

UNIDAD VI

¿Qué son las Variables Ficticias?

¿Qué son las Variables Ficticias?



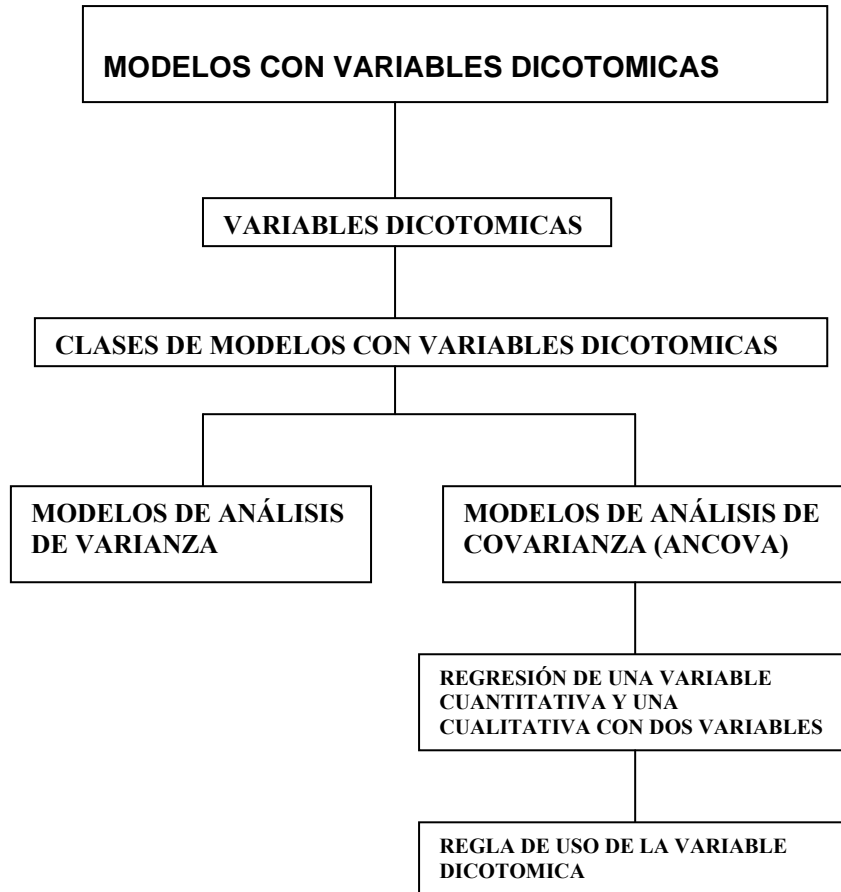
“Un modelo económico es un conjunto de suposiciones que describen en forma aproximada la conducta de un sector económico”

G.S. Maddala, 1996

- **¿Cómo se definen los modelos con variables dicotómicas o dummy?**
- **¿Cuántas y Cuáles son las clases de modelos con variables dicotómicas?**
- **¿Qué es un modelo de análisis de varianza?**
- **¿Cómo se definen los modelos de análisis de covarianza (ANCOVA)?**
- **¿Qué es una regresión de una variable cuantitativa y una cualitativa con dos variables?**
- **¿Cuáles son las reglas de uso de la variable dicotómica?**

MODELOS CON VARIABLES DICOTOMICAS

ESQUEMA CONCEPTUAL



COMPETENCIAS A LOGRAR

CONCEPTUAL	PROCEDIMENTAL	ACTITUDINAL
Explica que es una variable dicotómica, clases de modelo con variables dicotómicas, modelos de análisis de varianza y de covarianza.	Aplica las técnicas de variables dicotómicas en un análisis de regresión, como medio de introducir regresoras cualitativas.	Tiene una actitud crítica respecto a la naturaleza cualitativa de la variable.

CONCEPTOS –CLAVE

Dicotómicas, varianza, covarianza.

LECCIÓN 1

LOS MODELOS CON VARIABLES DICOTÓMICAS O DUMMY

1. VARIABLES DICOTÓMICAS

En los modelos de regresión se encuentran situaciones en las cuales las variables explicativas cambian bruscamente su impacto en aquella que es la variable endógena y que la naturaleza de ese cambio, no se puede atribuir a una variable que sea medible. Por ejemplo, un cambio de gobierno, para la realización de modelos macroeconómicos; un cambio sistemático en el comportamiento de una serie histórica mensual debido a factores estacionales; un cambio brusco en la demandas de un producto debido a una especulación con su precio futuro. Esta variable explicativas de naturaleza cualitativa puede modelizarse, aplicando dos valores: cero (0) y uno (1).

La introducción de estas variables en los modelos se origina cuando junto a variables explicativas cuantitativas (Ingreso, precio, costo, producción, etc.) se incluyen variables explicativas de naturaleza cualitativa tales como: educación, raza, sexo, religión, etc.; las cuales son conocidas como variables dummy.

En efecto, una variable de naturaleza cualitativa indica la presencia o la ausencia de una “cualidad” o “atributo”, tales como estacionalidad; cambio en la política económica; cambio en las expectativas; de tal manera que junto con variables explicativas cuantitativas (Ingreso, precio, costo, producción, etc.) se agregan éstas variables asignando el valor uno, que indica presencia del atributo o el valor cero, que indica ausencia del atributo.

2. CLASES DE MODELOS CON VARIABLE DICOTOMAS

- a. Modelos de Análisis de Varianza
- b. Modelos de Análisis de Covarianza

a. Modelos de Análisis de Varianza

Son aquellos modelos de regresión que contienen exclusivamente variables dicótomas, o cualitativas por naturaleza, por ejemplo:

$$Y_i = \alpha + \beta D_i + \mu_i$$

Donde:

Y: Salario anual de un empleado en el sector privado

$$D_i = \begin{cases} 1 & \text{Si es hombre} \\ 0 & \text{(Si no es hombre). Si es mujer} \end{cases}$$

