El Modelo de Regresión Lineal

Dante A. Urbina

CONTENIDOS

1. Regresión Lineal Simple
2. Regresión Lineal Múltiple
3. Multicolinealidad
4. Heterocedasticidad
5. Autocorrelación
6. Variables Dummy
7. Diagnóstico y Selección de Modelos

REGRESIÓN LINEAL SIMPLE

Definición

Sean las variables X (independiente) e Y (dependiente), el modelo de regresión lineal simple vendrá dado por:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

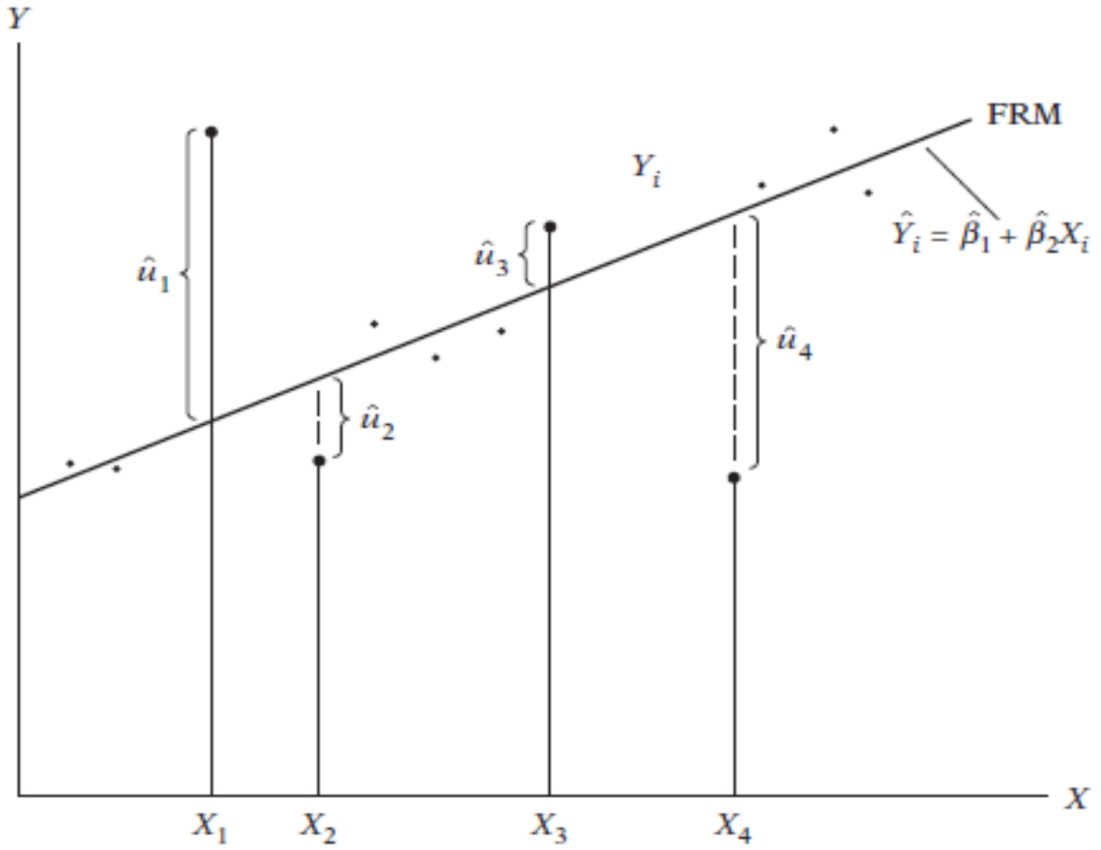
Donde:

β_0: Coeficiente de intercepto.

β_1: Coeficiente de pendiente.

ε_i: Error aleatorio o residuo.

i: Subíndice que indica los diferentes valores que puede tomar la variable.



Supuestos del modelo (1)

1. Linealidad: Implica que el efecto marginal de la variable independiente (X) en la variable dependiente (Y) no dependa de la primera, es decir, que sea constante. Matemáticamente:

$$\frac{dY_i}{dX_i} = \beta_1$$

2. Exogeneidad estricta: Los términos de error son independientes de los valores de X, es decir, la variable regresora está contemporáneamente no correlacionada con el término de error. Ello implica que:

$$E(\varepsilon_i \mid X_i) = 0 \qquad E(\varepsilon_i) = 0$$

3. Homocedasticidad: Se da cuando la varianza de los términos de error (incluidos los condicionados a los valores de la variable independiente) es la misma, es decir, es constante.

$$Var(\varepsilon_i) = \sigma^2 \qquad Var(\varepsilon_i \mid X_i) = \sigma^2$$

Supuestos del modelo (2)

4. No autocorrelación: Nos dice que los términos de error no están correlacionados unos con otros, es decir, no hay correlación entre las observaciones. Esto significa que:

$$Cov(\varepsilon_i, \varepsilon_j \mid X_i) = 0 \qquad ; \qquad Cov(\varepsilon_i, \varepsilon_j) = 0$$

5. Más observaciones que regresoras: Sea n el numero de observaciones con que contamos, para la regresión simple deberá cumplirse que:

$$n > 1$$

6. Variabilidad de los valores de X: Los valores de la variable independiente no deben ser todos iguales y tampoco deben haber valores atípicos.

