UNIVERSIDAD NACIONAL AUTONOMA DE NICARAGUA, LEÓN
Facultad de Ciencias y Tecnologías
Departamento de Agroecología
Centro de Investigación en Ciencias Agrarias y Economia Aplicada

Manual #4:Manual de Econometria Básica

Basic Econometric Handbook

Dr Carlos Zúniga G

2013

CICAEA HTTP://CICAEA.UNANLEON.EDU.NI/INDEX.HTML

BASIC ECONOMETRY HANDBOOK

This guide is designed econometrics laboratory for economists users allowing them to know the tool to contribute to the socio-economic analysis to improve sector policies.

It is a guide used in Econometrics components provided in the school for Business Economics Christian Autonomous University of Nicaragua, UCAN.

The aim of this series of guides is that the user knows and applies the Eviews software tool and fundamentally that it develops skills for socio-economic analysis applied to the economy of primary, secondary and tertiary sectors.

The manual is organized in 8 laboratories which correspond to the first part of the component of Econometrics Economics schools. These labs contain a theoretical supported from Mandala basic text, and a practical explanation Eviews software usage. It is important to note that over time the versions of this software vary so that in practice some commands also vary. However, the user depending on the version you will need to find or implement that change based on the basic explanation offered in this manual.

We add an appendix that defines the interpretation of key test parameters Eviews and a calculation for tweens.

It is important to note that at present the applied science researchers work with software that allows to go beyond the econometric analysis to enveloped data analysis, however it is essential to start with this tool.

References

Mandala, G.S. Introduction to Econometric. ISBN 978-0-470-01512-4, 2009

Zúniga, G. Carlos A., (2011), Texto básico de economía agrícola: Su Importancia para el Desarrollo Local Sostenible. Brought to you by the University of Minnesota Department of Applied Economics and the University of Minnesota Libraries with cooperation from the Agricultural and Applied Economics Association. ISBN: 978-99964-0-049-0. Registro de propiedad intelectual No OL-019-2011. Disponible On Line en: http://purl.umn.edu/111604

MANUAL DE ECONOMETRIA BÁSICA

La presente guía de laboratorio de econometría está diseñada para usuarios no economistas que les permita conocer la herramienta para contribuir al análisis socio-económico para mejorar las políticas del sector.

Es una guía utilizada en los componentes de Econometría impartida en la carrera de Economía Empresarial de la Universidad Cristiana Autónoma de Nicaragua, UCAN.

El objetivo de esta serie de guías es que el usuario conozca y aplique la herramienta del software Eviews y fundamentalmente que desarrolle habilidades para el análisis socio económico aplicado a la economía del sector primario, secundario y terciario.

El manual está organizado en 8 laboratorios que corresponden a la primera parte del componente de Econometría de las carreras de Economía. Estos laboratorios contienen una parte teórica soportada del texto básico de Mandala, y una parte práctica explicativa del uso del software Eviews. Es importante aclarar que con el tiempo las versiones de este software varían haciendo que en la práctica algunos comandos también varíen. Sin embargo, el usuario en dependencia de la versión tendrá que encontrar o aplicar esa variación en base a la explicación básica ofrecida en este manual.

Agregamos un anexo que define la interpretación de los principales parámetros del Test de Eviews y una memoria de cálculo para hacer interpolaciones.

Es importante aclarar que en la actualidad los investigadores de las ciencias aplicadas trabajamos con software que permiten ir más allá del análisis econométrico para análisis de datos envolventes, sin embargo es fundamental partir de esta herramienta.

Fuente bibliográfica

Mandala, G.S. Introduction to Econometric. ISBN 978-0-470-01512-4, 2009

Zúniga, G. Carlos A., (2011), Texto básico de economía agrícola: Su Importancia para el Desarrollo Local Sostenible. Brought to you by the University of Minnesota Department of Applied Economics and the University of Minnesota Libraries with cooperation from the Agricultural and Applied Economics Association. ISBN: 978-99964-0-049-0. Registro de propiedad intelectual No OL-019-2011. Disponible On Line en: http://purl.umn.edu/111604

LABORATORIO # 1: INTRODUCCION A LA ECONOMETRIA

Carrera :
Asignatura : Econometría
Año lectivo : IV
Clase No : 1
Fecha :
Unidad I : Introducción a la Econometría
Docente : Dr. Carlos A. Zúniga González, Economista Agrícola

En esta primera unidad abordaremos los elementos básicos de la econometría, una exposición de los objetivos y metodología de la Econometría. Abordaremos mediante un análisis crítico de los niveles de significancia convencionales.

El material lo ordenaremos en dos bloques. Iniciaremos con el primer bloque planteando los aspectos teóricos y generales de la Econometría. Un segundo Bloque abordará los aspectos básicos para operar desde la computadora bajo el programa para economistas Eviews.

PRIMER BLOQUE

1. ¿ Qué es la Econometría?

Es la aplicación de métodos estadísticos y matemáticos al análisis de datos económicos, con el propósito de dar un contenido empírico a las teorías económicas y verificarlas o refutarlas.
Antes de poder realizar cualquier análisis estadístico con datos económicos es necesario una formulación matemática clara de la teoría económica pertinente. Bien, el problema al que nos vamos a enfrentar es el hecho de que la teoría económica rara vez informa sobre las formas funcionales.

2. Economía y modelos econométricos

La primer tarea que enfrenta un econometrista es la de formular un modelo econométrico. ¿Qué es un modelo? Un modelo es una representación simplificada de un proceso del mundo real. La elección de un modelo sencillo para explicar fenómenos complejos del mundo real conlleva dos criticas: a) El modelo está simplificado en exceso. b) Las suposiciones son poco realistas. En la práctica, en el modelo se incluirán todas las variables que se consideren relevantes para el propósito y todas las demás pasarán a un cesto llamado perturbación o distorsión. Esto lleva a distinguir entre un modelo económico y uno econométrico. Un modelo económico es un conjunto de suposiciones que describen en forma aproximada la conducta de una economía o de un sector económico. Un modelo econométrico consiste de lo siguiente:

- Un conjunto de ecuaciones de conducta que se derivan del modelo económico. Tales ecuaciones incluyen algunas variables observadas y ciertas distorsiones, es decir el conjunto de todas las variables consideradas como irrelevantes para el propósito de este modelo, así como todos los sucesos no previstos.
- Un enunciado con respecto a la existencia de errores de observación en las variables observadas.
- Una especificación de la distribución de probabilidad de las distorsiones y errores de medición.

Teniendo en cuenta estas consideraciones nuestro trabajo en este curso consistirá en probar la validez empírica, algunos economistas le denominan validación del modelo con el propósito de hacer pronósticos o bien para formular análisis de políticas. Un ejemplo sencillo es el modelo de la demanda de, por lo general el modelo econométrico consiste en:

- La ecuación de conducta $q = \alpha + \beta p + ц$ donde q es la cantidad demandada y p es el precio. En este caso, p y q son las variables observadas y ц es el término de distorsión.
- Una especificación de la distribución de probabilidades de ц, que dice que E (ц/p) = 0 y que los valores de ц es un término a partir de las diferentes observaciones son independientes y se distribuyen normalmente con una media cero y una varianza σ.
- Con estas especificaciones se procede a probar empíricamente la ley de la demanda, o la hipótesis de $\beta < 0$. Así mismo es posible utilizar una función de la demanda para hacer predicciones y formular políticas.

3. **Objetivos**

3.1 Formulación de modelos econométricos en una forma verificable empíricamente. Por lo general, existen varias maneras de formular un modelo econométrico a partir de un modelo económico, pues es preciso elegir la forma funcional, la especificación de la estructura estocástica de las variables, etc. Esta parte constituye el aspecto de especificación del trabajo econométrico.

3.2 Estimación y comprobación de estos modelos con los datos observados. Esta parte constituye el aspecto de inferencia del trabajo econométrico.

3.3 Uso de estos modelos para propósitos de predicción y formulación de políticas.

SEGUNDO BLOQUE

Eviews es un programa para economistas que le permite la aplicación de los procedimientos econométricos para validar un modelo económico. Es de mencionar que existen otros programas como SPSS que les brindan las mismas oportunidades para validar modelo econométrico.

En la ventana principal podemos observar los siguientes comandos FILE EDIT OBJECTS VIEW PROC QUICK OPTIONS WINDOW HELP si ubicamos el cursos en cada comando podemos observar un sub menú con sus respectivos comandos.

FILE ⟹ NEW ⟹ WORK FILE/DATA BASE/PROGRAM/TEX FILE
OPEN ⟹ WORK FILE/DATA BASE/PROGRAM/TEX FILE
SAVE
SAVE AS
CLOSE

De igual manera cada uno de los comando presentan un submenú.

¿Cómo crear un archivo de trabajo (WORK FILE)?
Paso 1

File/New/Workfile les aparecerá una ventana titulada Work file range, es decir rango del archivo de trabajo y observarán una lista de rangos: anuales, semi anuales, trimestrales, mensuales, semanales e irregulares. Esto significa que usted debe elegir un rango de los datos u observaciones con las que usted trabajará. En las casillas más abajo usted debe especificar la fecha en que inicia y la fecha en que finaliza; en el caso que los datos no tengan fecha deberá escribir el número de observaciones. Finalmente teclee enter y obtendrá la hoja de trabajo donde introducirá su base de datos.

PRACTIQUÉMOSLO

Mes	Ingresos por ventas (y) (miles de dólares)	Gastos en publicidad (x) (cientos de dólares)
1	3	1
2	4	2
3	2	3
4	6	4
5	8	5
6	9	6

Antecedentes históricos de producción y ventas de un determinado producto

Año	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987
Demanda	10	20	30	45	70	90	125	150	180	220	270

Demanda estacional de un producto determinado

Año	Invierno	Primavera	Verano	Otoño	Total
1977	2	3	4	1	
1978	5	6	7	2	
1979	7	10	10	3	
1980	10	17	16	2	
1981	13	20	28	9	
1982	19	34	34	3	
1983	27	39	48	11	
1984	26	44	58	22	
1985	38	51	70	21	
1986	44	67	81	28	
1987	51	79	107	33	

Paso 2: Declarar las variables

Una vez establecido el rango de la fecha y el número de observaciones con las que trabajará. Usted se ubica en la ventana principal de su archivo de trabajo, generalmente aparece sin nombre o untitled eso se debe a que no se ha guardado. De acuerdo a la base de datos que eligió para

practicar es el momento de declarar las variables. El nombre de la variable es opcional, sin embargo para efectos de homogeneizar de la primera base de datos elijamos (y) para declarar la variable ingreso de ventas y (x) gastos en publicidad con los siguientes comandos

Objects/New Objects les aparece una ventana titulada New Object en el cuadro de la parte izquierda seleccione con el mouse series, significa que el tipo de variable que estamos seleccionando es datos ocurridos en una fecha determinada del tiempo. En la parte de la derecha escribamos y para nombras a la variable ingresos de las ventas. Y repetimos los pasos para la siguiente variable. Cada vez que declaremos una variable en la ventana de Work File Untitle ira apareciendo junto con otras variables que aparecen por defecto como son resid (ц) y c (α) que es la constante.
Antes de iniciar a introducir los datos es conveniente guardar la base de datos en un diskette porque en los futuros trabajos utilizaremos estas mismas bases de datos. Para ello haga clip en el comando SAVE y ubique su diskette.

Paso 3: Introducción de los datos

Una vez que hemos declarado las variables procedemos a introducir los datos con los siguientes comandos:

SHOW y nos aparece una ventana donde escribiremos las variables en el orden que queremos en este caso (y) y (x) luego teclee aceptar y se ubicaran en la ventana del archivo de trabajo donde introducirán los datos con los que se supone validarán un modelo econométrico. Hay que tener cuidado que por defecto esta hoja de trabajo esta inactiva, es decir que no se puede introducir datos, para ello debe activarla con el comando Edit +/- , o sea haga clip con el Mouse.

EJERCITÉMONOS INTRODUCIENDO LAS BASES DE DATOS ANTERIORMENTE ENTREGADAS O PREVIAMENTE ENTREGADAS.

LABORATORIO # 2

Regresión Simple

Carrera :
Asignatura : Econometría
Año lectivo : IV
Clase No : 2
Fecha :
Unidad II : Análisis de Regresión Simple
Docente : Dr. Carlos Alberto Zúniga González

Objetivo

Analizar el modelo de regresión lineal simple con una variable explicada y una variable explicativa.

El análisis de regresión trata con la descripción y evaluación de la relación entre una variable determinada a menudo llamada explicada o dependiente y una o más variables adicionales muchas veces llamadas explicativas o independientes.

La regresión simple es cuando en el modelo solamente existe una sola variable explicativa. Supongamos el ejemplo de la base de datos de la primera tabla (y) representa las ventas y (x) los gastos en este caso se trata de determinar la relación entre los gastos consumidos por un lado, y el ingreso familiar, los activos financieros de la familia y el tamaño de ésta por otro.

El objetivo de estas relaciones supone varios objetivos. Es posible utilizarlas para:

- Analizar los efectos de políticas que suponen el cambio de las x´s individuales. En el ejemplo esto significa analizar el efecto de cambiar los gastos de publicidad sobre las ventas.
- Pronosticar el valor de y para un conjunto determinado de x´s
- Examinar si alguna de las x´s tiene un efecto importante sobre y.

Especificación de las relaciones

Las relaciones entre las variables pueden expresarse de dos maneras:

1. Determinística o matemática. Ejemplo
2.

$$y = 2{,}500 + 100_x - x^2$$

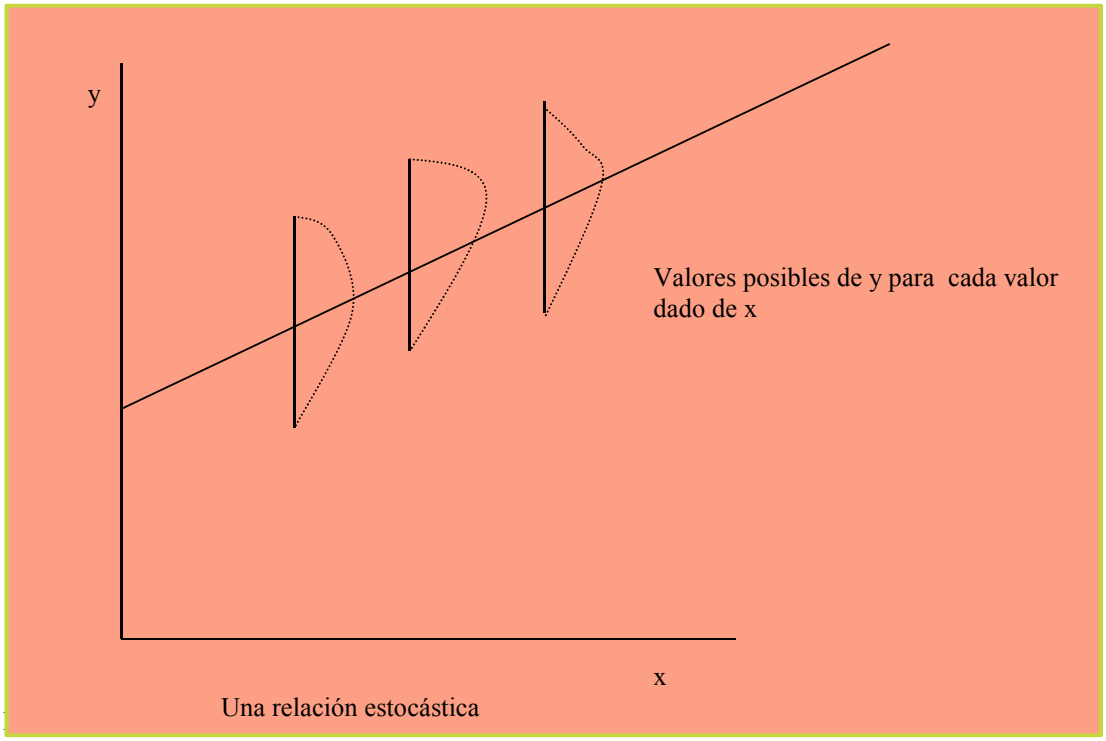
X	Y
0	2500
20	4100
50	5000
100	2500

3. Una relación Determinística. Por ejemplo $\boldsymbol{y = 2,500 + 100_x - x^2 + \mu}$ donde u=+500 con probabilidad de1/2 y u= - 500 con probabilidad de ½

X	Y		
1	2000	o	3000
21	3600	o	4600
51	4500	o	5500
100	2500	o	3000

El término error u tiene cuando tiene una distribución continua o normal con media cero y Varianza 1, entonces para cada valor de x se tiene una distribución normal para (y) y puede ser cualquier observación de esta distribución.

Para la relación $\boldsymbol{y = 2 + X + \mu}$ donde el término error es N(0,1), entonces para cada valor de x, y tendrá una distribución normal.



Una relación estocástica

La función $y = \alpha + \beta_x + \mu$. En esta función α y β son los parámetros, u se conoce como error o perturbación, tiene una distribución conocida de probabilidad. $y = \beta_x + \mu$ se conoce como parte determinística y u componente estocástico.

El objetivo en una ecuación de este tipo es obtener estimaciones de los parámetros desconocidos dadas n observaciones.. Para lograr esto es necesario hacer algunas suposiciones sobre los términos de error u.

1 Media cero. $E(u_i)=0$ para todo i
2 Varianza común. Varianza var = σ^2 para todo i
3 Independencia. u_i y u_j son independientes para todo $i \neq j$
4 Independencia. x_i y u_j son independientes para todo $i \neq j$
5 Normalidad.

