MODELOS DE REGRESIÓN CON VARIABLES DICOTÓMICAS

Cuando se habló de la estructura de los datos se hizo mención a las escalas de medición de las variables: escala de razón, escala de intervalo, escala ordinal y escala nominal. En los capítulos anteriores se ha tratado sólo la escala de razón, pero esto no implica que los modelos de regresión traten exclusivamente con ese tipo de variables.

Cuando se hace análisis de regresión la variable explicada está frecuentemente influenciada por variables de escala de razón (cuantitativas) y de escala nominal (cualitativas). Éstas últimas indican la presencia o ausencia de una cualidad o atributo (blanco o negro, alto o bajo, rico o pobre, católico o ateo, etc) por tanto son variables de escala nominal. Cuantificar directamente tales variables no se puede pero, haciendo uso de variables artificiales que tomen valores 0 y 1 para la ausencia y presencia de la cualidad respectivamente, es posible.

Las variables que adquieren valores 0 y 1 se llaman **VARIABLES DICÓTOMAS,** dichas variables se convierten en un recurso esencial para estimar modelos de regresión con variables cualitativas.

EN ESTE CURSO SOLAMENTE SE HARÁ USO DE VARIABLES DICÓTOMAS COMO VARIABLES EXPLICATIVAS

Al trabajar con variables dicótomas se debe tener en cuenta:

- Se construyen asignando números a la presencia o ausencia de la cualidad. La asignación se hace de manera arbitraria pero es de uso universal emplear el 0 y el 1 para tales fines.

- Se distingue entre dos modelos de regresión: los que sólo incluyen variables cualitativas en las variables regresoras se conocen como ***modelos de análisis de varianza (ANOVA)*** y los que incluyen variables cualitativas y cuantitativas del lado de las variables explicativas se llaman ***modelos de análisis de covarianza (ANCOVA).***

- Para evitar colinealidad perfecta, si una variable cualitativa tiene ***m*** categorías, sólo hay que agregar ***m-1*** variables dicótomas. Si esto no se respeta se caerá en la famosa ***trampa de la variable dicótoma.*** Para cada regresora cualitativa, el número de variables dicótomas introducidas deber ser una menos que las categorías de esa variable.

Intercepto	Ateo	Católico	Ateo + católico
1	1	0	**1**
1	1	0	**1**
1	1	0	**1**
1	0	1	**1**

- La categoría a la cual no se asigna variable dicótoma se conoce como ***categoría base, de comparación, de control, de referencia u omitida (benchmark).*** Todas las comparaciones se hacen respecto a dicha categoría.
- El valor de la intersección representa el valor medio de la categoría base.

- Los coeficientes anexos a las variables dicótomas se llaman ***coeficientes de intersección diferencial.*** Indican en qué medida el valor de la intersección varia del coeficiente de intersección de la categoría de comparación.

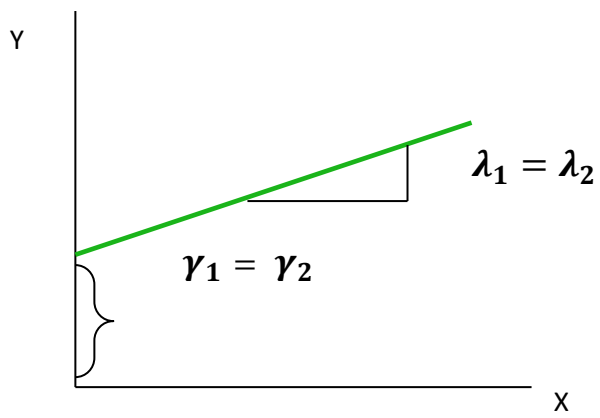
- La elección de la categoría de control queda a discreción del investigador.

- Existe una forma de eludir la trampa de la variable dicótoma, consiste en no incluir intercepto en el modelo. **SIN EMBARGO, AL MOMENTO DE TOMAR LA DECISIÓN NO OLVIDE LOS PROBLEMAS QUE PRESENTA LA REGRESIÓN A TRAVÉS DEL ORIGEN.**