Mínimos Cuadrados Ordinarios

Es un método de estimación de parámetros de modo tal que se minimiza la suma de cuadrados de los residuos (SCR). Luego el problema de optimización por Mínimos Cuadrados Ordinarios (MCO) puede plantearse como:

$$min\, SCR = min \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = min \sum_{i=1}^{n} \left[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i) \right]^2$$

De donde, resolviendo, finalmente resulta:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} Y_i \sum_{i=1}^{n} X_i}{n \sum_{i=1}^{n} {X_i}^2 - (\sum_{i=1}^{n} X_i)^2} = \frac{Cov(X,Y)}{Var(X)}$$

REGRESIÓN LINEAL MÚLTIPLE

Definición

Implica que hay más de una variable independiente, de modo que el modelo de regresión sería:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Donde:
Y_i: Variable dependiente o regresada.
X_i: Variable independiente o regresora.
β_i: Coeficientes estimados.
k: Cantidad de variables independientes.
ε_i: Error aleatorio o residuo.
i : Subíndice que indica los diferentes valores que puede tomar la variable.

Estimación

. Un modelo de regresión lineal múltiple se puede estimar por el método de Mínimos Cuadrados Ordinarios (MCO).

. Conforme al Teorema de Gauss – Markov, si se cumplen los supuestos clásicos del modelo de regresión lineal, los estimadores obtenidos por MCO serán MELI.

. Los supuestos en regresión lineal múltiple son básicamente los mismos que en regresión lineal simple agregándose el supuesto de que no debe haber relación lineal exacta o alta entre las regresoras.

Medidas de bondad de ajuste

1. Coeficiente de determinación: Mide el porcentaje total de variación de la variable dependiente que es explicada por el modelo de regresión.

$$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

2. Coeficiente de determinación ajustado: Penaliza el aumento artificial del R^2 por agregar variables regresoras en el modelo.

$$\bar{R}^2 = 1 - \frac{\frac{SCR}{n-k}}{\frac{SCT}{n-1}}$$

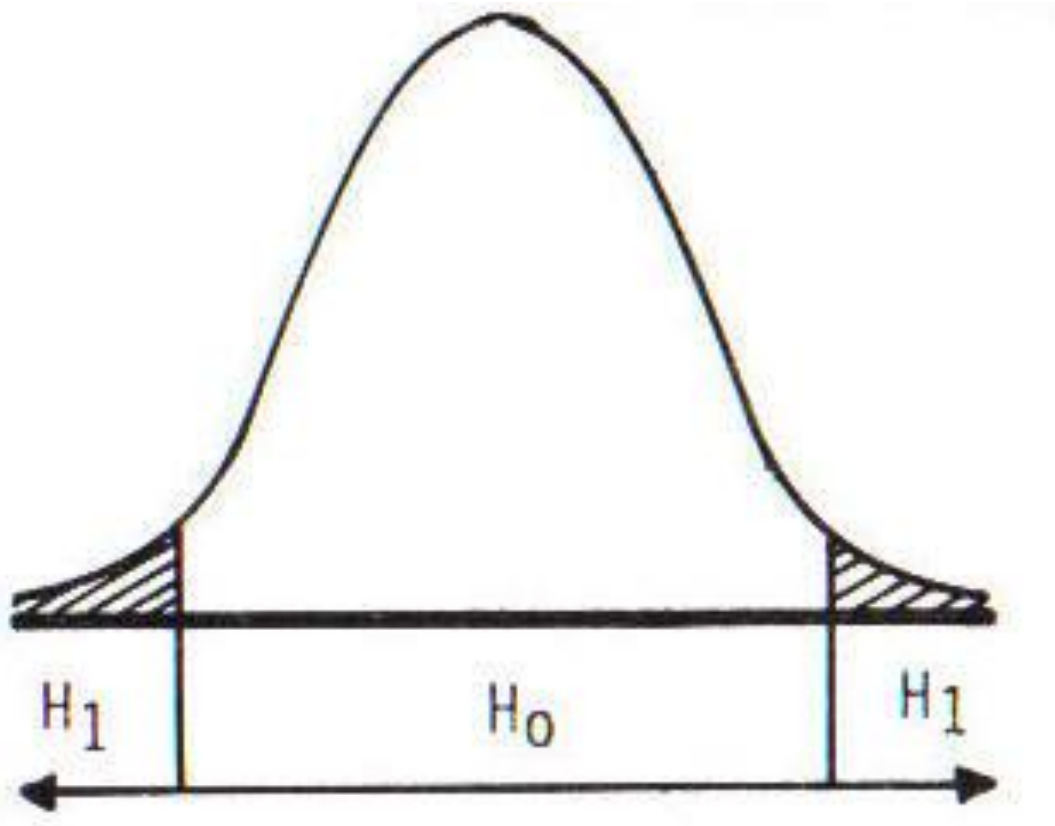
Contrastes de significación

1. Prueba t: Sirve para contrastar individualmente la significancia de las variables independientes conforme a la siguiente estructura de prueba:

$$H_0: \beta_i = 0$$
$$H_1: \beta_i \neq 0$$

2. Prueba F: Sirve para contrastar conjuntamente la significancia de las variables independientes conforme a la siguiente estructura de prueba:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1: \beta_1 \neq \beta_2 \neq \cdots \neq \beta_k \neq 0$$