Tabla 3.2

Observación	x	y
1	10	11
2	7	10
3	10	12
4	5	6
5	8	10
6	8	7
7	6	9
8	7	10
9	9	11
10	10	10

Los datos que contiene esta tabla son para 10 trabajadores x= horas de trabajo y= producción

LABORATORIO # 3

Regresión Simple

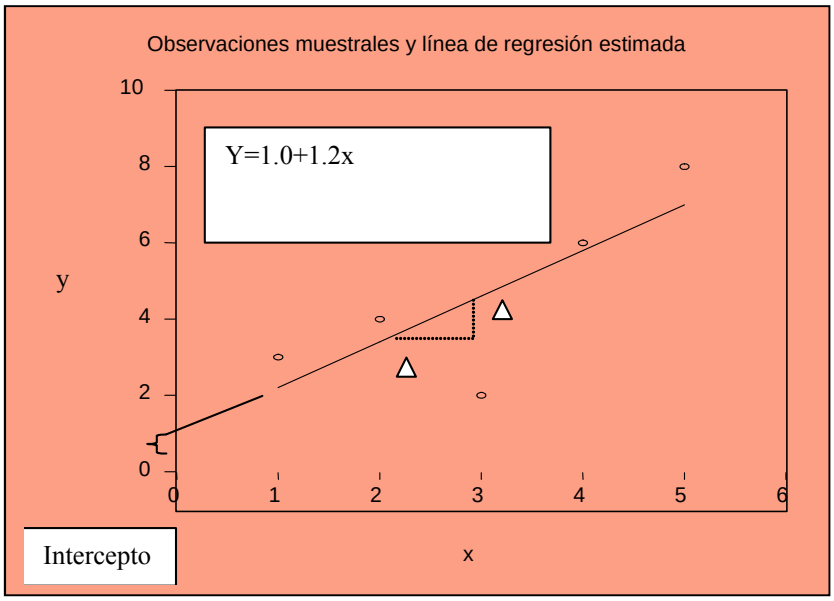
Carrera :
Asignatura : Econometría
Año lectivo : III
Laboratorio No : 3
Clase No : 3
Fecha :
Unidad II : Análisis de Regresión Simple
Docente : Dr. Carlos Alberto Zúniga González

Objetivo

Analizar el modelo de regresión lineal simple con una variable explicada y una variable explicativa.

Método de Mínimo Cuadrados

El método de mínimos cuadrados ordinarios (OLS) es uno de los métodos más usados para estimar los parámetros α y β. El método se basa en el principio de elegir α y β de tal modo que $\sum \mu_i^2$ sea mínimo. Es decir la suma de cuadrados de los errores de predicción es la mínima.



El gráfico nos da la idea intuitiva que hay detrás del procedimiento de mínimos cuadrados en ella observamos los puntos graficados la línea de regresión atraviesa los puntos de una manera tan próxima como sea posible. Ello implica hacer lo propio con la suma de cuadrados de las distancias verticales de los puntos a partir de la línea. Por tanto, Q es la suma de cuadrados de los errores de predicción dentro de la muestra cuando se predice y_i dada x_i y la ecuación de regresión estimada.

$$\mathbf{Q=} \sum (y_i - \alpha - \beta x_i)^2$$

Como economistas interesa analizar los resultados de los cálculos de mínimos cuadrados ordinarios y no la manera de calcular. Sin embargo, vamos analizar el procedimiento para tener una idea general del cálculo. Se define:

$S_{yy} = \sum (y_i - y)^2 = \sum y_i^2 - n\ y^2$

$S_{xx} = \sum (x_i - x)^2 = \sum x_i^2 - n\ x^2$

$S_{xy} = \sum (x_i - x)(y_i - y) = \sum x_i y_i - n\ x\ y$

Los estimadores de mínimos cuadrados son:
$\beta = S_{xy} / S_{xx}$ $\alpha = y - \beta x$

Los residuos estimados son:

$.u = y_i - \alpha - \beta x_i$

La suma de cuadrado residual que se denotará como RSS

$RSS = S_{yy} - (S^2_{xy} / S_{xx}) = S_{yy} - \beta S_{xy} = S_{yy} (1 - r^2_{xy})$

Por lo general S_{xy} se denota como TSS suma de cuadrados total y βS_{xy} por lo general se denota ESS, suma de cuadrados explicada. Por lo tanto

TSS	= ESS	+ RSS
Total	explicado	residual

La palabra residual se reservará para denotar $u = y_i - \alpha - \beta x_i$ y la palabra error para denotar la perturbación de u en la ecuación $y_i = \alpha + \beta x_i + u_i$ $i = 1,2,.....,n$. Por lo tanto, residuo es el error estimado.

La proporción explicada de la suma de cuadrados total se denota con r^2_{xy}, donde r_{xy} se conoce como coeficiente de correlación. Por tanto, $r^2_{xy} = ESS/TSS$ y $1 - r^2_{xy} = RSS/TSS$. Si r^2_{xy} es alto

(cerca de 1), entonces x es una buena variable explicativa para y. El término r^2_{xy} se conoce como coeficiente de determinación y debe quedar entre 0 y 1 para cualquier regresión dada.

Practiquemos un poco lo aprendido

Basados en el primer cuadro de datos donde y_i representa los ingresos por ventas en miles de dólares y x_i los gastos en publicidad en cientos de dólares.

Para obtener los α y β, es necesario calcular $\sum x$, $\sum x^2$, $\sum y^2$ $\sum xy$, $\sum u_i$

Los pasos son los sigueintes:

Ubicados en la ventana primaria declaramos las variable x2, xy. Luego Show x, y, x2, xy, resid.

Los resultados son:

	X	Y	Y2	X2	XY	RESID
	1	3	9	1	3	0.8
	2	4	16	4	8	0.6
	3	2	4	9	6	-2.6
	4	6	36	16	24	0.2
	5	8	64	25	40	1
Total	15	23	129	55	81	0

En la hoja de workfile View/descriptive/Common sample

	X	Y	C
Mean	3.000000	4.600000	1.000000
Median	3.000000	4.000000	1.000000
Maximun	5.000000	8.000000	1.000000
Minimun	1.000000	2.000000	1.000000
Std. Desv.	1.581139	2.408319	0.000000
Skewness	0.000000	0.403407	NA
Kurtosis	1.700000	1.763674	NA
Jarque-Bera	0.352083	0.454052	NA
Probability	0.838583	0.796900	NA
Observaciones	5	5	5

Ahora hagamos los cálculos:

$S_{yy} = \sum (y_i - y)^2 = \sum y_i^2 - n\ y^2 = (129) - 5\ (21.16) = 21$

$S_{xx} = \sum (x_i - x)^2 = \sum x_i^2 - n\ x^2 = (55) - 5\ (9) = 10$

$S_{xy} = \sum (x_i - x)\ (y_i - y) = \sum x_i\ y_i - n\ x\ y = (81) - 5\ (3)\ (4.6) = 12$

$\beta = S_{xy} / S_{xx} = 1.2 \qquad \alpha = y - \beta\ x = 4.6 - 1.2\ (3) = 1.0$

El coeficiente de determinación

$r^2_{xy} = S^2_{xy} / S_{xx} S_{yy} = 144 / 210 = 0.6857$ aprox. A 0.62 $r_{xy} = 0.8228$
$RSS = S_{yy} - (S^2_{xy} / S_{xx}) = S_{yy} - \beta S_{xy} = S_{yy} (1 - r^2_{xy}) = 6.6$ aproximado a 8.8 Por tanto, la regresión de y sobre x es

.y =1.0+1.2x

Dependent Variable: Y				
Method: Least Squares				
Date: 12/19/03 Time: 06:19				
Sample: 1 5				
Included observations: 5				
Variable	Coefficient	Standard Error	t-Statistic	Probability.
X	1.2	0.541602560309	2.21564683763	0.11350247495
C	1.0	1.79629247804	0.556702214269	0.616563390086
R-squared	0.620689655172	Mean dependent var		4.6
Adjusted R-squared	0.494252873563	S.D. dependent var		2.40831891576
S.E. of regression	1.71269767716	Akaike info criterion		4.20319087546
Sum squared resid	8.8	Schwarz criterion		4.04696604043
Log likelihood	-8.50797718865	F-statistic		4.90909090909
Durbin-Watson stat	2.13181818182	Prob(F-statistic)		0.11350247495

Cómo analizar la información

La intersección α =1.0 proporciona el valor de y cuando x=0. Esto significa que si los gastos de publicidad son de 0, los ingresos por ventas serán de 1000 dólares. El coeficiente de la pendiente es β =1.2, lo que quiere decir que si x varía en una cantidad Δx, el cambio en y es $\Delta y = 1.2\ \Delta x$. Por ejemplo, si los gastos de publicidad se elevan en una unidad (100 dólares), los ingresos por ventas aumentaran en 1.2 unidades (1200 dólares) en promedio.

r^2_{xy} = R-squared = Coeficiente de determinación = 0.620 nos indica que la regresión de y sobre x como esta alejada de 0 y cercana a 1 la variable gastos en publicidad, explica la mayor parte de la variación de los ingresos.
.r_{xy} = el coeficiente de correlación 0.7878 nos indica que los gastos de publicidad es una buena variable explicativa para los ingresos, porque está cerca de 1.

Inferencia estadística en el modelo de regresión lineal
La inferencia estadística es el área que describe los procedimientos según los cuales se utilizan los datos observados para extraer conclusiones sobre la población de la cual provienen los datos o del proceso que los generó. Se supone que existe un proceso desconocido que genera los datos que se tienen, el cual se describe por medio de una distribución de probabilidad que, a su vez, se caracteriza por algunos parámetros desconocidos. Por ejemplo, para una distribución normal los

parámetros desconocidos son μ y σ^2. Anteriormente se analizaron los procedimientos para obtener los estimadores de mínimos cuadrados. Para obtener los estimadores de mínimos cuadrados de α y β no es necesario asumir ninguna distribución de probabilidad particular para los errores μ_i. Pero para obtener estimadores de intervalo sobre los parámetros y para probar hipótesis sobre ellos, es necesario asumir que los errores μ_i tienen una distribución normal.

Intervalos de confianza para α , β, μ, σ^2

Los errores estándar o SE o Standard Error para SE(β) = 0.541602560309 y SE(α)= 1.79629247804.
S se conoce como error estándar de la regresión. En el test se identifica como:
S.E. of regression 1.71269767716

Como (α - α) / SE(α) y (β - β) / SE(β) tienen una distribución t student con (n – 2) grados de libertad, al utilizar la tabla de la distribución se obtienen los siguientes resultados:

$$\text{Prob}\left[-2.353<(\alpha - \alpha) / SE(\alpha)>2.353\right] = 0.95$$

$$\text{Prob}\left[-2.353<(\beta - \beta) / SE(\beta)>2.353\right] = 0.95$$

Sustituyendo obtenemos los intervalos

$$\text{Prob}\left[-2.353<(\alpha - 1.0) /1.7962)>2.353\right] = 0.95 \quad \text{intervalo } (-3.2264, 5.22645)$$

$$\text{Prob}\left[-2.353<(\beta - 1.2) / .5416>2.353\right] = 0.95 \text{ intervalo } (-0.07438, 2.47438)$$

Estos son los intervalos de confianza al 95 % para α y β, observe que los límites de confianza de 95 % para α son α=1.0 $\pm$ 2.353 SE (α) y para β=1.2 $\pm$ 2.353 SE(β).

Con relación a la media de la población (**μ),** relacionada con la media muestral **(ў).** Entendemos que tanto la media muestral como la varianza de la muestra S^2 está relacionada con la varianza de la población σ^2. Las distribuciones de muestreo para las muestras de una población normal consideraremos lo siguiente:

$$\breve{y} = 1 / n \sum y_i \quad \textbf{media de muestreo}$$

$$S^2 = 1 / n - 1 \sum (y_i - \breve{y})^2 \quad \textbf{media de varianza}$$

Tanto para la media de muestreo como la media de la varianza se trabaja con n –1 grados de libertad.

$\mu = \bar{Y} \pm$ valor de t (S) / $\sqrt{n}$

A partir de las tablas de la distribución de chi cuadrado X^2 con n –2 grados de libertad se encuentra que la probabilidad de obtener un valor de < 0.352 es 0.05 y de obtener uno > 7.81 es 0.95, por tanto

$$\text{Prob}\left[\ 0.352<3\ S^2 / \sigma^2.>7.81\right] = 0.95$$

$$\text{Prob}\left[3\ S^2 / 7.81 <\sigma^2.> 3\ S^2 / 0.352\right] = 0.95$$

Como S^2 es igual a 2.9241

S.E. of regression 1.71269767716

Luego podemos denotar al sustituir el valor se obtiene (1.1232 , 24.9213)

Prueba de hipótesis

Una hipótesis estadística es un enunciado sobre los valores de algunos parámetros de la población hipotética de la cual se toma la muestra.

Una prueba de hipótesis es un procedimiento que responde la pregunta si la diferencia observada entre el valor de la muestra y el valor de la población hipotético es verdadera o se debe a una variación aleatoria. La H_0 se conoce como hipótesis nula. H_1 se conoce hipótesis alternativa. La probabilidad de rechazar H_0 cuando de hecho es verdadera, se conoce como nivel de significancia. Para probar si la diferencia observada entre los datos y lo que se espera según la hipótesis nula es real o se debe a una variación aleatoria se utiliza una estadística de prueba. Para nuestro caso trabajaremos con las estadísticas de prueba con su distribución normal, chi cuadrado, t y F.

El nivel de significancia observado o valor P es la probabilidad de obtener un valor en la estadística de prueba que sea tan extremo o más que el valor observado en la propia estadística de prueba. Esta probabilidad se calcula sobre la base de que una hipótesis nula es verdadera. Por ejemplo, considere una muestra de n Observaciones independientes a partir de una población normal con media ц y varianza σ^2. Se desea probar H_0: ц=7 contra H_1: ц≠7

La prueba estadística que se utiliza es $t = \sqrt{n}\ (\hat{Y} - \mu) / S$ que tiene una distribución t con (n – 1) grados de libertad. Suponga que n = 25, $\hat{Y}$ = 10, S= 5. Entonces bajo la suposición de que H_0 es verdadera, el valor de t observado es t_0 = 3. Debido a que los valores grandes positivos de t son evidencia contra la hipótesis nula H_0, el valor P es (usando (n-1) = 24 grados de libertad)
P= Prob (t_{24} >3) este es el nivel observado de significancia.

Se dice estadísticamente significativo cuando la variación muestral es una explicación poco probable de la discrepancia entre la hipótesis nula y los valores de la muestra. Un nivel de 0.01 y 0.05 son significativos, ello va relacionado con la muestra. Estas cifras las sugirió Sir R.A Fisher (1890-1962), el padre de la estadística moderna.

Suele rechazarse la hipótesis nula H_0 cuando la estadística de prueba es estadísticamente significativa al nivel elegido de significancia y no hacerlo cuando dicha estadística no es estadísticamente significativa al nivel elegido de significancia,

Luego de esta referencia teórica pasemos a ver sus aplicaciones e acuerdo al ejemplo que estamos analizando.
Suponga que se desea probar la hipótesis de que el valor verdadero de β es 1.0. Se sabe que

$$t_0 = \beta - \beta / SE(\beta)$$

Tiene una distribución t con (n-2 grados de libertad). Sea t_0 el valor t observado. Sí la hipótesis alternativa $\beta \neq 1.0$, entonces es preciso considerar $|t_0|$ como la estadística de prueba. Por lo tanto, sí el valor verdadero de β es 1.0,

$$t_0 = 1.2 - 1.0 / 0.5416 = 0.36927 \text{ Por lo tanto, } |t_0| = 0.36927$$

Al observar en las tablas t para (5-2) 3 grados de libertad, se observa que Prob (t_3 >0.36927)
En este caso es necesaria interpolación lineal porque en la tabla t no existe estos valores, luego lo que se hace es

0.25	0.765
0.63	2.2116
0.1	1.638

Se divide (0.63-0.25)/(0.25-.0.10)*(1.638-0.765). Esta probabilidad no es demasiado baja y no se rechaza la hipótesis β es 1.0. Suele utilizarse 0.05 y 0.01 como probabilidad baja y rechazar la hipótesis sugerida.
Finalmente se observa que existe una correspondencia entre los intervalos de confianza de 0.95 %, que se derivaron antes y las pruebas de hipótesis. Por ejemplo, el intervalo para β es (-0.07438, 2.47438): Cualquier hipótesis $\beta = \beta_0$, donde β_0 se encuentre en este intervalo, no se rechazara en el nivel del 5 % para una prueba de hipótesis.

Análisis de Varianza
Se interpreta como la partición de la suma de cuadrados total TSS en las sumas de cuadrados explicada ESS y residual RSS. El propósito de presentar la tabla es probar la significancia de la suma de cuadrados explicada.

Fuente de variación	Suma de Cuadrados	Grados de libertad	Cuadrados medios
X	ESS= βS_{xy}	1	ESS/1
Residual	RSS= $S_{yy} - \beta S_{xy} = (1-r^2)^* S_{yy}$	.n - 2	RSS/(n-2)
Total	TSS= S_{yy}	.n -1	F=(ESS/1)/ RSS/(n-2)

Análisis de varianza para los datos del ejercicio Ingresos vrs Gastos de publicidad

Fuente de variación	Suma de Cuadrados	Grados de libertad	Cuadrados medios

X	14.4	1	14.4
Residual	8.8	.n - 2	2.93
Total	23.20	.n -1	4.9

r^2 = 0.620689655172

La estadística F es posible utilizarla para la hipótesis β = 0. Lo que la tabla de análisis de varianza proporciona es una prueba de significancia para todos los parámetros, excepto del termino constante de regresión juntas. Observe que en el test $F=t^2$.