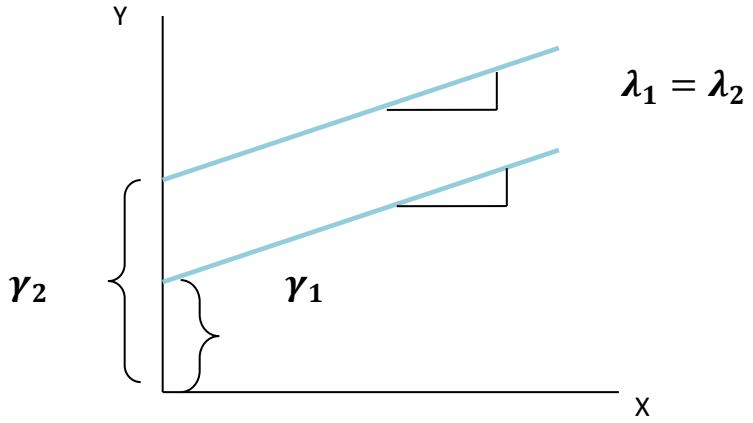
Aplicaciones frecuentes:

- Cambios en el intercepto de una regresión.
- Cambios en el intercepto y la pendiente.
- Estabilidad estructural o paramétrica de los parámetros del modelo de regresión. Alternativa al test de Chow.
- Análisis estacional.

Diferencia entre grupos: $Y_i = \gamma_1 + \lambda_1 X$ $Y_i = \gamma_2 + \lambda_2 X$

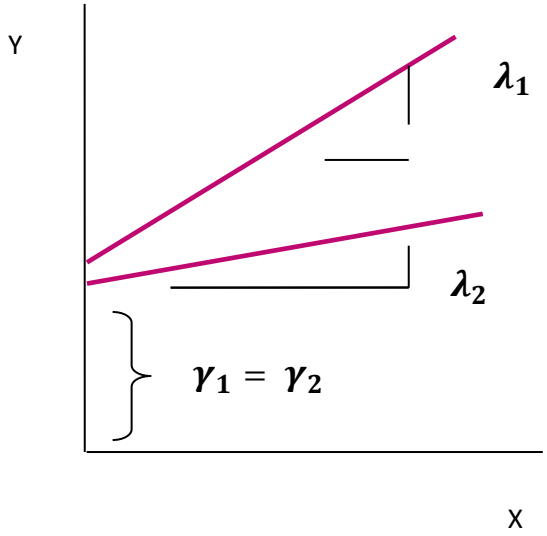


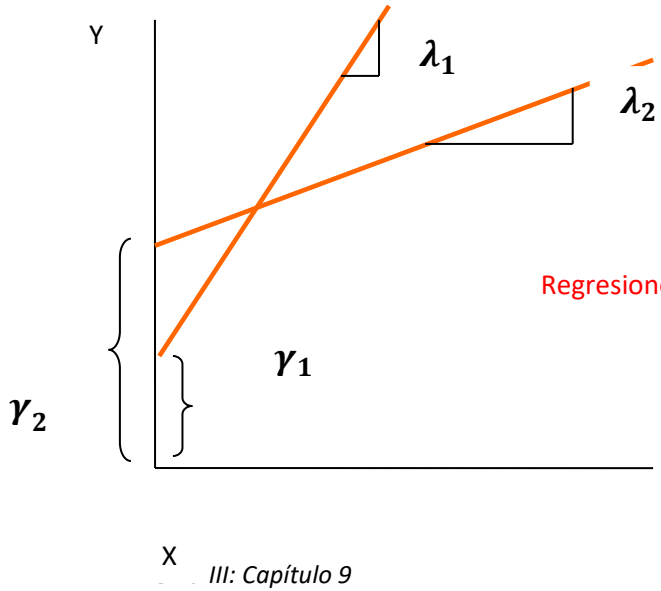
Regresiones coincidentes



Regresiones paralelas

Regresiones concurrentes





Regresiones no similares

MODELO ANOVA:

Ejemplo 1: Salario de los maestros de escuelas públicas por estado, EEUU 1986.

$$SALARIO = \hat{\beta}_1 + \hat{\beta}_2 E\_OESTE + \hat{\beta}_3 E\_SUR + \hat{u}_i$$

$$E\_NORTE = \begin{cases} 1 \text{ si el estado ésta en el Noreste o Nor-centro} \\ 0 \text{ otra región} \end{cases}$$

$$E\_SUR = \begin{cases} 1 \text{ si el estado ésta en el Sur} \\ 0 \text{ otra región} \end{cases}$$

$$E\_OESTE = \begin{cases} 1 \text{ si el estado ésta en el Oeste} \\ 0 \text{ otra región} \end{cases}$$

Dependent Variable: SALARIO
Method: Least Squares
Sample: 1 51
Included observations: 51

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24424.14	887.9170	27.50724	0.0000
E_OESTE	1734.473	1435.953	1.207890	0.2330
E_SUR	-1530.143	1327.516	-1.152636	0.2548