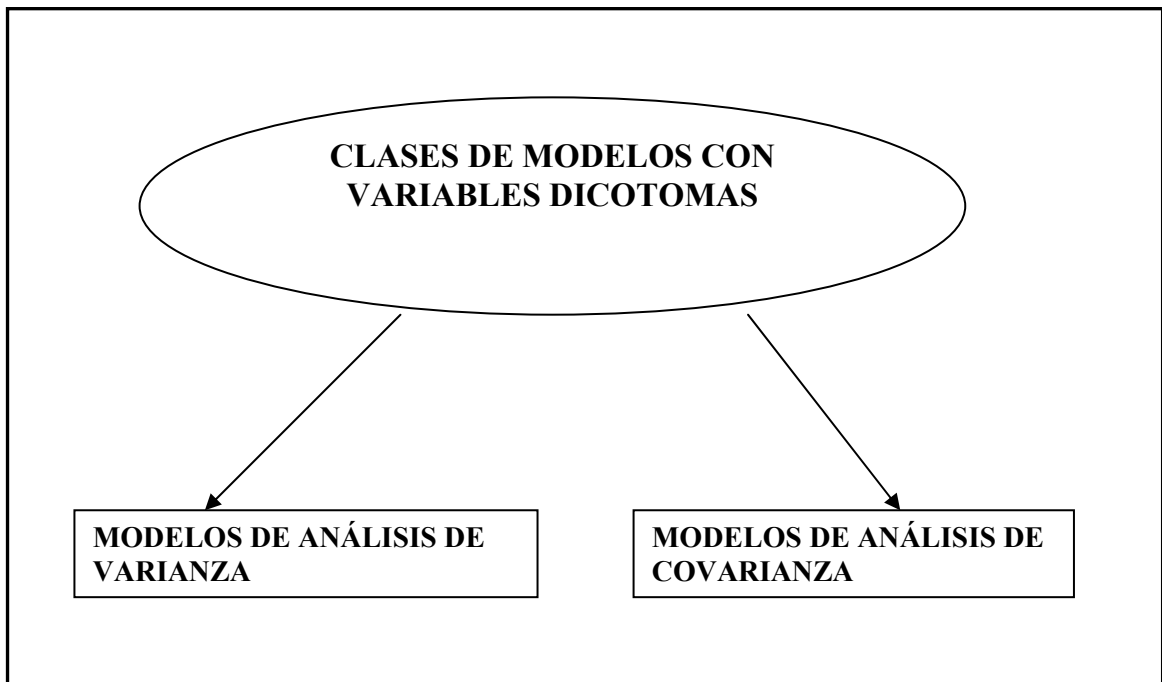
Bajo la hipótesis de que las perturbaciones satisfacen los supuestos del modelo se obtiene:

Salario si es mujer: $E[Y_i/D_i = 0] = \alpha$

Salario si el empleado es hombre: $E[Y_i/D_i = 1] = \alpha + \beta$

b. Modelos de Análisis de Covarianza

Son los más usados y contienen como variables explicativas las cuantitativas y las cualitativas.



LECCIÓN 2

MODELOS DE ANÁLISIS DE COVARIANZA (ANCOVA)

Utilizan Variables Cuantitativas y Cualitativas simultáneamente y es de mayor aplicación en los modelos de regresión de investigaciones económicas, tal es el caso del modelo:

$$Y_i = \alpha_1 + \alpha_2 D_i + \beta_1 X_i + \beta_2 D_i X_i + \mu_i \quad (1)$$

Donde:

Y_i = Salario anual de un profesor universitario

X_i = Años de experiencia del docente (Variable cuantitativa)

$$D_i = \begin{cases} 1 & \text{si es hombre} \\ 0 & \text{si es mujer} \end{cases}$$

De (1):

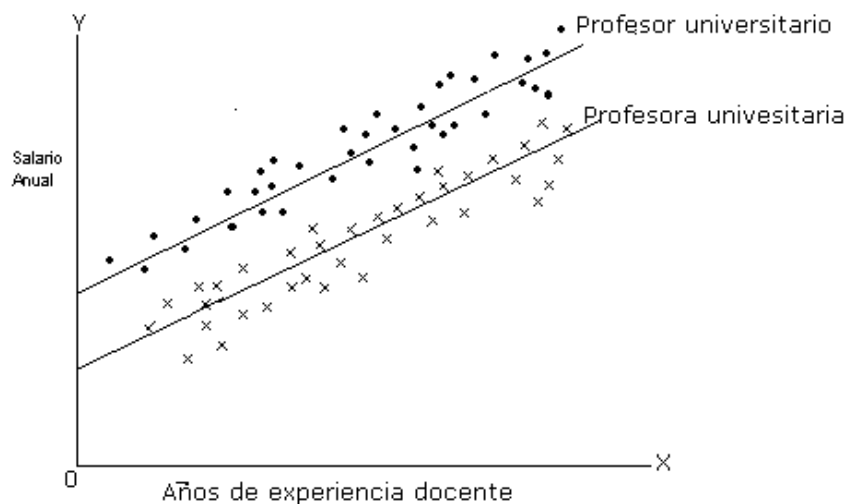
- El salario promedio de una profesora universitaria:

$$E\left[\frac{Y_i}{X_i}, D_i = 0\right] = \alpha_1 + \beta X_i$$

- El salario promedio de un profesor universitario:

$$E\left[\frac{Y_i}{X_i}, D_i = 1\right] = (\alpha_1 + \alpha_2) + \beta X_i$$

Se puede apreciar con el gráfico.



1. REGRESIÓN DE UNA VARIABLE CUANTITATIVA Y UNA CUALITATIVA CON DOS CLASES O CATEGORÍAS

Sea el siguiente modelo la función consumo

$$C_t = \alpha_0 + \alpha_1 D_t + \alpha_2 Y_t + \mu_t$$

Donde:

C_t = Consumo

Y_t = Ingreso

$$D_t = \text{Condición} \begin{cases} 1 \text{ recesión} \\ 0 \text{ normalidad} \end{cases}$$

Hipótesis

CASO 1

$$C_t = \alpha_0 + \alpha_2 Y_t + \mu_t$$

CASO 2 (Modelo Aditivo)

$$C_t = \alpha_0 + \alpha_1 D_t + \alpha_2 Y_t + \mu_t$$

Donde en época de recesión

$$E(C_t/D_t = 1) = (\alpha_0 + \alpha_1) + \alpha_2 Y_t$$

En época de normalidad

$$E(C_t/D_t = 0) = \alpha_0 + \alpha_2 Y_t$$

La asignación de los valores 0 y 1 a las categorías es arbitraria, es decir puede darse el caso que $D_t = 0$ la cual indicaría la condición de normalidad, aunque no necesariamente el valor de cero indica normalidad, todo dependerá de las consideraciones iniciales al formar el modelo.

En el modelo inicialmente planteado, se espera que el consumo promedio en época de recesión sea menor que en época de normalidad, es decir:

$C_t = \alpha_0 + \alpha_1 D_t + \alpha_2 Y_t + \mu_t$, en donde α_1 debe tener signo negativo, pero si se considera:

$$\text{con } D_t \begin{cases} 0 \text{ recesión} \\ 1 \text{ normalidad} \end{cases} \quad \alpha_1 \text{ debe tener signo positivo.}$$

Es decir, para una correcta interpretación es indispensable saber como se asignaron “0” y “1”.

El grupo al que se le asigna valor cero “0”, recibe el nombre de categoría base, de comparación omitida, en el sentido que todas las comparaciones se hacen con respecto a esa categoría.

En el modelo propuesto (caso2) el consumo en época de normalidad es la categoría base donde el intercepto será α_0 .