Estimación por Eviews

Dependent Variable: CONS
Method: Least Squares
Date: 04/19/16 Time: 05:54
Sample: 1980Q1 2015Q4
Included observations: 144
CONS=C(1)+C(2)*PBI+C(3)*IMP

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	8010.054	994.2818	8.056120	0.0000
C(2)	0.427313	0.036963	11.56059	0.0000
C(3)	0.402212	0.107783	3.731698	0.0003

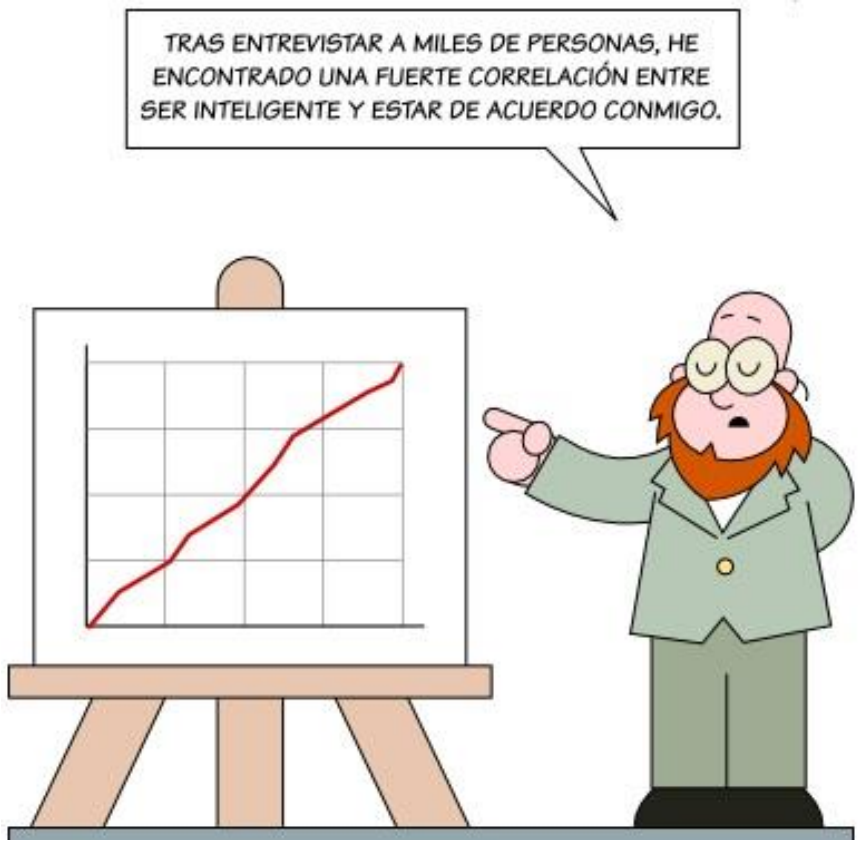
R-squared	0.981263	Mean dependent var	39804.68
Adjusted R-squared	0.980997	S.D. dependent var	14345.92
S.E. of regression	1977.609	Akaike info criterion	18.03778
Sum squared resid	5.51E+08	Schwarz criterion	18.09965
Log likelihood	-1295.720	Hannan-Quinn criter.	18.06292
F-statistic	3692.048	Durbin-Watson stat	1.752539
Prob(F-statistic)	0.000000		

MULTICOLINEALIDAD

Definición

Se refiere a la situación en que se da una relación lineal exacta (multicolinealidad perfecta) o casi exacta (multicolinealidad cuasi-perfecta) entre las variables regresoras del modelo.

Como criterio general se entiende que hay un problema de multicolinealidad cuando el coeficiente de correlación entre dos variables independientes toma un valor mayor a 0.8.



Causas

. Relación causal estrecha entre variables regresoras.

. Método erróneo de recolección de información.

. Restricciones en el modelo.

. Restricciones en la población de la que se extrae los datos.

. Mala especificación del modelo.

. Sobredeterminación del modelo.

Consecuencias

. Si hay multicolinealidad perfecta no se puede realizar la estimación quedando indeterminados los coeficientes de regresión.

. Si hay multicolinealidad cuasi-perfecta se puede estimar el modelo por MCO y los estimadores obtenidos son MELI pero tienen varianzas grandes generando intervalos de confianza artificialmente más amplios de modo que se introducen sesgos en las pruebas de hipótesis. Asimismo, el coeficiente de determinación presenta valores muy altos pese a que las variables no son individualmente significativas.

Detección y corrección

. Si hay multicolinealidad perfecta se sabrá en cuanto veamos que no se puede realizar la estimación dado que hay una matriz singular.

. Si hay multicolinealidad cuasi-perfecta ello se evidenciará en correlaciones mayores a 0.8 entre pares de regresores, un R^2 inusualmente alto y coeficientes de estimación que son significativos conjuntamente pero no individualmente.

. Para solucionar la multicolinealidad se pueden eliminar algunas de las variables que la causan o realizar una transformación de los datos.

