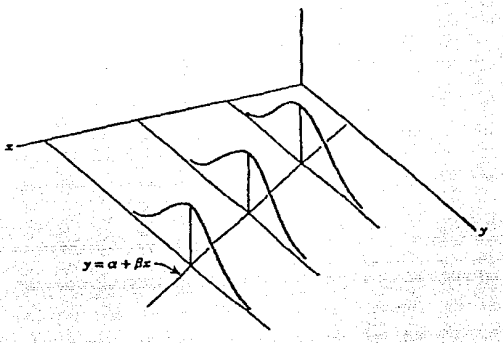


FIGURA 2.8 Ilustración de los supuestos 1 y 2

* SUPUESTO 2

$Var(Y_i|X_j) = \sigma^2$ para toda $i \neq j$

Esta condición dice que las poblaciones que corresponden a distintos valores de X tienen la misma varianza. Este supuesto se le conoce comúnmente por homoscedasticidad. fig.(2.8)

* SUPUESTO 3

$Cov(Y_i, Y_j) = 0$ para toda $i \neq j$.

Lo que dice ésta condición es que para cada muestra distinta, los valores ajustados correspondientes son independientes.

2.4.2 ERROR ESTANDAR DE ESTIMACION

Con base en las suposiciones anteriores, los coeficientes estimados a

y b, son valores de variables aleatorias que tienen distribuciones normales con media α y β, y desviaciones estándar

$$S_a = \sigma\left[\frac{1}{n} + \frac{\bar{X}^2}{Sxx}\right]^{1/2} \qquad y,$$

$$S_b = \frac{\sigma}{[Sxx]^{1/2}}$$

Estas fórmulas de error estándar requieren que se estime σ. La estimación de σ se denomina error estándar de la estimación y se representa por Se. Su expresión queda:

$$Se = \left[\frac{\sum (Y-\hat{Y})^2}{n-2}\right]^{1/2}$$

El error estándar de la estimación mide la variabilidad no explicada en la variable dependiente.

Sea MSE = $[Se]^2$. A MSE se le denomina error cuadrado medio. Mientras menor sea el valor de MSE, mejor será el pronóstico.

2.4.3 NIVEL DE SIGNIFICANCIA

Sabemos que en general, el coeficiente de regresión β de la variable independiente de una población es distinto del coeficiente b de una muestra, es decir, $\beta \neq b$. figura (2.9).

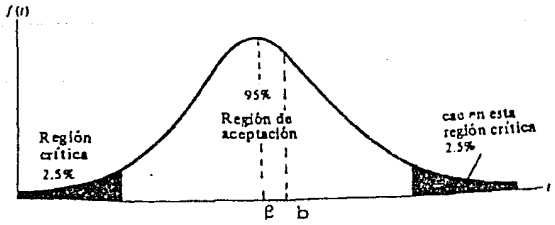


FIGURA 2.9 Interpretación Geométrica del Nivel de Significancia

La pregunta es ¿ qué tanta tolerancia le permitimos a b de que se aleje de β?, es decir, ¿ qué porcentaje de error de tipo 1 (aceptar b cuando en realidad no es representativo) estamos dispuestos a cometer?. A ese porcentaje suele denominarse nivel de significancia, denotado por α (no confundir con el coeficiente de regresión). Por lo general, las pruebas se realizan en un nivel de significación de 0.05 o 0.01; probar una hipótesis en un nivel de significación, de por ejemplo $\alpha = 0.05$, significa que fijamos la probabilidad de rechazar la hipótesis si ésta es verdadera en 0.05.

Retomando el problema de la prueba de hipótesis estadístico, esta puede expresarse en los siguientes términos: ¿es un hecho u observación compatible con alguna hipótesis, o no lo es?. La palabra "compatible" en este contexto quiere decir "suficientemente" cerca del valor hipotético, de suerte que nos lleve a aceptar la hipótesis que queremos analizar.

De este modo, si una teoría o una experiencia a priori nos lleva a creer que el verdadero coeficiente "β " de la pendiente en el ejemplo 2.1 es menos dos, se preguntará si el b observado igual a -1.104, obtenido a partir de los datos es consistente con la hipótesis en juicio. Si es así, se puede aceptar la hipótesis, si no, se rechaza.

En el lenguaje de la estadística, la hipótesis propuesta se conoce como hipótesis nula y se designa con el símbolo H_0. La hipótesis nula suele probarse contra la hipótesis alterna que se designa con H_1, la cual expresa por ejemplo, que β es diferente de menos dos.

La teoría de pruebas de hipótesis se preocupa por diseñar reglas o procedimientos que permitan decidir cuando aceptar o rechazar la hipótesis nula. Existen dos enfoques complementarios para lograr esas reglas, el del intervalo de confianza y el de la prueba de significancia. Ambos enfoques pretenden que la variable (estadístico o del estimador) que se considera tiene una distribución de probabilidad y que las pruebas de hipótesis encierran afirmaciones sobre los valores de los parámetros de dicha distribución.

2.4.4 PRUEBAS DE SIGNIFICANCIA

* PRUEBA t

Suponga que las hipótesis apropiadas son:

$$H_0 : b = 0$$
$$H_1 : b \neq 0 \qquad (2.15)$$

Ahora considere la función de distribución

$$t_0 = \frac{b}{(MSE\ /Sxx)^{1/2}}$$

t_0 está distribuido como t con n-2 grados de libertad si la hipótesis nula es verdadera. Los grados de libertad sobre t_0 es el número de grados de libertad asociados con MSE; t_0 se usa en la prueba H_0: b = 0 para comparar los valores observados de t_0 con la cota puntual porcentual $\alpha/2$ de la distribución t_{n-2} ($t_{\alpha/2,n-2}$) y rechazando la hipótesis nula si

$$|t_0| > t_{\alpha/2,n-2}$$

Si aceptamos H_0: b = 0 implica que no existe relación lineal entre X y Y, es decir, que X no implica la variación en Y y que el mejor estimador de Y para cualquier X es $\hat{Y} = \overline{Y}$, figura (2.10a), o que la relación entre X y Y es no lineal, figura (2.10b)

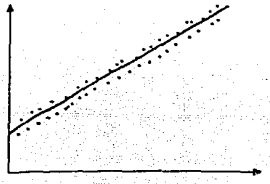
Del ejemplo 2.1 , tenemos que $t_0 = -5.52$ (tabla I). Por otro lado, $t_{.025,18} = 2.101$ (apéndice A). Ya que $|-5.52| > 2.101$, entonces la pendiente de la regresión lineal que relaciona el consumo y el precio del gas es altamente significativo (significativamente distinto de cero). Por tanto la variable P permanece en el modelo propuesto.

FIGURA 2.10 Situaciones donde la Hipótesis H_0: b = 0 no se rechaza

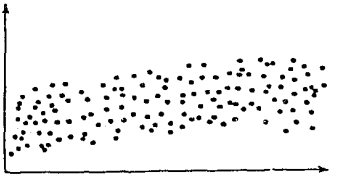
* LA PRUEBA F

El modelo de regresión simple $Y = a + b X + \epsilon$ tiene un coeficiente b de la pendiente. Si esta pendiente fuera cero, el modelo sería una línea $Y = a + \epsilon$. En otras palabras, conociendo los valores de X podría no tener consecuencia en Y. De igual manera, si en el modelo estimado b = 0.75, por ejemplo, es posible que el error ϵ fuera lo suficientemente grande para obscurecer la relación entre Y y X. Precisamente es esto lo que mide la prueba F: la posible relación significativa que pueden guardar las variables Y y X.

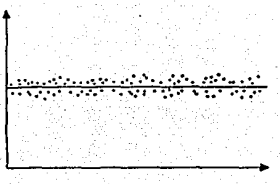
De la gráfica (2.11a) se observa que conforme transcurre el tiempo, las ventas aumentan, es decir, existe una relación estrecha entre el tiempo y las ventas. Caso contrario a lo que refleja la gráfica (2.11b), en el que el número de unidades producidas no explica el costo de producción por unidad de algún producto. La figura (2.11c) muestra la tendencia del consumo en el tiempo, pero la variación es sustancial.



a) Parámetro Significativo



b) Parámetro no Significativo



c) Parámetro bajo Incertidumbre

FIGURA 2.11

El estadístico Fo se define como

$$Fo = \frac{\sum(\hat{Y}-\overline{Y})^2/k-1}{\sum(Y-\hat{Y})^2/n-k}$$

donde k es el número de parámetros (coeficientes en la regresión).

Como el numerador y el denominador son variables aleatorias independientes, entonces si la hipótesis nula Ho : b = 0 es verdadero, entonces la prueba estadística Fo se distribuye como una $F_{1,n-2}$. Si F es grande entonces es muy probable que la pendiente b sea distinto de cero. Por consiguiente, Ho se rechaza si $Fo > F\alpha,_{1,n-2}$.

La Fo está estrechamente vinculada con la definición del coeficiente de determinación, de la siguiente manera:

Sabemos que $R^2_{XY} = \frac{SSR}{SST} = \frac{SSR}{SSE + SSR}$

Por consiguiente

$$\frac{SSR}{SSE} = \frac{R^2_{XY}}{1-R^2_{XY}}$$

de aquí que

$$Fo = \frac{SSR/k-1}{SSE/n-k} = \frac{R^2_{XY}/k-1}{1-R^2_{XY}/k-2}$$

Para los datos del consumo de gas natural, Fo = 29.89 (tabla II). Ahora viendo la tabla del apéndice B, $F_{0.01,1,18} = 8.28$. Como 29.89 > 8.28, entonces b es significativamente distinto de cero. En otras palabras, hay una relación lineal significativa entre el consumo y el precio del gas.

TABLA II Analisis de Varianza

Source	Sum of Squares	Df	Mean Square	F-Ratio	Prob. Level
Model	13005.703	1	13005.703	29.89	.00003
Residual	7832.2969	18	435.1276		
Lack-of-fit	7213.7969	16	450.8623	1.458	.48216
Pure error	618.50000	2	309.25000		
Total (Corr.)	20838.000	19			

Correlation Coefficient = -0.790021 R-squared = 62.41 percent
Stnd. Error of Est. = 20.8597

2.5 MODELOS INTRINSICAMENTE LINEALES

Todo lo expuesto anteriormente es válido solo cuando el modelo propuesto es lineal en los parámetros, entonces ¿ qué hacer en caso de que una línea recta no sea adecuada para el ajuste de datos?. En este apartado nos ocuparemos precisamente de este punto.

La no linealidad se puede detectar a través de la prueba del diagrama de puntos y de los residuales. En otras ocasiones, la experiencia a priori o algunas consideraciones téoricas puede indicar que la relación entre X y Y no es lineal. En algunos casos una función no lineal se puede expresar como una línea recta usando transformaciones adecuadas. Por ejemplo, la función exponencial

$$Y = ae^{bX}\epsilon$$

puede ser transformada en una línea recta por una transformación logarítmica

$$\text{Ln } Y = \text{Ln } a + bX + \text{Ln } \epsilon$$

o

$$Y' = a' + b'X + \epsilon'$$

A estos modelos no lineales se les denomina intrínsecamente lineales.

Vale la pena comentar que este tipo de transformaciones requiere que los errores trasformados $\epsilon' = \text{Ln } \epsilon$ sean $NID(o,\sigma^2)$. Esto implica que el error multiplicativo ϵ en el modelo original está distribuido en forma Ln-normal. Además, los estimadores a y b de las transformaciones anteriores tienen las propiedades de los mínimos cuadrados con respecto a los datos trasformados, no de los datos originales.

Mostraremos varios modelos intrínsicamente lineales (figura 2.12) con sus correspondientes transformaciones y la forma lineal resultante (Tabla III). Cuando el diagrama de puntos Y contra X indica curvatura (en cualquiera de las formas mostradas) entonces realizamos la transformación correspondiente.

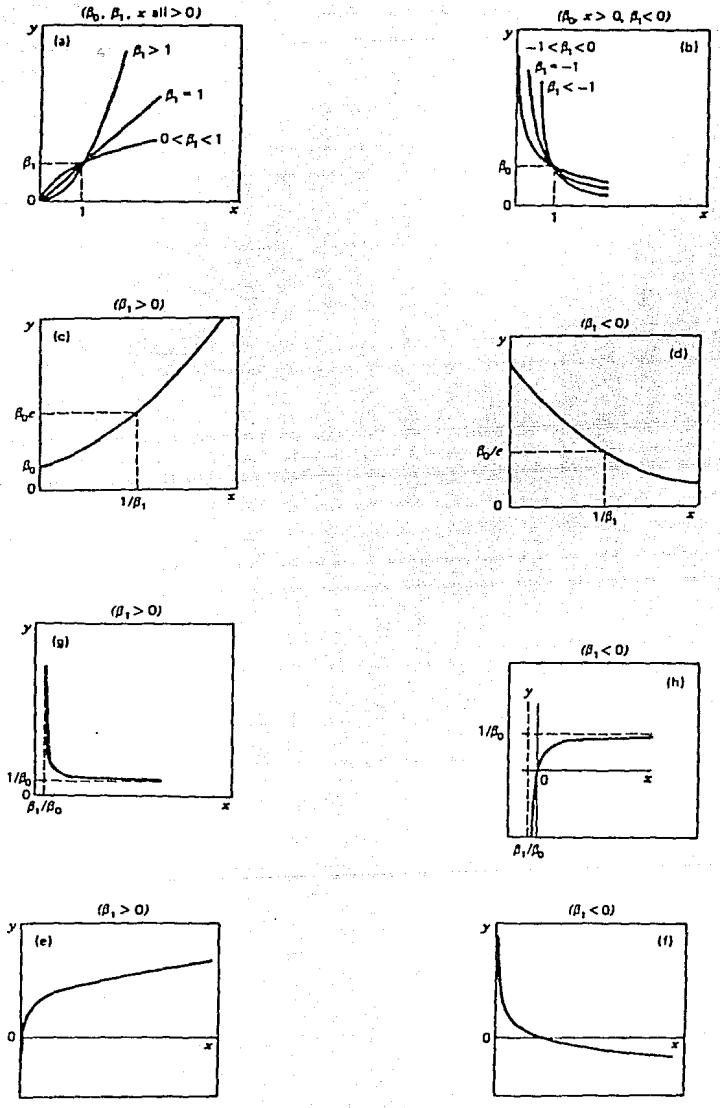


FIGURA 2.12 Funciones Linealizables

FIGURA	FUNCION LINEALIZABLE	TRANSFORMACION	FORMA LINEAL
2.10a,b	$Y = \beta_0 X^{\beta_1}$	$Y' = LnY,\ X' = LnX$	$Y' = Ln\beta_0 + \beta_1 X'$
2.10c,d	$Y = \beta_0 e^{\beta_1 X}$	$Y' = Ln\ Y$	$Y' = Ln\beta_0 + \beta_1 X$
2.10e,f	$Y = \beta_0 + \beta_1 LnX$	$X' = Ln\ X$	$Y' = \beta_0 + \beta_1 X'$
2.10g,h	$Y = \frac{X}{\beta_0 X - \beta_1}$	$Y' = 1/Y,\ X' = 1/X$	$Y' = \beta_0 - \beta_1 X'$

TABLA III. Funciones linealizables y su forma Lineal correspondiente

Usando las transformaciones tales como las anteriores, es posible convertir modelos no lineales a lineales. Los parámetros pueden ser estimados por el modelo lineal y la función transformada se puede expresar en términos de los parámetros del modelo no lineal original.

Transformando los valores de consumo y precio a Ln consumo y Ln precio del ejemplo 2.1, y aplicando el modelo lineal, tenemos

$$Ln\ C = Ln\ a + bLn\ X \qquad (2.16)$$

Este modelo está sugerido por la figura (2.12b) y su correspondiente forma lineal mostrado en la tabla III. Como vamos a comparar los modelos (2.16) y (2.10), repetiremos ésta última.

$$C = a + bP \qquad (2.10)$$

Observando la tabla IV, tenemos que la ecuación (2.16) queda

$$C' = 8.2 - 0.99P'$$

TABLA IV Modelo de Ln Consumo contra Ln PRECIO

Variable Dependiente: LNCONSUMO Variable Independiente: LNPRECIO

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	8.19736	0.577777	14.1878	.00000
Slope	-0.997272	0.141217	-7.06198	.00000

Analisis de Varianza

Source	Sum of Squares	Df	Mean Square	F-Ratio	Prob. Level
Model	2.804885	1	2.804885	49.87151	.00000
Residual	1.012360	18	.056242		
Lack-of-fit	.831536	16	.051971	.57483	.79281
Pure error	.1808237	2	.0904118		
Total (Corr.)	3.817245	19			

Correlation Coefficient = -0.857201 R-squared = 73.48 percent
Stnd. Error of Est. = 0.237154

Para comparar dos modelos por medio de los coeficientes de determinación, ya sea el ajustado o no, la variable dependiente debe ser la misma, mientras que las variables explicatorias pueden tomar cualquier forma. Entonces los modelos (2.16) y (2.10) no son comparables con los R^2_{XY} calculados, por la siguiente razón: por definición R^2_{XY} mide la proporción de variación en la variable dependiente debido a las variables explicatorias. Por tanto, la ecuación (2.16) mide la proporción de la variación en Ln C explicada por P, mientras que la ecuación (2.10) R^2_{XY} mide la proporción de la variación en C y las dos no son iguales.

Para comparar los valores de R^2_{XY} de las ecuaciones (2.16) y (2.10) podemos proceder así: estimar Ln C del modelo (2.16), obtener sus antilogaritmos y luego calcular R^2_{XY} entre antilog Ln C y C. Este R^2_{XY} es comparable con el valor R^2_{XY} del modelo (2.10). Alternamente, obtenga C del modelo (2.10), conviértalo en Ln C y finalmente calcule R^2_{XY} entre Ln $\hat{C}$ y Ln C. Este valor R^2_{XY} es comparable con el valor del modelo (2.10).

Ahora, observando las tablas I y IV, vemos que el R^2_{XY} del modelo (2.16) es mayor que el correspondiente a (2.10), lo que de una u otra manera nos puede sugerir que hemos mejorado el modelo. Para corroborar este sentir, comparamos los R^2_{XY} de las predicciones de ambos modelos

contra los valores reales. También, dado que la R^2 del segundo modelo (Tablas V y VI) es mayor que el del segundo, inferimos que hemos mejorado un poco el modelo. No se afirma que es el mejor modelo, sino que hay una mejora respecto al primero.

TABLA V Analisis de Varianza

Source	Sum of Squares	Df	Mean Square	F-Ratio	Prob. Level
Model	13005.694	1	13005.694	29.89	.00003
Residual	7832.3056	18	435.1281		
Lack-of-fit	7213.8056	16	450.8628	1.458	.48216
Pure error	618.50000	2	309.25000		
Total (Corr.)	20838.000	19			

Correlation Coefficient = 0.790021 R-squared = 62.41 percent
Stnd. Error of Est. = 20.8597

TABLA VI

Variable Dependiente:CONSUMO Variable Independinte: PRELNCONSUMO

Parameter	Estimate	Standard Error	T Value	Prob. Level
Intercept	-6.09007	10.8188	-0.562915	.58044
Slope	1.11278	0.150216	7.40784	.00000

Analisis de Varianza

Source	Sum of Squares	Df	Mean Square	F-Ratio	Prob. Level
Model	15691.131	1	15691.131	54.88	.00000
Residual	5146.8686	18	285.9371		
Lack-of-fit	4568.8686	17	268.7570	.465	.83924
Pure error	578.00000	1	578.00000		
Total (Corr.)	20838.000	19			

Correlation Coefficient = 0.867759 R-squared = 75.30 percent
Stnd. Error of Est. = 16.9097

En la figura (2.13) se muestran las gráficas de los valores reales contra los valores ajustados tanto para el modelo (2.10) como para (2.16)

FIGURA 2.13a Gráfica de Valores Reales contra los Ajustados del modelo 2.16

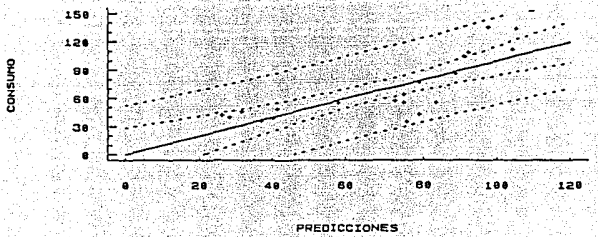
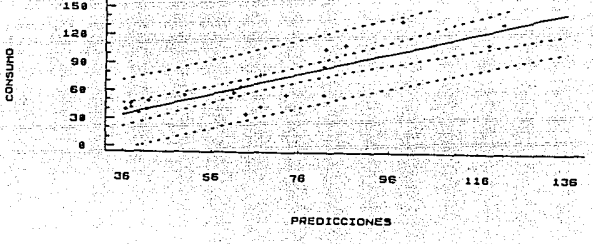


FIGURA 2.13b Gráfica de Valores Reales contra los Ajustados del modelo 2.16



A estas alturas es necesario hacer una advertencia: no se trata de maximizar R^2_{XY}, es decir, a escoger el modelo que de un mayor R^2_{XY}, pues en el análisis de regresión nuestro objetivo no es obtener un R^2_{XY} alto per-se, sino más bien obtener estimadores confiables de los verdaderos coeficientes de regresión poblacionales y hacer inferencias estadísticas de ellos. En el análisis empírico no es extraño obtener altos R^2_{XY} pero con coeficientes de regresión estadísticamente insignificantes o que tienen signos contrarios a los que se esperaba a priori. Por lo tanto, la preocupación debe centrarse en la relevancia teórica y lógica de la variable explicatoria en la dependiente y por su significancia estadística. Si en este proceso se encuentra un R^2_{XY} alto está muy bien. Pero por otra parte, si R^2_{XY} es bajo, esto no implica que el modelo sea malo.