CASO 3: (Modelo Multiplicativo)

D_t siempre es “0” y “1”

$$C_t = \alpha_0 + \alpha_2 Y_t + \alpha_3 (D_t Y_t) + \mu_t$$

Donde en época de recesión

$$E(C_t / D_t = 1) = \alpha_0 + (\alpha_2 + \alpha_3) Y_t$$

Época de normalidad

$$E(C_t / D_t = 0) = \alpha_0 + \alpha_2 Y_t$$

En este caso se supone que los consumos promedios no varían para cualquier realidad, sino más bien se dan variaciones en la Tasa de Cambio del Consumo al variar el Ingreso.

En efecto; aceptando que se consume menos en época de recesión, se tiene que: $(\alpha_2 + \alpha_3) Y_t < \alpha_2 Y_t$, es decir que las variaciones en el consumo en época de recesión, serán inferiores a la época de normalidad, cuando se presenten variaciones en el Ingreso.

COMBINACIÓN DEL CASO 2 Y CASO 3 (Modelo Aditivo y Multiplicativo)

$$C_t = \alpha_0 + \alpha_1 D_t + \alpha_2 Y_t + \alpha_3 (D_t Y_t) + \mu_t$$

Si fuera en época de recesión el modelo quedaría:

$$E(C_t / D_t = 1) = (\alpha_0 + \alpha_1) + (\alpha_2 + \alpha_3) Y_t$$

En época de normalidad:

$$E(C_t / D_t = 0) = \alpha_0 + \alpha_2 Y_t$$

Con el principio de que en época de recesión se consume menos se tendrá:

$$(\alpha_0 + \alpha_1) \langle \alpha_0$$

$$(\alpha_2 + \alpha_3) Y_t \langle \alpha_2 Y_t$$

Para efectos de predicción se elegirá el que presente mayor R^2 ajustado porque da una mayor bondad de ajuste, no sin antes de realizar las pruebas “t” para cada variable dummy ($\alpha_1 D_t$) y ($\alpha_3 D_t$) a fin de verificar si la situación de recesión o de normalidad influye en el consumo.

2. REGLAS DE USO DE LA VARIABLE DICÓTOMA

a. Cuando una variable tiene dos categorías, se debe introducir una sola variable dicótoma a fin de evitar la multicolinealidad. Por ejemplo:

El salario de un profesor en función de sus años de experiencia (X_i) y si es hombre o mujer (D_i).

$$Y_i = \alpha_0 + \alpha_1 D_i + \alpha_2 X_i + \mu_i$$

En consecuencia la regla general es si una variable tiene m categorías, se debe introducir m-1 variables dicótomas.

b. La técnica de las variables dicótomas se pueden extender fácilmente a más de una variable cualitativa.

Sea el estudio del ingreso en función del sexo y color de la persona y de sus años de experiencia:

El Ingreso del profesor (Y_i) en función de sus años de experiencia (X_i) y si es hombre o mujer (D_{1i}), o si es blanco o negro (D_{2i}), con lo cual el modelo será:

$$Y = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 X_i + \mu_i, \text{ siendo:}$$

$$D_{1i} \begin{cases} 1 \text{ Si es hombre} \\ 0 \text{ Si es mujer} \end{cases} \quad D_{2i} \begin{cases} 1 \text{ Si es blanco} \\ 0 \text{ Si es negro} \end{cases}$$

c. Del mismo modo en el caso de existir “n” variables cualitativas, deben existir “n” variables dummy; siempre y cuando cada variable cualitativa tuviese 2 categorías.

Si se presentan variables que tienen m categorías, la única precaución que debe tomarse es que el número de variables dicótomas para cada variable cualitativa debe ser una menos que el número de categorías de esa variable.

d. Cuando no se utiliza intercepto para cada variable habría que utilizar variables dummy, como categorías hubiere en el modelo.

Ejercicio Ilustrativo 1:

Se tiene un modelo con una variable cuantitativa y una variable cualitativa con más de 2 categorías:

Y: Gasto en educación y cultura

D: Educación

$$D \begin{cases} \text{Educación Superior} \\ \text{Secundaria} \\ \text{Menos de Educación secundaria} \end{cases}$$

$$Y = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \beta_1 X_i + \beta_2 D_{3i} X_i + \mu_i$$

Donde:

Y: Gasto promedio anual de un Trabajador en Educación y Cultura

α_0 : Gasto mínimo en educación

$$D_1 \begin{cases} 1 \text{ Tiene nivel secundaria} \\ 0 \text{ No tiene nivel secundaria} \end{cases}$$

$$D_2 \begin{cases} 1 \text{ Nivel Superior} \\ 0 \text{ No tiene nivel superior} \end{cases}$$

X_i : Ingreso promedio anual

Para cambiar la pendiente se utiliza la variable Dummy D_{3i} .

$$D_3 \begin{cases} 1 \text{ Nivel Superior} \\ 0 \text{ No tiene nivel superior} \end{cases}$$

Se obtiene los siguientes resultados:

$$Y = 60 + 10.2SEC + 20.3SUP + 0.3X_i + 0.1SUP.X_i$$

Interpretación:

$\hat{\alpha}_0 = 60$, el gasto mínimo en educación y cultura es de S/.60

$\hat{\alpha}_1 = 10.2$, por cada incremento en los gastos mínimos de educación, los que no tienen nivel superior (poseen hasta educación secundaria) le destinan S/.10.2.

$\hat{\alpha}_2 = 20.3$, por cada incremento en los gastos mínimos de educación, los que tienen nivel superior le destinan S/.20.3.

$\hat{\beta}_1 = 0.3$, por cada incremento de S/.1.00 en el Ingreso, 0.3 se destina a gastos en educación y cultura.

$\hat{\beta}_2 = 0.1$, los que tienen nivel de instrucción superior por cada aumento de S/.1.00 en el Ingreso, la propensión marginal a gastar en educación se incrementa en 0.1.

Ejercicio Ilustrativo 2¹

Una empresa utiliza dos tipos de procesos productivos (máquina A y máquina B) para obtener su producto final.

El proceso productivo puede expresarse mediante el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 D_i + \mu_i$$

Y_i = producto

$$D_i = \begin{cases} 1, & \text{si el producto es obtenido con la máquina A} \\ 0, & \text{si el producto es obtenido con la máquina B} \end{cases}$$

$$E(Y_i) = \begin{cases} \beta_1 & \text{si } D_i = 0 \\ (\beta_1 + \beta_2) & \text{si } D_i = 1 \end{cases}$$

O también

$$E(Y_i / D_i = 0) = \beta_1$$

$$E(Y_i / D_i = 1) = (\beta_1 + \beta_2)$$

Como la pendiente β_2 mide la diferencia del producto final asociado con el cambio de la máquina B a la máquina A, la prueba de la hipótesis nula: $\beta_2 = 0$ nos mostrará si hay o no diferencia en el producto asociado con la máquina A y B.

Ejercicio Ilustrativo 3:

Si en el ejemplo anterior hubiera tres máquinas A, B, C, se necesitan dos variables dummy. En general si se necesita diferenciar m asuntos, es necesario utilizar $(m-1)$ variables Dummy.