Predicción

Llamemos y_0 la predicción dado por x_0 con el siguiente enunciado $y_0 = \alpha + \beta x_0 + \mu_0$

El error de la predicción $E= (y_0 - y_0) = 0$

La varianza del error de predicción es igual a:

$$= \sigma^2 \left[1 + 1 / n + x_0 - x / S_{xx}\right]$$

σ^2 = RSS/grados de libertad

Error estándar de la predicción es la raíz cuadrado de la varianza del error de predicción.

Ejemplo

Considere la muestra de la tienda de ropa deportiva del laboratorio número 1, hasta cinco observaciones, es decir cuando corra el modelo elimine la observación número 6. Recuerde que la ecuación estimada fue y = 1.0+1.2x x = 3 RSS = 8.8

Suponga que el gerente de ventas desea que se prediga el ingreso por ventas que se lograría si se elevaran los gastos de publicidad en 600 dólares: También le gustaría un intervalo de confianza del 90% para esta predicción.

Se tiene $x_0 = 6$. Por lo tanto $y_0 = 1.0+1.2\ (6) = 8.2$ **la varianza del error de predicción es**

$$\sigma^2 \left[1 + 1 / 5 + 6 - 3 / 10\right] = 2.1\ \sigma^2$$

como σ^2 = RSS/grados de libertad = 8.8 /3 = 2.93

el error estándar de la predicción es $\sqrt{}$ 2.1(2.93) = $\sqrt{}$ 6.153 = 2.48

Observando la tabla en el punto superior de 5 % en la distribución de t con tres grados de libertad es 2.353. Por lo tanto, el intervalo de confianza de 90% para y_0 dado $x_0 = 6$ es

$y_0 \pm 2.353(SE(y_0))$.

$$8.2 \pm 2.353\ (2.48) = (2.36, 14.04)$$

así, el intervalo de confianza de 90 % para el ingreso de ventas al elevar los gastos de publicidad en 600 dólares es ($ 2,360,$14.040), es decir que multiplicamos por mil porque así estado en la tabla original de los datos.

Ahora considere el caso en el que el gerente de ventas desea una predicción de las ventas medias por mes durante los próximos dos años, cuando los gastos en publicidad son de 600 dólares mensuales. Asimismo el gerente quiere un intervalo de confianza de 90 % para la predicción.

Bueno ahora, el interés radica en predecir $E(y_0)$, no y_0. La predicción sigue siendo proporcionada por **$y_0 = 1.0+1.2\ (6) = 8.2$** La varianza del error de predicción es ahora

$$\sigma^2 \left[`1 / n + (x_0 - x_0)^2 / S_{xx} \right] =$$

$$\sigma^2 \left[1 / 5 + `(6 - 3)^2 / 10 \right] = 1.1\ \sigma^2$$

Al sustituir σ^2 por 2.93 como antes y extrae la raíz cuadrada, se obtiene ahora el error estándar como SE $(E(y_0)) = 1.795$
El intervalo de confianza de 90% ahora es $8.2 \pm 2.353\ (1.795) = (\$3,980;\$12,420)$
Observe que el intervalo de confianza es más estrecho que el que se obtuvo para la predicción de y.

Ejercicios
Realice todos los pasos efectuados en este ejercicio con los datos de la tabla 3.2 del segundo laboratorio.

Del libro de texto: Mandala, G.S. Introduction to Econometric. ISBN 978-0-470-01512-4, 2009. Trabaje las tablas 3.6, 3.7, 3.8, 3.11.

LABORATORIO # 4

Regresión Multiple

Carrera :
Asignatura : Econometría
Año lectivo : III
Laboratorio No : 4
Clase No : 4
Fecha :
Unidad III : Análisis de Regresión Múltiple
Docente : Dr. Carlos Alberto Zúniga González

Objetivo

Analizar e interpretar los resultados del Test. Analizar el modelo de regresión lineal múltiple con una variable explicada y varias variables explicativas.

Interpretación de los resultados del test

Los Coeficientes de la regresión: Coefficient

Cada coeficiente multiplica la variable correspondiente formando la predicción mejor de la variable dependiente. El coeficiente mide la contribución marginal, manteniendo las otras variables constantes (Ceteris Paribus) de su variable independiente a la predicción. El coeficiente de la serie llamado C es la constante o intercepta en la regresión--es el nivel bajo de la predicción cuando todas las otras variables independientes son el cero. Los otros coeficientes se interpretan como la pendiente de la relación entre la variable independiente correspondiente y la variable dependiente.

Los Errores normales: Standard Error / (SE)

Éstos miden la fiabilidad estadística de los coeficientes de la regresión--el más grande el error normal, el ruido más estadístico infecta el coeficiente. Según la teoría de la regresión, hay aproximadamente 2 oportunidades en 3 que el verdadero coeficiente de la regresión queda dentro de un error normal del coeficiente informado, y 95 oportunidades fuera de 100 que queda dentro de dos errores normales. La varianza del error es el cuadrado del SE

La t-estadística: t-Statistic

Ésta es una estadística de la prueba (Es usada para probar si la diferencia observada entre los datos y lo que se espera según la hipótesis nula H_0 es real o se debe a una variación aleatoria), para la hipótesis que un coeficiente tiene un valor particular. La t-estadística para probar si un coeficiente es el cero (es decir, si la variable no pertenece en la regresión) es la proporción del coeficiente a su error normal. Si la t-estadística excede uno en la magnitud que es por lo menos probablemente dos-tercero que el verdadero valor del coeficiente no es ningún cero, y si la t-estadística excede dos en la magnitud que tiene por lo menos probablemente 95 por ciento años que el coeficiente no es ningún cero.

La probabilidad: Probability / p-valor

La última columna muestra la probabilidad de dibujar una t-estadística de la magnitud del uno sólo a la izquierda de una distribución de t. Con esta información, usted puede decir de una ojeada si usted rechaza o acepta la hipótesis que el verdadero coeficiente es el cero. Normalmente, una probabilidad baja que se toman 0.01 ó 0.05 como la evidencia fuerte de rechazo de esa hipótesis.

R2: R-squared

Esto mide el éxito de la regresión prediciendo los valores de la variable dependiente dentro de la muestra. R2 tiene uno si la regresión encaja perfectamente, y ceros si encaja ningún bueno que la media simple de la variable dependiente. R2 es el fragmento de la variación de la variable dependiente explicado por las variables independientes. Puede ser negativo si la regresión no tiene un intercepto o constante, o si el dos-fase se usan los menores cuadrados. (Mínimos cuadrados de dos etapas).

R2 ajustó para los grados de libertad: Adjusted R-squared

Éste es un pariente íntimo de R2 en que se usan medidas ligeramente diferentes de las variaciones. Está menos de R2 (proporcionó hay más de una variable independiente) y puede ser negativo. Penaliza R2 si se agregan regresores con o no contribuciones, nunca es más grande que R2, pueden decrecer si se agregan más regresores.

El Error normal de la Regresión: S.E. of regression

Ésta es una medida sumaria del tamaño de los errores de la predicción. Tiene las mismas unidades como la variable dependiente. Sobre dos-tercero de todos los errores las magnitudes tienen de menos de un error normal. El error normal de la regresión es una medida de la magnitud de los residuos. Sobre dos-tercero de los residuos quedará en un rango del menos un error normal a la ventaja un error normal, y 95 por ciento de los residuos quedarán en un rango de menos dos a la ventaja dos errores normales.

La suma de Residuos Cuadrados: Sum squared resid

Esto es lo que dice. Usted puede querer acostumbrar este número como una entrada a ciertos tipos de pruebas.

Anote la Probabilidad: Log likelihood

Éste es el valor de la función de probabilidad de leño evaluado a los valores estimados de los coeficientes. Las pruebas de proporción de probabilidad pueden ser dirigidas mirando la diferencia entre las probabilidades del leño de restringió y versiones sin restricciones de una ecuación.

La Durbin-Watson Statistic: Durbin-Watson stat

Ésta es una estadística de la prueba para la correlación de serie. Si está menos de 2, hay evidencia de correlación de serie positiva. Una discusión de la Q-estadística y el Breusch-Godfrey LM prueba que para la correlación de serie los dos de que puede ser superior al Durbin-Watson la prueba.
Un hallazgo común en la regresión de la tiempo-serie es que los residuos se ponen en correlación con sus propios valores retrasados. La correlación es incoherente con la asunción que miente teoría de la regresión que las perturbaciones no se ponen en correlación entre sí en cualquier moda detrás.

Econometristas han desarrollado extensiones de análisis de la regresión para tratar con la correlación de serie.

La manera normal de descubrir la correlación de serie es basado en el Durbin-Watson la estadística. Una estadística mucho debajo de 2 es una indicación de correlación de serie positiva.
Si usted concluye que esa correlación del folletín está presente en los residuos en su regresión, la manera más simple de proceder es agregar una especificación de autoregresiva de primer-orden al modelo de la regresión. Aquí, nosotros presentaremos el acercamiento del libro de cocina a la especificación.

Suponga que su especificación de la regresión era

LAS VENTAS C NEW_ORDERS(0 A -4)

y que el Durbin-Watson la estadística de la menor estimación de los cuadrados ordinaria era 0.64. Usted concluye que hay correlación de serie positiva sustancial. Para incorporar la correlación de serie en su ecuación, usted debe agregar AR(1) al lado diestro. Empuje el botón de la Estimación en el toolbar de la ecuación y teclee en AR(1). Entonces el empujón OK y usted verán un nuevo juego de resultados de la regresión.
Los nuevos resultados se parecerán mucho el más temprano. Habrá una más línea en la mesa de coeficientes y los errores normales, AR(1 etiquetado). El coeficiente asoció con AR(1) es la estimación de la correlación de serie de los residuos. Si su Durbin-Watson la estadística era 0.64, el AR(1) el coeficiente probablemente estará alrededor de 0.7.

Hay dos mejoras en estos resultados. Primero, ambas las estimaciones de los coeficientes ellos y los errores normales estimados de los coeficientes son estadísticamente más fiables. Los menores resultados de los cuadrados antes de que usted agregara AR(1) es menos fiable que ellos parecen, porque los errores normales se subestiman. Segundo, la previsión a corto plazo se mejora considerablemente estimando y usando el coeficiente de la correlación de serie.
Luego el tema: La Teoría de la Correlación de serie

El Akaike Información Criterio: Akaike info criterion

El Akaike Inromation Criterio, o AIC, es una guía a la selección del número de condiciones en una ecuación. Es basado en la suma de residuos cuadrados pero lugares una multa en los coeficientes extras. Bajo ciertas condiciones, usted puede escoger la longitud de una distribución de retraso, por ejemplo, escogiendo la especificación con el valor más bajo del AIC.

El Criterio de Schwarz: Schwarz criterion

El criterio de Schwarz es una alternativa al AIC con básicamente la misma interpretación pero una multa más grande para los coeficientes extras.

La F-estadística: F-statistic / Prob(F-statistic)

Ésta es una prueba de la hipótesis que todos los coeficientes en una regresión son el cero (excepto el intercepte o constante). Si la F-estadística excede un nivel crítico, por lo menos uno de los

coeficientes probablemente es no-ceros. Por ejemplo, si hay tres variables independientes y 100 observaciones, una F-estadística sobre 2.7 indica que la probabilidad es por lo menos 95 por ciento que uno o más de los tres coeficientes es no-ceros. La probabilidad simplemente dada debajo del F La F-estadística le permite que lleve a cabo esta prueba de una ojeada.

Regresión múltiple

En la regresión múltiple se estudia la relación entre (y) y las diversas variables explicativas x_1, x_2, x_3, ... x_k. Por ejemplo, en los estudios de demanda, se estudia entre la cantidad demandada de un bien y el precio del mismo, los precios de los bienes sustitutos y el ingreso del consumidor.

El modelo supuesto es $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ...+ \beta_k x_{ki} + u_i$ $i = 1, 2, ...,n$

En el caso de la regresión simple también se definió lo siguiente:
Suma de cuadrados residual = $S_{yy} - \beta S_{xy}$
Suma de cuadrados de la regresión = βS_{xy}
$r^2_{xy} = S^2_{xy} / S_{xx} S_{yy} = \beta S_{xy} / S_{yy}$

Las expresiones análogas en la regresión múltiple, para el caso de dos variables explicativas son:

RSS = $S_{yy} – \beta_1 S_{1y} - \beta_2 S_{2y}$
Suma de cuadrados de la regresión = $\beta_1 S_{1y} - \beta_2 S_{2y}$
$R^2_{y.12} = \beta_1 S_{1y} - \beta_2 S_{2y} / S_{yy}$

$R^2_{y.12}$ se conoce como coeficiente de la determinación múltiple y su raíz cuadrada positiva es el coeficiente de correlación múltiple. El primer subíndice es la variable explicada. Los subíndices que siguen al punto son las variables explicativas. Para evitar una notación demasiado complicada, se escribió 12 en vez de x_1 x_2. Como sólo las X_s tienen subíndices, no habrá confusión en esta notación.

Para la interpretación de los coeficientes de la regresión es posible hablar del efecto conjunto de x_1 y x_2., y del efecto parcial de x_1 o de x_2 sobre y. El efecto parcial de x_1 se mide por y β_1 el efecto parcial de x_2 por β_2. Con efecto parcial queremos decir mantener constante la otra variable o bien después de eliminar el efecto de la otra variable. Por tanto, β_1 debe interpretarse como una medición del efecto de x_1 sobre y después de eliminar el efecto de x_2 sobre y después de eliminar el efecto de x_1 sobre x_2.

Correlaciones parciales y correlación múltiple

Si tenemos una variable explicada y tres variables explicativas x_1 ,x_2 ,x_3 y r^2_{y1} ,r^2_{y2} ,r^2_{y3} son los cuadrados de las correlaciones simples entre (y) y x_1 ,x_2 ,x_3, respectivamente, entonces r^2_{y1} ,r^2_{y2} ,r^2_{y3} miden la proporción de la varianza en y, que explican x_1 ,x_2 ,x_3 por sí solas. Por otra parte, $R^2_{y.123}$ mide la proporción de la varianza de (y) que explican x_1 ,x_2 ,x_3 en forma conjunta. Asimismo, sería bueno medir algo más. Por ejemplo, ¿qué tanto explica x_2 después de incluir x_1 en la ecuación de regresión? ¿Cuánto explica x_3 después de incluir x_1 y x_2 ? Esto se mide por medio de los coeficientes parciales de determinación $r^2_{y2.1}$,$r^2_{y3.12}$, respectivamente. Las variables que siguen al

punto son las ya incluidas Con tres variable explicativas, se tienen las siguientes correlaciones parciales: de $r_{y1.2}$,$r_{y1.3}$,$r_{y2.1}$,$r_{y2.3}$,$r_{y3.1}$,$r_{y3.2.}$ Estas se conocen como correlaciones parciales de primer orden. Asimismo, existen tres coeficientes de correlación parcial de segundo orden: $r_{y1..23}$,$r_{y2.13}$,y $r_{y3.12.}$ Las variables que van después del punto son siempre las que están ya incluidas en la ecuación de regresión. El orden del coeficiente de correlación depende del número de variables que hay después del punto. La convención usual es denotar las correlaciones simple y parcial con una r y las correlaciones múltiples con R mayúscula.

¿Cómo se calculan los coeficientes de correlación parcial? Para ello, se utiliza la relación entre r^2 y t^2 Por ejemplo:

$$r^2_{y2.3} = t^2_2 / t^2_{2+} \text{d.f.}$$

Los grados de libertad (d.f) = (número de observaciones, n) – (número de parámetros estimados en la regresión, α , β_1 ,β_2)

Las correlaciones parciales son muy importantes para decidir si se incluyen o no más variables explicativas.

Practiquemos con un ejemplo

El siguiente ejemplo de la tabla 4.1 se presenta los datos de una muestra de cinco personas elegidas al azar en una compañía grande, con respecto a sus salarios anuales, años de educación y años de experiencia con la compañía para la que trabajan.

Y= salario anual (miles de dólares)

X_1=años de educación después de la preparatoria

X_2=años de experiencia en la compañía.

Tabla 4.1

Y	X_1	X_2
30	4	10
20	3	8
36	6	11
24	4	9
40	8	12

Los resultados del Test al procesar los datos:

Dependent Variable: Y				
Method: Least Squares				
Date: 12/26/03 Time: 11:22				
Sample: 1 5				
Included observations: 5				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
X1	-0.25	0.684653196881	-0.36514837167	0.75
X2	5.5	0.866025403784	6.35085296109	0.0239078396423
C	-23.75	5.53342118404	-4.29210053059	0.0502279179528
R-squared	0.994485294118	Mean dependent var		30
Adjusted R-squared	0.988970588235	S.D. dependent var		8.24621125124
S.E. of regression	0.866025403784	Akaike info criterion		2.83390426208
Sum squared resid	1.5	Schwarz criterion		2.59956700954
Log likelihood	-4.08476065521	F-statistic		180.333333333
Durbin-Watson stat	2.08333333333	Prob(F-statistic)		0.00551470588235

Interpretando los coeficientes de regresión (Efecto parcial y Efecto conjunto)

Por tanto, la ecuación de regresión es $Y= -23.75 – 0.25X_1+5.5X_2$

Esta ecuación sugiere que los años de experiencia en la compañía,X2, son mucho más importantes que los de educación, X1, (tiene un signo negativo en realidad). La ecuación indica que es posible predecir que el resultado de un año más de experiencia, después de considerar los de escolaridad (o manteniendo esta constante o efecto parcial), da como resultado un incremento salarial de $5,500 dólares al año. Es decir, si consideramos a personas con el mismo nivel de educación, aquella que tenga un año más de experiencia se espera que reciba más alto. Del mismo modo, si se consideran dos personas con la misma experiencia, la que tiene un año más de escolaridad se espera que tenga un salario anual 250 dólares más bajo. Por supuesto, todos estos números están sujetos a cierta incertidumbre, que se discutirá más adelante. Entonces será evidente que la variable X1 deberá eliminarse. Con el término constante –23.75 se trata del salario que obtendría una persona sin experiencia y sólo con la escolaridad de preparatoria. Pero no es posible un salario negativo. ¿Qué ocurre cuando X2= 0, es decir, una persona de reciente incorporación a la compañía? Una vez más, la ecuación predice un salario negativo, luego que esta mal. Podemos concluir que la muestra que se tiene no es una muestra verdaderamente representativa de todas las personas que trabajan en la compañía. Dicha muestra debió haberse obtenido de un sub grupo. Se cuenta con datos de personas cuya experiencia en la firma va de 8 a 12 años. De modo que no es posible utilizar los resultados demasiado lejos de este rango muestral. No es posible utilizar la ecuación para predecir lo que ganaría una persona de nuevo ingreso.