2.6 AJUSTE POLINOMIAL

El modelo de regresión lineal $Y = a + bX$ es el modelo general para ajustar cualquier relación lineal entre las variables. Sin embargo podemos tener el caso en que la relación entre la variable dependiente y la independiente es no lineal (curvilínea), o aun más, que el modelo no sea intrínsicamente lineal, figura (2.14). En estos casos utilizamos los modelos denominados modelos polinomiales en una variable, cuya expresión algebraica es

$$Y = b_0 + b_1X + b_2X^2$$

El cálculo de los coeficientes se puede determinar por medio de la minimización de los errores al cuadrado. La bondad de ajuste se mide y se interpreta de igual manera que lo expuesto anteriormente (salvo el coeficiente de correlación). Para mayores detalles ver la referencia 1 de la página 181 a 201.

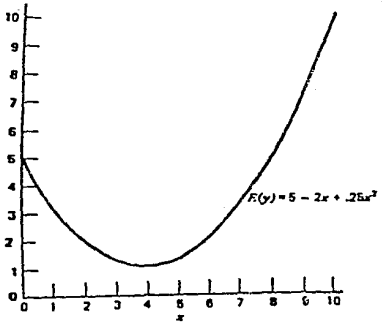


FIGURA 2.14 Ejemplo de polinomio cuadrático

2.7 ANALISIS DE RESIDUALES

2.7.1 DIAGRAMA DE RESIDUALES CONTRA $\hat{Y}_i$

Un método simple y efectivo para detectar modelos deficientes en el análisis de regresión es por medio del análisis de residuales. El i-ésimo residual se define como

$$\epsilon_i = Y_i - \hat{Y}_i$$

Para cada ϵ_i, definimos también el i-ésimo residual estandarizado como

$$\epsilon_i' = \frac{\epsilon_i}{[MSE]^{1/2}}$$

Los residuales estandarizados ϵ_i' tienen media cero y desviación estándar uno.

Un diagrama de residuales ϵ_i' contra los valores ajustados correspondientes $\hat{Y}_i$ es útil para detectar deficiencias comunes en un modelo de regresión. Si el diagrama se parece a la figura (2.15a), el cual indica que los residuales están contenidos en una banda horizontal, entonces no hay un defecto obvio en el modelo, es decir, que el modelo al menos cumple con todos los supuestos de la regresión mencionados en el punto (2.3.1).

Un diagrama de ϵ_i' contra $\hat{Y}_i$ que asemeje cualquiera de los patrones de la figura (2.15b), (2.15c) y (2.15d) es síntoma de modelos deficientes.

Los patrones de las figuras (2.15b) y (2.15c) indican que la varianza de los errores no es constante. La figura (2.15b) que tiene la forma de un embudo hacia afuera, implica que la varianza es una función creciente de Y (un embudo hacia adentro indica que $V(\epsilon)$ se incrementa cuando Y decrece). El doble arco que muestra la figura (2.15c) ocurre cuando Y es una proporción entre cero y uno. Para quitar ésta inestabilidad de la varianza existen técnicas para su estabilización que pueden ser consultados en la referencia 1 de la página 58 a la 70. Un diagrama de los residuales como el de la figura (2.15d) indica no linealidad. Esto podría indicar que en el modelo hacen falta otras variables explicatorias. Por ejemplo, un término cuadrado puede ser necesario, o posiblemente haga falta una transformación sobre el regresor

y/o sobre la variable dependiente.

Un diagrama de los residuales contra $\hat{Y}_i$ también puede revelar uno o más residuales no usuales (muy grandes). Estos puntos son los potenciales outliers (datos no comunes). Un dato de esta naturaleza puede modificar en forma alarmante el ajuste de datos. Lo conveniente en estos casos es suprimirlos, con la salvedad de analizar el nuevo modelo resultante ya no solo en términos puramente estadístico, sino tomar en cuenta el contorno del problema en cuestión.

Residuales muy grandes también puede indicar que la varianza no es constante o que la relación entre Y y X no es lineal. Estas posibilidades deben ser investigadas antes de considerar a los puntos como outliers.

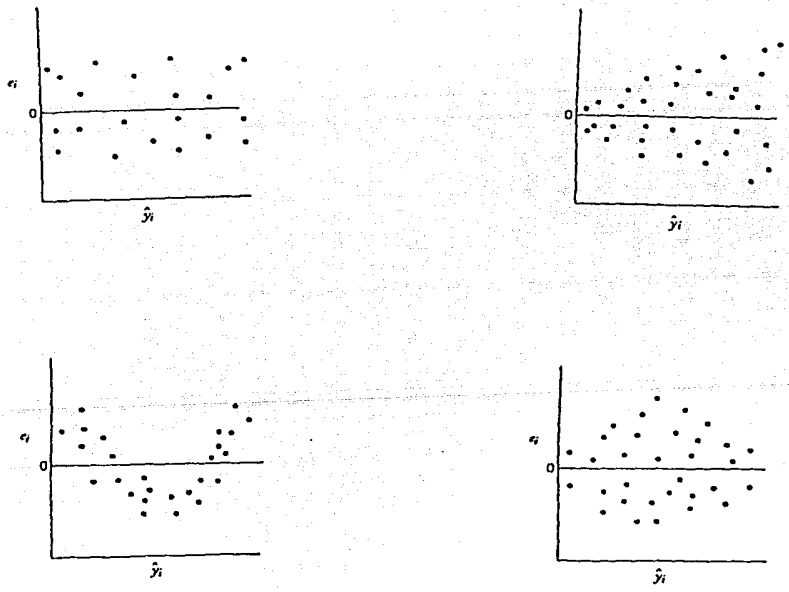


FIGURA 2.15 Patrones para diagramas de residuales.
a) Satisfactorio. (b) Embudo. (c) Doble Arco. (d) No Lineal

2.7.2 DIAGRAMA DE RESIDUALES CONTRA X_i

Haciendo los diagramas de los residuales contra los valores correspondientes de la variable dependiente es útil para analizar el modelo propuesto. Estos diagramas frecuentemente exhiben patrones tales como los de la figura (2.15), salvo que la escala horizontal X_i en lugar de $\hat{Y}_i$. En este caso también es deseable que los residuales se localicen en una banda horizontal. El embudo y el doble arco de la figura (2.15b) y (2.15c) indica varianza no constante. La banda curveada de la figura (2.15d) indica la posibilidad de incluir otros regresores o de la necesidad de hacer transformaciones.

El patrón de comportamiento de residuales del modelo (2.16) mostrado en la figura (2.16a) es menos acentuado que el de la figura (2.16b) correspondiente al modelo (2.10). De aquí que, en efecto, hemos mejorado las predicciones.

FIGURA 2.16a Grafica de los Residuales contra los Ajustados del Modelo 2.10

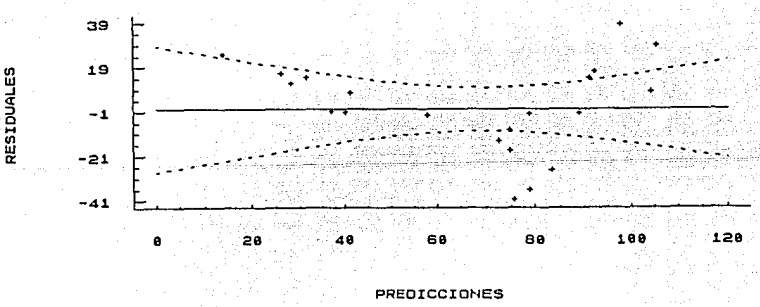
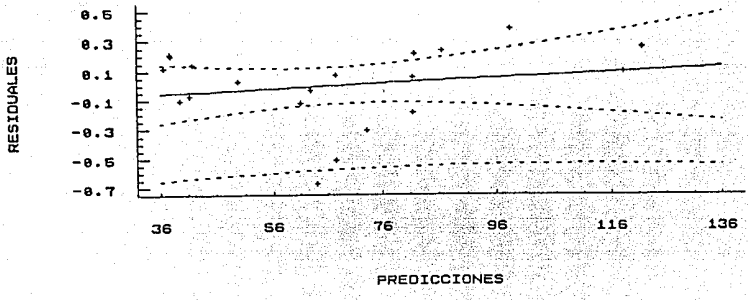
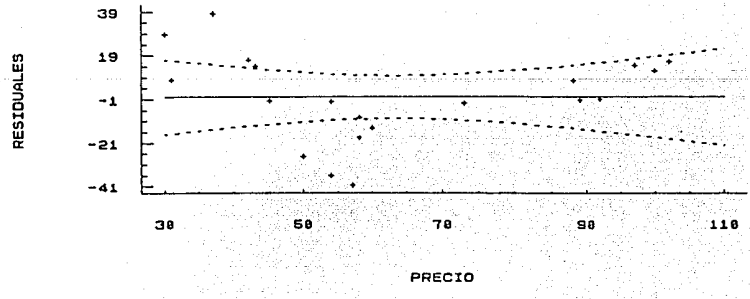


FIGURA 2.16b Grafica de los Residuales
contra los Ajustados del Modelo 2.16



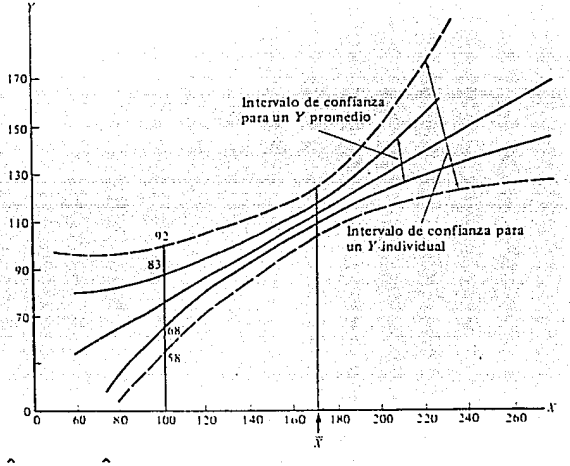
También presentamos la gráfica (2.17) que corresponde al modelo (2.10) de Precio contra los residuales. Obsérvese que las gráficas (2.16b) y (2.17) son semejantes.

FIGURA 2.17 Diagrama de los Residuales
contra Precio del Modelo 2.10



2.8 INTERVALO DE CONFIANZA

La estimación puntual de un parámetro no resulta de mucho valor si no se posee alguna medida del posible error cometido en la estimación, figura (2.18). Toda estimación $\hat{\theta}$ de un parámetro θ debería acompañarse de cierto intervalo que incluyera a $\hat{\theta}$, por ejemplo, de la forma $(\hat{\theta} - d, \hat{\theta} + d)$, junto con alguna medida de seguridad de que el parámetro verdadero θ fuera interior a dicho intervalo.



A $(\hat{\theta} - d, \hat{\theta} + d)$ se le llama intervalo de confianza, el cual es un rango estimado de valores con una probabilidad de cubrir el verdadero valor del parámetro en cuestión de la población.

Los intervalos de confianza del 100 por ciento de los parámetros (α y β) del modelo de regresión lineal están dados por:

$$b - t_{\alpha/2,n-2}\left[\, MSE/Sxx \,\right]^{1/2} \leq \beta \leq b + t_{\alpha/2,n-2}\left[\, MSE/Sxx \,\right]^{1/2}$$

y

$$a - t_{\alpha/2,n-2}\left[MSE\left(\frac{1}{n}+\frac{\bar{X}^{-2}}{Sxx}\right)\right]^{1/2} \leq \alpha \leq a + t_{\alpha/2,n-2}\left[MSE\left(\frac{1}{n}+\frac{\bar{X}^{-2}}{Sxx}\right)\right]^{1/2}$$

Los intervalos de confianza tienen la siguiente interpretación: se toma repetidamente muestras del mismo tamaño X, y construimos, por ejemplo, intervalos de confianza con un nivel del 95 por ciento sobre la pendiente para cada muestra, entonces el 95% de estos intervalos deberá contener el valor verdadero de β.

El intervalo de confianza del 95% para el ejemplo 2.1 de β es

$$-1.53 \leq \beta \leq -0.68$$

Si se elige un valor diferente para el nivel de sgnificación, la anchura del intervalo de confianza cambia. Siguiendo con el ejemplo, un intervalo de confianza del 90% sobre β es $-1.45 \leq \beta \leq -0.52$, el cual es

más angosto que el de 95%. Un intervalo del 99% es $-1.68 \leq \beta \leq -0.75$ que es más ancho que el de 95% . En general, entre mayor sea el nivel de confianza, mayor será la anchura del intervalo de confianza.

TABLA VII Intervalo de Confianza del 95% de los Coeficientes del Modelo 2.10

	Estimate	Standard error	Lower Limit	Upper Limit
CONSTANT	138.561	13.5515	110.083	167.039
EJEMPREC.var1	-1.10414	0.20196	-1.52855	-0.67974

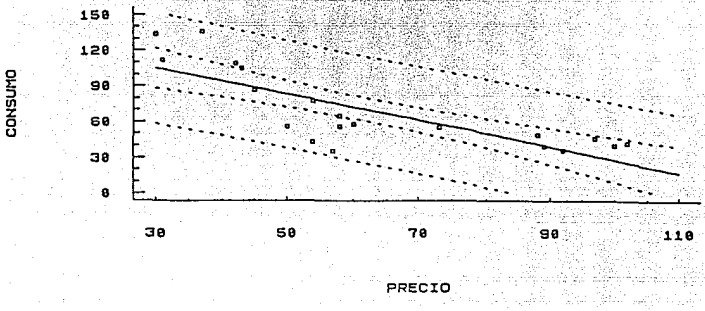
El cálculo del intervalo de confianza de $\hat{Y}_o = a + bX$ en el punto $X = X_o$ se logra mediante:

$$Y_o - t_{\alpha/2,n-2}\left[MSE\left(\frac{1}{n}+\frac{(X_o-\bar{X})^2}{Sxx}\right)\right]^{1/2} \leq \alpha \leq$$

$$\hat{Y}_o - t_{\alpha/2,n-2}\left[MSE\left(\frac{1}{n}+\frac{(X_o-\bar{X})^2}{Sxx}\right)\right]^{1/2}$$

Para el ejemplo 2.1, la figura (2.19) muestra el intervalo de confianza del 95% .

FIGURA 2.19 Intervalo de Confianza del 95% para el ejemplo 2.1



En resumen, la anchura de estos intervalos de confianza es una medida de la calidad de la línea de regresión. Una vez que hallamos establecido una hipótesis, la probabilidad de cometer un error de tipo 1 está absolutamente bajo nuestro control y podemos hacerla tan chica como deseemos. cuán chica la hagamos en un caso específico depende de las consecuencias (costo, incomodidad, desconcierto, etc.) de rechazar una hipótesis verdadera. Ordinariamente, cuanto más graves sean las consecuencias que resulten de cometer un error de tipo 1, tanto menor será el riesgo que estemos dispuestos a tomar para cometerlo. Sin embargo, en la práctica estamos limitados a fijar probabilidades muy bajas de α por el hecho de que, en relación con un tamaño de muestra fijo, cuanto menor hagamos la probabilidad de rechazar una hipótesis verdadera, tanto mayor será la probabilidad β de aceptar una falsa.

CAPITULO 3

REGRESION LINEAL MULTIPLE

3.1 INTRODUCCION

Un modelo de dos variables es adecuado sólo en determinados casos. En el ejemplo 2.1, referente al consumo-precio, se supone implícitamente que sólo el precio P afecta al consumo C. Pero esto rara vez es tan simple, pues además del precio, existen otras variables que pueden afectar al consumo. Un ejemplo obvio es la variación del ingreso del consumidor. Tenemos también la demanda de un bien que puede depender no solo de su precio, sino del precio de otros bienes sustitutos, de la categoría social del consumidor, etc. Por esta razón, necesitamos extender nuestro modelo simple con dos variables a uno que contenga más de dos variables. Esto nos conduce al estudio de los modelos de regresión múltiple, es decir, a los modelos en que la variable dependiente Y depende de dos o más variables explicatorias.

En general, la variable dependiente Y se puede expresar con k variables. El modelo

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k + \epsilon \qquad (3.1)$$

se denomina regresión lineal múltiple con k regresores. El parámetro b_j con $j = 1,2,\ldots,k$ se llaman los coeficientes de la regresión. Este modelo describe un hiperplano en el espacio k-dimensional de los X_k variables. El parámetro b_j representa el cambio esperado en Y cuando hay una unidad de cambio en X_j, manteniendo constante las variables restantes. Por esta razón los parámetros b_j, comúnmente se llaman los coeficientes parciales de regresión.

Los modelos de regresión lineal múltiple son regularmente usados como una función de aproximación. Esto es, la verdadera relación

funcional entre Y y X_1, X_2,..., X_k es desconocido.

En la práctica, la tarea del modelo de regresión lineal es estimar los parámetros desconocidos del modelo (3.1), es decir, b_0, b_1, ..., b_k a partir de un conjunto de datos conocidos, aplicándose el procedimiento de los mínimos cuadrados . Otra forma de expresar a la ecuación (3.1) es

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} + \epsilon_i \qquad (3.2)$$

donde

X_{ji} son medidas que se suponen sin error

b_j son los estimados por mínimos cuadrados de β_j y son variables aleatorias que además, tienen una distribución normal

ϵ_i es un término de error estimado para la i-ésima observación, y que son independientes y que tienen una distribución normal.

De muestra a muestra los coeficientes b_j fluctúan, dando lugar a una elevada familia de superficies de regresión.

En lo que sigue determinaremos la forma de calcular los b_j's, y también la bondad de ajuste correspondiente al modelo propuesto.

3.2 ESTIMACION DE LOS COEFICIENTES DE REGRESION POR MEDIO DE LAS DERIVADAS PARCIALES

El modelo de regresión múltiple más simple es el de la regresión de tres variables, una dependiente y dos explicatorias. En este apartado estudiaremos este modelo.

Para ilustrar la aplicación de regresión múltiple, consideremos el siguiente caso :

EJEMPLO 3.1 Una distribuidora de refrescos está analizando las rutas de servicio de las máquinas vendedoras. Está interesada en predecir el tiempo total requerido por el conductor en una salida.

Esta actividad incluye el surtido a la máquina de bebida y un mantenimiento menor. El ingeniero industrial responsable ha sugerido que

las dos variables más importantes que afectan el tiempo de entrega son: número de productos surtidos (número de casos) y la distancia caminada por el conductor de la ruta. El ingeniero ha reunido 25 observaciones sobre el tiempo de entrega, las cuales se muestran en el apéndice D.

El modelo de regresión lineal múltiple que deberá representar este ejemplo es:

$$Y = b_0 + b_1X_1 + b_2X_2 + \epsilon \qquad (3.3)$$

El término "lineal" se usa en (3.3) porque es una función lineal sobre los parámetros desconocidos b_0, b_1 y b_2. El modelo describe un plano en el espacio bidimensional de los regresores X1 y X2, figura (3.1). El parámetro b_0 es el intercepto, que no tiene una interpretación física. El parámetro b_1 indica el cambio esperado en Y por una unidad de cambio en X1 cuando X2 se mantiene constante. Similarmente, b_2 es el cambio esperado en Y por una unidad de cambio en X2 cuando X1 se mantiene constante.

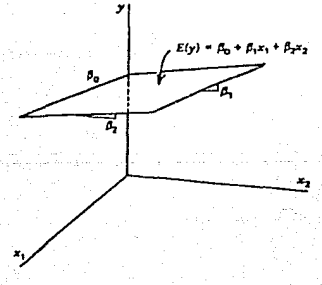


FIGURA 3.1 Ejemplo de modelo de Regresión Múltiple con dos regresores

Ahora se procederá a calcular los coeficientes de la ecuación (3.3). Primero reescribimos a tal ecuación en el formato de la ec. (3.2)

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \epsilon_i$$

$$= \hat{Y}_i + \epsilon_i$$

De aquí que

$$\epsilon_i = Y_i - \hat{Y}_i$$

y el método de los mínimos cuadrados se usa para encontrar la suma mínima de estos errores al cuadrado, es decir

$$\min S = \sum \epsilon_i^2$$

$$= \sum (Y_i - \hat{Y}_i)^2$$

$$= \sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})^2$$

Ahora, calculando las derivadas parciales de S con respecto a cada uno de los coeficientes desconocidos, b_0 , b_1 , b_2, e igualando a cero obtenemos

$$\frac{\partial S}{\partial b_0} = -2\sum (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}) = 0$$

$$\frac{\partial S}{\partial b_1} = -2\sum X_{1i}(Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}) = 0$$

$$\frac{\partial S}{\partial b_2} = -2\sum X_{2i}(Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i}) = 0$$

De este sistema de ecuaciones resultante, se determinan los b_i's. Del ejemplo 3.1 tenemos que (tabla VIII)

$$\hat{Y} = 2.37 + 1.622X_1 + 0.014X_2 \qquad (3.4)$$

donde $\hat{Y}$ indica que es el estimador de Y. Este estimado $\hat{Y}$ está basado sobre dos variables únicamente.