El modelo es: $Y_i = \beta_1 + \beta_2 D_{i2} + \beta_3 D_{i3} + \mu_i$

donde:

$$D_{i2} = \begin{cases} 1, & \text{si el producto es obtenido con la máquina A} \\ 0, & \text{en otro caso} \end{cases}$$

$$D_{i3} = \begin{cases} 1, & \text{si el producto es obtenido con la máquina B} \\ 0, & \text{en otro caso} \end{cases}$$

¹ García Gonzalez, Victor. Econometría para la Planificación, 1991

Los valores esperados son:

$$E(Y_i / D_2 = 1; D_3 = 0) = \beta_1 + \beta_2$$

$$E(Y_i / D_2 = 0; D_3 = 1) = \beta_1 + \beta_3$$

$$E(Y_i / D_2 = 0; D_3 = 0) = \beta_1$$

MAQ."A"	MAQ."B"	MAQ."C"
1	0	0
0	1	0
0	0	1

Observando el cuadro se aprecia que el parámetro β_1 expresa el valor esperado asociado a la máquina C, el parámetro β_2 representa la diferencia en la producción final asociada con el cambio de la máquina C por la máquina A; el parámetro β_3 mide el cambio con el producto final asociado con el cambio de la máquina C por la máquina B. La prueba de Hipótesis $\beta_2 = 0$ proporciona una prueba de que no hay diferencia en el producto final si se cambia la máquina C por la máquina A; la prueba de hipótesis $\beta_3 = 0$ muestra de que no hay diferencia en el producto final al cambiar la máquina C por la máquina B.

Si se desea probar si no hay cambio en el producto final al cambiar la máquina A por la B, se debe usar la prueba F, con la hipótesis nula $\beta_2 = \beta_3$. Sin embargo, si la ecuación de regresión se escribe de la forma:

$$Y_i = \alpha_1 + \alpha_2 D_{i2} + \alpha_3 (D_{i3} + D_{i2}) + \mu_i$$

Podemos probar lo mismo: no hay cambio en el producto final al cambiar la máquina A por la máquina B; utilizando la prueba t con la hipótesis nula $\alpha_2 = 0$ (García Gonzáles, 1991)

LABORATORIO DE LAS VARIABLES FICTICIAS O DICOTÓMICAS

Ejercicio aplicativo 1:

Modelo Inicial: $CP_t = \alpha_0 + \beta_1 PBI_t + \mu_t$

Periodo: 1970-2001

Donde:

CP_t : Consumo Privado

PBI_t : Producto Bruto Interno (Ingreso)

Se incorpora la variable D_t , el cual representa el período de hiperinflación que se registró en el Perú (1988-1991). Se determinará si este suceso afectó significativamente el consumo privado, ya sea afectando el consumo autónomo o por cambios en el nivel del ingreso.

$$D_t = \begin{cases} 1 & \text{periodo 1988-1991} \\ 0 & \text{periodo 1970-1987, 1992-2001} \end{cases}$$

1° Forma (intercepto)

$$CP_t = \alpha_0 + \alpha_1 D_t + \beta_1 PBI_t + \mu_t$$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	8080.718	1686.849	4.790421	0.0000
D1	-1368.771	840.4809	-1.628557	0.1142
PBI	0.663491	0.017074	38.86049	0.0000
R-squared	0.981204	Mean dependent var	67568.47	
Adjusted R-squared	0.979908	S.D. dependent var	11046.32	
S.E. of regresión	1565.768	Akaike info criterion	17.63920	
Sum squared resid	71097299	Schwarz criterion	17.77661	
Log likelihood	-279.2272	F-statistic	756.9578	
Durbin-Watson stat	1.034595	Prob(F-statistic)	0.000000	

La probabilidad asociada al estadístico t nos indica que a un nivel de 5% de significancia no se puede rechazar la hipótesis nula de significancia, es decir que la variable incorporada no es significativo en el modelo.

En el período de inflación no se alteró en una proporción mayor el consumo autónomo.

2° Forma (pendiente)

$$CP_t = \alpha_0 + \beta_1 PBI_t + \beta_2 D_t PBI_t + \mu_t$$

Dependent Variable: CP				
Method: Least Squares				
Date: 07/09/02 Time: 10:17				
Sample: 1970 2001				
Included observations: 32				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6700.907	1580.694	4.239220	0.0002
PBI	0.680595	0.020027	33.98387	0.0000
D1*PBI	-0.017166	0.009474	-1.811858	0.0804
R-squared	0.981572	Mean dependent var	67568.47	
Adjusted R-squared	0.980301	S.D. dependent var	11046.32	
S.E. of regresión	1550.400	Akaike info criterion	17.61947	
Sum squared resid	69708439	Schwarz criterion	17.75689	
Log likelihood	-278.9116	F-statistic	772.3282	
Durbin-Watson stat	1.044042	Prob(F-statistic)	0.000000	

En las salidas de la regresión podemos observar que la probabilidad asociada al estadístico “t” de la variable D1*PBI, nos indica que a un nivel de confianza del 95% no se puede rechazar la Ho de no significancia de la variable, es decir que D1*PBI no es significativa en el modelo.

Las tasas de cambio en el consumo privado ante variaciones en el ingreso no fueron diferentes en los dos períodos.

3° Forma (intercepto y pendiente)

$$CP_t = \alpha_0 + \alpha_1 D_t + \beta_1 PBI_t + \beta_2 D_t PBI_t + \mu_t$$

Dependent Variable: CP				
Method: Least Squares				
Date: 07/09/02 Time: 10:21				
Sample: 1970 2001				
Included observations: 32				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-16794.63	10238.90	-1.640276	0.1121
D1	23992.37	10346.59	2.318867	0.0279
PBI	0.947738	0.116707	8.120687	0.0000
D1*PBI	-0.289525	0.117785	-2.458077	0.0204
R-squared	0.984540	Mean dependent var	67568.47	
Adjusted R-squared	0.982884	S.D. dependent var	11046.32	
S.E. of regression	1445.167	Akaike info criterion	17.50631	
Sum squared resid	58478233	Schwarz criterion	17.68952	
Log likelihood	-276.1009	F-statistic	594.3926	
Durbin-Watson stat	1.184912	Prob(F-statistic)	0.000000	

En este tercer caso, podemos observar que las variables D1 y D1*PBI se vuelven significativas, bajo un nivel de confianza del 5%, es decir que el período de hiperinflación si influyeron en el consumo privado.

Si observamos los R^2 y el $R^2_{ajustado}$ de todos los modelos (incluyendo el modelo inicial), podemos concluir que en la última regresión: $CP_t = \alpha_0 + \alpha_1 D_t + \beta_1 PBI_t + \beta_2 D_t PBI_t + \mu_t$, se obtuvo la mayor bondad de ajuste: 98.5% y 98.2% respectivamente. Entonces para efectos de predicción se escogerá a este modelo.