HETEROCEDASTICIDAD

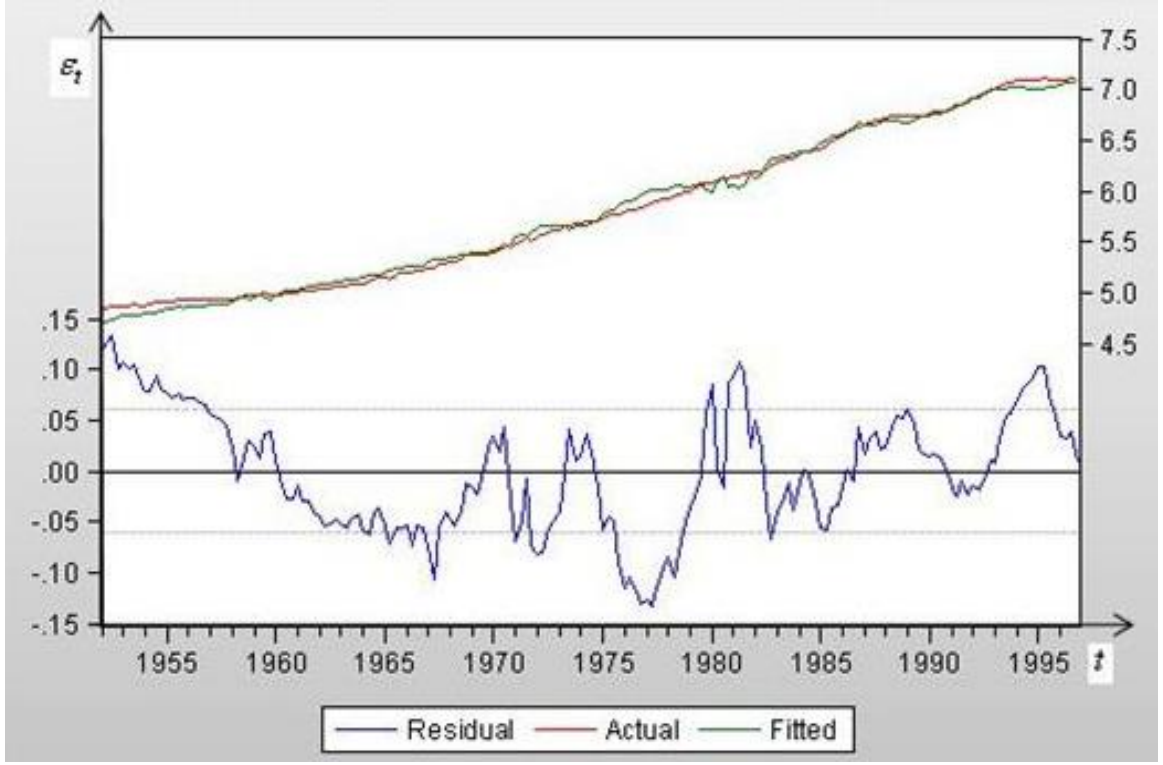
Definición

Se presenta cuando la varianza de los errores no es constante, es decir, es la violación del supuesto de homocedasticidad.

$$Var(\varepsilon_i) = \sigma^2$$

$$Var(\varepsilon_i \mid X_i) = \sigma^2$$

Se puede ver en primera instancia en la gráfica de los residuos.



Causas

. Modelos con dinámica de aprendizaje (los errores van decreciendo).

. Demasiada heterogeneidad entre los grupos de datos.

. Omisión de variables relevantes.

. Datos atípicos.

. Incorrecta transformación de datos.

Consecuencias

. Los estimadores son ineficientes, es decir, no tienen varianza mínima.

. El error de estándar de cada coeficiente es mayor que el que correspondería a la regresión que ajusta la heterocedasticidad.

. Los coeficientes tienen menor significancia estadística que los correspondientes a la regresión que ajusta la heterocedasticidad.

Detección: Test de White

Heteroskedasticity Test: White

F-statistic	8.655235	Prob. F(1,159)	0.0037
Obs*R-squared	8.311657	Prob. Chi-Square(1)	0.0039
Scaled explained SS	12.38302	Prob. Chi-Square(1)	0.0004

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 08/11/13 Time: 21:26
Sample: 2000M01 2013M05
Included observations: 161

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.001452	0.000365	3.981778	0.0001
LC^2	-1.23E-05	4.19E-06	-2.941978	0.0037

R-squared	0.051625	Mean dependent var	0.000391

AUTOCORRELACIÓN

Definición

Se refiere a la existencia de correlación entre los términos de error asociados a diferentes observaciones. Así, dado el modelo de regresión:

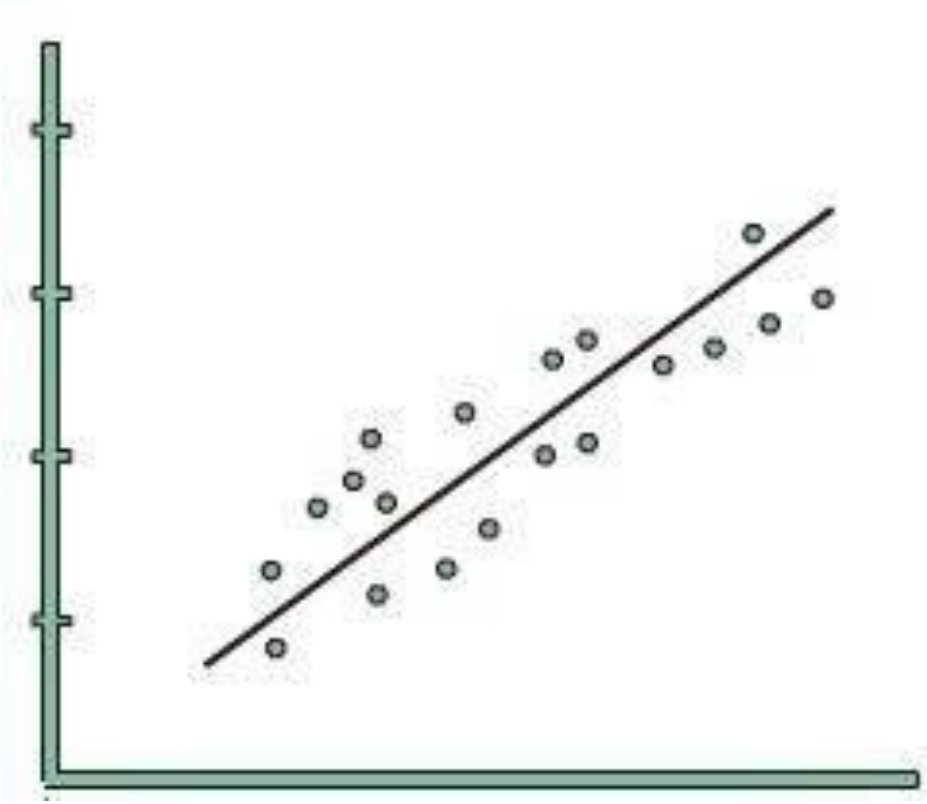
$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

Se dice que existe problema de autocorrelación si:

$$\varepsilon_t = f(\varepsilon_{t-i})$$

De modo que:

$$cov(\varepsilon_t, \varepsilon_{t-i}) \neq 0$$



Causas

. Omisión de variables relevantes.

. Especificación incorrecta de la forma funcional del modelo.

. Transformaciones de los datos.

. Existencias de ciclos o tendencias en las variables económicas.

. Inclusión en el modelo de valores retardados de la variable dependiente.

Consecuencias

. Los estimadores siguen siendo insesgados pero ya no son eficientes.

. Las pruebas t y F pierden validez.

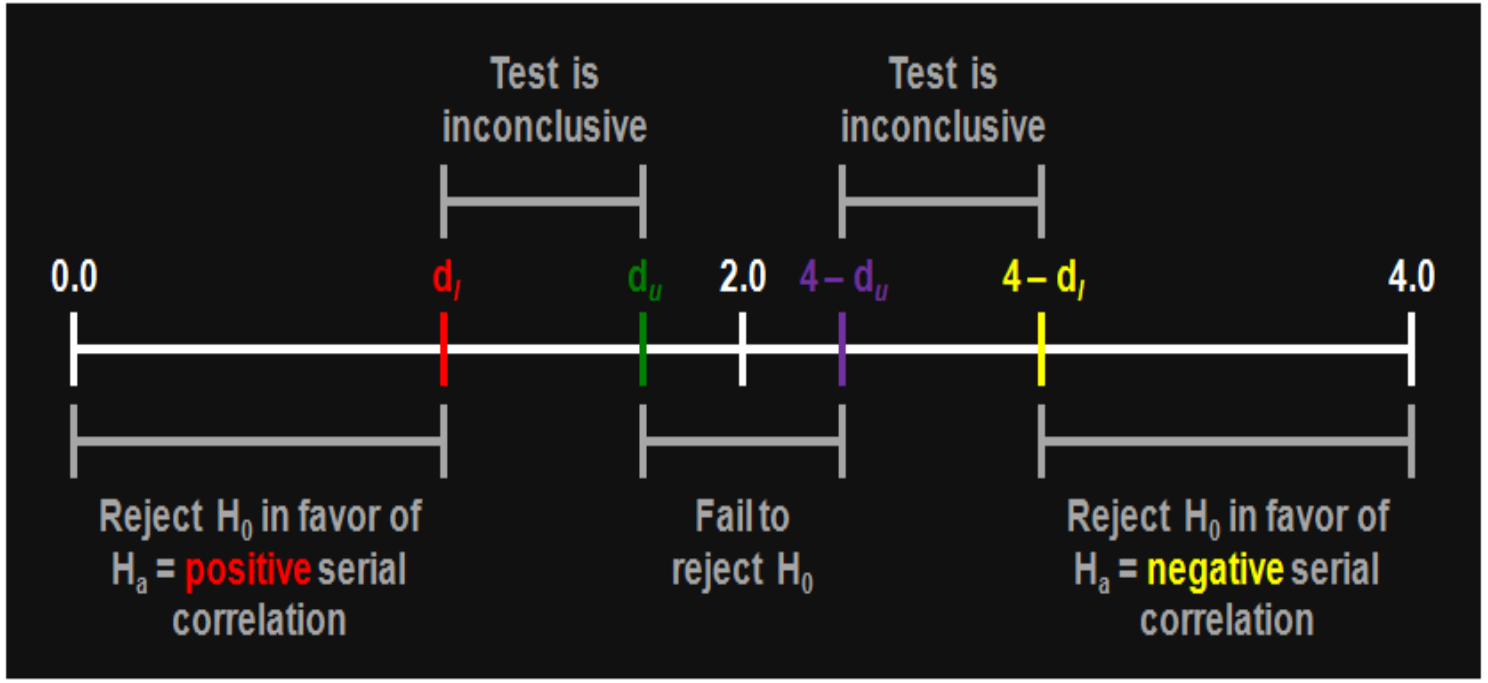
. Si la autocorrelación es positiva, la varianza de los residuos esta subestimada

. Si la autocorrelación es negativa, está sobrestimada. Lo mismo con la varianza de los estimadores.