TABLA VIII Modelo Lineal del Ejemplo 3.1

Independent variable: T. Entrega	coefficient	std. error	t-value	sig.level
CONSTANTE	2.366601	1.095858	2.1596	0.0420
NUMERO DE CASOS	1.622257	0.170599	9.5092	0.0000
DISTANCIA	0.014236	0.00361	3.9432	0.0007

R-SQ. (ADJ.) = 0.9560 SE= 3.256882 MAE= 2.264252 DurbWat= 1.147
Previously: 0.0000 0.000000 0.000000 0.000
25 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

Como vemos, hasta aquí no hay nada novedoso. El procedimiento para calcular los coeficientes de la ec. (3.3) es el mismo que el utilizado para el caso de la regresión simple. De hecho, para analizar la bondad de ajuste, sólo se harán los comentarios pertinentes.

3.3 CALCULO DE LOS COEFICIENTES DE REGRESION EN FORMA MATRICIAL

Como puede observarse, calcular los parámetros de la ecuación 3.3 por el método de las derivadas parciales es demasiado engorroso, razón por la cual optaremos por el uso del álgebra matricial para determinar b_0, b_1 , ..., b_k de la ecuación (3.2).

La expresión matricial de la ecuación general de regresión (3.2) es:

$$Y = Xb + \epsilon$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n1} & \dots & x_{nk} \end{bmatrix}$$

$$b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

En general, Y es el vector de las observaciones, X es una matriz que representa los niveles de los regresores, b es el vector de los coeficientes de regresión y ϵ es el vector de los errores.

Para obtener los valores de b, la suma de desviaciones cuadradas deberá ser minimizada:

$$\sum \epsilon_i^2 = \epsilon^t\epsilon = (Y-Xb)^t(Y-Xb)$$

donde

$\epsilon^t = (Y-Xb)^t$ es la transpuesta de ϵ

Entonces

$$\epsilon^t\epsilon = (Y^t - b^tX^t)(Y-Xb)$$

$$= Y^tY - Y^tXb - b^tX^tY + b^tX^tXb$$

$$= Y^tY - 2b^tX^tY + b^tX^tXb$$

ya que b^tX^tY es un escalar, $(b^tX^tY)^t = Y^tXb$ es el mismo escalar.

$$\frac{\delta \epsilon^t\epsilon}{\delta b} = -2X^tY + 2X^tXb = 0$$

que simplificado queda

$$X^tXb = X^tY ,$$

de aquí que

$$b = (X^tX)^{-1} X^tY$$

donde

$(X^tX)^{-1}$ es la inversa de X^tX

Hay que notar que $(X^tX)^{-1}$ tiene sentido cuando no existe "multicolinealidad" entre las variables.

Siguiendo el ejemplo 3.1, tenemos

$$X = \begin{pmatrix} 1 & 7 & 560 \\ 1 & 3 & 220 \\ 1 & 3 & 340 \\ 1 & 4 & 80 \\ 1 & 6 & 150 \\ 1 & 7 & 330 \\ 1 & 2 & 110 \\ 1 & 7 & 210 \\ 1 & 30 & 1460 \\ 1 & 5 & 605 \\ 1 & 16 & 688 \\ 1 & 10 & 215 \\ 1 & 4 & 255 \\ 1 & 6 & 462 \\ 1 & 9 & 448 \\ 1 & 10 & 776 \\ 1 & 6 & 200 \\ 1 & 7 & 132 \\ 1 & 3 & 36 \\ 1 & 17 & 770 \\ 1 & 10 & 140 \\ 1 & 26 & 810 \\ 1 & 9 & 450 \\ 1 & 8 & 635 \\ 1 & 4 & 150 \end{pmatrix} \qquad y = \begin{pmatrix} 16.68 \\ 11.50 \\ 12.03 \\ 14.88 \\ 13.75 \\ 18.11 \\ 8.00 \\ 17.83 \\ 79.24 \\ 21.50 \\ 40.33 \\ 21.00 \\ 13.50 \\ 19.75 \\ 24.00 \\ 29.00 \\ 15.35 \\ 19.00 \\ 9.50 \\ 35.10 \\ 17.90 \\ 52.32 \\ 18.75 \\ 19.83 \\ 10.75 \end{pmatrix}$$

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 25 & 219 & 10232 \\ 219 & 3055 & 133899 \\ 10232 & 133899 & 6725688 \end{bmatrix} \begin{bmatrix} 559.60 \\ 7375.44 \\ 337072.00 \end{bmatrix}$$

$$= \begin{bmatrix} .11321518 & -.00444859 & -.00008367 \\ -.00444859 & .00274378 & -.00004786 \\ -.00008367 & -.00004786 & .00000123 \end{bmatrix} \begin{bmatrix} 550.60 \\ 7375.44 \\ 337072.00 \end{bmatrix}$$

$$= \begin{bmatrix} 2.34123115 \\ 1.61590712 \\ 0.01438483 \end{bmatrix}$$

De aquí que la ecuación que representa el ejemplo 3.1 mediante este método (que coincide con la ecuación 3.4) es

$$Y = 2.34 + 1.61X_1 + 0.014X_2$$

A continuación generalizaremos los conceptos mencionados en el Capítulo 2 para tener los elementos necesarios en la evaluación del modelo (3.3) y decidir que tan "bueno" es.

3.4 CORRELACION MULTIPLE Y EL COEFICIENTE DE DETERMINACION

En el caso de dos variables vimos que R^2_{XY} mide la bondad de ajuste de la ecuación de regresión; es decir, nos da el porcentaje de la variación total en la variable dependiente Y explicada por la variable explicatoria X. Este sentido de R^2_{XY} puede extenderse a modelos de regresión de más de dos variables. Por consiguiente, en el modelo de tres variables estamos interesados en conocer la proporción de la variable Y explicada conjuntamente por las variables X_1 y X_2. El valor que nos da

esta información se conoce como el coeficiente de determinación múltiple y se denota con R^2.

De la misma forma que en la regresión simple, R^2 se puede calcular mediante

$$R^2 = \frac{\sum (\hat{Y}-\bar{Y})^2}{\sum (Y_i-\bar{Y})^2}$$

R^2 está comprendido entre 0 y 1. Si es 1 significa que la línea de regresión ajustada explica en un cien por ciento la variación en Y. Por otro lado, si es cero, el modelo no explica nada de las variaciones en Y. Sin embargo, cuando R^2 está entre estos valores extremos, se dice que el ajuste del modelo es "mejor" mientras más cerca de 1 esté el R^2.

Recuerde que en el caso de dos variables definimos el valor de r como el coeficiente de correlación e indicamos que medía el grado de asociación (lineal) entre dos variables. El análogo de r en el caso de tres o más variables es el coeficiente de correlación múltiple, denotado con R, y es una medida del grado de asociación entre Y y todas las variables explicatorias conjuntamente, y no es más que la raíz cuadrada de R^2. Aunque r puede ser positivo o negativo, R es siempre positivo. En la práctica se acostumbra emplear R^2.

Para el ejemplo 3.1 tenemos que $R^2 = 0.96$ (tabla IX), por lo que $R = 0.98$. Con este criterio podemos decir que la proporción de variación en el tiempo de entrega en función del número de casos y la distancia es alto.

TABLA IX Analisis de Varianza

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-value
Model	5550.04	2	2775.02	261.615	.0000
Error	233.360	22	10.6073		
Total (Corr.)	5783.40	24			

R-squared = 0.95965 Stnd. error of est. = 3.25688
R-squared (Adj. for d.f.) = 0.955982 Durbin-Watson statistic = 1.14681

3.5 CORRELACION ESPURIA

Si la variable dependiente se observa sobre el tiempo, entonces se pueden hacer predicciones de Y para un período futuro. Como la ecuación de regresión se basa sobre datos pasados, usamos la justificación de que la tendencia observada en el pasado deberá continuar en el futuro.

¿Cómo afecta la variable tiempo (t) en el análisis de correlación?. Si dos series de tiempo presentan una tendencia similar sobre t entonces, ambos están altamente correlacionados sobre un período largo (muestras grandes) pero no correlacionados en períodos cortos (muestras pequeñas). A este tipo de correlación se le denomina correlación espuria. Si logramos remover la tendencia de las dos series, entonces tal correlación espuria se elimina. Una vía para lograr esto es utilizar a t como otra variable independiente en la regresión múltiple . Esto muestra que en muchas ocasiones la correlación se debe a la tendencia (tiempo) y se atribuye a la asociación entre las variables.

La aplicación de este punto se llevará a cabo en el Capítulo 4 de manera muy extensa.

3.6 COMPARACION DE DOS O MAS VALORES DE R^2 : EL R^2 AJUSTADO

Una propiedad importante del R^2 es el hecho de ser una función no decreciente del número de variables explicatorias del modelo, esto es, a medida que aumenta el número de variables explicatorias R^2 crece. La justificación de ésta aseveración se logra a través de que

$$R^2 = 1 - \frac{\sum \epsilon_i^2}{\sum Y_i^2} \qquad (3.4)$$

Para comparar dos R^2, hay que tener en cuenta el número de variables X del modelo, lo cual puede hacerse mediante un coeficiente de determinación alterno, como sigue:

$$\bar{R}^2 = 1 - \frac{\sum \epsilon_i^2/(N-k)}{\sum Y_i^2/(N-1)} \qquad (3.5)$$

donde k = número de parámetros en el modelo incluyendo el término del intercepto. El $\bar{R}^2$ definido de esta forma se conoce como el R^2 ajustado. El término ajustado significa ajustado por los grados de libertad asociados con la suma de cuadrados que aparecen en (3.4): $\sum \epsilon_i^2$ tiene N-k grados de libertad en un modelo con k parámetros que incluyen el intercepto, y $\sum Y_i^2$ tiene N-1 grados de libertad.

La ecuación (3.5) puede escribirse como

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{S_{YY}^2}$$

donde la $\hat{\sigma}^2$ es la varianza residual, un estimador insesgado del verdadero σ^2 y S_{YY}^2 es la varianza muestral de Y.

Se puede ver que $\bar{R}^2$ y R^2 están relacionados, pues sustituyendo (3.4) en (3.5) obtenemos

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N-1}{N-k} \qquad (3.6)$$

De la ecuación (3.6) se deduce que

1) Para $k > 1$, $\bar{R}^2 < R^2$, lo que implica que a medida que el número de variables X aumenta, el $\bar{R}^2$ ajustado es cada vez menor que el R^2 no ajustado; y

2) $\bar{R}^2$ puede ser negativo, aunque R^2 es necesariamente no negativo. En el caso que R^2 sea negativo, se debe de tomar como cero.

Para el ejemplo 3.1, se tiene que $\bar{R}^2 = 0.96$ (Tabla VIII)

3.7 PRUEBAS DE HIPOTESIS

3.7.1 PRUEBA DE SIGNIFICANCIA F

La prueba de significancia de regresión es una prueba para determinar si hay una relación entre la respuesta Y con cualquiera de las variables independientes X_1, X_2, ..., X_k. Las hipótesis apropiadas son:

$$H_0 : b_1 = b_2 = \ldots = b_k = 0$$

$$H_1 : b_j \neq 0 \quad \text{para al menos una } j.$$

Rechazar $H_0 : b_j = 0$ implica que al menos uno de los regresores X_1, X_2, X_3, ..., X_k contribuye significativamente al modelo. El procedimiento de esta prueba es la misma que en el tratado para la regresión lineal simple. La suma total de cuadrados S_{YY} se particiona en una suma de cuadrados debido a la regresión y a una de cuadrados debido a los residuales, es decir:

$$S_{YY} = SSR + SSE$$

y si $H_0 : b_j = 0$ es verdadero, entonces $SSR/\sigma^2 \sim \chi^2_k$ donde el número de grados de libertad para χ^2 es igual al número de variables independientes en el modelo. También se puede mostrar que $(SSE/\sigma^2) \sim \chi^2_{n-k-1}$ y que SSE y SSR son independientes. El proceso de la prueba para $H_0 : b_j = 0$ es calcular

$$Fo = \frac{SSR/k}{SSE/n-k-1} = \frac{MSR}{MSE}$$

y se rechaza si $Fo > F_{\alpha,k,n-k-1}$.

Para el ejemplo 3.1, el análisis de varianza se muestra en la tabla IX. De la prueba $Ho : b_1 = b_2 = 0$, se tiene que $Fo = 261.61$. Ya que $Fo > F_{.05,2,22} = 3.44$, concluimos que la variable dependiente Y está relacionado con la primera variable explicatoria X_1 y/o con la segunda variable explicatoria X_2. Esto es, el tiempo de entrega está relacionado con el volumen y/o la distancia de entrega. Sin embargo, esto no necesariamente implica que la relación encontrada es apropiada como para predecir que el tiempo de entrega es una función del volumen y la distancia. El significado de éste último comentario se desarrolla en el siguiente apartado.

3.8 SELECCION DE VARIABLES Y CONSTRUCCION DE MODELOS

En los ejemplos 2.1 y 3.1 hemos supuesto que las variables explicatorias correspondientes tienen la suficiente influencia en los modelos para incluirlas. Después, con técnicas adecuadas corroboramos que en efecto los modelos propuestos eran correctos, y además que los supuestos de la regresión lineal no son violados.

En la regresión múltiple, bajo consideraciones teóricas o por experiencia se seleccionan los regresores que va a incluir nuestro modelo; sin embargo, en muchos problemas prácticos surge una cantidad considerable de posibles regresores. ¿ De todos estos posibles regresores existe un subconjunto que explique al modelo?. Encontrar un subconjunto apropiado de regresores para el modelo se le denomina problema de selección de variables.

Aquí analizaremos dos métodos de selección: 1) selección forward y 2) eliminación backward. Con el mismo lineamiento que los anteriores, existe otro método denominado regresión stepwise, que no es más que una combinación de los dos anteriores.

3.8.1 METODO FORWARD

Este procedimiento empieza con la suposición que no hay regresores en el modelo, salvo el intercepto. El primer regresor seleccionado para entrar en la ecuación es aquel que tenga la correlación simple más alta con la variable Y. Suponga que esta variable es X_1. Este regresor también debe de producir el valor más grande de F para la prueba de significancia de regresión. Este regresor entra si esta F excede un valor F preseleccionado, digamos F_{IN} (F a entrar).

Para determinar la siguiente variable a entrar en el modelo, calculamos $Y_0 = b_1 + b_2X_1$ y $X_j = a_{1j} + a_{2j}X_1$, $j = 2,3,\ldots,k$ y correlacionamos los residuales de Y_0 y X_j. Suponga que X_2 presenta la correlación más alta con Y. Entonces la segunda variable que entra en el modelo es X_2.

Esto implica que la F estadística parcial más alta es

$$F = \frac{SSR(X_2|X_1)}{MSE(X_1,X_2)}$$

Si este valor excede F_{IN}, entonces X_2 se anexa al modelo. En general, en cada paso la correlación más alta entre los residuales de cada regresor restante y Y se anexa al modelo, siempre y cuando su F estadística exceda a la F_{IN}. Este proceso termina cuando la F no es significativa ($F < F_{IN}$) o cuando el último regresor candidato es anexado al modelo.

Considerando el ejemplo 3.1, a continuación mostramos las tablas Xa, b, c que corresponden a las iteraciones para la selección de variables explicatorias por medio del método forward.

TABLA X a) Selección de Variables Mediante el Forward

Selection: Forward	Maximum steps: 500		F-to-enter: 2.00
Control: Manual	Step: 0		F-to-remove: 2.00
R-squared: .00000	Adjusted: .00000	MSE: 240.975	d.f.: 24

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
			1. NUMERO DE CASOS	.9650	310.9723
			2. DISTANCIA	.8910	88.5440

b)

Selection: Forward Maximum steps: 500 F-to-enter: 2.00
Control: Manual Step: 1 F-to-remove: 2.00

R-squared: .93113 Adjusted: .92814 MSE: 17.317 d.f.: 23

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. N. CASOS	2.17671	310.9723	2. DISTANCIA	.6435	15.5488

c)

Selection: Forward Maximum steps: 500 F-to-enter: 2.00
Control: Manual Step: 2 F-to-remove: 2.00

R-squared: .95965 Adjusted: .95598 MSE: 10.6073 d.f.: 22

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. N. CASOS	1.62226	90.4244			
2. DISTANCIA	0.01424	15.5488			

De la tabla XI, tenemos que Y está explicado por

$$Y = 2.4 + 1.6X_1 + 0.014X_2 \qquad (3.12)$$

TABLA XI Modelo Resultante del Ejemplo 3.1 con la Técnica Forward

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANTE	2.366601	1.095858	2.1596	0.0420
N. CASOS	1.622257	0.170599	9.5092	0.0000
DISTANCIA	0.014236	0.00361	3.9432	0.0007

R-SQ. (ADJ.) = 0.9560 SE= 3.256882 MAE= 2.264252 DurbWat= 1.147
Previously: 0.9560 3.256882 2.264252 1.147
25 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

3.8.2 ELIMINACION BACKWARD

La selección Forward empieza con ningún regresor en el modelo y con un cierto criterio se van insertando variables hasta que se obtiene el modelo deseado. La eliminación Backward trata de encontrar un buen modelo

pero en dirección opuesta. Esto es, empieza con un modelo que incluye todos los posibles k regresores. Entonces la F estadística parcial se calcula para cada regresor como si fuera la última variable para entrar al modelo. La más pequeña de estas F estadísticas parciales se compara con un valor preseleccionado, F_{OUT} (o F a remover). Si esta F es menor que F_{OUT}, el regresor correspondiente se remueve del modelo. Ahora un modelo de regresión con k-1 regresores es ajustado. Las F estadísticas para este nuevo modelo es calculado, y el proceso se repite. El algoritmo de la eliminación backward termina cuando el valor más pequeño de la F no es menor que el valor preseleccionado F_{OUT}.

Ahora aplicaremos el método backward al ejemplo 3.1. La tabla XII muestra los pasos seguidos (en este caso solo fué necesario un solo paso).

TABLA XII Selección de Variables Mediante Backward

Selection: Backward	Maximum steps: 500		F-to-enter: 2.00
Control: Manual	Step: 0		F-to-remove: 2.00
R-squared: .95965	Adjusted: .95598	MSE: 10.6073	d.f.: 22

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. N. CASOS	1.62226	90.4244			
2. DISTANCIA	0.01424	15.5488			

De la tabla XI, para este último método volvemos a obtener la ecuación 3.12 . Hay que notar que los coeficientes correspondientes a X_1 y X_2 en ambos métodos coinciden, esto en general no sucede.

3.9 FUNCIONES INTRINSICAMENTE LINEALES

Así como una línea recta puede no ser un modelo adecuado para el ajuste de datos (en el caso de tener una sola variable explicatoria), un plano o un hiperplano puede que no explique con satisfacción un modelo que presente dos o más variables explicatorias . Entonces lo que procede es proponer nuevos modelos que sean linealizables. La no linealidad se puede detectar a través la experiencia a priori o algunas consideraciones

téoricas. A estos modelos no lineales se les denomina intrínsecamente lineales.

Mostraremos dos modelos intrínsicamente lineales con sus correspondientes transformaciones, y la forma lineal resultante, que serán los posibles modelos propuestos en el capítulo 4.

$$Y = aX^{b_1}Z^{b_2} \tag{3.8}$$

$$Y = ae^{b_1X}e^{b_2Z} \tag{3.9}$$

cuyas transformaciones correspondientes son:

$$Ln\ Y = Ln\ a + b_1X + b_2\ Z \tag{3.10}$$

$$Ln\ Y = Ln\ a + b_1 Ln\ X + b_2 Ln\ Z \tag{3.11}$$

Para comparar un modelo lineal y su correspondiente transformación (para seleccionar el mejor) es necesario seguir las observaciones hechas en el capítulo 2 correspondiente a los modelos intrínsicamente lineales para el caso simple.