Correlaciones parciales (Observe los valores t del test)

$r^2_{y1.2}$ =0.133 / 0.133 + 2=0.06235 $r_{y1.2}$ = 0.2497
$r^2_{y2.1.}$ =40.333/42.3333=0.95275 $r_{y2.1}$ = 0.976
r^2_{y1} = 0.133/0.1333+3=0.042543 r_{y1} = 0.2062
$r^2_{y2.}$ =40.333/43.3333=0.930768 r_{y2} = 0.96476

Inferencia estadística en el modelo de regresión múltiple

Observemos que para los grados de libertad es n – k – 1 el uno es para el término constante alpha y k para los términos betas. Si utilizamos las distribución t-student con 2 grados de libertad, obtendremos intervalos de confianza de 95 % para β_1 ,β_2 como

α= -23.75 ± 2.920 SE (α)= 23.75 ± 2.920(-4.2921) = (-36.2829,11.2170)
β_1= -0.25 ± 2.920SE(β_1) = -0.25 ± 2.920(-0.3651) = (-1.316,-0.816)
β_2= 5.5 ± 2.920 SE(β_2) = 5.5 ± 2.920 (6.35085) = (13.0444,24.0444)

El intervalo para la media muestral
$\mu = \check{Y} \pm$ valor de t (S) / $\sqrt{n}$ = 30 ± 2.920 (0.8660) / 2.2360= (28.8669,31.1309). Este es el intervalo para la media poblacional hipotética μ con 95 % de confianza.

Supongamos que deseamos probar la hipótesis de β_1 =0 y β_2 = 1 Se acepta la hipótesis β_1 =0 con un nivel de significancia del 5%, ya que β_1 =0 está dentro del intervalo de confianza del 95 % para β_1 . Caso contrario es el de la hipótesis de β_2 = 1 con un nivel de significancia del 5 %, se rechaza ya que no se encuentra dentro del intervalo de confianza del 95 % para β_2.
Utilizando la distribución chi cuadrado para dos grados de libertad

$$\text{Prob}\left[\ 0.103<2\ S^2 / \sigma^2.>5.99\right] = 0.95$$

$$\text{Prob}\left[2\ (0.7499) / 5.99 <\sigma^2.> 2\ (0.7499) / 0.103\right] = 0.95$$

(0.25038, 14.561) Este es el intervalo de confianza con 95% para la varianza de la población hipotética.

En la regresión múltiple la inferencia estadística incluye la estadística F, que representa las regiones de confianza y las pruebas conjuntas de parámetros. Tiene una distribución F con k y (n – k - 1) grados de libertad. Se utiliza para construir regiones de confianza conjunta para β_1 ,β_2 , también para hacer pruebas sobre β_1 ,β_2 a la vez.

Para probar la hipótesis de $\beta_1 = 0$ y $\beta_2 = 1$ con un nivel de significancia de 5 % es posible verificar si el punto se encuentra en la región de confianza conjunta del 95%. Como no siempre se suele dibujar la región de aplica la prueba F.

$$F = 1/2\, S^2 \left[S_{11}(\beta_1 - \beta_1)^2 + 2\, S_{12}(\beta_1 - \beta_1)(\beta_2 - \beta_2) + S_{22}(\beta_2 - \beta_2)^2 \right] =$$

$$F_0 = 1/2\,(0.866) \left[16(-.25 - 0)^2 + 2\,12(-.25 - 0)(5.5 - 1) + 10(5.5 - 1)^2 \right] = 76.4243$$

En las tablas del F stadistic con 2 y 2 grados de libertad igual a 19.0 al nivel de significancia del 5 %. Como F_0 es mayor que 19, se rechaza la hipótesis con un nivel de significancia del 5 %. Observe que los resultados de la prueba de hipótesis donde se utiliza la t student como estadística de prueba son diferentes porque acá estamos valorando la región conjunta para β_1 ,β_2
Para trabajar con la F stadistic es necesario conocer las medias y las sumas de cuadrados de las desviaciones de las medias respectivas.
Luego podemos plantear el siguiente procedimiento en el caso de un modelo con dos variables explicativas:
Las notaciones que se calcularan son
$S_{11} = \sum x_{1i}^2 - n\, x^2_1$ El comando Show @sumsq(x1) calcula $\sum x_{1i}^2 = 141$, n=5 , el comando show @mean(x1)=5, entonces $(5)^2 = 25$
$S_{11} = 16$
$S_{12} = \sum x_{1i}\, x_{2i} - n\, x_1\, x_2$ El comando Show @cross(x1, x2) calcula $\sum x_1\, x_2 = 262$, n=5 , el comando show @mean(x1)=5, show @mean(x2)=10, entonces (5) (10)=50
$S_{12} = 12$

$S_{22} = \sum x_{2i}^2 - n\, x^2_2$ El comando Show @sumsq(x2) calcula $\sum x_{2i}^2 = 510$, n=5 , el comando show @mean(x2)=10, entonces $(10)^2 = 100$
$S_{22} = 10$

$S_{1y} = \sum x_{1i}\, y_i - n\, x_1\, y$ Show @cross(x1,y) calcula $\sum x_{1i}\, y_i = 812$, n=5 , el comando show @mean(x1)=5, show @mean(y)=30, entonces (5) (30)=150
$S_{1y} = 62$

$S_{2y} = \sum x_{2i}\, y_i - n\, x_2\, y$ Show @cross(x2 y) calcula $\sum x_{2i}\, y_i = 1552$, n=5 , el comando show @mean(x2)=10, show @mean(y)=30, entonces (10) (30)=300
$S_{1y} = 52$

$S_{yy} = \sum y_i^2 - n\, y^2_1$ El comando Show @sumsq(y) calcula $\sum y_i^2 = 4772$, n=5 , el comando show @mean(y)=30, entonces $(30)^2 = 900$
$S_{yy} = 272$

Tarea y ejercicios a resolver

Para efectos de realización práctica trabajemos la tabla 4.3, 4.7,4.8,4.9 del texto básico de Mandala. Se pide correr el programa, analice e interprete los coeficientes del Test y haga las inferencias estadísticas correspondiente, usted suponga los valores a probar en los parámetros.

LABORATORIO # 5

Regresión Multiple

Carrera :
Asignatura : Econometría
Año lectivo : III
Laboratorio No : 5
Clase No : 5
Fecha :
Unidad III : Análisis de Regresión Múltiple
Docente : Dr. Carlos Alberto Zúniga González

Objetivo

Analizar e interpretar los resultados de la tabla ANOVA y pruebas de hipótesis. Analizar el procedimiento para realizar pruebas de estabilidad.

ANALISIS DE VARIANZA Y PRUEBAS DE HIPOTESIS

Una expresión alternativa para la prueba F es F=[(RRSS – URSS)/r] / [URSS/(n-k-1)] donde URSS es igual a la suma de cuadrados residual no restringida, RRSS es igual a la suma de cuadrados residual restringida, obtenida mediante la imposición de restricciones a la hipótesis y r es igual al número de restricciones que se impone a la hipótesis.

Análisis de varianza para el modelo de regresión múltiple

Fuente de Variación	Suma de cuadrados, SS	Grados de libertad, d..f.	Cuadrados medios, SS / d .f.	F
Regresión	$R^2 S_{yy}$	K	$R^2 S_{yy}$ / k = MS_1	F = MS_1 / MS_2
Residuo	$(1 - R^2)S_{yy}$	.n – k - 1	$(1 - R^2)S_{yy}$ / n – k – 1 = MS_2	
Total	S_{yy}	.n -1		

Siempre apoyados en el ejemplo de la tabla 4.1 del laboratorio No 3.
Ya hemos discutido que $S_{yy} = \sum y_i^2 – n\ y^2{}_1$ El comando Show @sumsq(y) calcula $\sum y_i^2$=4772 , n=5 , el comando show @mean(y)=30, entonces $(30)^2$=900
S_{yy} =272

Análisis de varianza para el modelo de regresión de los datos de la tabla 4.1 sobre salarios, años de escolaridad y de experiencia

Fuente de Variación	Suma de cuadrados, SS	Grados de libertad, d..f.	Cuadrados medios, SS/d.f.	F
Regresión	270.5	2	135.25 = MS_1	F = 135.25 / 0.75
Residuo / RSS	1.5	2	0.75 = MS_2	F= 180.33333
Total	272	4		

Si observamos el F de la tabla ANOVA es igual al que se presenta en el Test del laboratorio No 3.

Consideremos un ejemplo para ilustrar este tipo de pruebas. Suponga que el problema es probar la hipótesis $H_0 : \beta_1 = \beta_2 = 0.0$ contra $H_1 : \beta_1 = \beta_2 \neq 0.0$

$F=[(RRSS - URSS)/r] / [URSS/(n-k-1)]$ Note que $URSS= (1 - R^2)S_{yy}$ que es la parte de la varianza en que se descompone la varianza de Y, es decir residual a lo no explicado por las variables explicativas o independientes.

$RRSS= S_{yy}$ es la parte de la varianza debido a las variables explicativa o regresión.

Otra forma de calcular RRSS es eliminar X2 y proceder a correr el modelo en el test al obtener RSS, pues ese será el valor buscado.

$F=[(272 - 1.5)/2] / [1.5/(2)]= 270.5/0.75= 3606.6667$ Como se tiene una distribución F con k y (n-k-1) grados de libertad. En las tablas F a 0.01 por ciento de nivel de confianza 99. Bueno como 3606.6667 es mayor que 99. Esta prueba nos indica que los años de educación y la experiencia en conjunto no influyen sobre los salarios anuales. Por supuesto se rechaza la hipótesis

PRUEBAS DE ESTABILIDAD

Al estimar una ecuación de regresión múltiple y utilizarla para predecir en puntos futuros de tiempo, se supone que los parámetros son constantes durante todo el tiempo de la estimación y predicción. Para probar esta hipótesis de constancia de parámetros (o de estabilidad), se propusieron algunas pruebas.
Estas pueden ser:

1. Pruebas de análisis de varianza
2. Pruebas predictivas

En el primer caso ya discutimos el análisis de varianza. El supuesto a agregar es que generalmente se trabaja con dos muestras, por tanto si corremos el modelo obtendremos dos RSS, bien a la suma de dos regresiones para las RSS se le denomina URSS. RRSS es la RSS de una regresión para el total de datos.

Practique este análisis con los datos de la tabla 4.3 de la función de alimentos los resultados tienen que ser URSS igual a RSS=0.1151 para los datos de 1927 a1941 y RSS=0.0544 para los datos de 1948 a 1962 para una URSS=0.1695 con 12+12=24 grados de libertad.
RRSS es el RSS de una regresión para el total de datos RSS=0.2866 con 27 grados de libertad.
Esta regresión que parte de los datos consolidados, impone la restricción de que los parámetros sean iguales en ambos períodos. Por lo tanto, F= 5.53 a partir de la tablas F con 3 y 24 grados de libertad, se observa que el punto de 5% está cerca de 3.01 y el punto de 1 % cerca de 4.72.
Así, aun con un nivel de significancia de 1%, se rechaza la hipótesis de estabilidad. Por lo tanto, la consolidación de los datos no tiene sentido.

2. Pruebas predictivas

Chow sugiere que puede utilizarse una prueba llamada predictiva para la estabilidad sí $n_2 < (k+1)$. La idea fundamental es la siguiente: se utilizan las primeras n_1 observaciones para calcular la ecuación de regresión y para realizar predicciones para las siguientes n_2 observaciones. Entonces, se prueba la hipótesis de que los errores de la predicción tienen media cero. Si $n_2 = 1$, simplemente se utilizará el método utilizado para las predicciones. Si $n_2 > 1$, la prueba F esta dada por:

$F=[(RSS - RSS_1)/ n_2] / [RSS_1/(n-k-1)]$ que tiene una distribución F con n_2 y n_1 -k-1 grados de libertad. En este caso, RSS= suma de cuadrados residual a partir de la regresión con base en $n_1 + n_2$ observaciones; ésta tiene $(n_1 + n_2) – (k + 1)$ grados de libertad. RSS_1 = suma de cuadrados residual de la regresión con base en n_1 observaciones; ésta tiene n_1 -k –1 grados de libertad.

Ejemplo

Considere el ejemplo de la demanda de alimentos de la tabla 4.3, podemos construir una tabla que consolide la información de los resultados si corremos el test con dos ecuaciones que definimos de la siguiente manera:

	Ecuación 1: log(q)=C(1)+C(2)*Log(p)+C(3)*Log(y)			Ecuación 2: log(q)=C(1)+C(2)*Log(p)+C(3)*Log(y)+C(4)*Lo		
Años	1927-41	1948-62	Todas	1927-41	1948-62	Todas
α	4.554918	5.052209	4.047253	4.057732	16.63212	8.117788
β_1 (Precio)	-0.235197	-0.237189	-0.118894	-0.122612	-2.745290	-1.023293
β_2 (Ingreso)	0.243239	0.140703	0.241154	0.367885	-2.415613	-0.721853
β_3 (Interacción)				-0.028217	0.553715	0.213722
η	15	15	30	15	15	30
R^2	0.906593	0.874147	0.971105	0.906630	0.876183	0.976710
F						
RSS	0.001151	0.000544	0.002869	0.001151	0.000535	0.002224
.d.f	12	12	27	11	11	26
S^2	0.009795	0.000045333	0.000106275	0.000104632	0.000048665	0.000092679

El último término de la ecuación 2 es un término de interacción que permite la variación en las elasticidades de precios e ingresos. La Ecuación supone que la elasticidad de precio e ingreso es constante.

Para la ecuación 1, observamos a partir del presente cuadro:

Para 1927 a 1941: $RSS_1 = 0.001151$

Para 1948 a 1962: $RSS_2 = 0.000544$

Datos combinados: RSS=0.002869

Al considerar las predicciones para 1948 a 1962, mediante la ecuación estimada para 1927 a 1941, tenemos que

$F= [(RSS - RSS_1) / n_2] / [RSS_1 / n_2 -k-1] = 1.19$ Este resultado los obtenemos con Eviews utilizando la ecuación 1 para todas las observaciones en la ventana del Test aplicando los comandos Views/Stability test/Chow Forecast test Enter en la ventana de dialogo escribimos las fechas de 1948 1962, es decir la predicción de la segunda muestra a partir de la primera los resultados son:

Chow Forecast Test: Forecast from 1948 to 1962			
F-statistic	1.19384356041	Probability	0.383690909338
Log likelihood ratio	27.3962349308	Probability	0.025668953978

A partir de la tablas F con 15 y 12 grados de libertad, se observa que el punto de 5% es de 2.62. Por tanto, con un nivel de significancia de 5%, no rechaza la hipótesis de estabilidad. La prueba de análisis de varianza llegó a la conclusión opuesta.

Para la segunda ecuación
$F= [(RSS - RSS_1) / n_2] / [RSS_1 / n_2 -k-1] = 0.80$

Chow Forecast Test: Forecast from 1948 to 1962			
F-statistic	0.803744	Probability	0.660030
Log likelihood ratio	22.20113	Probability	0.102638

Por lo tanto, con un nivel de significancia de 5% no se rechaza la hipótesis de la estabilidad. De este modo, las conclusiones de la prueba predictiva no parecen ser diferentes de la que se obtienen por medio de la prueba de análisis de varianza, para esta ecuación.

Grados de libertad y R cuadrado ajustado
Hemos venido discutiendo que el estimado de la varianza residual $<\sigma^2.>$ está dado por
σ^2=RSS/grados de libertad

A medida que el número de variables explicativas aumenta, la RSS se reduce, pero también se presenta una reducción en los grados de libertad. La idea conveniente es que la varianza puede llegar al mínimo en la medida que se agregue o eliminen variables explicativas. Esta es la razón porque en el Test aparece el R cuadrado ajustado, es simplemente el R cuadrado ajustado por los grados de libertad, se define mediante:

$1 – R^2$ ajustada $=(n –1 / n – k –1) * (1 - R^2)$, donde k es el número de regresores. Se resta (k –1) de n debido a que se estima un término constante además de los coeficientes de estos k regresores.

Existe una relación entre las pruebas t y las pruebas F descritas antes y el R^2 ajustada. Si el coeficiente t para el coeficiente de cualquier variable es menor que 1, entonces al eliminar esa variable, R cuadrado ajustado aumentará y la varianza disminuirá, esto último es lo que nos interesa del razonamiento. En forma más general, si el cociente F de cualquier conjunto de variables es menor que 1, entonces la eliminación de este conjunto de variables de la ecuación de regresión elevará R cuadrado ajustado.

Hay dos situaciones que ocasionan problemas:
Uno es el caso en que las variables explicativas tienen una alta intercorrelación (Multicolinealidad). Esto es cuando las relaciones t son menores que 1, pero el cociente de F es mayor que1. En este caso, el que todos los cocientes t sea menores que 1 no significa que sea posible elevar R cuadrado

ajustado mediante la eliminación de todas las variables. Una vez eliminada una variable, los otros cocientes t cambiarán.