Detección (1)

El test de Durbin-Watson sirve para detectar autocorrelación de primer orden en los errores (su hipótesis nula es que no la hay). Su estadístico viene dado por:

$$d = \frac{\sum_{i=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{i=1}^{n}(\hat{\varepsilon}_t)^2}$$

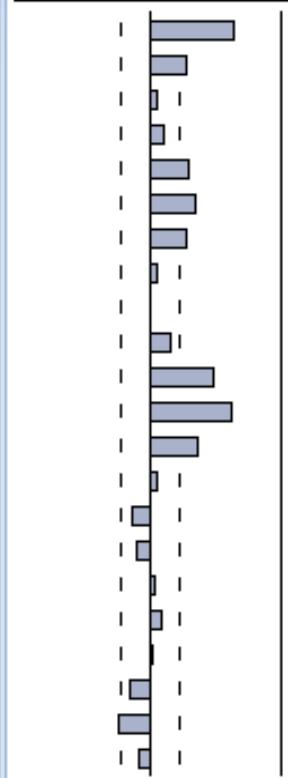


Detección (2)

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Correlogram of Residuals

Date: 08/03/10 Time: 17:25
Sample: 2001M01 2007M01
Included observations: 73



Autocorrelation	Partial Correlation		AC	PAC	Q-Stat	Prob
		1	0.670	0.670	34.128	0.000
		2	0.291	-0.285	40.678	0.000
		3	0.060	0.000	40.959	0.000
		4	0.102	0.263	41.780	0.000
		5	0.296	0.235	48.830	0.000
		6	0.363	-0.050	59.591	0.000
		7	0.282	0.006	66.198	0.000
		8	0.060	-0.152	66.499	0.000
		9	-0.003	0.168	66.500	0.000
		10	0.164	0.248	68.841	0.000
		11	0.506	0.426	91.407	0.000
		12	0.650	0.054	129.34	0.000
		13	0.384	-0.380	142.83	0.000
		14	0.058	-0.093	143.13	0.000
		15	-0.143	-0.064	145.05	0.000
		16	-0.112	-0.201	146.25	0.000
		17	0.038	-0.117	146.39	0.000
		18	0.089	-0.142	147.17	0.000
		19	0.010	0.008	147.18	0.000
		20	-0.168	-0.003	150.10	0.000
		21	-0.248	-0.140	156.57	0.000
		22	-0.091	-0.030	157.47	0.000

VARIABLES DUMMY

Definición

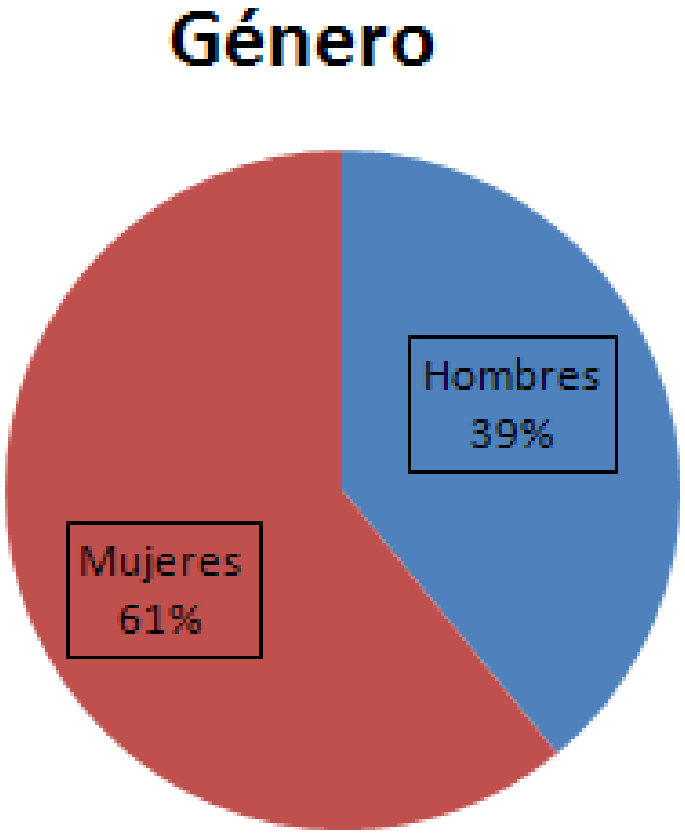
Una variable dummy, también conocida como variable binaria o dicotómica, es aquella que toma los valores de 0 o 1 para indicar la ausencia o presencia de alguna característica cualitativa que puede tener efecto sobre la variable dependiente. Tiene dos formas básicas:

1. **Modelos de Análisis de Varianza (ANOVA):**

$$Y_i = \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + \varepsilon_i$$

2. **Modelos de Análisis de Covarianza (ANCOVA):**

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \varepsilon_i$$



Interpretación

Dependent Variable: EARNINGS
Method: Least Squares
Date: 11/11/12 Time: 14:42
Sample: 1 540
Included observations: 540

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	16.12126	0.861967	18.70286	0.0000
MALE	7.195963	1.219006	5.903140	0.0000