3.10 EFECTO DEL RETRASO DE DATOS

Supongamos que la cantidad cosechada (CC) en un período, de algún producto, depende de variables tales como: créditos (C), fertilizantes (F), lluvias (LL), etc. logrados en ese mismo periodo. Lo anterior lo expresamos como

$$CC_t = f(C_t,\ F_t,\ LL_t)$$

Sin embargo, la cosecha de la etapa presente, también puede depender de las mismas variables, pero de uno, dos o tres periodos anteriores. A esto se le denomina retraso de datos, y lo podemos expresar como

$$CC_t = f(C_t, F_t, LL_t, C_{t-1}, F_{t-1}, LL_{t-1}, C_{t-2}, F_{t-2}, LL_{t-2})$$

donde t es el periodo actual
t-1 es el periodo anterior
t-2 representa los dos periodos anteriores

Para ilustrar esto, consideremos la variable dependiente X, y las variables independientes Y y Z. Tomaremos 10 datos y dos retrasos

t	X_t	Y_t	Z_t	Y_{t-1}	Z_{t-1}	Y_{t-2}	Z_{t-2}
1	4	5	3				
2	7	4	8	5	3		
3	3	7	5	4	8	5	3
4	2	9	3	7	5	4	8
5	9	15	18	9	3	7	5
6	11	16	22	15	18	9	3
7	13	8	27	16	22	15	18
8	2	12	52	8	27	16	22
9	25	20	9	12	52	8	27
10	32	15	7	20	9	12	52

entonces del modelo

$X_t = aY_t + bZ_t + cY_{t-1} + dZ_{t-1} + eY_{t-2} + fZ_{t-2}$ con $t = 3,4,\ldots, 10$.

seleccionamos las variables que tienen mayor influencia en X_t.

3.11 ESTACIONALIDAD

Un patrón estacional existe cuando los datos están influenciados por factores estacionales (un día del mes, un mes, un cuarto de año, etc.). La venta de productos como refrescos, helados, sweters, etc., exhiben este tipo de patrones. Un patrón de estacionalidad de un cuarto de año se muestra en la figura (3.2)

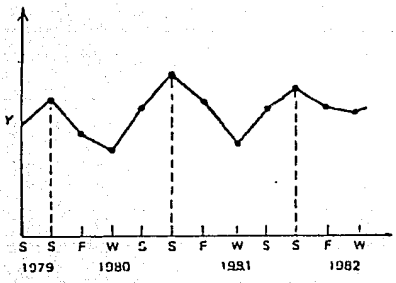


FIGURA 3.2 Tendencia Estacional

3.12 ANALISIS DE RESIDUALES

Los residuales ϵ_i del modelo de regresión múltiple también juegan un papel muy importante para juzgar dicho modelo, tal como se hizo en la regresión simple.

De igual manera que en el punto 2.9 del capítulo 2, analizaremos:

i) Los residuales contra cada regresor X_j , $j = 1, 2, \ldots, N$

ii) Los residuales contra $\hat{Y}$ ajustada.

Los diagramas correspondientes detectan una desviación de normalidad, outliers, varianza no constante, y la especificación funcional errónea de un regresor. Una inspección de los residuales estandarizados se usa con frecuencia para detectar outliers.

Una variante que podemos incluir en este segmento es el diagrama del regresor X_j contra X_i para todo $i \neq j$. Este diagrama ayuda a visualizar la relación de dependencia que pueden guardar entre las variables explicatorias. Por ejemplo, la figura (3.3) indica que X_i y X_j están correlacionadas. Consecuentemente no es necesario incluir alguno de los regresores en el modelo. Si dos o más regresores están correlacionadas, se dice que entre los datos está presente la multicolinealidad.

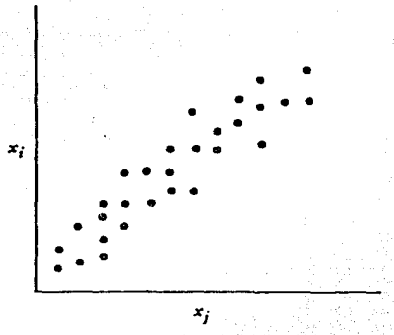
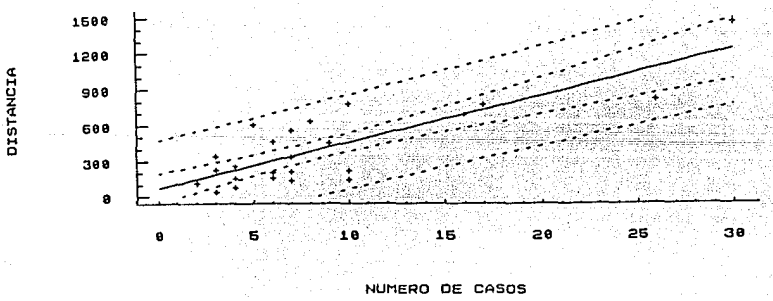


FIGURA 3.3 Dependencia entre variables

En la figura (3.4) exhibimos el diagrama entre las variables explicatorias del ejemplo 3.1, y no se observa una correlación fuerte entre ellas. Por consiguiente, los dos regresores permanecen en el modelo.

FIGURA 3.4 Diagrama de Puntos entre las Variables Explicatorias



Observando las figuras (3.5a) y (3.5b) (sin perder de vista los comentarios pertinentes hechos en el punto (2.5): análisis de residuales), y retomando los análisis estadísticos mencionandos con

anterioridad, concluimos que el modelo 3.3 es adecuado.

FIGURA 3.6 b) Diagrama de Residuales contra las Predicciones

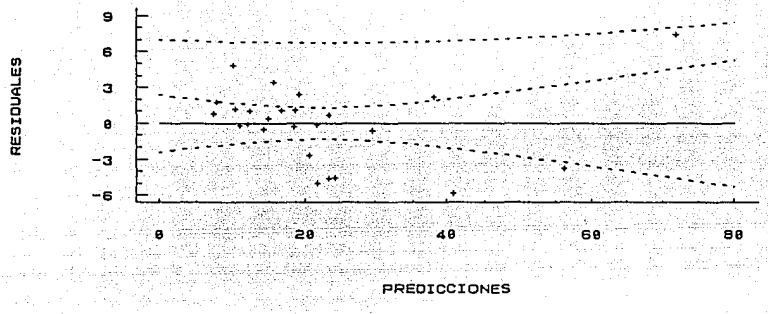
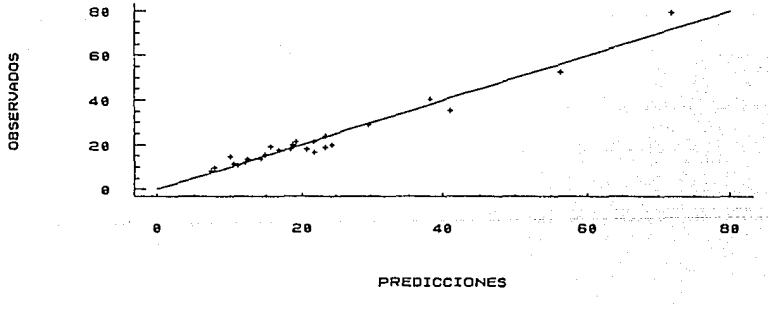


FIGURA 3.6 a) Diagrama de Valores Observados contra sus Predicciones



3.13 INTERVALOS DE CONFIANZA SOBRE LOS COEFICIENTES DE REGRESION

Para construir el intervalo de confianza para el coeficiente de

regresión β_j , debemos de suponer que los errores ϵ_i son NID$(0,\sigma^2)$. Por consiguiente las observaciones Y_i son NID$(\beta_0 + \sum X_{ij} , \sigma^2)$. Ya que el estimador b de los mínimos cuadrados es una combinación lineal de las observaciones, se sigue que b tiene una distribución normal con media β y matriz de covarianza $\sigma^2(X^tX)^{-1}$. Esto implica que la distribución marginal de cualquier coeficiente b_j es normal con media β_j y varianza $\sigma^2 C_{jj}$, donde C_{jj} es el j-ésimo elemento de la matriz $(X^tX)^{-1}$. Consecuentemente, cada uno de los estadísticos

$$\frac{b_j - \beta_j}{[\hat{\sigma}^2 C_{jj}]^{1/2}} , \quad j = 0,1, \ldots, k$$

está distribuido como t con n-p grados de libertad. Por consiguiente, un intervalo de confianza del 100 por ciento para el coeficiente de regresión β_j, j = 0, 1, ..., k es.

$$b_j - t_{\alpha/2,n-p}[\hat{\sigma}^2 C_{jj}]^{1/2} \leq \beta_j \leq b_j + t_{\alpha/2,n-p}[\hat{\sigma}^2 C_{jj}]^{1/2}$$

En la tabla XIII presentamos los intervalos de confianza del 95% de los coeficientes correspondientes al ejemplo 3.1.

TABLA XIII Intervalos de Confianza de los Coeficientes del Modelo 3.1

	Estimate	Standard error	Lower Limit	Upper Limit
CONSTANTE	2.36660	1.09586	0.09338	4.63982
NUMERO DE CASOS	1.62226	0.17060	1.26837	1.97614
DISTANCIA	0.01424	0.00361	0.00675	0.02172

3.14 MULTICOLINEALIDAD

Si dos vectores (columnas de datos) tienen la misma dirección,

entonces ellos son colineales. En el análisis de regresión, un conjunto de variables son multicolineales si cumplen al menos una de las siguientes condiciones:

a) Si la correlación entre dos variables independientes es uno. A esto se le denomina correlación perfecta.

b) Si la correlación entre dos variables independientes es casi perfecta, esto es, si la correlación entre ellos está cerca de ± 1.

c) Si una combinación lineal de variables independientes está casi correlacionado con otra variable independiente.

Si en un problema de regresión está presente la multicolinealidad perfecta, entonces no es posible solucionarlo con mínimos cuadrados. Si la multicolinealidad es casi perfecta, entonces la solución encontrada presenta un margen de error.

Una manera de evitar la multicolinealidad es aplicando la técnica de selección de variables (forward o backward)

CAPITULO 4

STATGRAF

En este capítulo describiremos brevemente el paquete utilizado en este trabajo: statgraf. El statgraf es un paquete netamente estadístico, (ver tabla 4.1), en el que los datos son alimentados por medio de una hoja de cálculo (LOTUS) o bien capturados dentro de la opción FILE OPERATIONS que se localiza en el menú de DATA MANAGEMENT.

TABLA 4.1

STATGRAPHICS Statistical Graphics System	
DATA MANAGEMENT AND SYSTEM UTILITIES	TIME SERIES PROCEDURES
A. Data Management	L. Forecasting
B. System Environment	M. Quality Control
C. Report Writer and Graphics Replay	N. Smoothing
D. Graphics Attributes	O. Time Series Analysis
PLOTTING AND DESCRIPTIVE STATISTICS	ADVANCED PROCEDURES
E. Plotting Functions	P. Categorical Data Analysis
F. Descriptive Methods	Q. Multivariate Methods
G. Estimation and Testing	R. Nonparametric Methods
H. Distribution Functions	S. Sampling
I. Exploratory Data Analysis	T. Experimental Design
ANOVA AND REGRESSION ANALYSIS	MATHEMATICAL AND USER PROCEDURES
J. Analysis of Variance	U. Mathematical Functions
K. Regression Analysis	V. Macros and User Functions

Dado que el presente trabajo es de Análisis de Regresión , entonces elegimos la opción REGRESSION ANALYSIS. De donde obtenemos el menú que aparece en la tabla 4.2. Aquí tenemos tres opciones a elegir (dado el lineamiento del trabajo) : regresión simple (1), regresión múltiple (3) o selección de variables (4).

```
+----------------------------------+
|       REGRESSION ANALYSIS        |
+----------------------------------+

1.  Simple Regression
2.  Interactive Outlier Rejection
3.  Multiple Regression
4.  Stepwise Variable Selection
5.  Ridge Regression
6.  Nonlinear Regression
```

TABLA 4.2

Las dos primeras opciones lo describiremos brevemente, y la última lo haremos con más detalle. La salida de la regresión simple se muestra en la tabla 4.3.

TABLA 4.3 Simple Regression

```
Dependent variable: VARIABLE 1

Independent variable: VARIABLE 2

Model: Linear

Confidence limits:  95.00

Prediction limits:  95.00

Point labels:
```

La salida correspondiente al análisis múltiple se muestra en la siguiente tabla (4.4a y b)

TABLA 4.4a Multiple Regression

```
Dep. var.: VARIABLE 1

Ind. vars.: VARIABLE 2
            VARIABLE 3
            VARIABLE 4
            VARIABLE 5

Weights:

Constant: Yes     Vertical bars: No     Conf. level: 95
```

b) Multiple Regression

```
Dep. var.: PRNAA.var1

Ind. vars.: TIEMPO.X
            PREGARA.var1
            FERTOT.var1
            CREDIT.var1

Weights:

Constant: Yes     Vertical bars: No     Conf. level: 95
```

```
+-----------------------------+
| Analysis of variance        |
| Conditional sums of squares |
| Plot residuals              |
| Summarize residuals         |
| Plot predicted values       |
| Probability plot            |
| Component effects plot      |
| Influence measures          |
| Correlation matrix          |
| Generate reports            |
| Confidence intervals        |
| Interval plots              |
| Save results                |
+-----------------------------+
```

Cabe mencionar que también es posible realizar análisis de modelos no lineales tales como:

$$Y = a + bX + cX^2$$

$$Y = a + b\ (1/X)$$

Ahora describiremos el método de selección de variables. La tabla 4.5 permite el acceso a las variables de interés. En METHOD aparecen las dos opciones para llevar a cabo la selección de variables: FORWARD y BACKWARD. En CONTROL tenemos opciones: Manual o Automatic. El primero permite visualizar paso a paso las variables que son introducidas al modelo; el segundo solo permite ver el modelo final.

TABLA 4.5 Stepwise Variable Selection

Dep. var.: VARIABLE 1

Ind. vars.: VARIABLE 2
VARIABLE 3
VARIABLE 4
VARIABLE 5

Weights:

Constant: Yes	Vertical bars: No	Conf. level: 95	Method: Forward
F-enter: 4	F-remove: 4	Max. steps: 500	Control: Manual

Para ilustrar lo expuesto anteriormente consideremos el siguiente ejemplo: supongamos que la producción de arroz depende de variables tales como: lluvia, crédito, fertilizantes, precio de garantía y tiempo.

La tabla 4.6 describe las variables que van entrando al modelo (crédito - precio de garantía - fertilizantes)

TABLA 4.6 a) Stepwise Selection for PRODUCCION DE ARROZ

Selection: Forward	Maximum steps: 500	F-to-enter: 2.00
Control: Manual	Step: 0	F-to-remove: 2.00

R-squared: .00000 Adjusted: .00000 MSE: 1.69181E10 d.f.: 25

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
			1. CREDIT.var1	.7729	35.6146
			2. FERTOT.var1	.7695	34.8434
			3. PREGARA.var1	.5615	11.0517
			4. TIEMPO.X	.7501	30.8852

b) Stepwise Selection

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 1 — F-to-remove: 2.00

R-squared: .59741 — Adjusted: .58064 — MSE: 7.09479E9 — d.f.: 24

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. CREDIT.var1	9.21238	35.6146	2. FERTOT.var1	.2297	1.2814
			3. PREGARA.var1	.4270	5.1292
			4. TIEMPO.X	.1411	.4672

c) Stepwise Selection

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 2 — F-to-remove: 2.00

R-squared: .67082 — Adjusted: .64220 — MSE: 6.05331E9 — d.f.: 23

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. CREDIT.var1	7.77708	24.8411	2. FERTOT.var1	.2977	2.1387
3. PREGARA.var1	3.22438	5.1292	4. TIEMPO.X	.2591	1.5834

d) Stepwise Selection

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 3 — F-to-remove: 2.00

R-squared: .69999 — Adjusted: .65908 — MSE: 5.76776E9 — d.f.: 22

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. CREDIT.var1	15.8167	7.6876	4. TIEMPO.X	.0696	.1022
2. FERTOT.var1	-0.21613	2.1387			
3. PREGARA.var1	6.64666	5.9639			

Esta última tabla dice que la producción del arroz depende solamente del crédito, su precio de garantía y fertilizantes. Ahora regresamos a la tabla 4.6a y nos quedamos solo con las variables seleccionadas, que son las que aparecen en la tabla 4.7a

TABLA 4. 7a Stepwise Variable Selection

Dep. var.: PRNAA.var1

Ind. vars.: CREDIT.var1
FERTOT.var1
PREGARA.var1

Weights:

Constant: Yes	Vertical bars: No	Conf. level: 95	Method: Forward
F-enter: 2	F-remove: 2	Max. steps: 500	Control: Manual

conforme a lo anterior, hacemos una corrida para obtener el modelo de regresión final (tabla 4.7b):

TABLA 4.7 b) Model fitting results for: PRODUCCION DE ARROZ

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	2.723926E5	3.17572E4	8.5774	0.0000
CREDIT.var1	15.816725	5.704552	2.7726	0.0111
FERTOT.var1	-0.216129	0.147788	-1.4624	0.1578
PREGARA.var1	6.646657	2.721679	2.4421	0.0231

R-SQ. (ADJ.) = 0.6591 SE= 75945.755423 MAE= 50182.984562 DurbWat= 2.100
Previously: 0.0000 0.000000 0.000000 0.000
26 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

A partir de este momento, estamos en posibilidad de manejar una amplia variedad de opciones (tabla 4.8) que nos permiten tanto revisar si se cumplen los supuestos teóricos de la regresión (análisis de varianza, análisis de residuales) como elaborar los reportes finales de resultados (intervalo de confianza, ajuste de la curva)

TABLA 4.8 Stepwise Variable Selection

```
Dep. var.: PRNAA.var1

Ind. vars.: CREDIT.var1
            FERTOT.var1
            PREGARA.var1

Weights:

Constant: Yes     Vertical bars: No     Conf. level: 95     Method: None
F-enter: 2        F-remove: 2           Max. steps: 500     Control: Manual
```

```
+----------------------------+
| Analysis of variance       |
| Conditional sums of squares|
| Plot residuals             |
| Summarize residuals        |
| Plot predicted values      |
| Probability plot           |
| Component effects plot     |
| Influence measures         |
| Correlation matrix         |
| Generate reports           |
| Confidence intervals       |
| Interval plots             |
| Save results               |
+----------------------------+
```

Para el análisis de varianza (útil para medir la significancia de los coeficientes y en la bondad de ajuste) obtenemos la tabla 4.9

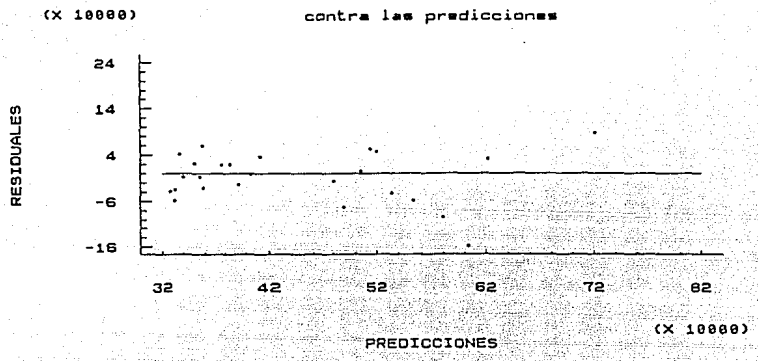
TABLA 4.9 Analysis of Variance for the Full Regression

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-value
Model	296062019357.	3	98687339786.	17.1102	.0000
Error	126890670868.	22	5767757767.		
Total (Corr.)	422952690225.	25			

R-squared = 0.699989 Stnd. error of est. = 75945.8
R-squared (Adj. for d.f.) = 0.659078 Durbin-Watson statistic = 2.09994

Oprimiendo ESC regresamos a la tabla 4.8. Si deseamos hacer el análisis de residuales tecleamos PLOT RESIDUALS y obtenemos la figura 4.1

FIGURA 4.1 Diagrama de los Residuales contra las predicciones

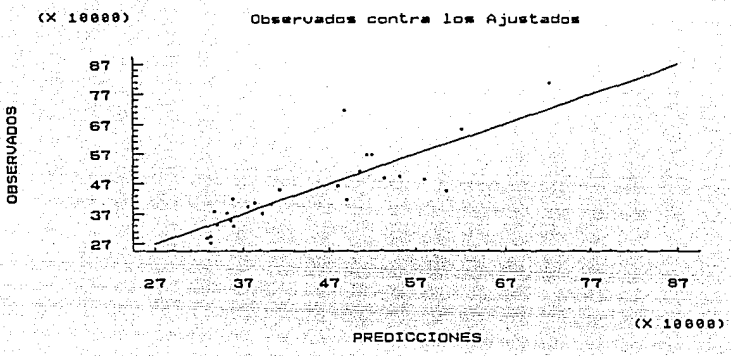


A continuación (tabla 4.10) se muestra el intervalo de confianza para los coeficientes del modelo propuesto, y el ajuste de la curva (figura 4.2).