El segundo caso es cuando los cocientes t son todos mayores que 1 pero el cociente F para un conjunto de variables es menor que 1. En este caso, al eliminar cualquier variable no es posible elevar el R cuadrado ajustado, es posible obtener una R cuadrado ajustado más elevado si se elimina un conjunto de variables explicativa.

Esta situación nos hace reflexionar en la pregunta ¿Cómo se sabe si es posible elevar R cuadrado ajustado eliminando algunos conjuntos de variables sin buscar en todos los conjuntos y sub conjuntos?

Para responder a esta pregunta, se establecerá un sencilla regla que proporciona la relación entre los cocientes t y F. La regla dice que si $F \leq 1$, el valor t absoluto de cada una de la variable k es menor que la raíz cuadrada de k.
Ejemplo, considere una ecuación de regresión con cinco variables independientes y relaciones t de 1.2, 1.5, 1.6, 2.3, y 2.7. Observe que $\sqrt{1}=1$, $\sqrt{2}=1.414$, $\sqrt{3}=1.732$, $\sqrt{4}=2$, $\sqrt{5}=2.236$. Consideremos k=1, 2, 3, 4, 5, y verifiquemos si existe< k relaciones $t < \sqrt{k}$. Se observa que la regla es válida para k=3. Por lo tanto, lo que se debe hacer efectuar la eliminación de las tres variables x_1, x_2, x_3, esto para lograr elevar R cuadrado ajustado, entonces lo que se debe hacer es correr el programa excluyendo estas tres variables y verificar si la R cuadrada aumenta.

Tarea y ejercicios a resolver

Para efectos de realización práctica trabajemos la tabla 4.3, 4.7,4.8,4.9 del texto básico de Mandala. Los datos de las bases de datos ya fueron introducidas en el laboratorio anterior en él de pedía correr el programa, analice e interprete los coeficientes del Test y haga las inferencias estadísticas correspondiente, usted suponga los valores a probar en los parámetros, ahora agregamos hacer los análisis de varianza y las pruebas de estabilidad.

LABORATORIO # 6

Heterocedasticidad

Carrera :
Asignatura : Econometría
Año lectivo : III
Laboratorio No : 6
Clase No : 6
Fecha :
Unidad IV : Heterocedasticidad
Docente : Dr. Carlos Alberto Zúniga González

Objetivo

Dar a conocer en que consiste el problema de Heterocedasticidad, como se detecta y como se soluciona.

Introducción

Al inicio del curso hemos planteado la situación de validación del modelo para poder hacer predicciones económicas. A partir de este momento comenzaremos a abordar los temas sobre la validación del modelo econométrico. Para obtener estimadores de intervalo sobre los parámetros y para probar cualquier hipótesis sobre ellos, es necesario asumir que los errores u_i tienen una distribución normal. Los supuestos son:

1. $E(u_i)= 0$. Media
2. $V(u_i)= \sigma^2$ para todo i. Varianza común.
3. (u_i) y (u_j) son independientes para todo $i \neq j$. Independencia.
4. (x_j) es no estocástico. Independencia. (u_i) y (x_j) son independientes para todo i y j. Esta suposición es consecuencia automática si las (x_j) se consideran variables no aleatorias.
5. (u_i) está normalmente distribuido para todo i. Normalidad. Junto con las suposiciones 1, 2, y 3, esto implica que los (u_i) son independientes y tienen una distribución normal con media 0 y varianza común σ^2. Esto se escribe como $u \sim IN(0, \sigma^2)$

Para el primer supuesto, éste se obtiene una vez que usted corre el modelo y en la ventana principal se le presenta el RESID= (u_i). Generalmente esto se logra escribiendo en la ventana de comando show @sum(resid).

El segundo supuesto es el que abordaremos en el presente laboratorio. A este supuesto se le conoce como homocedasticidad. Si los errores no tienen una varianza constante, se dice que son heterocedásticos.

Cómo se detecta el problema de Heterocedasticidad

Cuando la regresión es un modelo simple, es posible detectar este problema observando el gráfico de los RESID o residuo. Además se puede ver en la ventana de Workfile nombrando para Show las variables X y el Resid, entonces valoramos la relación comparativa entre ambas variables.

Un ejemplo, es los datos de la tabla 5.1, donde se presenta los gastos de consumo (y), y el ingreso (x) de 20 familias.

Tabla 5.1

Familias	Y	X	Familias	Y	X
1	19.9	22.3	11	8.0	8.1
2	31.2	32.3	12	33.1	34.5
3	31.8	36.6	13	33.5	38.0
4	12.1	12.1	14	13.1	14.1
5	40.7	42.3	15	14.8	16.4
6	6.1	6.2	16	21.6	24.1
7	38.6	44.7	17	29.3	30.1
8	25.5	26.1	18	25.0	28.3
9	10.3	10.3	19	17.9	18.2
10	38.8	40.2	20	19.8	20.1

Para observa el gráfico desde la ventana del Test aplicamos Quick en la caja de dialogo nombrar las series x y resid, enter /Graph/Scatter.

Otra forma, es haciendo doble clic en la ventana principal en el icono de RESID, valorando los datos para determinar un patrón sistemático en los residuos. Agregamos que desde la ventana del Test en pulsando el icono de Resids usted puede observar el gráfico, y desde esta misma ventana View/Actual,fitted,Residual/Actual, fitted Residual Table usted podrá observar más detalladamente cuál es el residuo asimétrico, eso se detecta identificando el punto que se ubica fuera de las líneas punteadas.

En el caso de regresión múltiple se utiliza un procedimiento de regresar la potencia de y, el valor predicho de y o bien las potencias de todas las variables explicativas.

Las pruebas sugeridas para detectar la heterocedásticidad son: Para trabajar con estas pruebas y correr los modelos es necesario crear una serie que sea igual al Resid, por ejemplo cree la variable Miu,luego Genere Miu=Resid.

1. Las pruebas sugeridas por Anscombe y una prueba de RESET, sugerida por Ramsey, involucra tanto la regresión de μ_i sobre y_t^2, y_t^3 ,y_t^4,.....como la prueba de si los coeficientes son significativos o no. La prueba Ramsey RESET es aplicable solamente a una ecuación estimada por mínimos cuadrados ordinarios. Para aplicar la prueba, pulse View/Stability Tests/Ramsey RESET Test. En el cuadro de dialogo, usted debe especificar cuantas variables debería agregar en la ecuación. Cada una de estas variables serán calculadas como un valor ajustado de la regresión original, iniciando con el cuadrado del segundo valor. El valor de F estadístico nos indica si rechazamos o no, es decir si los coeficientes son significativos o no. En el test

observará que Ninguno de los coeficientes tuvo una relación t > 1, lo que revela que es imposible rechazar la hipótesis de que los errores son homocedásticos. Sin embargo, la guía en el análisis deber ser el F estadístico. Los resultados son $\mu = -0.379 + 0.236\ 10^{-2}\ x^2 - 0.549\ 10^{-4}\ x^3$ $R^2 = 0.034$

2. La prueba sugerida por White involucra la regresión de μ_t^2 sobre todas las variables explicativas, sus cuadrados y productos cruzados. Por ejemplo, con tres variables explicativas x_1, x_2, x_3 supone regresar μ_i^2 sobre $x_1, x_2, x_3, x_1^2, x_2^2 x_3^2, x_1 x_2, x_2 x_3, x_3 x_1$. Para ejecutar la prueba Heterocedasticidad de White, teclee Views/Residual Tests/White Heteroskedasticity (cross terms) desde la ventana del Test.

Hasta el momento con estas pruebas hemos valorado el criterio de t > 1 en el caso de la prueba de Ramsey, y el R cuadrado para la prueba, en ambos caso se concluye que son homocedásticos, es decir no detectan el problema de Heterocedasticidad. Los resultados del ejemplo son: $\mu^2 = -1.370 + 0.116x$ $R^2 = 0.7911$

$\mu^2 = 0.493 - 0.071\ x + 0.0037\ x^2$ $R^2 = 0.878$

3. Sin embargo, con la prueba de Glejser si se logra detectar este problema. Glejser sugirió estimar las regresiones del tipo $/ \mu_t / = \alpha + \beta\ x_{i.}$, $/ \mu_t / = \alpha + \beta / x_{i.}$, $/ \mu_t / = \alpha + \beta \sqrt{x_{i.}}$ etc y demostrar la hipótesis $\beta = 0$. Los resultados fueron para el ejemplo de la tabla 5.1:

a) En la ventana de comando escribimos ls abs(miu) c x para obtener:
$/ \mu_t / = \alpha + \beta\ x_i = / \mu_t / = -0.209 + 0.0512\ x$ $R^2 = 0.927$
b) En la ventana de comando escribimos ls abs(miu) c (1/x) para obtener:
$/ \mu_t / = 1.826 - 13.78\ 1/x$ $R^2 = 0.649$
c) En la ventana de comando escribimos ls abs(miu) c sqr(x) para obtener:
$/ \mu_t / = \alpha + \beta \sqrt{x_{i.}}$ / $\mu /$ =-1.232 + 0.475 $\sqrt{x}$ $R^2 = 0.902$

Todas las pruebas rechazan la hipótesis de homocedásticidad, si bien sobre la base de R^2, el primer modelo es preferible a los demás.

Se deja que el estudiante realice estas pruebas utilizando el logaritmo y verifique si existe o no el problema de heterocedásticidad, es decir determinar cual de las pruebas detecta este problema.

¿Cómo soluciona el problema?

La solución al problema de la heterocedasticidad depende de las suposiciones acerca de sus fuentes de origen. Cuando no se tiene seguridad al respecto, al menos es posible tratar de hacer correcciones para los errores estándar, ya que se ha venido observando que el estimador de mínimos cuadrados es insesgado pero ineficiente y que, además, los errores estándar también están insesgado.

Una manera de solucionar depende de suposiciones determinadas acerca de σ^2_i. Hay dos métodos de estimación para solucionar la problemática de la heterocedásticidad: mínimo cuadrados ponderados (WLS) y máxima verosimilitud (ML).
Para usar mínimos cuadrados ponderados, presione el botón Options desde la ventana del Test Estimate entonces en la caja de dialogo elija la opción Weighted LS/TSLS. Llene el espacio en blanco después de Weight: con el nombre de la serie que servirá de ponderación. Como se describe abajo, cada peso es interpretado como una desviación estándar reciproca de los disturbios para esa observación.

Otra forma de solucionar es transformar los datos en logaritmos. Este método suele reducir la Heterocedasticidad en las varianzas del error, si bien existen otros criterios por medio de los cuales es preciso decidir entre las formas funcionales lineal y logarítmica.

Deflactar las variables a partir de alguna medida de tamaño, éstos índices de deflación es preciso tener cuidado para estimar la ecuación con las variables explicativas correctas, por ejemplo si la ecuación original involucra un término constante, no se deberá estimar una ecuación similar en las variables deflactadas.

.
En el ejemplo de la tabla 5.1 obtenemos los siguientes resultados:
Aplicando logaritmo y Mínimo cuadrados ponderados

Dependent Variable: LOG(Y)
Method: Least Squares
Date: 12/28/03 Time: 06:08
Sample: 1 20
Included observations: 20
Weighting series: X

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.0831698727826	0.123246251252	0.67482679544	0.508361773348
LOG(X)	0.954019620123	0.0352162699701	27.0903085685	4.84279730571e-16
Weighted Statistics				
R-squared	0.999185434634	Mean dependent var		3.27191284009
Adjusted R-squared	0.999140181003	S.D. dependent var		1.91371488128
S.E. of regression	0.0561152306407	Akaike info criterion		-2.82817947401
Sum squared resid	0.0566805439775	Schwarz criterion		-2.72860624665
Log likelihood	30.2817947401	F-statistic		733.884818339
Durbin-Watson stat	2.51764574826	Prob(F-statistic)		4.84279730571e-16
Unweighted Statistics				
R-squared	0.993474958509	Mean dependent var		3.03391101415
Adjusted R-squared	0.993112456204	S.D. dependent var		0.550836011215
S.E. of regression	0.0457145565701	Sum squared resid		0.0376167722832
Durbin-Watson stat	2.17369281649			

Examine el gráfico tecleando el icono resid y observará que la curva está dentro de la pista punteada indicando que el problema de Heterocedasticidad mejoró. Calcule las pruebas de Glejser para verificar como R cuadrado se redujo. Examine los resultados de las otras pruebas.

Tarea

De la tabla 3.11 del texto básico de Mandala realice las pruebas que le permitan detectar un problema de Heterocedasticidad y corrija el problema. Verifique si cumple el primer supuesto.

UNIVERSIDAD NACIONAL AUTONOMA DE NICARAGUA, LEÓN
FACULTAD DE CIENCIAS Y TECNOLOGÍA
DEPARTAMENTO DE AGROECOLOGÍA
CENTRO DE INVESTIGACIÓN EN CIENCIAS AGRARIAS Y ECONOMIA APLIDADA
http://cicaea.unanleon.edu.ni/index.html

LABORATORIO # 7

Autocorrelación

Carrera :
Asignatura : Econometría
Año lectivo : III
Laboratorio No : 7
Clase No : 7
Fecha :
Unidad V : Auto correlación
Docente : Dr. Carlos Alberto Zúniga González

Objetivo

Exponer las diversas limitaciones de la prueba de Durbin-Watson, que demuestra su casi inutilidad en la práctica. El procedimiento como identificar este problema y como resolverlo.

Introducción

Nuevamente planteamos los supuestos acerca de los errores del modelo que hemos venido discutiendo en la regresión. Los errores u_i se deben a los errores de medición en (y) y a la especificación de la relación entre (y) y las x. Las suposiciones son:

6. $E(u_i)= 0$. Media
7. $V(u_i)= \sigma^2$ para todo i. Varianza común.
8. (u_i) y (u_j) son independientes para todo $i \neq j$. Independencia.
9. (x_j) es no estocástico. Independencia. (u_i) y (x_j) son independientes para todo i y j. Esta suposición es consecuencia automática si las (x_j) se consideran variables no aleatorias.
10. (u_i) está normalmente distribuido para todo i. Normalidad. Junto con las suposiciones 1, 2, y 3, esto implica que los (u_i) son independientes y tienen una distribución normal con media 0 y varianza común σ^2. Esto se escribe como $u \sim IN(0, \sigma^2)$

Ahora discutiremos sobre el tercer supuesto. Existen dos situaciones bajo las cuales los términos de error en el modelo de regresión pueden estar correlacionados. En los datos de corte transversal, la correlación puede surgir entre unidades contiguas. Por ejemplo, si se estudian los patrones de consumo de las familias, es posible que los términos de error de éstas en la misma zona estén correlacionados. Esto se debe a que el término de error de éstas recoge el efecto de las variables omitidas y éstas tienden a correlacionarse para las familias de una misma zona (debido al efecto de mantenerse al nivel de los vecinos).

Lo que se discutirá en este laboratorio es la correlación entre los términos de error que surge en los datos de series de tiempo[1]. Este tipo de correlación se conoce como Autocorrelación o correlación serial. El término de error u_i par un período de tiempo t, está correlacionado con los términos de

[1] Una serie de tiempo es una secuencia de datos numéricos cada uno de los cuales se asocia con un instante específico de tiempo.

error u_{t+1}, u_{t+2}, ...y u_{t-1}, u_{t-2}, etc. Muchas veces, tal correlación en los términos de error surge a partir de la correlación de la variables omitidas cuyo efecto captura el término de error.

La correlación entre u_t y u_{t-k} se conoce como Autocorrelación de orden k. La correlación entre u_t y u_{t-1} es de primer orden y, por lo general, se denota como ρ_1. La correlación entre u_t y u_{t-2} se conoce como Autocorrelación de segundo orden y se denota por ρ_2., etc. Existe (n-1) de tales autocorrelaciones si se tienen n observaciones. Sin embargo, no es posible estimar todas éstas a partir de los datos. Por tanto, muchas veces se supone que estas (n-1) autocorrelaciones pueden representarse en términos de uno o dos parámetros.

Prueba de Durbin y Watson

El modelo más sencillo y de uso más común es aquel en el que los errores u_t y u_{t-1} tienen una correlación ρ. Para este modelo, es posible pensar en probar la hipótesis en torno a ρ con base en ρ_1, la correlación entre residuos de los mínimos cuadrados u_t y u_{t-1}. Una estadística de uso común para este propósito (que se relaciona con ρ) es la estadística de Durbin Watson (DW) como aparece en el Test, se denotará como d y se define como:

. $d=\sum(u_t \text{ y } u_{t-1})^2 / \sum u_t^2$ donde u_t es el residuo estimado par el periodo t.

Considerando que $d \approx 2 (1 - \rho)$

Si ρ= +1$\Rightarrow$ d = 0	Si ρ= -1$\Rightarrow$ d = 4	Si ρ= 0 $\Rightarrow$ d = 2

Si d está próximo a 0 o a 4, los residuos tienen una alta correlación.

En las tablas que se les entregaron una de ellas muestra la distribución de muestreo de d, ésta depende de los valores de las variables explicativas, por tanto, la Durbin Watson calcularon los límites superior (d_U) e inferior (d_L) para niveles de significancia de 5%. Estas tablas prueban la hipótesis de Autocorrelación cero contra las hipótesis de Autocorrelación positiva de primer orden (para la autocorrelación negativa, se intercambia d_L y d_U.

Si d < d_L se rechaza la hipótesis nula de no Autocorrelación	Si d < d_U no se rechaza la hipótesis nula de no Autocorrelación	Si d_L< d < d_U La prueba no es concluyente

Hannan y Terrell demuestran que el límite superior de la estadística de DW es una buena aproximación a su distribución, cuando los regresores cambian con lentitud. Afirman que la serie económica de tiempo cambian con lentitud y que, por tanto, es posible utilizar d_U como punto correcto de significancia.