R-squared	0.060831	Mean dependent var	19.71924
Adjusted R-squared	0.059086	S.D. dependent var	14.60151
S.E. of regression	14.16357	Akaike info criterion	8.142920
Sum squared resid	107926.4	Schwarz criterion	8.158815
Log likelihood	-2196.588	Hannan-Quinn criter.	8.149137
F-statistic	34.84706	Durbin-Watson stat	1.921263
Prob(F-statistic)	0.000000		

DIAGNÓSTICO Y SELECCIÓN DE MODELOS

Criterios para comparación

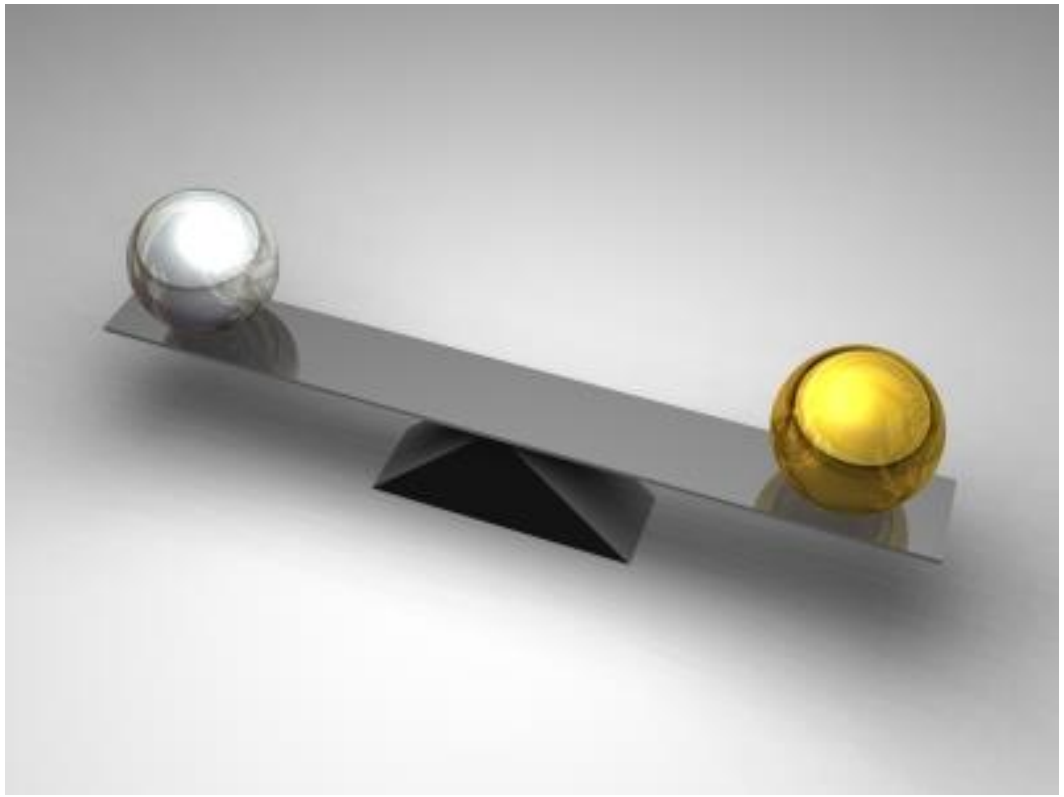
1. Significación económica.

2. Coeficiente de determinación directo y ajustado.

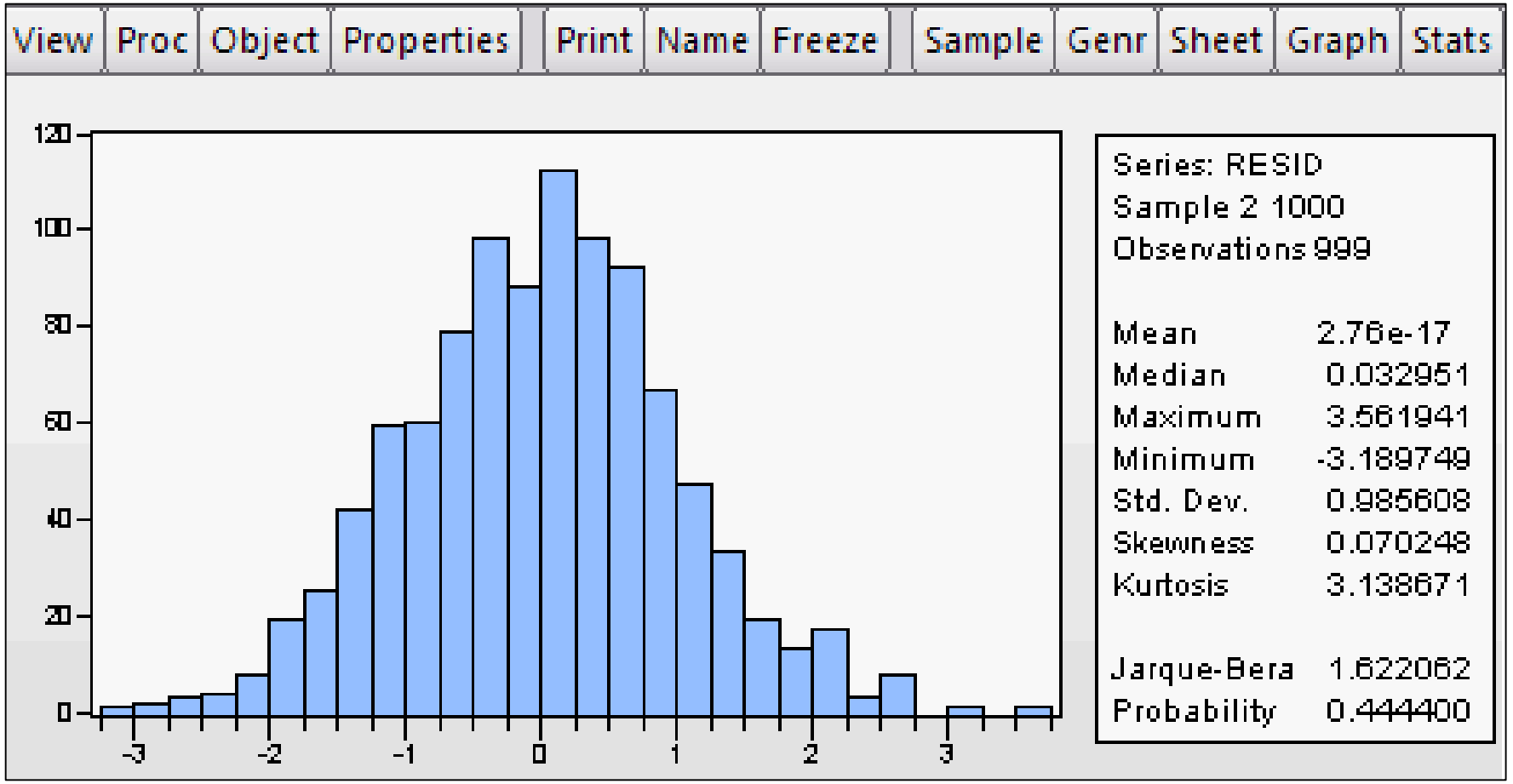
3. Problemas de multicolinealidad, heterocedasticidad o autocorrelación.