Ejercicio aplicativo 2:

Se tiene los siguientes datos de utilidades y ventas de una compañía:

PERIODO	UTILIDADES	VENTAS	PERIODO	UTILIDADES	VENTAS
1993:1	1050	11486	1996:1	1253	14886
1993:2	1009	12396	1996:2	1484	15891
1993:3	1083	12145	1996:3	1320	15572
1993:4	1220	13191	1996:4	1494	16841
1994:1	1224	12991	1997:1	1415	16278
1994:2	1400	14097	1997:2	1594	17606
1994:3	1221	13782	1997:3	1402	17242
1994:4	1282	14546	1997:4	1431	18333
1995:1	1134	13698	1998:1	1238	17042
1995:2	1261	14512	1998:2	1399	18131
1995:3	1101	14153	1998:3	1217	17671
1995:4	1273	15177	1998:4	1098	18037

Regresionar el modelo inicial: utilidades = $\alpha_1 + \alpha_2$ ventas + μ

Dependent Variable: UTILIDADES				
Method: Least Squares				
Date: 07/26/02 Time: 11:26				
Sample: 1993:1 1998:4				
Included observations: 24				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
VENTAS	0.046739	0.012317	3.794775	0.0010
C	562.9322	189.3175	2.973482	0.0070
R-squared	0.395610	Mean dependent var	1275.125	
Adjusted R-squared	0.368137	S.D. dependent var	153.2592	
S.E. of regresión	121.8254	Akaike info criterion	12.52271	
Sum squared resid	326511.4	Schwarz criterion	12.62088	
Log likelihood	-148.2725	F-statistic	14.40032	
Durbin-Watson stat	0.954864	Prob(F-statistic)	0.000994	

$$UTILIDADES = 0.04673896719 * VENTAS + 562.9321976$$

¿El segundo trimestre de cada año presenta algún efecto estacional?

Para determinar si existe un efecto estacional se incorpora variables dummy al modelo, de tal forma que el modelo inicial se transforma en:

$$UTILIDADES = \alpha_1 + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \alpha_2 VENTAS + \mu$$

PERÍODO	D1	D2	D3	PERÍODO	D1	D2	D3
1993:1	0	0	0	1996:1	0	0	0
1993:2	1	0	0	1996:2	1	0	0
1993:3	0	1	0	1996:3	0	1	0
1993:4	0	0	1	1996:4	0	0	1
1994:1	0	0	0	1997:1	0	0	0
1994:2	1	0	0	1997:2	1	0	0
1994:3	0	1	0	1997:3	0	1	0
1994:4	0	0	1	1997:4	0	0	1
1995:1	0	0	0	1998:1	0	0	0
1995:2	1	0	0	1998:2	1	0	0
1995:3	0	1	0	1998:3	0	1	0
1995:4	0	0	1	1998:4	0	0	1

Donde:

$$D_1 = \begin{cases} 1, & \text{para el segundo trimestre} \\ 0, & \text{otros trimestres} \end{cases}$$

$$D_2 = \begin{cases} 1, & \text{para el tercer trimestre} \\ 0, & \text{otros trimestres} \end{cases}$$

$$D_3 = \begin{cases} 1, & \text{para el cuarto trimestre} \\ 0, & \text{otros trimestres} \end{cases}$$

Entonces primero se genera las variables D1, D2 y D3, luego regresionamos el modelo obteniéndose

Dependent Variable: UTILIDADES				
Method: Least Squares				
Date: 07/26/02 Time: 11:48				
Sample: 1993:1 1998:4				
Included observations: 24				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	571.0152	191.0858	2.988266	0.0076
D1	91.93412	71.29894	1.289418	0.2127
D2	-26.38616	70.60469	-0.373717	0.7128
D3	7.572313	73.06518	0.103638	0.9185
VENTAS	0.045009	0.012819	3.511030	0.0023
R-squared	0.482461	Mean dependent var	1275.125	
Adjusted R-squared	0.373506	S.D. dependent var	153.2592	
S.E. of regresión	121.3068	Akaike info criterion	12.61757	
Sum squared resid	279591.4	Schwarz criterion	12.86300	
Log likelihood	-146.4109	F-statistic	4.428055	
Durbin-Watson stat	0.612021	Prob(F-statistic)	0.010699	

El Modelo final será:

$$UTILIDADES = 571.02 + 91.93 \cdot D_1 - 26.39 \cdot D_2 + 7.57 \cdot D_3 + 0.045 \cdot VENTAS$$

De las salidas de la regresión podemos observar que sólo las ventas y el intercepto son significativos, ya que sus probabilidades asociadas son inferiores al 5%. Por lo tanto se concluye que no hay ningún factor estacional operando.

Ejercicio aplicativo 3:

a. Variable Cualitativa con dos categorías:

Coefficiente del Intercepto Diferencial: dice qué tanto difiere el valor del término de intercepto de la categoría que recibe el valor de 1 del coeficiente de la categoría base (valor 0). Se prueba si las diferencias son significativas.

Con información de la Encuesta de Hogares del año 1999 se corrió un modelo de los años de estudios en función del área urbana y rural.

Dependent Variable: ANOEST				
Method: Least Squares				
Included observations: 7305				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.614264	0.083701	55.12819	0.0000
AREA	4.585217	0.105169	43.59842	0.0000

Los que están en el área urbana tienen 4.58 años más de estudios de los que habitan en el área rural

b. Variable Cualitativa con más de dos categorías

Asimismo con la misma información anterior se corrió un modelo de años de estudios en función de si la persona es de la costa sierra, selva o Lima.

Es suficiente una variable dicotómica para diferenciar dos categorías. Si una variable cualitativa tiene m categorías, se crean $m-1$ variables dicotómicas

Dependent Variable: ANOEST				
Method: Least Squares				
Included observations: 7305				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.750905	0.124379	54.27669	0.0000
LIMA	3.497196	0.170270	20.53907	0.0000
COSTA	0.674606	0.160847	4.194097	0.0000
SIERRA	-0.569246	0.157479	-3.614737	0.0003

La interpretación sería los que están en la sierra tienen 0.5 menos años de estudios de los que habitan en la selva. Los que están en la costa tienen 0.67 años de estudio más que los que habitan en la selva. Los que están en Lima tienen 3.5 años de estudio más que los habitantes de la selva.

Ejercicio de autoconocimiento

¿Porqué hacer uso de las variables dummy?

	SI	NO	NO SÉ
1. Considero útil para resolver problemas que se presentan cuando se hace un estudio que no es sólo cuantitativo.			
2. Porque permitirá hacer un análisis de los modelos en cuanto a las variables que intervienen			
3. Porque en un modelo de regresión se puede encontrar variables no medibles.			
4. Porque la teoría aconseja incluir variables explicativas de naturaleza cualitativa.			
5. Para utilizar estas variables como recurso para clasificar datos en categorías excluyentes.			
6. Para dar soluciones alternativas a los problemas.			
7. Porque permite aplicar en los modelos de regresión en forma tan fácil como las variables cuantitativas.			
8. Para aplicar estas técnicas en modelos de regresión lineal.			
9. Para utilizarlas como medio de introducir regresoras cualitativas en el análisis de regresión.			
10. Para predecir sucesos futuros en modelos con variables cualitativas			

CALIFICACION

Puntuar con un punto cada respuesta “SI”.