R-squared	0.090083	Mean dependent var	24356.22
Adjusted R-squared	0.052170	S.D. dependent var	4179.426
S.E. of regression	4068.947	Akaike info criterion	19.51718
Sum squared resid	7.95E+08	Schwarz criterion	19.63082
Log likelihood	-494.6880	Hannan-Quinn criter.	19.56060
F-statistic	2.376027	Durbin-Watson stat	1.162044
Prob(F-statistic)	0.103764		

CUANTO MÁS GANA UN MAESTRO EN LOS ESTADOS DEL OESTE

CUANTO MENOS GANA UN MAESTRO EN LOS ESTADOS DEL SUR

$$\widehat{SALARIO} = 24424.14 + 1734.47OESTE - 1530.14E\_SUR$$

RESPONDA:

¿Cuál es el salario promedio de los maestros de las escuelas públicas del Norte?

¿Cuál es el salario promedio de los maestros de las escuelas públicas del Sur?

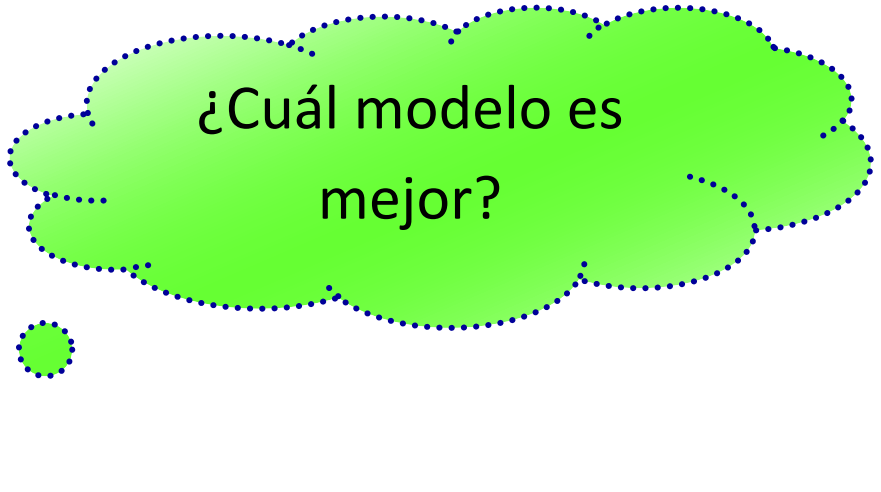
¿Cuál es el salario promedio de los maestros de las escuelas públicas del Oeste?

Eliminando el intercepto:

Dependent Variable: SALARIO
Method: Least Squares
Sample: 1 51
Included observations: 51

Variable	Coefficient	Std. Error	t-Statistic	Prob.
E_NORTE	24424.14	887.9170	27.50724	0.0000
E_OESTE	26158.62	1128.523	23.17952	0.0000
E_SUR	22894.00	986.8645	23.19873	0.0000

R-squared	0.090083	Mean dependent var	24356.22
Adjusted R-squared	0.052170	S.D. dependent var	4179.426
S.E. of regression	4068.947	Akaike info criterion	19.51718
Sum squared resid	7.95E+08	Schwarz criterion	19.63082
Log likelihood	-494.6880	Hannan-Quinn criter.	19.56060
Durbin-Watson stat	1.162044		



Ejemplo 2: Salario según el sexo (efecto de discriminación). Sexo toma valor 1 si es hombre, cero si es mujer.

SALARIO PROMEDIO DE UNA MUJER

Dependent Variable: SALARIO
Method: Least Squares
Sample: 1 159376
Included observations: 159376

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	36230.43	1021.628	35.46342	0.0000
SEXO	23857.89	636.1683	37.50248	0.0000
EDAD	1058.339	24.56819	43.07761	0.0000

R-squared	0.019849	Mean dependent var	24356.22
Adjusted R-squared	0.019836	S.D. dependent var	4179.426
S.E. of regression	4068.947	Akaike info criterion	9.51718
Sum squared resid	7.95E+08	Schwarz criterion	9.63082
Log likelihood	-494.6880	F-statistic	1613.692
Durbin-Watson stat	1.362044	Prob(F-statistic)	0.000000

EFECTO DE LA EDAD SOBRE EL SALARIO CONSTANTE PARA HOMBRES Y MUJERES

CUANTO MÁS GANA UN HOMBRE QUE UNA MUJER

$$\widehat{SALARIO} = 36230.43 + 23857.89SEXO + 1058.339EDAD$$

MODELOS ANCOVA:

a) Efecto de una variable cualitativa en el origen:

Ejemplo 3: Salario de los maestros respecto a la región y al gasto en escuelas públicas por alumno, EEUU 1986.