TABLA 4.10 95 percent confidence intervals for coefficient estimates

	Estimate	Standard error	Lower Limit	Upper Limit
CONSTANT	272393.	31757.2	206516.	338269.
CREDIT.var1	15.8167	5.70455	3.98337	27.6501
FERTOT.var1	-0.21613	0.14779	-0.52270	0.09044
PREGARA.var1	6.64666	2.72168	1.00089	12.2924

FIGURA 4.2 Diagrama de los Valores Observados contra los Ajustados



No hay Hoja.

77

CAPITULO V

CASO DE APLICACION

La Agricultura Nacional, para efectos de análisis, se integra por los siguientes grupos de cultivos: alimentos básicos, oleaginosas, hortalizas, fibras, forrajes, otros granos, frutas de ciclo corto, frutas de ciclo largo y agrícolas industrializables.

El principal grupo lo forman los granos básicos (maíz, frijol, trigo y arroz), que por su significación social y por su participación en el valor de la producción, resultan de particular interés. El propósito de este trabajo consiste en la predicción de la producción anual de estos granos en función de variables que son estadísticamente manejables (lluvia, fertilizante, crédito, precios de garantía, superficie cosechada, etc.).

5.1 CONCEPTUALIZACION

Una forma posible de explicar la producción respecto a su entorno (variables) se muestra en el diagrama 5.1. El origen de una flecha indica que esa variable es independiente, y el extremo terminal se toma como la variable dependiente.

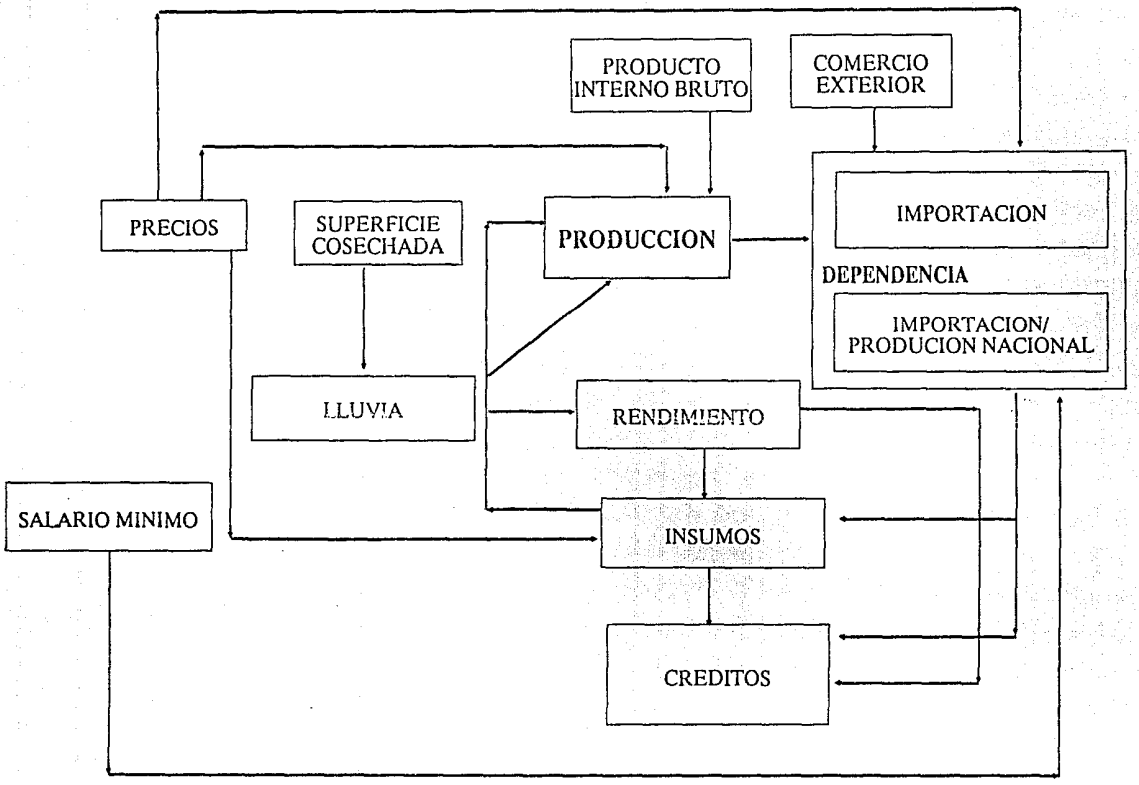


DIAGRAMA 5.1

Sin embargo, la modelación de un sistema de esta naturaleza se torna muy complicada, por lo que se decidió analizar previamente tres submodelos (Diagramas 5.2, 5.3, 5.4 y 5.5):

- RENDIMIENTO,
- PRODUCCION, e
- INFLUENCIA DE FACTORES ECONOMICOS Y DE AUTOSUFICIENCIA

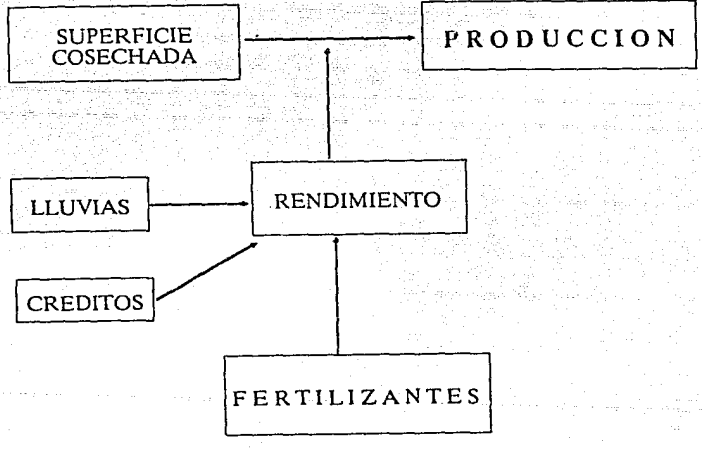


DIAGRAMA 5.2

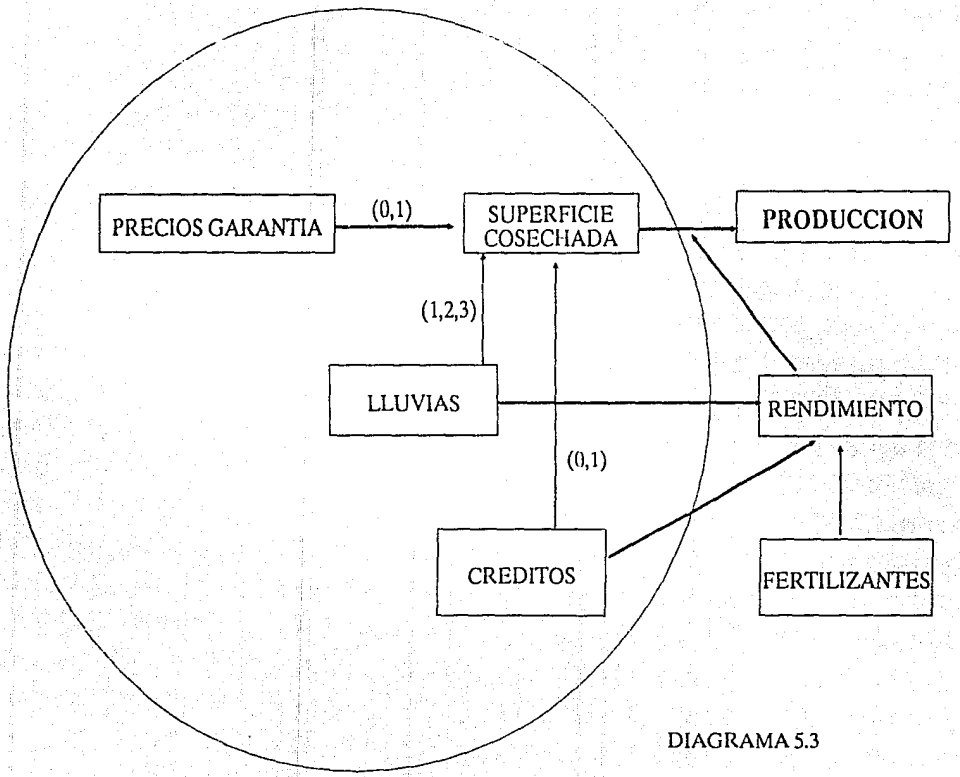

PRECIOS GARANTIA
(0,1)
SUPERFICIE COSECHADA
PRODUCCION
(1,2,3)
LLUVIAS
RENDIMIENTO
(0,1)
CREDITOS
FERTILIZANTES

DIAGRAMA 5.3

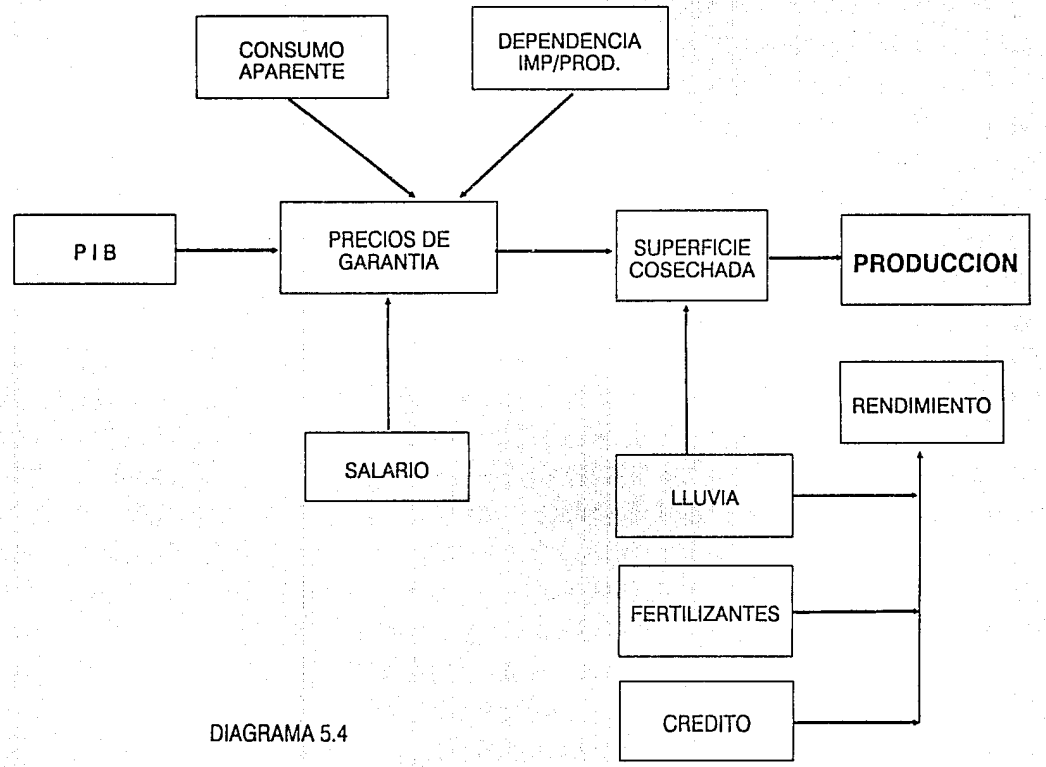

CONSUMO APARENTE
DEPENDENCIA IMP/PROD.
PIB
PRECIOS DE GARANTIA
SUPERFICIE COSECHADA
PRODUCCION
RENDIMIENTO
SALARIO
LLUVIA
FERTILIZANTES
CREDITO

DIAGRAMA 5.4

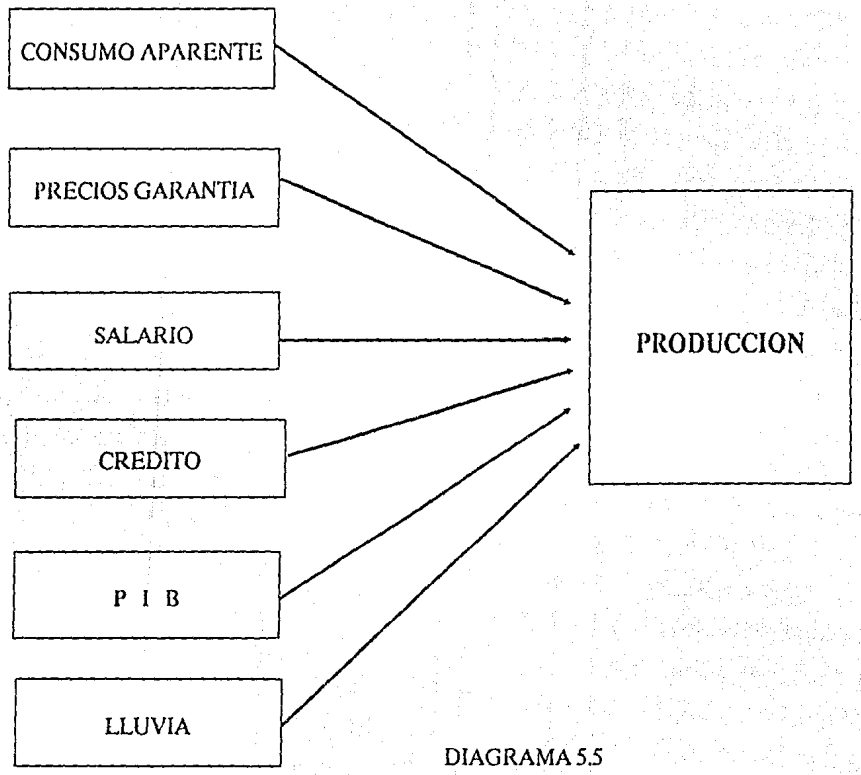

CONSUMO APARENTE
PRECIOS GARANTIA
SALARIO
CREDITO
P I B
LLUVIA
PRODUCCION

DIAGRAMA 5.5

5.2 MODELACION

Los modelos propuestos para los pronósticos serán del tipo lineal

$$Y = a + bX + cZ \qquad (5.0)$$

o alguno de los siguientes modelos que son intrínsicamente lineales

$$Y = aX^{b}Z^{c}W^{d} \qquad (5.1)$$

$$Y = ae^{bX+cZ+dW} \qquad (5.2)$$

cuya forma lineal es respectivamente:

$$Ln\ Y = Ln\ a + bLnX + cLnZ + dLnW \qquad (5.3)$$

$$Ln\ Y = Ln\ a + bX + cZ + dW \qquad (5.4)$$

5.2.1 RENDIMIENTO

Para llevar a cabo la selección de variables en el diagrama 5.2, tomamos al rendimiento (RTO) como la variable dependiente y la lluvia (LL), el crédito (C) y los fertilizantes (F) como las independientes, quedando entonces

$$RTO = ao + aLL + bC + cF \qquad (5.5)$$

Por otro lado, al añadir la superficie cultivada (SC), (diagrama 5.2) se buscará explicar la producción (P) final:

$$P = ao + aLL + bC + cF + fSC \qquad (5.6)$$

$$LnP = Ln\ a + bLnLL + cLnC + dLnF + dLnSC \qquad (5.7)$$

$$LnP = Ln\ a + bLL + cC + dF + eSC \qquad (5.8)$$

5.2.2 SUPERFICIE COSECHADA

Para estudiar la superficie cosechada (diagrama 5.3), se propone el siguiente modelo para la selección de variables:

$$SC = aPG_0 + bPG_1 + cLL_1 + dLL_2 + eLL_3 + fC_0 + gC_1$$

donde los subíndices indican los retrasos correspondientes.

5.2.3 INFLUENCIA DE FACTORES ECONOMICOS Y LA AUTOSUFICIENCIA

Para estudiar la influencia de factores económicos y la autosuficiencia (diagrama 5.4) primero seleccionamos a las variables significativas utilizando al precio de garantía como la variable dependiente, y las variables independientes, al consumo aparente (CA), la dependencia (DEP) que no es más que el cociente de la importación con la producción y el salario (S), todos ellos con los retrasos correspondientes (0, 1, 2 ó 3 años).

5.2.4 PRODUCCION

Para explicar la producción (P) en términos de factores económicos y la autosuficiencia, tomamos a P como la variable dependiente, y las independientes, a las seleccionadas más la superficie cosechada. Entonces el modelo general puede quedar como:

$$P = aSC + \sum b_i(\text{variables seleccionadas}) \qquad (5.9)$$

5.3 RESULTADOS

5.3.1 DATOS

A continuación presentaremos los datos numéricos de las variables que ocuparemos a lo largo del resto del capítulo.

No Hay Hojas.

86. 87. 88. 89. 90. 91. 92.

PRODUCCION NACIONAL DE GRANOS BASICOS
(ton.)

A\O	MAIZ	TRIGO	ARROZ	FRIJOL	TOTAL
1960	5419783	1189979	327513	528175	7465450
1961	6246106	1404910	332944	723340	8707300
1962	6337359	1455256	288973	655608	8737196
1963	6870201	1702989	296373	677280	9546843
1964	8454046	1526613	274430	891526	11146615
1965	8936381	1658673	377531	859584	11832169
1966	9271485	1611947	372227	1013169	12268828
1967	8603279	2061433	417888	980169	12062769
1968	9061823	1780057	347249	856939	12046068
1969	8410891	2326056	394937	834597	11966481
1970	8879384	2676451	405386	925040	12886261
1971	9785733	1830844	368589	921059	12906225
1972	9222837	1809018	403192	869504	12304551
1973	8609132	2090844	450574	1008887	12159437
1974	7847763	2788577	491608	971576	12099524
1975	844708	2798219	716628	1027303	5386858
1976	8017294	3363299	463432	739812	12583837
1977	10137914	2455774	567338	770093	13931119
1978	10930077	2784660	567338	948744	15230819
1979	8457849	2286525	493794	640514	11878682
1980	12374400	2784914	445364	935174	16539852
1981	14550074	3193954	651947	1331305	19727280
1982	10026305	4464101	511366	953544	15955316
1983	13157797	3558294	415703	1242175	18373969
1984	12788809	4505245	484024	930692	18708770
1985	14103454	5214315	807529	911908	21037206

SUPERFICIE NACIONAL COSECHADA DE GRANOS BASICOS
(toneladas)

A\O	MAIS	TRIGO	FRIJOL	ARROZ
1960	5558429	839814	1325760	142587
1961	6287747	836538	1617107	146341
1962	6371704	747728	1673694	133904
1963	6963077	819210	1710767	134757
1964	7460627	742680	2091025	132594
1965	7718371	773791	2116858	138065
1966	8286935	726595	2240022	152642
1967	7610932	750913	1929967	168363
1968	7675845	704551	1790669	138712
1969	7103509	841279	1655520	152980
1970	7439684	886169	1746947	149973
1971	7691656	614180	1932236	153442
1972	7292180	686665	1686746	156145
1973	7606341	640456	1869686	150400
1974	6717234	774149	1551877	172949
1975	6694267	778237	1752641	256661
1976	6783184	894140	1315819	159410
1977	7469649	708863	1630732	180464
1978	7191128	759526	1580228	121316
1979	5581158	584206	1051431	151228
1980	6766479	723804	1551352	127477
1981	7668692	859830	1990669	174792
1982	5606409	1012222	1588638	156315
1983	7343490	879948	1942735	133318
1984	6892682	1033054	1679426	125896
1985	7589537	1217082	1782341	216466

CONSUMO NACIONAL DE FERTILIZANTES
(miles de toneladas)

A\O	NITROGENADOS	FOSFATADOS	POTASICOS
1960	97119	24786	7596
1961	108319	27857	9342
1962	141075	42461	14752
1963	196587	63352	13082
1964	231863	68829	15386
1965	215236	72527	559
1966	264033	89816	10388
1967	95033	101538	16557
1968	363078	116460	24220
1969	397328	139000	22523
1970	404271	111123	21693
1971	434606	152723	27619
1972	484709	156462	35462
1973	555422	178013	40468
1974	593333	230976	34012
1975	732620	276398	63851
1976	830214	239256	67347
1977	779333	218022	34255
1978	730636	258677	72209
1979	824498	245139	75363
1980	922144	288952	110230
1981	1106513	369823	66691
1982	1178600	411300	76900
1983	1049600	334700	66400
1984	1195000	381400	99900
1985	1298900	413900	110900

IMPORTACION
(ton.)