Si obtienes de 1 - 3 puntos tienes pocas expectativas de hacer un buen uso de las variables dummy.

Si tienes entre 4 – 7, tienes buenas expectativas de hacer un buen uso de las variables dummy.

Y si tienes entre 8 – 10, denotas excelentes expectativas de hacer un buen uso de las variables dummy.

RESUMEN

Las variables explicativas de naturaleza cualitativa tales como: educación, raza, sexo, religión, etc.; son conocidas como variables dummy.

Esta variable explicativas de naturaleza cualitativa puede modelizarse, aplicando dos valores: cero (0) y uno (1).

Una variable de naturaleza cualitativa indica la presencia o la ausencia de una “cualidad” o “atributo”, tales como estacionalidad; cambio en la política económica; cambio en las expectativas; de tal manera que junto con variables explicativas cuantitativas (Ingreso, precio, costo, producción, etc.) se agregan éstas variables asignando el valor uno, que indica presencia del atributo o el valor cero, que indica ausencia del atributo.

Las clases de modelos con variable dicotómicas son:

- Modelos de Análisis de Varianza
- Modelos de Análisis de Covarianza

Modelos de Análisis de Varianza, son aquellos modelos de regresión que contienen exclusivamente variables dicótomas, o cualitativas por naturaleza.

Modelos de Análisis de Covarianza (ANCOVA), utilizan Variables Cuantitativas y Cualitativas simultáneamente y es de mayor aplicación en los modelos de regresión de investigaciones económica.

EXPLORACIÓN ON LINE

Demostración del método de ponderación de referenciales por medio del Análisis de Regresión Logística Múltiple para el caso de variables dicotómicas.

<http://www.monografias.com/trabajos5/ponde/ponde.shtml>.

Intervalos de confianza para variables dicotómicas

<http://www.medprev.uma.es/libro/node107.htm-5K>

Asociación de variables cualitativas (variables dicotómicas).

<http://www.fisterra.com/mbe/investiga/chi/chi.asp>.

LECTURA:

NATURALEZA DE LAS VARIABLES DICÓTOMAS

En el análisis de regresión, la variable dependiente o regresada, está influida frecuentemente no sólo por variables de razón de escala (por ejemplo: ingreso, producción, precios, costos, estatura y temperatura), sino también variables que son esencialmente cualitativas por naturaleza, o de escala nominal (por ejemplo, sexo, raza, color, religión, nacionalidad, región geográfica, trastornos políticos y afiliación a un partido).

Por ejemplo, manteniendo los demás factores constantes, se ha encontrado que las trabajadoras ganan menos que sus colegas masculinos y que las personas de color ganan menos que las blancas. Este patrón puede resultar de la discriminación sexual o racial, pero cualquiera que sea la razón, las variables cualitativas tales como sexo y raza sí influyen sobre la variable dependiente y es claro que deben ser incluidas dentro de las explicativas o regresoras.

Puesto que tales variables usualmente indican la presencia o ausencia de una “cualidad” o atributo, tal como femenino o masculino, negro o blanco, católico o no católico, demócrata o republicano son variables de escala nominal esencialmente. Se podrán “cuantificar” tales atributos mediante la elaboración de variables artificiales que tomarán los valores 0 y 1, donde 1 indicará la presencia (o la posesión) de ese atributo y 0 la ausencia de tal atributo. Por ejemplo, el 1 puede indicar que una persona es de sexo masculino y 0 puede designar una de sexo femenino; o el 1 puede indicar que una persona se ha graduado en la universidad y 0 que no lo ha hecho y así sucesivamente.

Las variables que adquieren tales valores 0 y 1 se llaman variable dicótomas. Tales variables son por tanto, esencialmente un recurso para clasificar datos en categorías mutuamente excluyentes, como masculino o femenino.

Las variables dicótomas pueden utilizarse en los modelos de regresión en forma tan fácil como las variables cuantitativas.

De hecho, un modelo de regresión puede contener variables explicativas que son exclusivamente dicótomas o cualitativas por naturaleza. Tales modelos se denominan modelos de varianza (ANOVA).

Damodar N. Gujarati, 2003

ACTIVIDADES

1. Se quiere estudiar la dependencia de los años de educación respecto de la renta familiar y la procedencia socio-geográfica, para lo que se dispone de la siguiente muestra:

Y	X ₁	X ₂
10	urbana	800
12	urbana	1300
8	urbana	900
12	urbana	1200
9	urbana	1600
6	urbana	1300
5	rural	900
4	rural	300
6	rural	700
4	rural	500
5	rural	600
5	rural	100

Siendo:

Y : Años de educación
X₁ : Procedencia socio-geográfica
X₂ : Renta

- a. Especifique a un modelo que explique la duración de la educación en base a la información disponible.
- b. Estima los parámetros del modelo propuesto con la muestra dada y calcule el coeficiente de determinación
- c. Contraste las siguientes hipótesis:

La renta familiar no es una variable significativa.
La variable procedencia socio-geográfica no es significativa.

2. Considere el siguiente modelo econométrico:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

Donde:

Y_i : Sueldos de Lima Metropolitana en nuevos soles reales .
X : PBI en valores constantes base 1979.
X₂ : Índice de precios al consumidor anual de Lima Metropolitana, base 1979
X₃ : Variable ficticia de cambio de pendiente debido al incremento del ritmo inflacionario.
Toma el valor: 0, antes de 1987.
Toma el valor: 1, luego de 1987.

Los valores de las observaciones se muestran a continuación:

Año	Y _i	X ₁	X ₂
1980	181.40	3646.637	60.82
1981	184.94	3807.715	72.67
1982	198.96	3815.751	72.93
1983	170.71	3334.220	125.07
1984	157.21	3494.779	111.46
1985	144.81	3573.928	158.26
1986	180.75	3904.219	62.90
1987	189.55	4234.711	114.51
1988	131.20	3881.284	722.32
1989	82.90	3428.614	775.63
1990	57.09	3243.760	7649.64
1991	64.38	3334.495	139.23
1992	67.63	3287.198	56.73
1993	67.07	3497.230	39.48
1994	94.94	3954.738	15.38
1995	96.32	4233.069	10.23

Se pide:

- Hallar la ecuación de regresión e interprete los parámetros estimados.
 - Probar la significancia estadística de cada uno de los parámetros estimados, con un nivel de confianza de 95% y establezca sus conclusiones.
 - Derivar las regresiones individuales e interprételas.
 - Comparar las regresiones antes y después de 1987 realizando la prueba de Chow.
 - Establecer cuáles son las ventajas de utilizar la prueba de Chow en este caso, si existen o no.
3. Un empresario textil se plantea dos posibles modelos con el fin de estimar la productividad de sus fábricas, en función de la antigüedad de la maquinaria que alquila (A_i), la experiencia de los trabajadores contratados (E_i) y la calidad del combustible utilizado (D_{ji}), siendo esta última una variable ficticia que toma los valores:

$D_{1i}=1$ si el combustible es de calidad baja, cero en caso contrario

$D_{2i}=1$ si el combustible es de calidad media, cero en caso contrario

$D_{3i}=1$ si el combustible es de calidad alta, cero en caso contrario $i=1, \dots, 20$

Los modelos son:

Modelo a:

$$P_i = \beta_0 + \beta_1 A_i + \beta_2 E_i + \beta_3 D_{1i} + \beta_4 D_{2i} + \beta_5 D_{3i} + \mu_i$$

Modelo b:

$$P_i = \beta_1 A_i + \beta_2 E_i + \beta_3 D_{1i} + \beta_4 D_{2i} + \beta_5 D_{3i} + \mu_i$$

- Responda qué modelo sería más apropiado, si la estimación se lleva a cabo por MCO.
- Interprete los coeficientes de las variables en el modelo b
- ¿Se podría plantear un modelo alternativo?

4. Considere el siguiente modelo econométrico:

$$Y_i = \alpha_0 + \alpha_1 D_i + \beta_1 X_1 + \beta_2 X_2 + \beta_3 D_i X_2$$

Donde:

Y_i : Sueldos de Lima Metropolitana en nuevos soles reales, bimensual desde 93-96

X_1 : PBI bimensual a precios constantes con base en 1979

X_2 : Índice de precios al consumidor anual de Lima Metropolitana, base 1979

D_i : Variable ficticia de cambio de pendiente debido al incremento del ritmo inflacionario.

Toma el valor: 0, antes de Febrero 1994

Toma el valor: 1, luego de Febrero 1994

Se pide:

- Hallar la ecuación de regresión e interprete los parámetros estimados.
- Probar la significancia estadística de cada uno de los parámetros estimados, con un nivel de confianza de 95% y establezca sus conclusiones.
- Derivar las regresiones individuales e interprételas.
- Comparar las regresiones antes y después de dic 1995 utilizando la prueba de Chow.

Los valores de las observaciones se muestran a continuación:

Año	Y_i	X_1	X_2
Febrero 93	877.21	262.53	1145.82
Abril 93	914.52	290.02	1247.24
Junio 93	960.46	310.98	1308.41
Agosto 93	1084.43	288.26	1378.31
Octubre 93	1186.91	262.63	1421.78
Diciembre 93	1300.13	301.79	1480.87
Febrero 94	1407.03	287.53	1535.54
Abril 94	1513.04	332.65	1595.50
Junio 94	1556.66	356.85	1625.25
Agosto 94	1563.41	318.12	1664.85
Octubre 94	1591.61	307.12	1678.22
Diciembre 94	1685.42	333.66	1708.69
Febrero 95	1597.12	329.38	1734.53
Abril 95	1618.23	358.20	1775.65
Junio 95	1619.92	380.43	1804.95
Agosto 95	1675.94	341.56	1834.08
Octubre 95	1677.83	318.78	1850.63
Diciembre 95	1674.12	328.93	1883.46
Febrero 96	1777.16	342.18	1936.10
Abril 96	1778.25	363.04	1979.93
Junio 96	1781.18	398.96	2003.61
Agosto 96	1785.21	401.52	2049.92
Octubre 96	1790.10	405.28	2071.67
Diciembre 96	1795.25	408.35	2106.46

AUTOEVALUACIÓN

1. Con la finalidad de tratar de explicar el número de personas empleadas en la industria química se ha efectuado un muestreo a 100 empresas grandes y medianas, siendo el modelo econométrico propuesto:

$$Y = \beta_1 + \beta_2 T + \beta_3 E + \beta_4 F$$

Donde:

Y: Número de empleados de una empresa (medido en cientos de personas)

T: Permite saber si la empresa incorpora los últimos tecnológicos (toma el valor de 1 si la empresa lo hace y cero en caso contrario).

E: Determina la presencia de empresas competidoras a los alrededores (toma el valor de 1 si es que existen empresas competidoras en un radio de 50Km, será cero en cualquier otro caso).

F: Determina la presencia de alguna empresa complementaria en los alrededores (toma el valor de 1 si es que existe una empresa de ese tipo, como por ejemplo la farmacéutica, en radio de 30Km)

Dependent Variable: Y				
Method: Least Squares				
Date: 11/11/04 Time: 10:28				
Sample: 1 100		Included observations: 100		
Variable	Coefficient	Std. Error	t-Statistic	Prob.
T	0.05002	0.037261	0.13562	0.1321
E	-2.40971	0.053112	-4.52226	0.0210
F	1.91120	0.612219	3.114754	0.0000
C	2.31987	0.991251	1.010578	0.1966
R-squared	0.639630	Mean dependent var	443.14	
Adjusted R-squared	0.645999	S.D. dependent var	315.68	
S.E. of regression	522.62	Akaike info criterion	2.22915	
Sum squared resid	261.63	Schwarz criterion	2.26287	
Log likelihood	-130.288	F-statistic	11.0525	
Durbin-Watson stat	2.141804	Prob(F-statistic)	0.00000	

- a. El valor estimado del coeficiente del número de empresas competidoras es 2.4.
 - b. Una empresa con tecnología de punta tiene en promedio 5 empleados más que una que no está en la vanguardia de la innovación.
 - c. Por cada empresa de la competencia existente en un radio de 50Km, una empresa de la industria química contrata 240 trabajadores menos.
 - d. Cuando existen empresas competidoras en general, la empresa en cuestión emplea a 240 trabajadores menos.
2. Se ha realizado una encuesta de presupuestos familiares en el departamento de Lima, en la que se ha considerado aquellas variables referidas al gasto de las familias en vivienda (GTO), ingresos familiares (Y), número de miembros que componen el hogar (TAM) y el lugar de residencia (URBANO).
Para el lugar de residencia se ha considerado:
URBANO =1 familia reside en una zona urbana
URBANO =0 familia reside en una zona rural

Dependent Variable: GTO				
Method: Least Squares				
Date: 11/11/04 Time: 10:28				
Sample: 1 500		Included observations: 500		
Variable	Coefficient	Std. Error	t-Statistic	Prob.
URBANO	14655.83	6789.100	2.158730	0.0314
Y	0.050108	0.003048	16.43694	0.0000
TAM	-8692.680	2268.410	-3.832058	0.0001
C	17955.59	9397.991	1.910578	0.0566
R-squared	0.399630	Mean dependent var	94443.14	
Adjusted R-squared	0.395999	S.D. dependent var	93315.68	
S.E. of regression	72522.62	Akaike info criterion	25.22915	
Sum squared resid	2.61E+12	Schwarz criterion	25.26287	
Log likelihood	-6303.288	F-statistic	110.0525	
Durbin-Watson stat	2.141804	Prob(F-statistic)	0.000000	