Los puntos de significancia de las tablas de DW se tabulan para probar $\rho = 0$ contra $\rho > 0$. Si de $d > 2$ y se desea probar la hipótesis $\rho = 0$ contra $\rho < 0$, se considera $4 - d$ y se hace referencia a las tablas como si se probara una Autocorrelación positiva.

$d \approx 2 (1 - \rho)$ es válida para muestras grandes. Se ha demostrado que el valor esperado de, cuando $\rho = 0$, está dado aproximadamente por:

$E(d) \approx 2 + (2k - 1) / n - k$ donde k es el número de parámetros de regresión estimados (incluyendo el término constante) y n el tamaño de la muestra.

Ejemplo considere los datos de la tabla 3.11. Los resultados del test utilizando logaritmo fueron:

Dependent Variable: LOG(X)
Method: Least Squares
Date: 01/06/04 Time: 10:08
Sample: 1929 1967
Included observations: 39

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-3.9376059933	0.236950596849	-16.6178353026	1.82283634117e-18
LOG(L1)	1.45066276393	0.0832113464758	17.4334730223	3.91078643748e-19
LOG(K1)	0.383926174337	0.0480079584384	7.9971360338	1.69255899841e-09

R-squared	0.994629745953	Mean dependent var	5.68746569812
Adjusted R-squared	0.994331398506	S.D. dependent var	0.460975514433
S.E. of regression	0.0347068997316	Akaike info criterion	-3.80995303195
Sum squared resid	0.0433644800032	Schwarz criterion	-3.68198675148
Log likelihood	77.2940841231	F-statistic	3333.79673838
Durbin-Watson stat	0.857435565098	Prob(F-statistic)	0

Paso 1

El primer paso es verificar si existe el problema de Autocorrelación, es decir si rechazamos la hipótesis de correlación serial cero. Es importante estar claros, que para dar este paso se debe haber resuelto el problema de heterocedasticidad y verificado la media cero de los errores. Para este caso es conveniente probar si existe un problema de heterocedasticidad (compruébelo, se recomienda utilizar la prueba de Glejser).

Si nos referimos a las tablas DW, donde para tamaños intermedios o números de variables, es posible utilizar la interpolación, con k=2 y n=39, para un nivel de significancia de 5%, se observa que $d_L = 1.38$. Dado que la d observada es igual a 0.8574 y menor que d_L, se rechaza la hipótesis de $\rho = 0$ al nivel de 5 %, en esta caso $\rho = 0.5713$. De igual manera y de forma práctica cuando DW no es igual a 2 rechazamos la hipótesis.

En el ejemplo se evidencia una Autocorrelación positiva DW=0.86 con un $\rho = 0.57$.

Paso 2

Como segundo paso es considerar una estrategia cuando la estadística de prueba de Durbin Watson es significativa (no es clara en cuanto a la fuente de origen). Recordemos que la estrategia de DW está diseñada como una prueba de la hipótesis $\rho = 0$ si los errores siguen un proceso autorregresivo de primer orden $u_t = \rho\ u_{t-1} + e_{t..}$

La estrategia a seguir es:

a) Suponer que la estadística de DW indica correlación serial pero que no puede deberse a errores AR(1). Se recomienda verificar autorregresiones de orden más alto, calculando ecuaciones de la forma $u_t = \rho\ u_{t-1} + \rho\ u_{t-2} + e_{t..}$ Una vez determinado el orden, es posible estimar el modelo con suposiciones apropiadas sobre la estructura de los errores por medio primeras diferencias que se explican en el paso número 3. De los resultados del Test descrito arriba y para aplicar esta estrategia: aplicamos Estimate y al final de la ecuación probamos AR(2) AR(3), etc. Observe el comportamiento de DW de esta forma ud. Verifica el orden de la Autocorrelación. Estas pruebas descartan en cierta forma esta fuente.

b) Probar si la correlación serial se debe a variables omitidas. Es bastante difícil de atacar. Muchas veces se afirma que la fuente de la correlación en los errores es la omisión de algunas variables que debieron incluirse en la ecuación, y que estas variables están autocorrelacionadas por sí solas. Es posible realizar pruebas sobre las variables omitidas mediante la prueba RESET de Ramsey o la prueba de White descritas en la unidad de Heterocedasticidad. Si la estadística de prueba de DW es significativa, pero estas pruebas muestran también significancia, la estrategia apropiada es utilizar mínimos cuadrados ponderado o máxima verosimilitud. Si aplicamos estas pruebas al Test podemos observar que existe realmente un problema de variables omitidas.

c) Probar si la correlación serial se debe a una dinámica mal especificada. Aquí se aplica la prueba del factor común de Sargan. Hendry y Mizon también hicieron énfasis en ello. El argumento es el siguiente. $y_t = \beta\ x_t + u_t$ con $u_t = \rho\ u_{t-1} + e_{t..}$ y e_t es independiente con una varianza común σ^2 Es posible escribir este modelo como
$y_t = \rho\ y_{t-1} + \beta\ x_t + \beta\ \rho\ x_{t-1} + e_t$. Consideremos un modelo dinámico alternativo:
$y_t = \beta_1\ y_{t-1} + \beta_2\ x_t + \beta_3\ x_{t-1} + e_t$. $/\beta_1\ / < 1$ La ecuación anterior y esta es la misma con la restricción $\beta_1\ \beta_2 + \beta_1 = 0$. Una prueba para $\rho = 0$ es una prueba $\beta_1 = 0$ y ($\beta_3 = 0$). Pero antes de probar esto, Sargan dice primero debería probarse la restricción $\beta_1\ \beta_2 + \beta_1 = 0$ y probar para $\rho = 0$, sólo si no se rechaza la hipótesis $H_0 = \beta_1\ \beta_2 + \beta_1 = 0$. Si se rechaza, no se cuenta con un modelo de correlación serial, y éste, en los errores de $y_t = \beta\ x_t + u_t$ con $u_t = \rho\ u_{t-1} + e_{t..}$ se debe a una dinámica mal especificada, es decir la omisión de variables
y_{t-1} y x_{t-1}
La restricción $\beta_1\ \beta_2 + \beta_1 = 0$ no es lineal en β y, por tanto, sólo es preciso utilizar la prueba de Wald, la razón de verosimilitud o la del multiplicador de Lagrange.

Sargan propone comenzar con el modelo $y_t = \beta_1\ y_{t-1} + \beta_2\ x_t + \beta_3\ x_{t-1} + e_t$. $/\beta_1\ / < 1$ y probar la restricción $\beta_1\ \beta_2 + \beta_1 = 0$

Utilizando la base de datos de la tabla 3.11 y la estimación de función de producción discutida en la unidad anterior. En el paso 3 se explican los procedimientos para calcular ρ en primeras diferencias y cuasi diferencias para resolver el problema de Autocorrelación. En la sección anterior al paso 1 observará el Test con una DW = 0.86 evidenciando el problema de Autocorrelación. Ahora supondremos $y_t = \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 x_{t-1} + e_t$ propuesto por Sargan.

Apliquemos el multiplicador de Lagrange (LR). En primer lugar va a organizar los datos en dos grupos de 19 observaciones cada uno y determinar dos ecuaciones primera y la segunda. Bien, primero haga clip en el comando Object y elija Equation y nómbrela Primera en la ventana de formula escriba LOG(X) C LOG(L1) LOG(K1) en la Sample escriba 1929 1948, repita el proceso para la segunda ecuación con la observación de 1949 1967. Desde la ventana de comandos escribimos scalar lr = -2*(primera. @logl-segunda.@logl) y luego escribimos scalar lr_pval=1-@cchisq(lr,1), observaremos que esta prueba nos demuestra que no existe un problema de mala específicación.

Paso 3

Como parte de los procedimientos de estimación con errores autocorrelacionados. Empezaremos con estimación en niveles contra primeras diferencias, en tales caso se calcula una regresión mediante la transformación de todas las variables, que resulta de aplicar ρ-diferencias con respecto a ρ, es decir, se regresa $y_t - \rho y_{t-1}$ sobre $x_t - \rho x_{t-1}$, donde ρ es ρ estimados. Sin embargo, dado que está sujeto a errores de muestreo, otra alternativa en caso de que la estadística d de Durbin Watson, sea demasiado pequeña es utilizar una ecuación de primeras diferencias. De hecho una regla básica dice: calcule una ecuación en las primeras diferencias, siempre que la estadística de Durbin Watson sea < R cuadrado. En las ecuaciones de primeras diferencias, se regresa $(y_t - y_{t-1})$ sobre $(x_t - x_{t-1})$ con todas las diferencias en variables explicatorias de modo similar.

Resolvamos para el ejemplo de la tabla 3.11 ρ - diferencias o cuasi primeras diferencias.
Desde la ventana de comando aplique **ls log(x-(x11(-1))) log(l1-(l11(-1))) log(k1-(k11(-1)))** **y** genere x11=x*0.5713, k11=k1*0.5713, y l11=l1*0.5713

En los resultados abajo presentados observamos que la DW=0.3312 esto se debe al error en el muestreo, por lo que se recomienda trabajar mejor con las primeras diferencias.

Dependent Variable: LOG(X-(X11(-1)))
Method: Least Squares
Date: 01/07/04 Time: 19:02
Sample(adjusted): 1930 1967
Included observations: 38 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOG(L1-(L11(-1)))	0.228776753418	0.07919409885	2.88881061519	0.00650862128522
LOG(K1-(K11(-1)))	0.960028439182	0.089164824121	10.7668965721	8.31916076775e-13
R-squared	0.942101252476	Mean dependent var		4.88721961438
Adjusted R-squared	0.940492953934	S.D. dependent var		0.492821963015
S.E. of regression	0.12021931539	Akaike info criterion		-1.34780214903
Sum squared resid	0.520296616543	Schwarz criterion		-1.26161340378

Log likelihood	27.6082408315	Durbin-Watson stat	0.331221023009

Primeras diferencias

Partimos del hecho de la recomendación 0.33 < 0.94, entonces el modelo que regresa es $(x_t - x_{t-1})$ sobre $(li_t - l1_{t-1})$ $(ki_t - k1_{t-1})$, en la ventana de Estimate escribimos log(x - x(-1)) log(l1 - l1(-1)) log(k1 - k1(-1)), los resultados son:

Dependent Variable: LOG(X-X(-1))
Method: Least Squares
Date: 01/08/04 Time: 10:36
Sample(adjusted): 1935 1967
Included observations: 24
Excluded observations: 9 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
LOG(L1-L1(-1))	1.06783384311	0.152706206179	6.99273375869	5.10858259553e-07
LOG(K1-K1(-1))	0.417434467655	0.15919915037	2.62208979561	0.0155638442541
R-squared	-0.304326352448	Mean dependent var		2.92635936331
Adjusted R-squared	-0.363613913923	S.D. dependent var		0.476928129985
S.E. of regression	0.556927484012	Akaike info criterion		1.74689188046
Sum squared resid	6.82370089385	Schwarz criterion		1.84506303299
Log likelihood	-18.9627025655	Durbin-Watson stat		1.62182603667

La discusión de las cuasi diferencias y las primeras diferencias sólo es importante si se cree que es posible describir toda la estructura de correlación entre los errores en términos de ρ, el coeficiente de correlación entre $u_t - u_{t-1}$. Analizamos que las R cuadrados son más altas en regresiones en niveles, y más bajos en regresiones en primeras diferencias. Sin embargo, debemos observar la estadística de DW, pues en niveles es más baja e implica una mala especificación del modelo, en todo caso debe regresar en primeras diferencias y el R cuadrado bajo debe entenderse solo como que las variables explicativas no tienen relación entre sí. Por eso cuando se regresa en primeras diferencias se denota la verdadera naturaleza de la relación entre las variables explicativas.

Paso 4

Luego que hemos discutidos sobre lo que pasa al trabajar con modelos en niveles, cuasi diferencias y primeras diferencias, es importante valorar el procedimiento de estimación con errores autocorrelacionados. Nos centraremos en estimaciones con cuasi diferencias, es decir regresar $y_t - \rho y_{t-1}$ sobre $x_t - \rho x_{t-1}$. Como se mencionó antes, se discutirá el caso más sencillo, en el que es posible resumir la estructura completa de correlación de los errores de u_t en un solo parámetro ρ Ëste sería el caso si los errores u_t fueran autorregresivos de primer orden, es decir $u_t = \rho u_{t-1} + e_t$, donde e_t tiene media cero, varianza común σ^2_e y no tiene una correlación serial. A esta ecuación se le llama de autorregresión, debido a que el modelo usual de regresión es u_t sobre u_{t-1} y se conoce como autorregresión de primer orden debido a que u_t se regresa con un solo rezago. Si aparecen dos rezagos, se conoce como autorregresión de segundo orden. Si hubiera tres rezagos, se llamaría de tercer orden, etc. Entonces si los errores satisfacen la ecuación descrita para un rezago AR(1), y le llamaremos autorregresivo de primer orden, AR(2) autorregresiva de segundo orden, etc.

El problema para regresar en cuasi diferencias es que no se conoce el parámetro ρ entonces habrá que calcularlo. Existen dos tipo de procedimientos para calcularla:

a) Procedimientos iterativos
b) Procedimientos de búsqueda reticular

En los procedimientos iterativos, el más antiguo es el de Cochrane y Orcutt. En éste se calcula $y_t = \alpha + \beta x_t + \mu_t$ mediante OLS, se obtienen los residuos (resid) y se calcula $\rho = \sum u_t u_{t-1} / \sum u_t^2$. Una vez que se obtiene el estimado de ρ se construyen las variables trasformadas y se estima $y_t - \rho y_{t-1}$ sobre $x_t - \rho x_{t-1}$ para ello desde la ventana principal deberá generar las nuevas series y regresar el modelo especificado.

Si hay demasiadas variables explicativas en la ecuación, el método de DW supone regresar demasiadas variables (dos veces el número de variables explicativas más y_{t-1})

El programa Evews implica utilizar al final de la ecuación AR(1) para este caso obteniendo los resultados que aparecen abajo en el cuadro.

El procedimiento de búsqueda reticular sugerido por Hildreth y Lu en 1960. El procedimiento es el siguiente. Se calcula $y_t - \rho y_{t-1}$ sobre $x_t - \rho x_{t-1}$ para los diferentes valores de ρ en intervalos de longitud 0.1 en el rango $-1 \leq \rho \leq 1$. Se calcula la regresión de $y_t - \rho y_{t-1}$ sobre $x_t - \rho x_{t-1}$ y la suma de cuadrados residual en cada caso. Se elige el valor de ρ para el cual la suma de cuadrados residual es mínima. Este procedimiento se repite para intervalos cada vez más pequeños de ρ, en torno a este valor. Por ejemplo, si el valor de ρ para el cual la RSS mínima es –0.4, repetir este procedimiento de búsqueda para valores de ρ a intervalos de 0.01 en el rango de $-0.5 < \rho < -0.3$.

Apliquemos los dos procedimientos estudiados:

Estimado de	OLS	Hildreth-Lu	Cochrane-Orcutt
α	-3.93771449855	-2.909	-2.47309987628
L_1	1.45078600604	1.092	1.03121670705
K_1	0.383808125307	0.570	0.548251791541
R^2	0.99462715616		0.996698204423
RSS	0.0433823043027	0.02635	0.025999439648
DW	0.858080234731		1.17143333524
ρ	0.5710	0.77	0.846297390436

Para el caso que estudiamos lo importante es considerar que estimar el modelo en niveles y detectamos el problema de Autocorrelación, este se puede corregir aplicando AR(1) si los errores son de primer orden.

Pruebas para la correlación serial en los modelos con variables dependientes rezagadas

La prueba h es la indicada para estos casos. Esta prueba utiliza $h= \rho\sqrt{n} / 1 - n\ V(\alpha)$ como una variable normal estándar. En este caso, ρ es la correlación serial de primer orden estimada a partir de los residuos OLS, $V(\alpha)$ es la varianza del estimador OLS de α y n el tamaño de la muestra. Si $n\ V(\alpha) > 1$, la prueba no aplicable. En este caso, Durbin sugiere la prueba siguiente.

A partir de la estimación de OLS de la ecuación en niveles, se calculan los residuos (μ). Después se regresa μ_t sobre μ_{t-1}, y, y_{t-1} x_t. La prueba para $\rho= 0$ se efectúa al probar la significancia del coeficiente de μ_{t-1} en esta última regresión.

Aplique la fórmula al Test de la ecuación en niveles considerando el rezago de la variable dependiente.

Prueba de LM

Breusch y Godfrey discuten algunas pruebas generales de fácil aplicación, válidas para hipótesis muy generales en torno a la correlación general del Multiplicador de Lagrange(LM). Analizamos en que consiste esta prueba:

Consideremos el modelo de regresión $y_t=\sum x_{it} \beta_i + u_t,$ $t= 1,2,.....,n$ y

$$\mu_t = \rho_1 \mu_{t-1} + \rho_2 \mu_{t-2} +.....\rho_p \mu_{t-p} +e_t \quad e_t \sim IN(0,\sigma^2)$$

Nos interesa probar $H_0=\rho_1=\rho_2=.....\rho_p=0$. Las xs en la primera ecuación incluyen también las variables dependientes rezagadas. La prueba del multiplicador de Lagrange es la siguiente: Primero, se estima la primera ecuación y se obtienen los residuos de mínimos cuadrados. A continuación, se calcula estimar la ecuación de regresión:

$$u_t=\sum x_{it} \gamma_i + \sum \mu_{t-1} \rho_i + \eta_t,$$

y se prueba se los coeficientes de μ_{t-1} son todos ceros. Se tomará la estadística F convencional y se utilizará $p \bullet F$ como χ^2 con ρ grados de libertad. Se utilizará la prueba χ^2 en lugar de F, debido a que la prueba del multiplicador de Lagrange es para muestras grandes.