4. Normalidad en la distribución de los residuos o errores.

5. Uso de información (criterios de Akaike, Schwarz y Hannan-Quin).



Test de normalidad de los residuos



Ejemplo de comparación de modelos

MODELO 1

Dependent Variable: CONS
Method: Least Squares
Date: 04/19/16 Time: 05:54
Sample: 1980Q1 2015Q4
Included observations: 144
CONS=C(1)+C(2)*PBI+C(3)*IMP

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	8010.054	994.2818	8.056120	0.0000
C(2)	0.427313	0.036963	11.56059	0.0000
C(3)	0.402212	0.107783	3.731698	0.0003

R-squared	0.981263	Mean dependent var	39804.68
Adjusted R-squared	0.980997	S.D. dependent var	14345.92
S.E. of regression	1977.609	Akaike info criterion	18.03778
Sum squared resid	5.51E+08	Schwarz criterion	18.09965
Log likelihood	-1295.720	Hannan-Quinn criter.	18.06292
F-statistic	3692.048	Durbin-Watson stat	1.752539
Prob(F-statistic)	0.000000		

MODELO 2

Dependent Variable: CONS
Method: Least Squares
Date: 04/19/16 Time: 05:56
Sample: 1980Q1 2015Q4
Included observations: 144
CONS=C(1)+C(2)*INV+C(3)*GOB

	Coefficient	Std. Error	t-Statistic	Prob.
C(1)	16386.36	835.6242	19.60973	0.0000
C(2)	1.638409	0.090477	18.10865	0.0000
C(3)	1.024197	0.214470	4.775485	0.0000

R-squared	0.961259	Mean dependent var	39804.68
Adjusted R-squared	0.960709	S.D. dependent var	14345.92
S.E. of regression	2843.625	Akaike info criterion	18.76416
Sum squared resid	1.14E+09	Schwarz criterion	18.82603
Log likelihood	-1348.020	Hannan-Quinn criter.	18.78930
F-statistic	1749.277	Durbin-Watson stat	0.625925
Prob(F-statistic)	0.000000		

CONCLUSIONES

- El modelo de regresión lineal nos permite captar en términos estadístico-matemáticos ciertas relaciones entre variables en términos de coeficientes estimados que pueden permitir interpretaciones con significado económico para la dilucidación y/o contraste de ciertas teorías económicas.

- Cuando no se cumplen los supuestos clásicos del modelo de regresión lineal ello puede introducir distorsiones en determinados resultados. En específico, se pueden hallar situaciones de multicolinealidad (correlación perfecta o alta entre regresores), heterocedasticidad (varianza no constante de los errores) y/o autocorrelación (correlación entre los errores).

- Bajo determinados criterios se puede comparar modelos para ver cuál es "mejor" en términos de significación económica, cumplimiento de supuestos econométricos y eficiencia en el uso de la información.

REFERENCIAS

. Gujarati, D. y Porter, D. (2011). *Econometría*. México: McGraw-Hill.

. Larios, J., Álvarez, V. y Quineche, R. (2014). *Fundamentos de Econometría*. Lima: Universidad San Ignacio de Loyola.

. Novales, A. (1993). *Econometría*. Madrid: McGraw-Hill.

. Sosa, W. (2015). *El Lado Oscuro de la Econometría*. Buenos Aires: Temas.

Profesor Dante A. Urbina:

. Página Web: http://www.danteaurbina.com
. Facebook: http://www.facebook.com/danteaurbina.oficial
. Canal YouTube: http://www.youtube.com/channel/UCCwVIDA-8wV4D_GpYNVecrg