Con respecto a la solución propuesta al modelo planteado, puede afirmar que :

- No existe una diferencia considerable en los gastos de vivienda de acuerdo al lugar de residencia.
 - No interesa el considerar el lugar de residencia para estimar el modelo adecuado.
 - El modelo planteado, no considera que las regresoras en su conjunto permitan explicar el consumo familiar.
 - El gasto anual de las familias urbanas en Lima es S/.14655.83, el cual es superior al de las familias rurales.
3. Continuando con el ejercicio anterior considere adicionalmente que se recogió información acerca del nivel de estudios del jefe de familia (ESTU), la cual está dada por:

ESTU =1 , si el individuo es analfabeto o sin estudios

ESTU =2 , individuos con estudios primarios

ESTU =3 , individuos que poseen otro tipo de estudios

ESTU =4 , titulados universitarios o equivalentes

Dependent Variable: GTO				
Method: Least Squares				
Sample: 1 500		Included observations: 500		
Variable	Coefficient	Std. Error	t-Statistic	Prob.
Y*URBANO	0.0073	0.0030	2.4730	0.0137
Y	0.0456	0.0039	11.7060	0.0000
TAM	-8551.0640	2266.4290	-3.7729	0.0002
C	25377.4100	8803.8550	2.8825	0.0041
R-squared	0.40137	Mean dependent var	94443.1400	
Adjusted R-squared	0.39775	S.D. dependent var	93315.6800	
S.E. of regression	72417.43	Akaike info criterion	25.2263	
Sum squared resid	2.60E+12	Schwarz criterion	25.2600	
Log likelihood	-6302.562	F-statistic	110.8531	
Durbin-Watson stat	2.146351	Prob(F-statistic)	0.0000	

Dependent Variable: GTO				
Method: Least Squares				
Sample: 1 500		Included observations: 500		
Variable	Coefficient	Std. Error	t-Statistic	Prob.
URBANO	28538.9900	17577.0800	1.6236	0.1051
Y*URBANO	0.0128	0.0062	2.0569	0.0402
Y	0.0424	0.0049	8.7040	0.0000
TAM*URBANO	-10295.7100	4525.0670	-2.2753	0.0233
TAM	-3228.2860	3250.3750	-0.9932	0.3211
C	10027.9300	12652.5100	0.7926	0.4284
R-squared	0.407592	Mean dependent var	94443.1400	
Adjusted R-squared	0.401596	S.D. dependent var	93315.6800	
S.E. of regression	72185.82	Akaike info criterion	25.2238	
Sum squared resid	2.57E+12	Schwarz criterion	25.2744	
Log likelihood	-6299.95100	F-statistic	67.9770	
Durbin-Watson stat	2.14797	Prob(F-statistic)	0.0000	

- a. Existe una influencia significativa del ingreso sobre el gasto en vivienda en las zonas urbanas.
 - b. La influencia del ingreso en las zonas rurales es mayor que para los residentes en las urbanas.
 - c. El efecto producido por la zona de residencia es el único que es significativo en la explicación de los gastos en vivienda.
 - d. Se tiene una influencia no significativa del número de miembros que componen un hogar sobre el gasto en la zona urbana
4. Se posee información acerca de los ingresos del sector público (INGPUB), el PBI, el ahorro privado (AHORRO), las importaciones (IMPORT) y las exportaciones (EXPORT) del Perú durante los años de 1975-1998.

Con los datos disponibles se estimó un modelo el cual relaciona los ingresos públicos con el PBI, las variables del sector exterior y el nivel de ahorro. Además se desea determinar si el efecto producido a partir del año 1987 originó un cambio estructural en dicha relación.

Para poder el cambio estructural a partir de 1987, será necesario considerar una variable ficticia (F) la cual será igual a 1 a partir de 1987 siendo cero en los años anteriores.

De los resultados mostrados a partir del uso de E-Views, indicar la alternativa correcta:

Primer modelo: (no se considera la variable ficticia)

Dependent Variable: INGPUB				
Method: Least Squares				
Sample: 1975 1998		Included observations: 24		
Variable	Coefficient	Std. Error	t-Statistic	Prob.
PBI	0.0003	0.0000	25.1587	0.0000
EXPORT	0.5765	0.0791	7.2917	0.0000
IMPORT	-0.2578	0.1128	-2.2861	0.0339
AHORRO	-0.8769	0.1511	-5.8026	0.0000
C	38.4145	4.2424	9.0548	0.0000
R-squared	0.9917	Mean dependent var	31.6875	
Adjusted R-squared	0.9900	S.D. dependent var	6.9170	
S.E. of regression	0.6917	Akaike info criterion	2.2836	
Sum squared resid	9.0894	Schwarz criterion	2.5290	
Log likelihood	-22.4032	F-statistic	570.3155	
Durbin-Watson stat	1.3842	Prob(F-statistic)	0.0000	

Segundo modelo:(considerando la variable ficticia)

Dependent Variable: PBI				
Method: Least Squares				
Sample: 1975 1998		Included observations: 24		
Variable	Coefficient	Std. Error	t-Statistic	Prob.
EXPORT	2326.8662	492.6973	4.7227	0.0003
IMPORT	-323.1763	303.6055	-1.0645	0.3040
AHORRO	-1787.7445	437.1629	-4.0894	0.0010
F	-18356.1052	30801.9324	-0.5959	0.5601
C	18356.1052	16225.4954	1.1313	0.2757
F*PBI	1.0000	0.0446	22.4324	0.0000
F*EXPORT	-2326.8662	713.7399	-3.2601	0.0053
F*IMPORT	323.1763	773.1371	0.4180	0.6819
F*AHORRO	1787.7445	1103.9495	1.6194	0.1262
R-squared	0.995142459	Mean dependent var	23850.1667	
Adjusted R-squared	0.992551771	S.D. dependent var	19213.3651	
S.E. of regression	1658.173407	Akaike info criterion	17.9448	
Sum squared resid	41243085.72	Schwarz criterion	18.3866	
Log likelihood	-206.337807	F-statistic	384.1228	
Durbin-Watson stat	1.279136966	Prob(F-statistic)	0.0000	

- Para explicar el ingreso del sector público no es necesario hacer uso de una variable ficticia.
- El modelo estimado con el uso de la variable ficticia es más adecuado en comparación con el otro ya que posee un coeficiente de determinación ajustado mayor.
- El segundo obtenido es el mejor que se puede obtener.
- El único efecto que se debe eliminar del segundo modelo para explicar el gasto son las exportaciones.

RESPUESTAS DE CONTROL

Rpta 1.d 2.d 3.a 4.b