$$SALARIO = \hat{\beta}_1 + \hat{\beta}_2 E\_NORTE + \hat{\beta}_3 E\_SUR + \hat{\beta}_4 GASTO + \hat{u}_i$$

Dependent Variable: SALARIO
Method: Least Squares
Sample: 1 51
Included observations: 51

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13269.11	1395.056	9.511530	0.0000
E_NORTE	**-1673.514**	**801.1703**	**-2.088837**	**0.0422**
E_SUR	**-1144.157**	**861.1182**	**-1.328687**	**0.1904**
GASTO	**3.288848**	**0.317642**	**10.35393**	**0.0000**

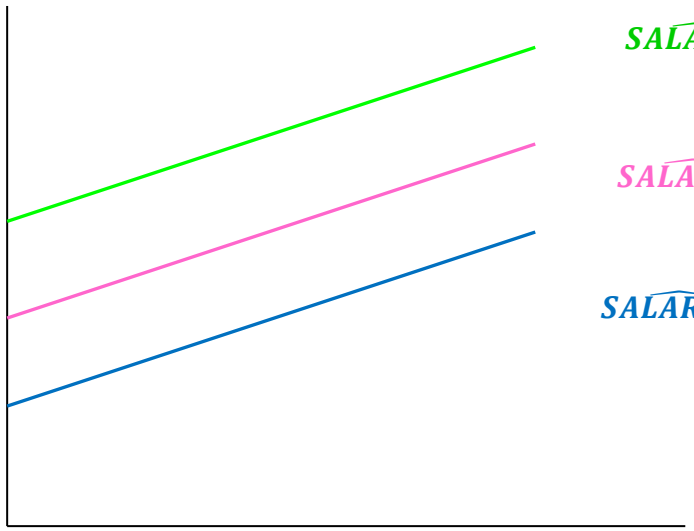
R-squared	0.722665	Mean dependent var	24356.22
Adjusted R-squared	0.704963	S.D. dependent var	4179.426
S.E. of regression	2270.152	Akaike info criterion	18.36827
Sum squared resid	2.42E+08	Schwarz criterion	18.51978
Log likelihood	-464.3908	Hannan-Quinn criter.	18.42616
F-statistic	40.82341	Durbin-Watson stat	1.414238
Prob(F-statistic)	0.000000		

$$\widehat{SALARIO} = 13269.11 - 1673.51 E\_NORTE - 1144.15 E\_SUR + 3.28 GASTO$$

¿Cuál es el salario promedio de los maestros de las escuelas públicas del Noreste o Nor-centro que no depende del gasto?

¿Cuál es el salario promedio de los maestros de las escuelas públicas del Sur que no depende del gasto?

¿Cuál es el salario promedio de los maestros de las escuelas públicas del Oeste que no depende del gasto?



$$\widehat{SALARIO}_O = 13269.11 + 3.28GASTO$$

$$\widehat{SALARIO}_S = 12124.95 + 3.28GASTO$$

$$\widehat{SALARIO}_N = 11595.59 + 3.28GASTO$$

b) Efecto de una variable cualitativa en el origen y en la pendiente:

Ejemplo 4: Volvamos al ejemplo de los salarios en función del sexo y la edad (cómo proxy de la experiencia).

CUANTO MÁS GANA UN HOMBRE QUE UNA MUJER

$$\widehat{SALARIO} = \widehat{\beta_1} + \widehat{\beta_2}\text{SEXO} + \widehat{\beta_3}\text{EDAD} + \widehat{\beta_4}\text{EDAD} * \text{SEXO} + \widehat{u_1}$$

SALARIO PROMEDIO DE LAS MUJERES QUE NO DEPENDE DE LA EDAD

Dependent Variable: SALARIO
Method: Least Squares
Sample: 1 159376
Included observations: 159376

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	**59827.03**	1673.050	35.75925	0.0000
SEXO	**10166.17**	2014.393	5.046766	0.0000
EDAD	**403.7232**	44.21486	9.130940	0.0000
EDAD*SEXO	**946.1699**	53.15690	17.79957	0.0000

R-squared	0.021793	Mean dependent var	24356.22
Adjusted R-squared	0.021775	S.D. dependent var	123041.0
S.E. of regression	121694.0	Akaike info criterion	26.25643
Sum squared resid	2.36E+15	Schwarz criterion	26.25668
Log likelihood	-2092319	F-statistic	1183.535
Durbin-Watson stat	1.358432	Prob(F-statistic)	0.000000

EFECTO PROMEDIO DE LA EDAD SOBRE EL SALARIO. IGUAL PARA HOMBRES Y MUJERES

CUANTO MÁS GANA UN HOMBRE POR CADA AÑO DE EDAD

VARIABLE DICÓTOMA COMO ALTERNATIVA AL TEST DE CHOW

Estabilidad paramétrica ⇒ **Coeficientes de regresión estables en el tiempo o en grupos.**

Recuerde: el test de Chow indica la existencia o no de estabilidad pero no indica en qué parámetro (intercepto, pendiente o ambos).