A\O	MAIZ	FRIJOL	TRIGO	ARROZ
1960	28484	24864	4363	22304
1961	34060	9764	7605	236
1962	17902	3267	27127	100
1963	475833	8656	46163	2065
1964	46496	8202	62411	41
1965	12033	458	12535	17834
1966	4502	583	1122	11514
1967	5080	409	1172	28
1968	5500	303	1599	9107
1969	8442	381	762	4844
1970	761791	8647	1130	16301
1971	18308	466	177107	801
1972	204213	2686	641499	662
1973	1145184	18088	719558	37866
1974	1282132	39478	976643	71274
1975	2660839	104400	88526	9
1976	913786	179	5331	18
1977	1985619	29256	456373	92
1978	1344404	1220	458501	112
1979	746278	6786	1169006	35679
1980	4187072	443066	923469	95002
1981	2844369	490167	1027930	93255
1982	233038	1461994	398460	21651
1983	233038	1428	422555	223
1984	2497815	119119	345037	170445
1985	1703500	144566	319983	156172

RENDIMIENTO MEDIO POR HECTAREA DE GRANOS BASICOS
(Kilogramos)

A\O	MAIZ	FRIJOL	TRIGO	ARROZ	TOTAL
1960	975	398	1417	2297	5087
1961	993	447	1676	2275	5391
1962	995	392	1946	2158	5491
1963	987	396	2079	2199	5661
1964	1133	426	2692	2070	6321
1965	1158	406	2505	2734	6803
1966	1119	452	2254	2439	6264
1967	1130	508	2727	2482	6847
1968	1181	479	2632	2503	6795
1969	1184	504	2765	2582	7035
1970	1194	530	3020	2703	7447
1971	1272	485	2981	2404	7142
1972	1265	515	2634	2582	6996
1973	1132	540	3264	2996	7932
1974	1168	626	3602	2843	8239
1975	1262	586	3596	2792	8236
1976	1182	562	3761	2907	8412
1977	1357	472	3464	3143	8436
1978	1520	600	3666	3212	8998
1979	1517	616	3881	3278	9292
1980	1770	551	3771	3456	9548
1981	1897	668	3714	3729	10008
1982	1788	600	4410	3271	10069
1983	1792	639	4043	3118	9592
1984	1855	554	4357	3844	10610
1985	1858	511	4284	3730	10383

DEPENDENCIA
IMPOT/PROD
(ton.)

A\O	MAIZ	FRIJOL	TRIGO	ARROZ
1960	0.0053	0.0471	0.0037	0.0681
1961	0.0055	0.0135	0.0054	0.0007
1962	0.0028	0.0050	0.0186	0.0003
1963	0.0693	0.0128	0.0271	0.0070
1964	0.0055	0.0092	0.0409	0.0001
1965	0.0013	0.0005	0.0076	0.0472
1966	0.0005	0.0006	0.0007	0.0309
1967	0.0006	0.0004	0.0006	0.0001
1968	0.0006	0.0004	0.0009	0.0262
1969	0.0010	0.0005	0.0003	0.0123
1970	0.0858	0.0093	0.0004	0.0402
1971	0.0019	0.0005	0.0967	0.0022
1972	0.0221	0.0031	0.3546	0.0016
1973	0.1330	0.0179	0.3441	0.0840
1974	0.1634	0.0406	0.3502	0.1450
1975	0.1500	0.1016	0.0316	0.0000
1976	0.1140	0.0002	0.0016	0.0000
1977	0.1959	0.0380	0.1858	0.0002
1978	0.1230	0.0013	0.1647	0.0002
1979	0.0882	0.0106	0.5113	0.0723
1980	0.3384	0.4738	0.3316	0.2133
1981	0.1955	0.3682	0.3218	0.1430
1982	0.0232	1.5332	0.0893	0.0423
1983	0.0177	0.0011	0.1188	0.0005
1984	0.1953	0.1280	0.0766	0.3521
1985	0.1208	0.1585	0.0614	0.1934

PRECIOS DE GARANTIA
(base 1978)

A\O	MAIZ	TRIGO	FRIJOL	ARROZ
1960	800	913	1500	850
1961	800	913	1750	900
1962	800	913	1750	900
1963	940	913	1750	1050
1964	940	913	1750	1100
1965	940	800	1750	1100
1966	940	800	1750	1100
1967	940	800	1750	1100
1968	940	800	1750	1100
1969	940	800	1750	1100
1970	940	800	1750	1100
1971	940	800	1750	1100
1972	940	800	1750	1100
1973	940	870	2150	1100
1974	1500	1300	5300	3000
1975	1750	1750	5100	2500
1976	1900	1750	4750	2875
1977	2900	2050	5000	2925
1978	2900	2600	6250	2925
1979	3480	3000	7750	3150
1980	4450	3550	12000	4500
1981	6550	4600	16000	6500
1982	9976.5	6997	21100	9344
1983	19008	14504	32440	19882
1984	32973	25299	51051	33440
1985	52712	35739	150100	53076

CREDITO AGROPECUARIO OPERADO
LOS BANCOS OFICIALES ESPECIALIZADOS
BASE = 1978

A\O	TOTAL	CONSTANTES
1960	1768.1	6879.8
1961	1567.1	5891.3
1962	1559	5860.9
1963	1873.5	7069.8
1964	2119.6	7651.9
1965	2114.2	7523.8
1966	2757.4	9443.1
1967	3173.8	10579.3
1968	3597.2	11641.4
1969	4317.5	13750
1970	4512.3	13969.9
1971	5298.5	15583.8
1972	6131.8	17175.9
1973	7622.1	19055.2
1974	13375.1	27020.4
1975	15540	27263.1
1976	18133	27474.2
1977	24268.8	28517.9
1978	28549.7	28549.7
1979	37856.2	32027.2
1980	55420.1	37119.9
1981	77186.6	40390.6
1982	101921.8	33571.1
1983	154510.2	25209.6
1984	286910	28292.1
1985	495147.4	30952.5

PRODUCTO INTERNO BRUTO
(millones de pesos)

A\O	CORRIENTES	CONSTANTES 1978
1960	159703	621412.42
1961	173236	651263.16
1962	186781	702184.21
1963	207952	784724.53
1964	245501	886285.2
1965	267420	951672.6
1966	297196	1017794.5
1967	325025	1083416.7
1968	359858	1164589
1969	397796	1266866.2
1970	444271	1375452
1971	440011	1294150
1972	564727	1581868.3
1973	690891	1727227.5
1974	899707	1817589.9
1975	1100050	1929912.3
1976	1370968	2077224.2
1977	1849263	2173047
1978	2347454	2347454
1979	3067526	2595199.7
1980	4276490	2864360.3
1981	5874386	3073985.3
1982	9417089	3101808
1983	17141694	2796817.4
1984	28748889	2834916.6
1985	45588462	2849813.2

SALARIO
(pesos)
(base 1978)

A\O	DATOS
1960	39.56
1961	38.04
1962	47.48
1963	47.3
1964	63.25
1965	62.11
1966	63.58
1967	66.27
1968	65.54
1969	72.69
1970	69.05
1971	77.67
1972	88.8
1973	76.34
1974	80.87
1975	105.37
1976	89.87
1977	88.5
1978	90.36
1979	89.85
1980	93.6
1981	93.6
1982	104.83
1983	74.83
1984	70.9
1985	69.24

CREDITO AGROPECUARIO OPERADO
LOS BANCOS OFICIALES ESPECIALIZADOS
BASE = 1978

A\O	TOTAL	CONSTANTES
1960	1768.1	6879.8
1961	1567.1	5891.3
1962	1559	5860.9
1963	1873.5	7069.8
1964	2119.6	7651.9
1965	2114.2	7523.8
1966	2757.4	9443.1
1967	3173.8	10579.3
1968	3597.2	11641.4
1969	4317.5	13750
1970	4512.3	13969.9
1971	5298.5	15583.8
1972	6131.8	17175.9
1973	7622.1	19055.2
1974	13375.1	27020.4
1975	15540	27263.1
1976	18133	27474.2
1977	24268.8	28517.9
1978	28549.7	28549.7
1979	37856.2	32027.2
1980	55420.1	37119.9
1981	77186.6	40390.6
1982	101921.8	33571.1
1983	154510.2	25209.6
1984	286910	28292.1
1985	495147.4	30952.5

CREDITO AGROPECUARIO OPERADO
LOS BANCOS OFICIALES ESPECIALIZADOS
BASE = 1978

LLUVIAS
(m.m.)

A\O	CANTIDAD
1960	692.2
1961	1007.3
1962	804.3
1963	749.2
1964	774.4
1965	708.9
1966	765.8
1967	751.8
1968	769.5
1969	835
1970	839.1
1971	862
1972	792.2
1973	773
1974	806.1
1975	804.8
1976	844.9
1977	798.2
1978	765.6
1979	850.5
1980	697.1
1981	891.4
1982	734.5
1983	781.9
1984	937.3
1985	706.2

5.3.2 ANALISIS DEL RENDIMIENTO

Consideremos a la ecuación 5.5. La pretensión es determinar de qué depende el rendimiento de cada grano básico; utilizando el método forward, obtenemos (tabla 5.1):

TABLA 5.1.a Stepwise Selection for RENDIMIENTO DEL ARROZ

Selection: Forward		Maximum steps: 500		F-to-enter:	2.00
Control: Automatic		Step: 1		F-to-remove:	2.00
R-squared: .55744	Adjusted: .85150		MSE: 38756.6	d.f.:	24

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
3. FERTILIZANTES	0.00090	144.3460	1. TIEMPO	.2194	1.1636
			2. LLUVIA	.0591	.0807
			4. CREDITO	.2639	1.7212

b) Stepwise Selection for RENDIMIENTO DEL FRIJOL

Selection: Forward		Maximum steps: 500		F-to-enter:	2.00
Control: Automatic		Step: 2		F-to-remove:	2.00
R-squared: .73167	Adjusted: .70834		MSE: 1890.85	d.f.:	23

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
2. LLUVIA	0.19234	2.6336	1. TIEMPO	.1625	.5968
4. CREDITO	0.00612	58.9883	3. FERTILIZANTES	.0275	.0167

c) Stepwise Selection for RENDIMIENTO DEL MAIZ

Selection: Forward		Maximum steps: 500		F-to-enter:	2.00
Control: Manual		Step: 1		F-to-remove:	2.00
R-squared: .86185	Adjusted: .85609		MSE: 13516.2	d.f.:	24

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
3. FERTILIZANTES	0.00054	149.7216	1. TIEMPO	.0242	.0135
			2. CREDITO	.1289	.3888
			4. LLUVIA	.0589	.0800

d) Stepwise Selection for RENDIMIENTO DEL TRIGO

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Automatic — Step: 1 — F-to-remove: 2.00

R-squared: .92906 — Adjusted: .92610 — MSE: 52042.7 — d.f.: 24

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. TIEMPO	105.758	314.3110	2. LLUVIA	.0411	.0390
			3. FERTILIZANTES	.0254	.0143
			4. CREDITO	.0135	.0042

Esto es, los rendimientos de cada producto y su dependencia son:

TABLA I

PRODUCTIVIDAD	VAR. EXPLICATIVAS	R^2
Arroz	Fertilizantes	0.86
Frijol	Lluvia y crédito	0.73
Maíz	Fertilizantes	0.86
Trigo	Tiempo	0.92

De los resultados obtenidos es justo mencionar los siguientes hechos:

a) El rendimiento del arroz, trigo y maíz no dependen de la precipitación pluvial. Hecho que es un poco díficil de aceptar, dado que la siembra por temporal es predominante para estos productos. Para corroborar esta última afirmación compárese los datos de la siguientes tablas:

SUPERFICIE COSECHADA EN RIEGO DE GRANOS BASICOS (1984-85)
(hectáreas)

	ARROZ	FRIJOL	MAIZ	TRIGO
1984	72700	161508	883060	898969
1985	432776	140150	978190	1050210

SUPERFICIE COSECHADA EN TEMPORAL DE GRANOS BASICOS (1984-85)
(hectáreas)

	ARROZ	FRIJOL	MAIZ	TRIGO
1984	53196	1517918	6009622	134885
1985	83690	1546634	6487910	166872

b) Sólo el trigo presenta una tendencia. Significa que para los otros productos el rendimiento en un periodo no depende de lo logrado anteriormente.

c) La participación de los fertilizantes (que puede considerarse como un símbolo de la tecnificación) en el rendimiento es notorio. Esto da hincapié a pensar a que una política basada en asistencia técnica y económica tiene amplias posibilidades de éxito; sin embargo, se requiere hacer un estudio más profundo para solidificar esta idea.

5.3.3 PRODUCCION DE GRANOS BASICOS EN TERMINOS DEL RENDIMIENTO.

Para explicar, por ejemplo, la producción del arroz, tomamos las variables más significativas en el rendimiento, y la superficie cosechada. Sólo en este caso mostraremos los tres modelos propuestos, correspondientes a las ecuaciones 5.6, 5.7 y 5.8. Los resultados son los siguientes (tabla 5.2)

TABLA 5.2 a) Model fitting results for: PRODUCCION DE ARROZ

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	-5.853632E4	5.311717E4	-1.1020	0.2819
SUP. COSECHADA DE ARROZ	2.464856	0.352527	6.9920	0.0000
FERTILIZANTES	0.151491	0.019575	7.7389	0.0000

R-Squared	= 0.87	SE=	48986.436989	MAE=	31453.515761	DurbWat= 1.738
Previously:	0.0000		0.000000		0.000000	0.000

26 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

b) Model fitting results for: LnPRODUCCION DEL ARROZ

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	11.995187	0.116968	102.5507	0.0000
SUP. COSECHADA ARROZ	4.517977E-6	7.762929E-7	5.8199	0.0000
FERTILIZANTES	3.411592E-7	4.310651E-8	7.9143	0.0000

R-Squared = 0.85 SE= 0.107872 MAE= 0.072977 DurbWat= 1.566
Previously: 0.0 48986.436989 31453.515761 1.738
26 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

c) Model fitting results for: LnPRODUCCION ARROZ

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	0.827519	1.661962	0.4979	0.6233
LnSUP. COSECHADA ARROZ	0.759621	0.145613	5.2167	0.0000
LnFERTILIZANTES	0.230394	0.031097	7.4089	0.0000

R-Squared = 0.83 SE= 0.115103 MAE= 0.080423 DurbWat= 1.621
Previously: 0.0 0.107872 0.072977 1.566
26 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

Para ilustrar el método de selección del modelo, observemos la R^2 de las tablas 5.2a, b y c. Un primer intento es escoger la mayor. Entonces

$$PA = -5.9 + 2.5SCA + 0.15F \qquad (5.10)$$

es el elegido.

donde

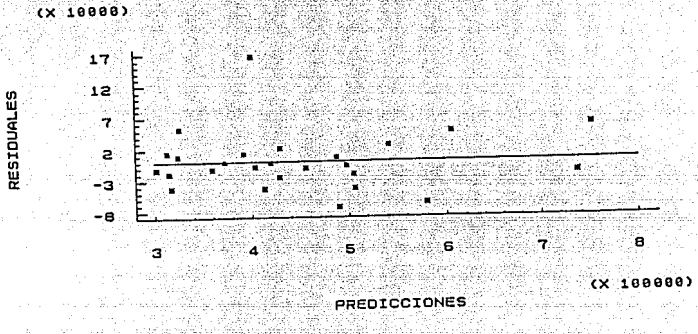
PA es la producción del arroz
SCA es la superficie cosechada de arroz

Lo que refleja la ec. 5.10 es que la producción del arroz está en función de su superficie cosechada y los fertilizantes.

Por otro lado, observando la gráfica correspondiente al análisis de

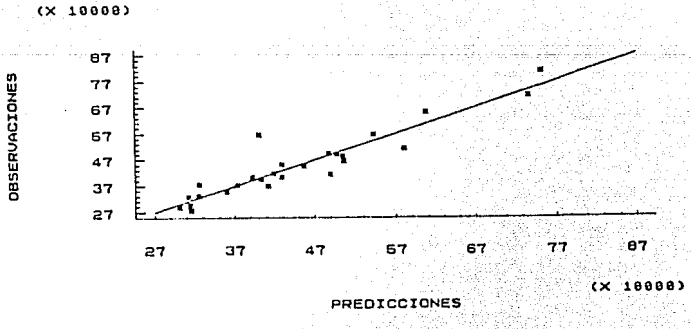
residuales (figura 5.1) corroboramos que en efecto la ec. 5.10 es un buen modelo (no se presenta una tendencia marcada).

FIGURA 5.1 Produccion del arroz



Finalmente en la figura 5.2 mostramos el diagrama de los valores ajustados contra los reales.

FIGURA 5.2 Produccion del Arroz



Para cada uno de los otros productos proponemos los tres modelos correspondientes y elegimos aquel que tenga la R^2 más alta (tabla 5.3):

TABLA 5.3 a) Model fitting results for: LnPRODUCCION DEL FRIJOL

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	-5.312109	1.912184	-2.7780	0.0110
LnLLUVIA	0.271922	0.166251	1.6356	0.1161
LnCREDITO	0.225127	0.02408	9.3491	0.0000
LnSUP. COSECHADA FRIJOL	1.04457	0.095818	10.9016	0.0000

R-Squared = 0.88	SE=	0.073679	MAE=	0.049449	DurbWat=	1.845
Previously: 0.0		0.182327		0.129046		1.452

26 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

b) Model fitting results for: LnPRODUCCION DEL MAIZ

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	-0.93138	3.333912	-0.2794	0.7825
LnFERTILIZANTES	0.248592	0.029009	8.5694	0.0000
LnSUP. COSECHADA MAIZ	0.864872	0.212194	4.0759	0.0005

R-SQ. (ADJ.) = 0.80	SE=	0.112096	MAE=	0.083866	DurbWat=	0.555
Previously: 0.0		0.073679		0.049449		1.845

26 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

c) Model fitting results for: LnPRODUCCION DEL TRIGO

ndependent variable	coefficient	std. error	t-value	sig.level
ONSTANT	-4.662305	1.272441	-3.6641	0.0013
nSUP. COSECHADA TRIGO	1.363286	0.093812	14.5325	0.0000
nTIEMPO	0.339471	0.017227	19.7060	0.0000

-Squared = 0.98	SE=	0.075502	MAE=	0.061249	DurbWat=	1.990
reviously: 0.0000		0.000000		0.000000		0.000

6 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

Esto es,

$$LnPFR = -5.3 + 0.27LnLL + 0.22LnC + LnSCFR \quad (5.11)$$

$$LnPTR = -3.9 + 0.40LnF + 0.98LnSCTR \quad (5.12)$$

$$LnPM = 0.93 + 0.25LnF + 0.86LnSCM \quad (5.13)$$

donde

PFR	es la prod. del frijol
SCFR	es la superficie cosechada de frijol
SCTR	es la superficie cosechada de trigo
PM	es la producción de maíz
SCM	es la superficie cosechada de maíz

Las tablas 5.2a, 5.3a, b y c, muestran las variables explicativas de la producción de cada grano básico en términos del rendimiento, que resumido queda:

TABLA II

PRODUCCION	VAR. EXPLICATIVAS	R^2
Arroz	Sup. Cosechada Arroz Fertilizantes	0.85
Frijol	Lluvia, Crédito, Sup. Cosechada Frijol	0.88
Maíz	Fertilizantes, Sup. Cosechada Maíz	0.80
Trigo	Tiempo Sup. Cosechada Trigo	0.98

Lo que dicen las ecuaciones 5.10, 5.11, 5.12, 5.13 y la tabla II es:

a) La relación entre la producción de cada uno de los productos respecto a las variables seleccionadas es intrínsicamente lineal.

b) No se observa alguna tendencia en la producción.

c) La lluvia sigue teniendo poca influencia en la producción.

d) Las variables explicatorias son consistentes.

5.3.4 INFLUENCIA EN LA SUPERFICIE COSECHADA DEL PRECIO

Consideremos el diagrama 5.3. Primero trataremos de explicar la superficie cosechada en términos de precios de garantía, lluvia y crédito con los retrasos de datos correspondientes. Los resultados son los siguientes (tabla 5.4):

TABLA 5.4 a) Model fitting results for: SUPERFICIE COSECHADA ARROZ

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	1.455187E5	7394.317571	19.6798	0.0000
PRECIO GARANTIA 0 RETRASO	-9.306271	4.300331	-2.1641	0.0427
PRECIO GARANTIA 1 RETRASO	16.451079	7.157087	2.2986	0.0324

R-Squared = 0.24	SE=	27711.417721	MAE=	17791.068582	DurbWat= 1.744
Previously: 0.0		515877.688499		395257.092978	1.203

23 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

b) Model fitting results for: SUPERFICIE COSECHADA DE FRIJOL

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	2.034709E6	1.135917E5	17.9125	0.0000
CREDITO CON 0 RETRASO	-13.224314	4.74166	-2.7890	0.0110

R-Squared = 0.27	SE=	229238.703746	MAE=	164210.930653	DurbWat= 1.791
Previously: 0.0		27711.417721		17791.068582	1.744

23 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

c) Model fitting results for: SUPERFICIE COSECHADA DE MAIZ

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	9.883663E6	1.339544E6	7.3784	0.0000
CREDITO CON 1 RETRASO	-31.196495	10.870279	-2.8699	0.0095
LLUVIA CON 3 RETRASOS	-2585.513533	1656.083108	-1.5612	0.1342

R-Squared = 0.35	SE=	540932.419680	MAE=	390773.912236	DurbWat= 2.270
Previously: 0.0		229238.703746		164210.930653	1.791

23 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

d) Model fitting results for: SUPERFICIE COSECHADA DE TRIGO

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	7.371465E5	2.174035E4	33.9068	0.0000
PRECIO GARANTIA DEL TRIGO	12.990525	2.20015	5.9044	0.0000

R-Squared = 0.62 SE= 90662.232217 MAE= 66756.697162 DurbWat= 1.834
Previously: 0.0 540932.419680 390773.912236 2.270
23 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

La siguiente tabla (resumen de las anteriores) muestra las variables más significativas de la superficie cosechada:

TABLA III

SUPERFICIE COSECHADA DE	VAR. SIGNIFICATIVAS	R^2
Arroz	Precio de Garantía (0) Precio de Garantía (1)	0.24
Frijol	Crédito (0)	0.27
Maíz	Crédito (1) Lluvia (3)	0.35
Trigo	Precio de Garantía (0)	0.62

donde los números entre paréntesis indican retrasos.