Practique esta prueba estimando log(x) log(k1) log(l1) log(x(-1)) escriba esto en la ventana de comando y cuando aparezca el Test, escriba auto y aparecerán los resultados de la prueba.

Modelos ARCH y correlación serial

El modelo ARCH es sugerido por Engel útil en el análisis de precios especulativos. ARCH son las iniciales en inglés de heterocedasticidad condicional autorregresiva (Autoregressive Conditional Heteroskedasticity).

Cuando se escribe el modelo autorregresivo simple $y_t=\lambda\ y_{t-1} + \varepsilon_t,$ $\varepsilon_t \sim IN(0,\sigma^2)$

Se dice que la media condicional $E(y_t / y_{t-1}) = \lambda y_{t-1}$ depende de t, pero la varianza condicional var $(y_t / y_{t-1}) = \sigma^2$ es una constante. La media condicional de y_t es cero y la varianza incondicional es $\sigma^2/ (1 - \lambda^2)$.

El modelo ARCH es una generalización de esto, en el sentido de que la variable condicional también se vuelve una función del pasado. Si la densidad condicional $f(y_t / z_{t-1})$ es normal, una expresión general del modelo ARCH es $y_t / z_{t-1} \sim IN[g(z_{t-1}), h(z_{t-1})]$
Para hacer esto operacional, Engel especifica que la media condicional $g(z_{t-1})$ es una función lineal de las variables z_{t-1} y h como $h_t = \alpha_0 + \alpha_1 \varepsilon^2_{t-1} + \alpha_2 \varepsilon^2_{t-2} + \ldots\ldots\ldots\ldots \alpha_2 \varepsilon^2_{t-p}{}^{\varepsilon}$ donde $\varepsilon_t = y_t - g_t$. En el caso ,más sencillo, es posible considerar el modelo $y_t = \lambda y_{t-1} + \beta x_t, + \varepsilon_t$ $\varepsilon_t, \sim IN(0,\sigma^2)$ $h_t = \text{var}\, \varepsilon_t = \alpha_0 + \alpha_1 + \alpha_1 \varepsilon^2_{t-1}$

Por otro lado, la manera más rápida de comprobar si el problema de Autocorrelación es debido a un efecto ARCH es desde la ventana del TEST ESTIMATE en el cuadros Estimation Setting / Method hacemos clip en la esquina y buscamos ARCH en la caja de dialogo: en la parte superiror usted debe reflejar la variable dependiente seguida de los regresores estandares los cuales describen la variable dependiente. Elija el boton de OPTION para requerir Heteroskedasticity Consistent Covariance.

ARCH opciones son:
E or EGARCH EGARCH estimation
T or TARCH Asymmetric or Threshold ARCH
C Component (permanent and transitory) ARCH
A Asymmetric component (permanent and transitory) ARCH
V ARCHM--ARCH in mean; variance in the mean equation.
M ARCHM--ARCH in mean; standard deviation (volatility) in the mean equation.

La idea es observar cómo se comporta la estadística DW después de aplicar estas pruebas, pues la prueba ARCH permitirá juzgar si la correlación serial observada es espuria.

Revise los tests que ha corrido en laboratorios anteriores y observe la estadística DW y realice las pruebas que hemos discutido.

LABORATORIO # 8

Multicolinearidad

Carrera : I
Asignatura : Econometría
Año lectivo : III
Laboratorio No : 7
Clase No : 7
Fecha :
Unidad VI : Multicolinearidad
Docente : Dr. Carlos Alberto Zúniga González

Objetivo

Interpretar los estimados de coeficientes individuales si las variables tienen una alta intercorrelación. Así como analizar las diferentes soluciones a este problema.

Introducción

El problema de intercorrelación de las variables explicativas se conoce como Multicolinealidad. Esta situación se ubica en los supuestos anteriormente estudiados, a saber:

11. $E(\mu_i)= 0$. Media
12. $V(\mu_i)= \sigma^2$ para todo i. Varianza común. (**HETEROCEDASTICIDAD)**
13. (μ_i) y (μ_j) son independientes para todo $i \neq j$. Independencia. **(AUTOCORRELACION)**
14. (x_j) es no estocástico. Independencia. (μ_i) y (x_j) son independientes para todo i y j. Esta suposición es consecuencia automática si las (x_j) se consideran variables no aleatorias. **(MULTICOLINEARIDAD)**
15. (μ_i) está normalmente distribuido para todo i. Normalidad. Junto con las suposiciones 1, 2, y 3, esto implica que los (μ_i) son independientes y tienen una distribución normal con media 0 y varianza común σ^2. Esto se escribe como $\mu \sim IN(0, \sigma^2)$

En este modulo nos referimos al punto cuatro. Con frecuencia, los datos que se utilizan en el análisis de regresión múltiple no proporcionan repuestas definitivas a las preguntas prácticas ¿Qué tan alta es esta intercorrelación entre las variables independientes? ¿Qué tan alta debe ser para ocasionar problemas en la inferencia sobre los parámetros individuales? , y ¿Qué es posible hacer al respecto?

Raganar Frisch introdujo por primera vez el término Multicolinealidad en 1934, en un libro sobre análisis de confluencia, refiriéndose a una situación en la que las variables que se trataban sujetas a dos o más relaciones. En dicho análisis, no hubo dicotomía entre las variables explicativas y explicadas. Frisch supuso que todas ellas estaban sujetas a error y, dadas las varianzas y covarianzas muestrales, el problema consistía en estimar las diferentes relaciones lineales entre las variables verdaderas. El problema, entonces, era de errores en las variables. Sin embargo, se

discutirá el problema de Multicolinealidad como suele discutirse en el análisis de regresión múltiple, es decir, el problema de intercorrelaciones latas entre variables explicativas.

Posibles Causas

Uno de los síntomas de la multicolinearidad es que los estimados del parámetro de las variables explicativas son sensibles a la inclusión a la eliminación de observaciones. Es posible, en la práctica, verificar este aspecto de la Multicolinealidad mediante la eliminación o la inclusión de algunas observaciones y examinar la sensibilidad de las estimaciones a tales perturbaciones.

Otro síntoma del problema de la Multicolinealidad, que muchas veces se menciona, es que los errores estándar de los coeficientes de regresión estimados son muy elevados. Se sugiere que la varianza de los parámetros es alta sí:
La varianza es alta, la suma de cuadrados del error es bajo y el coeficiente de determinación es alto. Ello implica que los coeficientes de determinación no indican nada sobre la existencia de un problema de Multicolinealidad.

Cuando se tiene más de dos variables explicativas, las correlaciones simples entre ellas carecen de sentido. Luego, se deberá utilizar los valores de R^2_i para medir el grado de intercorrelación entre la variable explicativa y no las correlaciones simples entre éstas.

Cómo detectar el problema

Klein dice: La intercorrelación de variables no es necesariamente un problema, a menos que sea muy elevada en relación con el grado global de la regresión múltiple. En virtud de la regla de Klein, la Multicolinealidad se considera problema sólo si $R^2_y < R^2_i$, donde R^2_y es el cuadrado del coeficiente de correlación entre (y) y las variables explicativas, R^2_i es el cuadrado del coeficiente de correlación múltiple entre x_i y las demás variables explicativas.

Algunos autores resumen este problema de la siguiente manera:

1. Si se tienen más de dos variables explicativas, deberían utilizarse los valores de R^2_i para el grado de intercorrelación entre las variables explicativas, y no las correlaciones simples entre éstas.
2. Sin embargo, si la multicolinearidad es o no un problema para hacer inferencias en los parámetros, dependerá de otros factores además de R^2_i ,.Lo que importa son los errores estándar y las relaciones t. Por supuesto, sería mucho mejor si R^2_i fuera baja. Pero esto no es un consuelo.
3. Si la correlación entre las variables explicativas es alta se sugiere que la multicolinearidad es seria.
4. Si los errores estándar o las relaciones t son significativas, se sugiere que la multicolinearidad no es muy seria.
5. La estabilidad de los coeficientes estimados cuando se eliminan algunas observaciones, luego de probar esta estabilidad con resultado significativo, también sugiere una multicolinearidad no muy seria.

Soluciones al problema de la multicolinearidad

Ha habido diferentes soluciones al problema de la multicolinearidad:

(a) La regresión por cordillera, tratada en gran cantidad de textos.
(b) La regresión por componentes principales.
(c) La eliminación de variables.

Todas estas llamadas soluciones son, en realidad, procedimientos ad hoc. Cada una supone el uso de cierta información previa y es mejor examinarla antes de proceder con una solución mecánica sugerida por otros.

El problema principal es la falta de información suficiente para responder a las preguntas expuestas. Las únicas soluciones son obtener más datos, averiguar qué preguntas es posible responder con los datos disponibles, y examinar la información previa que sería de mayor utilidad.

Discutamos un poco la situación de estas soluciones:

Tabla 7.1 Datos sobre consumo, ingreso y activos líquidos

Año	Trimestre	C	Y	L
1952	I	220	238.1	182.7
	II	222.7	240.9	183.0
	III	223.8	245.8	184.4
	IV	230.2	248.8	187.0
1953	I	234.0	253.3	189.4
	II	236.2	256.1	192.2
	III	236.0	255.9	193.8
	IV	234.1	255.9	194.8
1954	I	233.4	254.4	197.3
	II	236.4	254.8	197.0
	III	239.0	257.0	200.3
	IV	243.2	260.9	204.2
1955	I	248.7	263.0	207.6
	II	253.7	271.5	209.4
	III	259.9	276.5	211.1
	IV	261.8	281.4	213.2
1956	I	263.2	282.0	214.1
	II	263.7	286.2	216.5
	III	263.4	287.7	217.3
	IV	266.9	291.0	217.3
1957	I	268.9	291.1	218.2
	II	270.4	294.6	218.5
	III	273.4	296.1	219.8
	IV	272.1	293.3	219.5
1958	I	268.9	291.3	220.5
	II	270.9	292.6	222.7

	III	274.4	299.9	225.0
	IV	278.7	302.1	229.4
1959	I	283.8	305.9	232.2
	II	289.7	312.5	235.2
	III	290.8	311.3	237.7
	IV	292.8	313.2	237.7
1960	I	295.4	315.4	238.4
	II	299.5	320.3	238.4
	III	298.6	321.0	240.1
	IV	299.6	320.1	243.3
1961	I	297.0	318.4	246.1
	II	301.6	324.8	250.0
	III			
	IV			

(a) La regresión por cordillera, tratada en gran cantidad de textos.

Es una de las soluciones que se sugiere con más frecuencia para el problema de la multicolinearidad, introducida por Hoerl y Kennard. En términos simples consiste en añadir una constante λ a las varianzas de las variables explicativas antes de resolver las ecuaciones normales. Supongamos $S_{11}=200$ y $S_{22}=113$ le sumamos 5, es fácil observar que el cuadrado de las relaciones es: $r^2_{12}= (150)^2/(205)*(118) = 0.930$, por lo tanto las intercorrelaciones se reducen. Este método no es recomendable porque se le critica de utilizar argumentos subjetivos, no es invariante ante las unidades de medida de las variables explicativas y transformaciones lineales de las variables.

(b) La regresión por componentes principales.

Supongamos que tenemos k variables explicativas. Entonces es posible considerar algunas funciones lineales de estas variables:

$Z_1= a_1 x_1 +a_2 x_2 +.....+a_k x_k$

$Z_2= b_1 x_1 +b_2 x_2 +.....+b_k x_k$ etc. Supongamos que las a se eligen de modo que la varianza de Z_1 se maximice, sujeta a la condición de que $a^2_1 +a^2_2 +.....+a^2_k = 1$ Esto se conoce como condición de normalización. Es necesaria, o de otro modo la varianza de Z_1 se elevará en forma indefinida. Se dice entonces, que Z_1 es el primer componente. Es la función lineal de las xs que tiene la mayor varianza sujeta a la regla de normalización.

El proceso de maximizar la varianza de la función lineal z sujeta a la condición de que el cuadrado de la suma de los coeficientes de las xs es igual a uno, produce k soluciones. Correspondiendo a esto, so construyen k funciones lineales Z_1 ,Z_2 ,Z_3..... Z_k. Estas se conocen como componentes principales de las xs. Estos componentes principales tienen las siguientes propiedades:

1. $Var(Z_1)+var(Z_2)$ +.....+ $var(Z_k)= Var(x_1)+var(x_2)$ +.....+ $var(x_k)$
2. A diferencia de las x, que están correlacionadas, las z son ortogonales o no correlacionadas. Por tanto existe multicolinearidad cero entre ellas.

Trabajemos con la tabla 7.3 para ilustrar el método. El conjuntos de datos de Malinvaud refiere la Importaciones, producción, formación de activos y consumo en Francia (millones de francos nuevos a precios de 1959) Las variables son y=importaciones, x1= producción doméstica, x2=fromación de activos, x3=consumno.

IMPORTACIONES y	PRODUCCION X1	ACTIVOS X2	CONSUMO X3
15.9	149.3	4.2	108.1
16.4	161.2	4.1	114.8
19	171.5	3.1	123.2
19.1	175.5	3.1	126.9
18.8	180.8	1.1	132.1
20.4	190.7	2.2	137.7
22.7	202.1	2.1	146
26.5	212.4	5.6	154.1
28.1	226.1	5	162.3
27.6	231.9	5.1	164.3
26.3	239	0.7	167.6
31.1	258	5.6	176.8
33.3	269.8	3.9	186.6
37	288.4	3.1	199.7
43.3	304.5	4.6	213.9
49	323.4	7	223.8
50.3	336.8	1.2	232
56.6	353.9	4.5	242.9

Los resultados de la regresión de y sobre x1,x2, x3 son los siguientes:

Dependent Variable: IMPORTACIONES
Method: Least Squares
Date: 02/01/04 Time: 23:48
Sample: 1949 1966
Included observations: 18

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-19.725107117	4.12525254838	-4.7815514046	0.000292539266951
PRODUCCION	0.0322044690892	0.18688431604	0.172323016567	0.86564988583
ACTIVOS	0.414199097131	0.322259758142	1.28529574874	0.219545083771
CONSUMO	0.242747006499	0.285360658516	0.850667389685	0.409267983952
R-squared	0.973043059413	Mean dependent var		30.0777777778

Adjusted R-squared	0.967266572145	S.D. dependent var	12.4813062832
S.E. of regression	2.25816558416	Akaike info criterion	4.6601126731
Sum squared resid	71.3903652769	Schwarz criterion	4.85797306374
Log likelihood	-37.9410140579	F-statistic	168.448923053
Durbin-Watson stat	0.240325049818	Prob(F-statistic)	3.21165592624e-11

La R^2 es muy elevada y la relación F es altamente significativa, pero las relaciones t individual son no significativas. Esto demuestra la existencia del problema de la multicolinearidad. Chatterjee y Price afirman que, antes de hacer algún análisis posterior, debería observarse los residuos de esta ecuación, ellos descubren (le queda como tarea observar la gráfica de los residuos) un patrón distintivo: los residuos disminuyen hasta 1960 y después aumentan. Chatterjee y Price establecen que la dificultad de este modelo es que el mercado común europeo comenzó a operar en ese año, lo que provocó cambios en las relaciones de importación y exportación. Por lo tanto, se eliminan los años posteriores a 1959 y se consideran sólo los 11 que van de 1949 a 1959. Los resultados son los siguientes:

Dependent Variable: IMPORTACIONES
Method: Least Squares
Date: 02/02/04 Time: 00:12
Sample: 1949 1959
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-10.1279881552	1.2121599585	-8.35532314382	6.89918291601e-05
PRODUCCION	-0.0513961597135	0.0702799870766	-0.731305765004	0.488344308322
ACTIVOS	0.586949042664	0.094618420693	6.20332741094	0.000443813469459
CONSUMO	0.28684867564	0.102208114323	2.80651568165	0.0262770992253
R-squared	0.991896552056	Mean dependent var		21.8909090909
Adjusted R-squared	0.988423645794	S.D. dependent var		4.54366692121
S.E. of regression	0.48886887832	Akaike info criterion		1.68184273402
Sum squared resid	1.67294946133	Schwarz criterion		1.82653192413
Log likelihood	-5.25013503712	F-statistic		285.60994086
Durbin-Watson stat	2.73975522934	Prob(F-statistic)		1.11171497799e-07

Observe la gráfica de los resid de modo que no existen patrones sistemáticos, aún si R^2 es muy elevada, el coeficiente de x1 no es significativo. Por tanto, existe un problema de multicolinearidad. Para saber que se debe hacer al respecto, primero se observará las correlaciones simples entre las variables explicativas. Estas son r^2_{12} =0, 026, r^2_{13} =0.99, y r^2_{23} =0.036. Se sospecha que la alta correlación entre x1 y x3 podría ser el origen del problema.

Se sospecha que la alta correlación entre x1 y x3 podría sé el origen del problema.

Explicaremos en que consiste de manera ilustrativa, porque Eviews no calcula los componentes principales, los resultados a partir de un programa de componentes principales son:

Z1=0.7063X1+0.0435X2+0.7065X3

Z2=-0.0357X1+0.9990X2-0.0258X3

Z3=-0.7070X1-0.0070X2+0.7072X3

X1, X2 y X3 son los valores normalizados de x1, x2, y x3.

(c) La eliminación de variables

El problema de la multicolinearidad es, en esencia, la falta de información suficiente en la muestra, que permita una estimación precisa de los parámetros; entonces, es posible obtener estimadores

para aquellos parámetros en los que tenemos interés y que tengan errores cuadráticos medios más pequeños que los estimadores de mínimo cuadrado ordinarios.