Ejemplo 5: Datos sobre ahorro e ingreso, EEUU 1970-1995

En 1982 EEUU sufrió la peor recesión en tiempos de paz y tomando como referencia ese acontecimiento se propuso esa fecha como punto de ruptura de la estabilidad. Las variables dicótomas pueden decirnos realmente si eso es cierto o no para ello:

$$\text{Periodo} = \begin{cases} 0 \text{ antes de } 1982 \\ 1 \text{ después de } 1982 \end{cases}$$

$$\boldsymbol{Ahorro = \hat{\beta}_1 + \hat{\beta}_2 Ingreso + \hat{\beta}_3 Periodo + \hat{u}_i}$$

Si $\hat{\beta}_3$ resulta ser estadísticamente significativo existe un cambio estructural en el origen.

Dependent Variable: AHORRO
Method: Least Squares
Sample: 1970 1995
Included observations: 26

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	71.70587	13.54567	5.293639	0.0000
INGRESO	0.026468	0.007925	3.339604	0.0028
PERIODO	37.83347	22.90507	1.651751	**0.1122**

Esta regresión confirma lo expuesto por el test de Chow de la existencia de un cambio estructural.

Chow Forecast Test
Equation: UNTITLED
Specification: AHORRO INGRESO C
Test predictions for observations from 1982 to 1995

	Value	Df	Probability
F-statistic	8.588587	(14, 10)	0.0008
Likelihood ratio	66.73668	14	0.0000

$F_{2,22} = 3.44$

$F_c > F_\alpha$ SE RECHAZA LA Ho, ES DECIR QUE LOS PARÁMETROS NO SON CONSTANTES A TRAVÉS DEL TIEMPO.

Probemos con una interacción:

$$\boldsymbol{Ahorro = \widehat{\beta}_1 + \widehat{\beta}_2 Ingreso + \widehat{\beta}_3 Periodo + \widehat{\beta}_4 Ingreso * Periodo + \widehat{u}_i}$$

Si $\hat{\beta}_3$ y $\hat{\beta}_4$ resultan ser estadísticamente significativos existe un cambio estructural en el origen y en pendiente.

Dependent Variable: AHORRO
Method: Least Squares
Sample: 1970 1995
Included observations: 26

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.016117	20.16483	0.050391	0.9603
INGRESO	0.080332	0.014497	5.541347	0.0000
PERIODO	152.4786	33.08237	4.609058	0.0001
INGRESO*PERIODO	-0.065469	0.015982	-4.096340	0.0005

R-squared	0.881944	Mean dependent var	162.0885
Adjusted R-squared	0.865846	S.D. dependent var	63.20446
S.E. of regression	23.14996	Akaike info criterion	9.262501
Sum squared resid	11790.25	Schwarz criterion	9.456055
Log likelihood	-116.4125	Hannan-Quinn criter.	9.318238
F-statistic	54.78413	Durbin-Watson stat	1.648454
Prob(F-statistic)	0.000000		

Efectivamente existe un cambio estructural en la regresión tanto en intercepto como en pendiente.

$$\widehat{Ahorro} = 1.016 + 0.080Ingreso + 152.478Periodo - 0.065Ingreso * Periodo$$

Antes de 1982:

$$\boldsymbol{\widehat{Ahorro} = 1.016 + 0.080Ingreso}$$

Después de 1982:

$$\boldsymbol{\widehat{Ahorro} = 153.494 + 0.015Ingreso}$$

TAREA:
¿CÓMO SE INTERPRETA UN MODELO EN LOGARITMO CON UNA VARIABLE DUMMY?
¿CÓMO SE HACE ANÁLISIS ESTACIONAL CON VARIABLES DICOTÓMICAS?
¿CÓMO SE HACE ANÁLISIS ESTACIONAL CON VARIABLES DICOTÓMICAS?

Lecturas recomendadas:

- ✓ Gujarati, D. y Porter, D. (2010). Econometría. 5ta. Edición. McGraw Hill Interamericana.
- ✓ Novales, A. (1993). Econometría. 2da. Edición. McGraw Hill Interamericana.
- ✓ Wooldridge, J. (2010). Introducción a la econometría. Un enfoque moderno. 4ª. Edición CENGAGE Learning Editores.