De esta tabla concluimos que:

a) En los cuatro casos las R^2 son muy pequeñas,

b) Antes (tabla I) habíamos explicado bien la productividad con asistencia técnica, y ahora no lo podemos hacer con la superficie cosechada.

c) La influencia del precio en la superficie cosechada no es clara, al grado de que para el maíz y frijol no aparecen.

5.3.5 INFLUENCIA DE FACTORES ECONOMICOS EN LA SUPERFICIE COSECHADA

Por último analicemos el diagrama 5.4. Utilizando al precio de garantía como un intermediario para la selección de variables, tenemos (tabla 5.5):

TABLA 5.5 a) Stepwise Selection for PRECIO DE GARANTIA ARROZ

election: Forward	Maximum steps: 500			F-to-enter: 2.00	
ontrol: Manual	Step: 1			F-to-remove: 2.00	
-squared: .77349	Adjusted: .76364		MSE: 3.49381E7	d.f.: 23	
ariables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. CONS. APARENTE(0)	0.05912	78.5404	2. CONAPA1.var1	.1461	.4798
			3. DEPENA1.var1	.2343	1.2781

b) Stepwise Selection for PRECIO DE GARANTIA FRIJOL

election: Forward	Maximum steps: 500			F-to-enter: 2.00	
ontrol: Manual	Step: 1			F-to-remove: 2.00	
-squared: .41903	Adjusted: .39377		MSE: 8.29263E7	d.f.: 23	
ariables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. PIB	0.00922	16.5889			

c) Stepwise Selection for PRECIO DE GARANTIA MAIZ

election: Forward	Maximum steps: 500			F-to-enter: 2.00	
ontrol: Manual	Step: 1			F-to-remove: 2.00	
-squared: .31157	Adjusted: .28164		MSE: 3.76909E7	d.f.: 23	
ariables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. PIB	0.00492	10.4094	2. DEPENMO.var1	.1267	.3589

d) Stepwise Selection for PRECIO DE GARANTIA TRIGO

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 2 — F-to-remove: 2.00

R-squared: .63861 — Adjusted: .60575 — MSE: 1.18245E7 — d.f.: 22

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
2. CONS. APARENTE(0)	0.00398	38.7656	1. CONAPT1.var1	.0569	.0683
3. DEPENDENCIA(1)	-13140.3	6.9490			

Resumiendo la tabla 5.5 obtenemos:

TABLA IV

PRECIO DE GARANTIA	VAR. SIGNIFICATIVAS	R^2
ARROZ	Consumo Aparente (0) Consumo Aparente (1)	0.76
FRIJOL	PIB	0.42
MAIZ	PIB	0.31
TRIGO	Consumo Aparente (0) Dependencia (1)	0.64

Lo que se puede decir de esta tabla es:

a) El comportamiento del maíz y del frijol es distinto que el del arroz y el trigo.

b) Es más fácil predecir el comportamiento del arroz y el trigo (obsérvese las R^2 correspondientes)

c) El deficit no lleva a una política de estímulo a la producción vía precios.

c) El precio de garantía del arroz y el trigo se estimula por la cantidad de población total, mientras que el del maíz y el frijol queda explicado por la economía.

De igual manera que la tabla 5.5 obtenemos la tabla 5.6 utilizando a la superficie cosechada como intermediario:

TABLA 5.6 a) Stepwise Selection for SUP. COSECHADA ARROZ

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 9 — F-to-remove: 2.00

R-squared: .82710 — Adjusted: .78160 — MSE: 1.90545E8 — d.f.: 19

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. DEPENDENCIA(1)	89529.4	2.1113	3. CONAPA1.var1	.1752	.5701
2. DEPENDENCIA(0)	-101230.	4.0510			
4. CONS. APARENTE(0)	0.17590	17.2424			
5. PIB	-0.04188	26.3375			
6. SALARIO	1424.95	26.1947			

b) Stepwise Selection for SUP. COSECHADA FRIJOL

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 2 — F-to-remove: 2.00

R-squared: .51693 — Adjusted: .47302 — MSE: 3.36617E10 — d.f.: 22

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
3. CONS. APARENTE	0.69647	18.2475	1. DEPENF0.var1	.0087	.0016
5. PIB	-0.24159	18.8131	2. DEPENF1.var1	.2717	1.6738
			4. CONAPF1.var1	.1603	.5540
			6. SALAR25.var1	.1809	.7104

c) Stepwise Selection for SUP. COSECHADA MAIZ

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 1 — F-to-remove: 2.00

R-squared: .08715 — Adjusted: .04747 — MSE: 4.10716E11 — d.f.: 23

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
4. CONS. APARENTE(1)	-0.04403	2.1959	1. DEPENMO.var1	.1358	.4133
			2. DEPENM1.var1	.0814	.1466
			3. CONAPMO.var1	.1114	.2764
			5. PIB25.var1	.0319	.0224
			6 SALAR25.var1	.0555	.0679

d) Stepwise Selection for SUP. COSECHADA TRIGO

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 4 — F-to-remove: 2.00

R-squared: .90465 — Adjusted: .88559 — MSE: 2.20838E9 — d.f.: 20

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. DEPENDENCIA(0)	-582264.	58.7454	2. DEPENT1.var1	.0474	.0428
3. CONS. APARENTE(1)	-0.03508	2.3360	5. PIB25.var1	.1520	.4494
4. CONS. APARENTE(0)	0.14895	55.8088			
6. SALARIO	-2187.12	7.3237			

Resumiendo la última tabla, tenemos:

TABLA V

SUP. COSECHADA DE	VAR. SIGNIFICATIVAS	R^2
Arroz	Salario, PIB, Consumo Aparente (0) Dependencia (0) Dependencia (1)	0.82
Frijol	Consumo Aparente (0) PIB	0.52
Maíz	Consumo Aparente (1)	0.08
Trigo	Salario Consumo Aparente (0) Consumo Aparente (1) Dependencia (0)	0.90

Lo que podemos decir de esta tabla es:

a) El consumo contribuye en la cantidad de la superficie cosechada pero no mediante precios, sino a través de la totalidad de la población.

b) El salario influye en la superficie cosechada del trigo y arroz.

c) El maíz no responde a la afirmación del inciso b.

5.3.6 SINTESIS

Retomando cada una de las variables más significativas en los diagramas 5.2, 5.3, 5.4, y tomando la producción como la variable dependiente para cada uno de los productos, obtenemos la tabla 5.7:

TABLA 5.7 a) Stepwise Selection for PRODUCCION DE ARROZ

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 4 — F-to-remove: 2.00

R-squared: .80170 — Adjusted: .76204 — MSE: 4.0414E9 — d.f.: 20

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. PRECIO GARANTIA	-5.79658	6.3732	3. CONAPA1.var1	.0336	.0214
2. CONS. APARENTE(0)	0.91847	22.6082	5. DEPENA1.var1	.2428	1.1906
4. DEPENDENCIA(0)	-434180.	4.6488	6. LLUV25.var1	.0174	.0057
7. TIEMPO	4584.07	2.5721			

b) Stepwise Selection for PRODUCCION DE FRIJOL

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 3 — F-to-remove: 2.00

R-squared: .56866 — Adjusted: .50704 — MSE: 1.27917E10 — d.f.: 21

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
2. SUP. COS. DE MAIZ	0.14706	16.1634	1. CREDIT25.var1	.2616	1.4689
3. CONS. APARENTE(1)	0.21771	4.9313	4. LLUV25.var1	.1309	.3488
5. TIEMPO	7252.67	3.6943			

c) Stepwise Selection for PRODUCCION DE MAIZ

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 3 — F-to-remove: 2.00

R-squared: .72502 — Adjusted: .68574 — MSE: 1.58869E12 — d.f.: 21

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
1. PROD. TRIGO	-1.13064	2.7427	3. FERT25.var1	.1481	.4482
2. P. GARANTIA MAIZ	94.4522	6.5058	4. LLUV25.var1	.1424	.4141
5. TIEMPO	282746.	15.3448			

d) Stepwise Selection for PRODUCCION DE TRIGO

Selection: Forward — Maximum steps: 500 — F-to-enter: 2.00
Control: Manual — Step: 1 — F-to-remove: 2.00

R-squared: .88676 — Adjusted: .88183 — MSE: 1.23597E11 — d.f.: 23

Variables in Model	Coeff.	F-Remove	Variables Not in Model	P.Corr.	F-Enter
6. CONS. APARENTE(0)	0.81364	180.1035	1. PRNAA25.var1	.0533	.0826
			2. CONAPT1.var1	.1543	.5364
			3. FERT25.var1	.0382	.0321
			4. LLUV25.var1	.0630	.0878
			5. TIEM25.var1	.1872	.7994
			7. DEPENA1.var1	.1853	.7822

Como un resumen final, tenemos entonces que:

TABLA VI

PRODUCCION	VAR. SIGNIFICATIVAS	R^2
Arroz	Consumo Aparente Arroz (0) Dependencia (0) Precio de Garantía Arroz Tiempo	0.80
Frijol	Tiempo Sup. Cosechada de Maíz Consumo Aparente Frijol	0.56
Maíz	Tiempo Precio de Garantía Maíz Producción de Trigo	0.72
Trigo	Consumo Aparente Trigo	0.56

De aquí concluimos que:

a) No se nota la influencia de la tecnificación en la producción.
b) La producción del arroz y el trigo responden al consumo.
c) Existe una tendencia en la producción del maíz y frijol.

Las figuras 5.3 a la 5.14 muestran los residuales de cada producto, y sus respectivas gráficas de ajustes contra los datos reales. También mostramos los diagramas de la posible tendencia de producción que pueden tener los granos básicos.

FIGURA 5.3 Produccion del Arroz

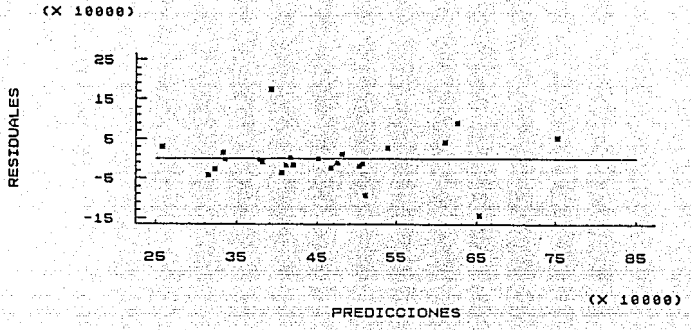


FIGURA 5.4 Produccion del Arroz

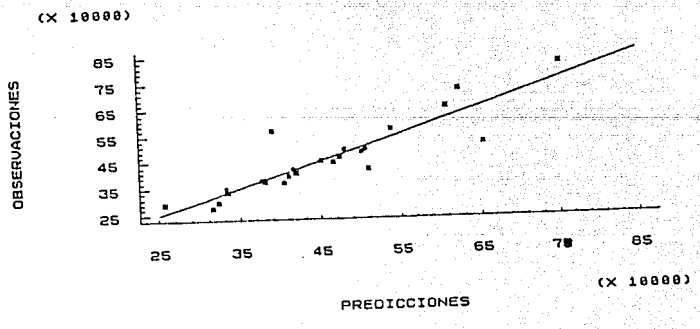


FIGURA 5.5 Produccion del Arroz

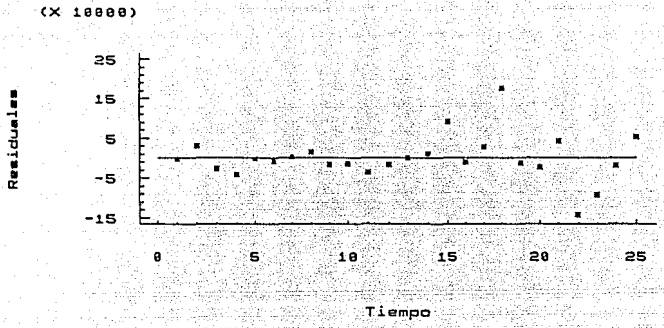


FIGURA 5.6 Produccion del frijol

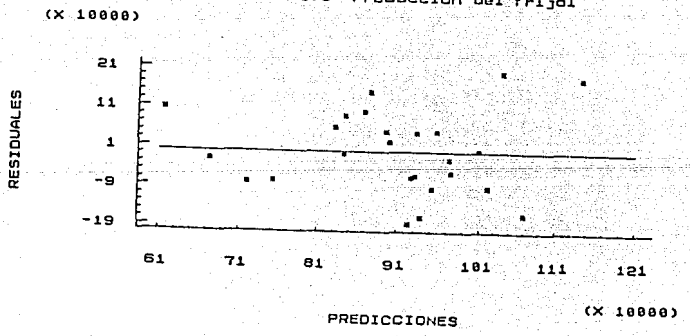


FIGURA 5.7 Produccion del Frijol

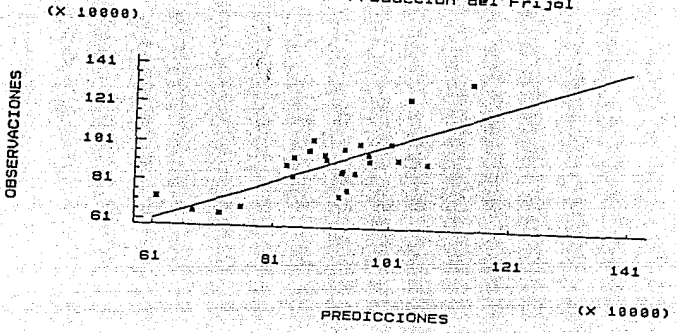


FIGURA 5.8 Produccion del frijol

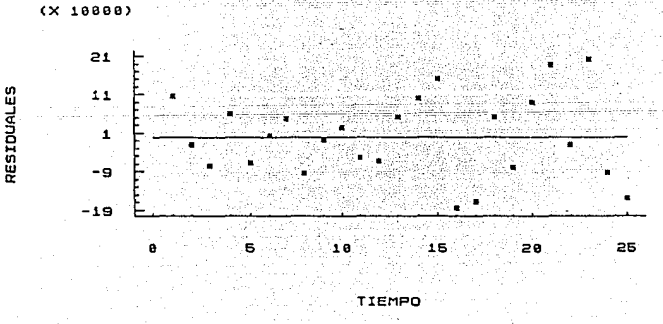


FIGURA 5.9 Produccion del maiz

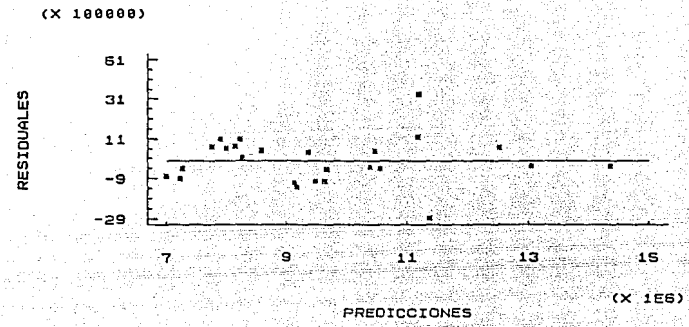


FIGURA 5.10 Produccion del maiz

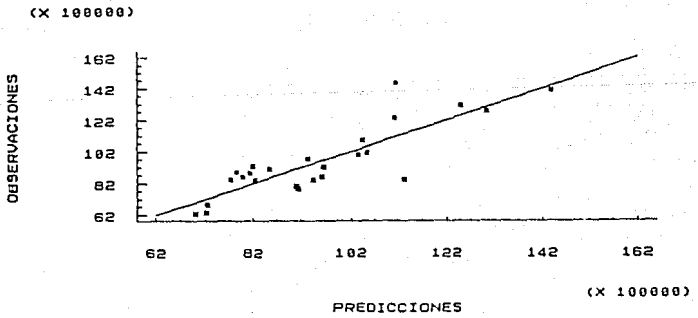


FIGURA 5.11 Produccion del Maiz

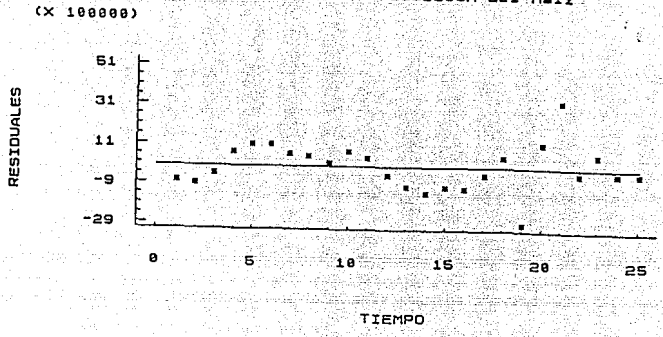


FIGURA 5.12 Produccion del Trigo

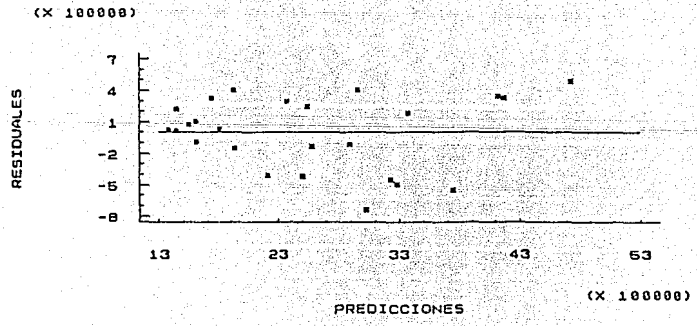


FIGURA 5.13 Produccion del Trigo

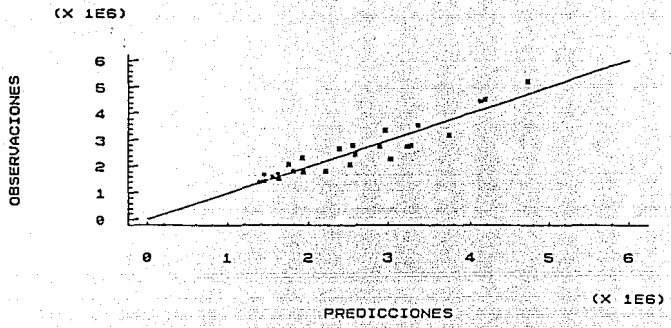
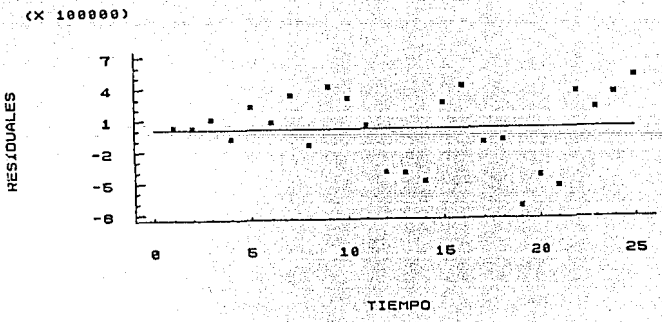


FIGURA 5.14 Produccion del Trigo



El reflejo de la tabla VI es que la explicación de la producción del frijol y el trigo es muy pobre. Lo que haremos a continuación es incluir dentro de las posibles variables independientes a la producción de forrajeros y oleaginosas. Los resultados son:

TABLA VII

PRODUCCION	VAR. EXPLICATIVAS	R^2
Arroz	Consumo, Dependencia Prod. de Forrajeros Tiempo	0.94
Frijol	Consumo Aparente	0.74
Maíz	Consumo Aparente Dependencia Precio garantía Prod. de Frijol Prod. de Trigo Salario	0.92
Trigo	Consumo Aparente Crédito, Dependencia, Prod. de Arroz Prod. de Maíz Prod. de Oleaginosas Tiempo	0.99

Lo que se puede deducir de esta tabla es lo siguiente:

a) La producción responde a una demanda (a mayor demanda, mayor producción).

b) No existe influencia del precio de garantía en la producción.

c) No existe una tendencia.

d) No influye el apoyo técnico.

e) La lluvia no afecta.

f) No es posible apoyar el diseño de una política para la toma de decisiones.

g) Deben de existir variables desconocidas para inducir la producción.

i) La R^2 no implica nada.