Ejemplo:

Consideremos el modelo Importaciones = β_1 produccion + β_2 activos + β_3 consumo + μ de la tabla 7.3 y el problema de que produccion y activos tienen una correlación muy alta. Supongamos que el interés principal radica en produccion. Entonces eliminemos activos y estimamos la ecuación, mediante View/Coefficient Tests/Redundant Variables.

Observemos que el coeficiente de producción en el modelo completo fue de -0.05 ahora fue de 0.04, observemos que el coeficiente de determinación a disminuido favorablemente

Redundant Variables: ACTIVOS				
F-statistic	1.651985	Probability		0.219545
Log likelihood ratio	2.007728	Probability		0.156500
Test Equation:				
Dependent Variable: IMPORTACIONES				
Method: Least Squares				
Date: 02/17/04 Time: 13:17				
Sample: 1949 1966				
Included observations: 18				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-18.63332	4.123649	-4.518649	0.0004
PRODUCCION	0.042764	0.190718	0.224227	0.8256
CONSUMO	0.230341	0.291330	0.790653	0.4415
R-squared	0.969862	Mean dependent var		30.07778
Adjusted R-squared	0.965844	S.D. dependent var		12.48131
S.E. of regression	2.306720	Akaike info criterion		4.660542
Sum squared resid	79.81435	Schwarz criterion		4.808937
Log likelihood	-38.94488	F-statistic		241.3567
Durbin-Watson stat	0.540904	Prob(F-statistic)		0.000000

El F stadistic nos indica que la variable omitida activos es una variable demás en tanto no sea estadísticamente significativo.

En el caso contrario se aplica el test importaciones C producción consumo y supongamos que activos es una variable omitida, es importante estar claros que la variable omitida debe tener el mismo número de observaciones. Haga estas pruebas con cada una de las variables.

ANEXOS

Interpretación de variables del test

TRADUCCION

Seguir son definiciones informales de estos resultados. Pulse el botón para las definiciones matemáticas.

Los Coeficientes de la regresión

Cada coeficiente multiplica la variable correspondiente formando la predicción mejor de la variable dependiente. El coeficiente mide la contribución de su variable independiente a la predicción. El coeficiente de la serie llamado C es la constante o intercepta en la regresión--es el nivel bajo de la predicción cuando todas las otras variables independientes son el cero. Los otros coeficientes se interpretan como la cuesta de la relación entre la variable independiente correspondiente y la variable dependiente.

Los Errores normales

Éstos miden la fiabilidad estadística de los coeficientes de la regresión--el más grande el error normal, el ruido más estadístico infecta el coeficiente. Según la teoría de la regresión, hay aproximadamente 2 oportunidades en 3 que el verdadero coeficiente de la regresión queda dentro de un error normal del coeficiente informado, y 95 oportunidades fuera de 100 que queda dentro de dos errores normales.

la t-estadística

Ésta es una estadística de la prueba para la hipótesis que un coeficiente tiene un valor particular. La t-estadística para probar si un coeficiente es el cero (es decir, si la variable no pertenece en la regresión) es la proporción del coeficiente a su error normal. Si la t-estadística excede uno en la magnitud que es por lo menos probablemente dos-tercero que el verdadero valor del coeficiente no es ningún cero, y si la t-estadística excede dos en la magnitud que tiene por lo menos probablemente 95 por ciento años que el coeficiente no es ningún cero.

La probabilidad

La última columna muestra la probabilidad de dibujar una t-estadística de la magnitud del uno sólo a la izquierda de una distribución de t. Con esta información, usted puede decir de una ojeada si usted rechaza o acepta la hipótesis que el verdadero coeficiente es el cero. Normalmente, una probabilidad baja que se toman .05 como la evidencia fuerte de rechazo de esa hipótesis.

R2

Esto mide el éxito de la regresión prediciendo los valores de la variable dependiente dentro de la muestra. R2 tiene uno años si la regresión encaja perfectamente, y ceros si encaja ningún bueno que la media simple de la variable dependiente. R2 es el fragmento de la variación de la variable dependiente explicado por las variables independientes. Puede ser negativo si la regresión no tiene un intercepte o constante, o si el dos-fase se usan los menores cuadrados.

R2 ajustó para los grados de libertad

Éste es un pariente íntimo de R2 en que se usan medidas ligeramente diferentes de las variaciones. Está menos de R2 (proporcionó hay más de una variable independiente) y puede ser negativo.

El Error normal de la Regresión

Ésta es una medida sumaria del tamaño de los errores de la predicción. Tiene las mismas unidades como la variable dependiente. Sobre dos-tercero de todos los errores las magnitudes tienen de menos de un error normal. El error normal de la regresión es una medida de la magnitud de los residuos. Sobre dos-tercero de los residuos quedará en un rango del menos un error normal a la ventaja un error normal, y 95 por ciento de los residuos quedarán en un rango de menos dos a la ventaja dos errores normales.

La suma de Residuos Cuadrados

Esto es lo que dice. Usted puede querer acostumbrar este número como una entrada a ciertos tipos de pruebas.

Anote la Probabilidad

Éste es el valor de la función de probabilidad de leño evaluado a los valores estimados de los coeficientes. Las pruebas de proporción de probabilidad pueden ser dirigidas mirando la diferencia entre las probabilidades del leño de restringió y versiones sin restricciones de una ecuación.

Durbin-Watson Statistic

Ésta es una estadística de la prueba para la correlación de serie. Si está menos de 2, hay evidencia de correlación de serie positiva. Pulse el botón aquí para una discusión más extensa y también una discusión de la Q-estadística y el Breusch-Godfrey LM pruebe para correlación de serie los dos de que puede ser superior al Durbin-Watson la prueba.
Un hallazgo común en la regresión de la tiempo-serie es que los residuos se ponen en correlación con sus propios valores retrasados. La correlación es incoherente con la asunción que miente teoría de la regresión que las perturbaciones no se ponen en

correlación entre sí en cualquier moda detrás. Econometricians han desarrollado extensiones de análisis de la regresión para tratar con la correlación de serie.

La manera normal de descubrir la correlación de serie es basado en el Durbin-Watson la estadística. Una estadística mucho debajo de 2 es una indicación de correlación de serie positiva.
Si usted concluye que esa correlación del folletín está presente en los residuos en su regresión, la manera más simple de proceder es agregar una especificación de autoregressive de primero-orden al modelo de la regresión. Aquí, nosotros presentaremos el acercamiento del libro de cocina a la especificación. Las secciones restantes de este capítulo echan una mirada más profunda al asunto.

Suponga que su especificación de la regresión era

LAS VENTAS C NEW_ORDERS(0 A -4)

y que el Durbin-Watson la estadística de la menor estimación de los cuadrados ordinaria era 0.64. Usted concluye que hay correlación de serie positiva sustancial. Para incorporar la correlación de serie en su ecuación, usted debe agregar AR(1) al lado diestro. Empuje el botón de la Estimación en el toolbar de la ecuación y teclee en AR(1). Entonces el empujón OK y usted verán un nuevo juego de resultados de la regresión.
Los nuevos resultados se parecerán mucho el más temprano. Habrá una más línea en la mesa de coeficientes y los errores normales, AR(1 etiquetado). El coeficiente asoció con AR(1) es la estimación de la correlación de serie de los residuos. Si su Durbin-Watson la estadística era 0.64, el AR(1) el coeficiente probablemente estará alrededor de 0.7.

Hay dos mejoras en estos resultados. Primero, ambas las estimaciones de los coeficientes ellos y los errores normales estimados de los coeficientes son estadísticamente más fiables. Los menores resultados de los cuadrados antes de que usted agregara AR(1) es menos fiable que ellos parecen, porque los errores normales se subestiman. Segundo, la previsión a corto plazo se mejora considerablemente estimando y usando el coeficiente de la correlación de serie.
Luego el tema: La Teoría de la Correlación de serie

El Akaike Información Criterio

El Akaike Inromation Criterio, o AIC, es una guía a la selección del número de condiciones en una ecuación. Es basado en la suma de residuos cuadrados pero lugares una multa en los coeficientes extras. Bajo ciertas condiciones, usted puede escoger la longitud de una distribución de retraso, por ejemplo, escogiendo la especificación con el valor más bajo del AIC.

El Criterio de Schwarz

El criterio de Schwarz es una alternativa al AIC con básicamente la misma interpretación pero una multa más grande para los coeficientes extras.

La F-estadística

Ésta es una prueba de la hipótesis que todos los coeficientes en una regresión son el cero (excepto el intercepte o constante). Si la F-estadística excede un nivel crítico, por lo menos uno de los coeficientes probablemente es non-ceros. Por ejemplo, si hay tres variables independientes y 100 observaciones, una F-estadística sobre 2.7 indica que la probabilidad es por lo menos 95 por ciento que uno o más de los tres coeficientes es non-ceros. La probabilidad simplemente dada debajo del F
La -estadística le permite que lleve a cabo esta prueba de una ojeada.

Pulse el botón para las definiciones matemáticas

Luego el tema: Los Valores reales y En buen salud y Residuos

****

VERSION EN ENGLISH

Following are informal definitions of these results. Click for mathematical definitions.

Regression Coefficients

Each coefficient multiplies the corresponding variable in forming the best prediction of the dependent variable. The coefficient measures the contribution of its independent variable to the prediction. The coefficient of the series called C is the constant or intercept in the regression--it is the base level of the prediction when all of the other independent variables are zero. The other coefficients are interpreted as the slope of the relation between the corresponding independent variable and the dependent variable.

Standard Errors

These measure the statistical reliability of the regression coefficients--the larger the standard error, the more statistical noise infects the coefficient. According to regression theory, there are about 2 chances in 3 that the true regression coefficient lies within one standard error of the reported coefficient, and 95 chances out of 100 that it lies within two standard errors.

t-Statistic

This is a test statistic for the hypothesis that a coefficient has a particular value. The t-statistic to test if a coefficient is zero (that is, if the variable does not belong in the regression) is the ratio of the coefficient to its standard error. If the t-statistic exceeds one in magnitude it is at least two-thirds likely that the true value of the coefficient is not zero, and if the t-statistic exceeds two in magnitude it is at least 95 percent likely that the coefficient is not zero.

Probability

The last column shows the probability of drawing a t-statistic of the magnitude of the one just to the left from a t distribution. With this information, you can tell at a glance if you reject or accept the hypothesis that the true coefficient is zero. Normally, a probability lower than .05 is taken as strong evidence of rejection of that hypothesis.

R2

This measures the success of the regression in predicting the values of the dependent variable within the sample. R2 is one if the regression fits perfectly, and zero if it fits no better than the simple mean of the dependent variable. R2 is the fraction of the variance of the dependent variable explained by the independent variables. It can be negative if the regression does not have an intercept or constant, or if two-stage least squares is used.

R2 adjusted for degrees of freedom

This is a close relative of R2 in which slightly different measures of the variances are used. It is less than R2 (provided there is more than one independent variable) and can be negative.

Standard Error of the Regression

This is a summary measure of the size of the prediction errors. It has the same units as the dependent variable. About two-thirds of all the errors have magnitudes of less than one standard error. The standard error of the regression is a measure of the magnitude of the residuals. About two-thirds of the residuals will lie in a range from minus one standard error to plus one standard error, and 95 percent of the residuals will lie in a range from minus two to plus two standard errors.

Sum of Squared Residuals

This is just what it says. You may want to use this number as an input to certain types of tests.

Log Likelihood

This is the value of the log likelihood function evaluated at the estimated values of the coefficients. Likelihood ratio tests may be conducted by looking at the difference between the log likelihoods of restricted and unrestricted versions of an equation.

Durbin-Watson Statistic

This is a test statistic for serial correlation. If it is less than 2, there is evidence of positive serial correlation. Click here for a more extensive discussion and also a discussion of the Q-statistic and the Breusch-Godfrey LM test for serial correlation, both of which may be superior to the Durbin-Watson test.
A common finding in time-series regression is that the residuals are correlated with their own lagged values. The correlation is inconsistent with the assumption lying behind regression theory that the disturbances are not correlated with each other in any fashion. Econometricians have developed extensions of regression analysis to deal with serial correlation.

The standard way to detect serial correlation is based on the Durbin-Watson statistic. A statistic much below 2 is an indication of positive serial correlation.
If you conclude that serial correlation is present in the residuals in your regression, the simplest way to proceed is to add a first-order autoregressive specification to the regression model. Here, we will present the cookbook approach to the specification. The remaining sections of this chapter take a deeper look at the subject.

Suppose that your regression specification was

SALES C NEW_ORDERS(0 TO -4)

and that the Durbin-Watson statistic from ordinary least squares estimation was 0.64. You conclude that there is substantial positive serial correlation. To incorporate the serial correlation in your equation, you should add AR(1) to the right-hand side. Push the Estimate button on the equation's toolbar and type in AR(1). Then push OK and you will see a new set of regression results.
The new results will look much like the earlier ones. There will be one more line in the table of coefficients and standard errors, labeled AR(1). The coefficient associated with AR(1) is the estimate of the serial correlation of the residuals. If your Durbin-Watson statistic was 0.64, the AR(1) coefficient will probably be around 0.7.

There are two improvements in these results. First, both the estimates of the coefficients themselves and the estimated standard errors of the coefficients are statistically more reliable. The least squares results before you added AR(1) are less reliable than they seem, because the standard errors are understated. Second, short-term forecasting is considerably improved by estimating and using the serial correlation coefficient.
Next topic: Serial Correlation Theory
Akaike Information Criterion

The Akaike Inromation Criterion, or AIC, is a guide to the selection of the number of terms in an equation. It is based on the sum of squared residuals but places a penalty on extra coefficients. Under certain conditions, you can choose the length of a lag distribution, for example, by choosing the specification with the lowest value of the AIC.

Schwarz Criterion

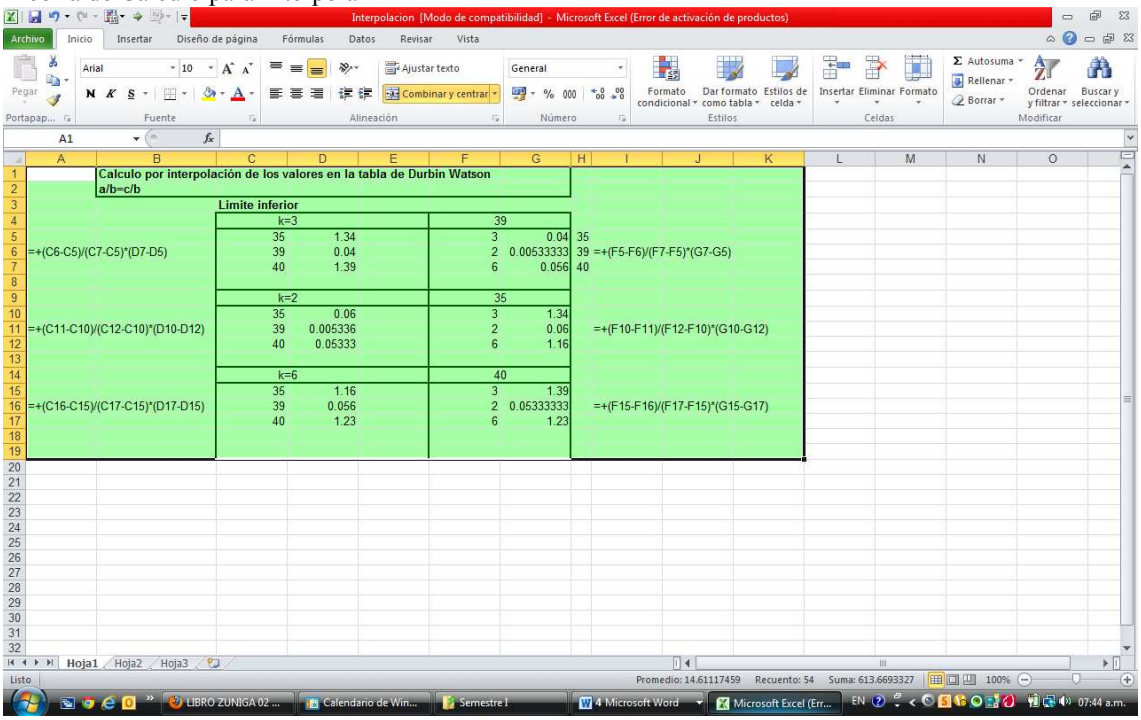
The Schwarz criterion is an alternative to the AIC with basically the same interpretation but a larger penalty for extra coefficients.

F-Statistic

This is a test of the hypothesis that all of the coefficients in a regression are zero (except the intercept or constant). If the F-statistic exceeds a critical level, at least one of the coefficients is probably non-zero. For example, if there are three independent variables and 100 observations, an F-statistic above 2.7 indicates that the probability is at least 95 percent that one or more of the three coefficients is non-zero. The probability given just below the F
-statistic enables you to carry out this test at a glance.

Click for mathematical definitions `1Next topic: Actual and Fitted Values and Residuals

Meoria de Cálculo para Interpolar



	Calculo por interpolación de los valores en la tabla de Durbin Watson a/b=c/b									
		Limite inferior								
		k=3			39					
		35	1.34		3	0.04	35			
=+(C6-C5)/(C7-C5)*(D7-D5)		39	0.04		2	0.00533333	39	=+(F5-F6)/(F7-F5)*(G7-G5)		
		40	1.39		6	0.056	40			
		k=2			35					
		35	0.06		3	1.34				
=+(C11-C10)/(C12-C10)*(D10-D12)		39	0.005336		2	0.06		=+(F10-F11)/(F12-F10)*(G10-G12)		
		40	0.05333		6	1.16				
		k=6			40					
		35	1.16		3	1.39				
=+(C16-C15)/(C17-C15)*(D17-D15)		39	0.056		2	0.05333333		=+(F15-F16)/(F17-F15)*(G15-G17)		
		40	1.23		6	1.23				