Para seguir con otra sorpresa presentamos las siguientes matrices de correlación (tabla 5.8):

Correlation matrix for coefficient estimates de la prod. del arroz

	CONSTANT	CONS. APARENTE	CREDITO	DEPENDENCIA
ONSTANT	1.0000	-.5211	-.3282	.3691
ONS. APARENTE	-.5211	1.0000	-.5227	-.5256
REDITO	-.3282	-.5227	1.0000	-.0811
EPENDENCIA	.3691	-.5256	-.0811	1.0000

Correlation matrix for coefficient estimates para la prod. del frijol

	CONSTANT	CONS. APARENTE	CREDITO	DEPENDENCIA
CONSTANT	1.0000	-.7907	-.0275	.2389
CONS. APARENTE	-.7907	1.0000	-.5358	-.1056
CREDITO	-.0275	-.5358	1.0000	-.3348
DEPENDENCIA	.2389	-.1056	-.3348	1.0000

Correlation matrix for coefficient estimates para la prod. del maiz

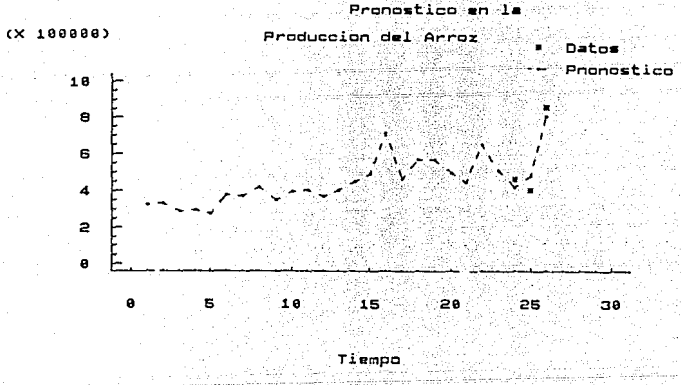
	CONSTANT	CONS. APARENTE	CREDITO	DEPENDENCIA
CONSTANT	1.0000	-.4816	-.1004	.0284
CONS. APARENTE	-.4816	1.0000	-.7801	.0513
CREDITO	-.1004	-.7801	1.0000	-.1911
DEPENDENCIA	.0284	.0513	-.1911	1.0000

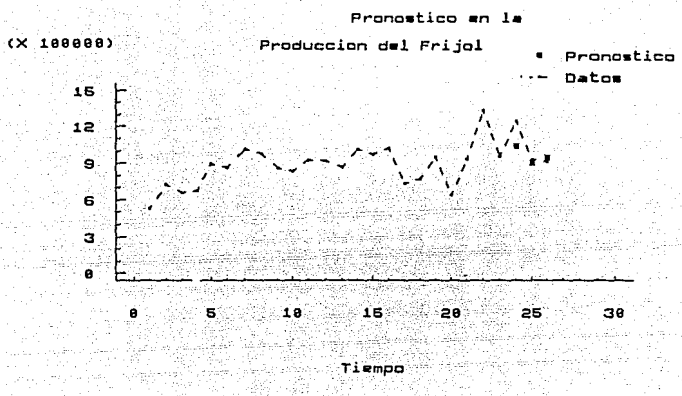
Correlation matrix for coefficient estimates para la prod. del trigo

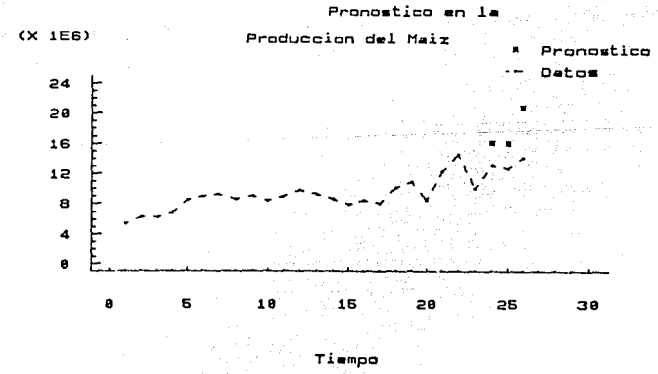
	CONSTANT	CONS. APARENTE	CREDITO	DEPENDENCIA
CONSTANT	1.0000	-.5773	.1920	-.0634
CONS. APARENTE	-.5773	1.0000	-.8763	.3276
CREDITO	.1920	-.8763	1.0000	-.5560
DEPENDENCIA	-.0634	.3276	-.5560	1.0000

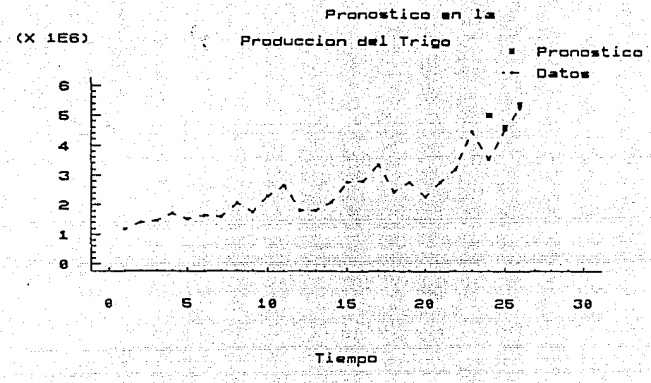
El resultado más sorprendente que se puede observar es la correlación negativa existente entre el crédito y la dependencia. Esto dice, a mayor dependencia, menor crédito.

Finalmente, utilizando la tabla VII, elaboraremos los prónosticos de cada uno de los productos:




Pronostico en la
(X 100000)
Produccion del Frijol
Pronostico
Datos
15
12
9
6
3
0
0
5
10
15
20
25
30
Tiempo


Pronostico en la
(X 1E6)
Produccion del Maiz
Pronostico
Datos
24
20
16
12
8
4
0
0
5
10
15
20
25
30
Tiempo


Pronostico en la
Produccion del Trigo
(X 1E6)
Pronostico
Datos
6
5
4
3
2
1
0
0
5
10
15
20
25
30
Tiempo

CONCLUSIONES

El énfasis de este trabajo ha sido .no tanto revisar los fundamentos matemáticos de la regresión, sino más bien el hacer un estudio en el que partiendo de los fundamentos básicos se fuera avanzando hacia el terreno de la aplicación. En este caso encaminado a la predicción de la producción de los granos básicos (arroz, frijol, maíz y trigo).

Entre los resultados o puntos tratados que se consideran de mayor importancia, destacan los siguientes:

i) Se explica lo que son las relaciones de causalidad y de asociación entre variables.

ii) La inclusión de algunos tópicos tales como la correlación espuria , efecto de retraso de datos, método forward y backward, que en ocasiones los libros de tipo matemático no plantean.

iii) El paquete utilizado (statgraphics) es más que suficiente para el desarrollo del trabajo.

iv) Los pronósticos que se han obtenido son buenos, pero que desgraciadamente no ayudan en mucho para una política de decisiones.

v) La R^2 no es de mucha utilidad si no se le da la interpretación adecuada.

Para mejorar este trabajo se proponen dos alternativas posibles:

a) Estudiar la producción por riego y temporal, y luego integrarlos.

b) Particionar el periodo de estudio para este trabajo (1960-1985) por periodos más cortos, y posteriormente integrarlos.

BIBLIOGRAFIA

1. DAMODAR, GUAJARATI (1989), Econometría Básica, Mc Graw Hill, México.

2. DOUGLAS C. MONTGOMERY, ELIZABETH A. PECK (1982), Introduction to Linear Regression Analysis, Jhon Wiley and Sons.

3. ERWIN, KREYZING (1970), Introducción a la Estadística Matemática Principios y Métodos, Limusa, México.

4. SYPROS MACKRIDAKIS, STEVEN C. WHEELWRIGHT, VICTOR E. McGEE (1986), Forecasting: methods and Aplications, John Wiley and Sons.

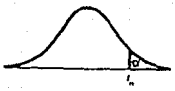
5. WILLIAM, L. HAYAS y ROBERT, L. WINKLER (1971), Statistics: Probability, Inference and Decision, International Series in Decission Processes.

6. SECRETARIA de Agricultura y Recursos Hidraúlicos, Estadísticas Básicas 1960-1986 para la Planeación del Desarrollo Rural Integral, V.1.

7. NACIONAL Financiera (1988), La Economía Mexicana en Cifras, Segunda Edición.

Apéndice A

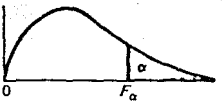
Puntos porcentuales de las distribuciones t



$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	g.l.
3.078	6.314	12.706	31.821	63.657	1
1.886	2.920	4.303	6.965	9.925	2
1.638	2.353	3.182	4.541	5.841	3
1.533	2.132	2.776	3.747	4.604	4
1.476	2.015	2.571	3.365	4.032	5
1.440	1.943	2.447	3.143	3.707	6
1.415	1.895	2.365	2.998	3.499	7
1.397	1.860	2.306	2.896	3.355	8
1.383	1.833	2.262	2.821	3.250	9
1.372	1.812	2.228	2.764	3.169	10
1.363	1.796	2.201	2.718	3.106	11
1.356	1.782	2.179	2.681	3.055	12
1.350	1.771	2.160	2.650	3.012	13
1.345	1.761	2.145	2.624	2.977	14
1.341	1.753	2.131	2.602	2.947	15
1.337	1.746	2.120	2.583	2.921	16
1.333	1.740	2.110	2.567	2.898	17
1.330	1.734	2.101	2.552	2.878	18
1.328	1.729	2.093	2.539	2.861	19
1.325	1.725	2.086	2.528	2.845	20
1.323	1.721	2.080	2.518	2.831	21
1.321	1.717	2.074	2.508	2.819	22
1.319	1.714	2.069	2.500	2.807	23
1.318	1.711	2.064	2.492	2.797	24
1.316	1.708	2.060	2.485	2.787	25
1.315	1.706	2.056	2.479	2.779	26
1.314	1.703	2.052	2.473	2.771	27
1.313	1.701	2.048	2.467	2.763	28
1.311	1.699	2.045	2.462	2.756	29
1.282	1.645	1.960	2.326	2.576	inf.

Apéndice B

Puntos porcentuales de las distribuciones F



g.l. del denominador		*g.l. del numerador*								
	α	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	.100	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86
	.050	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5
	.025	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3
	.010	4052	4999.5	5403	5625	5764	5859	5928	5982	6022
	.005	16211	20000	21615	22500	23056	23437	23715	23925	24091
2	.100	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.050	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38
	.025	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39
	.010	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39
	.005	198.5	199.0	199.2	199.2	199.3	199.3	199.4	199.4	199.4
3	.100	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.050	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.025	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47
	.010	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35
	.005	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88
4	.100	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.050	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.025	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.010	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66
	.005	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14
5	.100	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.050	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.025	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.010	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16
	.005	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77
6	.100	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.050	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	.025	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	.010	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.005	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39
7	.100	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.050	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	.025	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	.010	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.005	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51

(*Continuación*)

F_α

g.l. del numerador											*g.l. del denominador*
10	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	∞	α	
60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33	.100	1
241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3	.050	
968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018	.025	
6056	6106	6157	6209	6235	6261	6287	6313	6339	6366	.010	
24224	24426	24630	24836	24940	25044	25148	25253	25359	25465	.005	
9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49	.100	2
19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50	.050	
39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50	.025	
99.40	99.42	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50	.010	
199.4	199.4	199.4	199.4	199.5	199.5	199.5	199.5	199.5	199.5	.005	
5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13	.100	3
8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53	.050	
14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90	.025	
27.23	27.05	26.87	26.69	26.60	26.50	26.41	26.32	26.22	26.13	.010	
43.69	43.39	43.08	42.78	42.62	42.47	42.31	42.15	41.99	41.83	.005	
3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76	.100	4
5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63	.050	
8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26	.025	
14.55	14.37	14.20	14.02	13.93	13.84	13.75	13.65	13.56	13.46	.010	
20.97	20.70	20.44	20.17	20.03	19.89	19.75	19.61	19.47	19.32	.005	
3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10	.100	5
4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36	.050	
6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02	.025	
10.05	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02	.010	
13.62	13.38	13.15	12.90	12.78	12.66	12.53	12.40	12.27	12.14	.005	
2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72	.100	6
4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67	.050	
5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85	.025	
7.87	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88	.010	
10.25	10.03	9.81	9.59	9.47	9.36	9.24	9.12	9.00	8.88	.005	
2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47	.100	7
3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23	.050	
4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14	.025	
6.62	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65	.010	
8.38	8.18	7.97	7.75	7.65	7.53	7.42	7.31	7.19	7.08	.005	

(*Continuación*)

F_α

g.l. del denominador	*g.l. del numerador*									
	α	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
8	.100	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.050	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.025	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.010	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.005	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34
9	.100	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.050	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.025	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.010	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.005	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54
10	.100	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.050	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.025	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.010	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.005	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97
11	.100	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27
	.050	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
	.025	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59
	.010	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
	.005	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54
12	.100	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.050	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	.025	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.010	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.005	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20
13	.100	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16
	.050	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
	.025	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31
	.010	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
	.005	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94
14	.100	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12
	.050	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
	.025	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21
	.010	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
	.005	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72

(*Continuación*)

F_α

g.l. del numerador											g.l. del denominador
10	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	α		
2.54	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29	.100	8
3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93	.050	
4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67	.025	
5.81	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86	.010	
7.21	7.01	6.81	6.61	6.50	6.40	6.29	6.18	6.06	5.95	.005	
2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16	.100	9
3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71	.050	
3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33	.025	
5.26	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31	.010	
6.42	6.23	6.03	5.83	5.73	5.62	5.52	5.41	5.30	5.19	.005	
2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06	.100	10
2.98	2.91	2.85	2.74	2.77	2.70	2.66	2.62	2.58	2.54	.050	
3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08	.025	
4.85	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91	.010	
5.85	5.66	5.47	5.27	5.17	5.07	4.97	4.86	4.75	4.64	.005	
2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97	.100	11
2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40	.050	
3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88	.025	
4.54	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60	.010	
5.42	5.24	5.05	4.86	4.76	4.65	4.55	4.44	4.34	4.23	.005	
2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90	.100	12
2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30	.050	
3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72	.025	
4.30	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36	.010	
5.09	4.91	4.72	4.53	4.43	4.33	4.23	4.12	4.01	3.90	.005	
2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85	.100	13
2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21	.050	
3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60	.025	
4.10	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17	.010	
4.82	4.64	4.46	4.27	4.17	4.07	3.97	3.87	3.76	3.65	.005	
2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80	.100	14
2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13	.050	
3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49	.025	
3.94	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00	.010	
4.60	4.43	4.25	4.06	3.96	3.86	3.76	3.66	3.55	3.44	.005	

(*Continuación*)

F_α

g.l. del numerador											*g.l. del denominador*
10	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	∞	α	
1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57	.100	22
2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78	.050	
2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00	.025	
3.26	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31	.010	
3.70	3.54	3.36	3.18	3.08	2.98	2.88	2.77	2.66	2.55	.005	
1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55	.100	23
2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76	.050	
2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97	.025	
3.21	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26	.010	
3.64	3.47	3.30	3.12	3.02	2.92	2.82	2.71	2.60	2.48	.005	
1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53	.100	24
2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73	.050	
2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94	.025	
3.17	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21	.010	
3.59	3.42	3.25	3.06	2.97	2.87	2.77	2.66	2.55	2.43	.005	
1.87	1.82	1.77	1.72	1.69	1.66	1.63	1.59	1.56	1.52	.100	25
2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71	.050	
2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91	.025	
3.13	2.99	2.85	2.70	2.62	2.54	2.45	2.36	2.27	2.17	.010	
3.54	3.37	3.20	3.01	2.92	2.82	2.72	2.61	2.50	2.38	.005	
1.86	1.81	1.76	1.71	1.68	1.65	1.61	1.58	1.54	1.50	.100	26
2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69	.050	
2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88	.025	
3.09	2.96	2.81	2.66	2.58	2.50	2.42	2.33	2.23	2.13	.010	
3.49	3.33	3.15	2.97	2.87	2.77	2.67	2.56	2.45	2.33	.005	
1.85	1.80	1.75	1.70	1.67	1.64	1.60	1.57	1.53	1.49	.100	27
2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67	.050	
2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85	.025	
3.06	2.93	2.78	2.63	2.55	2.47	2.38	2.29	2.20	2.10	.010	
3.45	3.28	3.11	2.93	2.83	2.73	2.63	2.52	2.41	2.29	.005	
1.84	1.79	1.74	1.69	1.66	1.63	1.59	1.56	1.52	1.48	.100	28
2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65	.050	
2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83	.025	
3.03	2.90	2.75	2.60	2.52	2.44	2.35	2.26	2.17	2.06	.010	
3.41	3.25	3.07	2.89	2.79	2.69	2.59	2.48	2.37	2.25	.005	

(*Continuación*)

F_α

g.l. del denominador	α	*g.l. del numerador*								
		1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
22	.100	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93
	.050	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
	.025	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76
	.010	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
	.005	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81
23	.100	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92
	.050	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32
	.025	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73
	.010	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30
	.005	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75
24	.100	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.050	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.025	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.010	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.005	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69
25	.100	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89
	.050	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
	.025	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68
	.010	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22
	.005	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64
26	.100	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88
	.050	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
	.025	5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65
	.010	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
	.005	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60
27	.100	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87
	.050	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
	.025	5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63
	.010	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15
	.005	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56
28	.100	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87
	.050	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
	.025	5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61
	.010	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
	.005	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52

(*Continuación*)

F_α

g.l. del denominador	*g.l. del numerador*									
	α	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
15	.100	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	.050	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	.025	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	.010	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	.005	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54
16	.100	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06
	.050	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
	.025	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05
	.010	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
	.005	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38
17	.100	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03
	.050	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
	.025	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98
	.010	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
	.005	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25
18	.100	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00
	.050	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
	.025	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93
	.010	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
	.005	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14
19	.100	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98
	.050	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
	.025	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88
	.010	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
	.005	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04
20	.100	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.050	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.025	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.010	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.005	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96
21	.100	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95
	.050	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37
	.025	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80
	.010	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40
	.005	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88

(*Continuación*)

F_α

g.l. del numerador											*g.l. del denominador*
10	*12*	*15*	*20*	*24*	*30*	*40*	*60*	*120*	∞	α	
2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76	.100	15
2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07	.050	
3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40	.025	
3.80	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87	.010	
4.42	4.25	4.07	3.88	3.79	3.69	3.58	3.48	3.37	3.26	.005	
2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72	.100	16
2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01	.050	
2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32	.025	
3.69	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75	.010	
4.27	4.10	3.92	3.73	3.64	3.54	3.44	3.33	3.22	3.11	.005	
2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69	.100	17
2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96	.050	
2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25	.025	
3.59	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65	.010	
4.14	3.97	3.79	3.61	3.51	3.41	3.31	3.21	3.10	2.98	.005	
1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66	.100	18
2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92	.050	
2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19	.025	
3.51	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57	.010	
4.03	3.86	3.68	3.50	3.40	3.30	3.20	3.10	2.99	2.87	.005	
1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63	.100	19
2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88	.050	
2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13	.025	
3.43	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49	.010	
3.93	3.76	3.59	3.40	3.31	3.21	3.11	3.00	2.89	2.78	.005	
1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61	.100	20
2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84	.050	
2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09	.025	
3.37	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42	.010	
3.85	3.68	3.50	3.32	3.22	3.12	3.02	2.92	2.81	2.69	.005	
1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59	.100	21
2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81	.050	
2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04	.025	
3.31	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36	.010	
3.77	3.60	3.43	3.24	3.15	3.05	2.95	2.84	2.73	2.61	.005	

APENDICE C

PRECIO Y CONSUMO PER CAPITA DE GAS NATURAL DE 20 CIUDADES DE TEXAS

CIUDAD	PRECIO PROMEDIO (cent. por miles de pies cúbicos)	CONSUMO (miles de pies cúbicos)
AMARILLO	30	134
BORGER	31	112
DALHART	37	136
SHAMROCK	42	109
ROYALTY	43	105
TEXARKANA	45	87
CORPUS CHRISTI	50	56
PALESTINE	54	43
MARSHALL	54	77
IOWA PARK	57	35
PALO PINTO	58	65
MILLSAP	58	56
MEMPHIS	60	58
GRANGER	73	55
LLANO	88	49
BROWNSVILLE	89	39
MERCEDES	92	36
KARNES CITY	97	46
MATHIS	100	40
LA PYROR	102	42

APENDICE D

DATOS PARA EL EJEMPLO 3.1

OBSERVACION NUMERO	TIEMPO DE ENTREGA (Minutos)	NUMERO DE CASOS	DISTANCIA (Pies)
1	16.68	7	560
2	11.5	3	220
3	12.03	3	340
4	14.88	4	80
5	13.75	6	150
6	18.11	7	330
7	8	2	110
8	17.83	7	210
9	79.24	30	1460
10	21.5	5	605
11	40.33	16	688
12	21	10	215
13	13.5	4	255
14	19.75	6	462
15	24	9	448
16	29	10	776
17	15.35	6	200
18	19	7	132
19	9.5	3	36
20	35.1	17	770
21	17.9	10	140
22	52.32	26	810
23	18.75	9	450
24	19.83	8	635
25	10.75	4	150