Econometría I (EC402)
Clase #15 - Análisis de Varianza y Predicción

Prof. Andrés M. Castaño

Ingeniería Comercial
Universidad Católica del Norte
Lunes 28 de octubre de 2013

Análisis de varianza

- $$STC = SEC + SRC$$
  $$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{\mu}_i^2$$
  $$= \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{\mu}_i^2$$

- El estudio de los componentes de la STC se conoce como el análisis de varianza.

- Asociado con cada uno de estos componentes está, sus respectivos grados de libertad. STC (n-1) gl, SRC (n-2) gl, SEC (1) gl.

Tabla análisis de varianza

TABLA 5.3 TABLA ANOVA PARA EL MODELO DE REGRESIÓN CON DOS VARIABLES

Fuente de variación	SC*	g de l	SPC†
Debido a la regresión (SEC)	$\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$	1	$\hat{\beta}_2^2 \sum x_i^2$
Debido a los residuos (SRC)	$\sum \hat{u}_i^2$	$n-2$	$\frac{\sum \hat{u}_i^2}{n-2} = \hat{\sigma}^2$
STC	$\sum y_i^2$	$n-1$	

* SC significa suma de cuadrados.
† Significa suma de promedio de cuadrados, la cual se obtiene dividiendo SC por el número de g de l.

El estadístico F

- Si se satisface que μ_i esté normalmente distribuida, y si $H_0 : \beta_2 = 0$, existe un estadístico F que satisface la distribución F, con 1 y n-2 gl.

- $$F = \frac{SPC(SEC)}{SPC(SRC)}$$
$$= \frac{\hat{\beta}_2^2 \sum x_i^2}{\frac{\sum \hat{\mu}_i}{n-2}}$$
$$= \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2}$$

El estadístico F

- $$E(\hat{\beta}_2^2 \sum x_i^2) = \sigma^2 + \beta_2^2 \sum x_i^2$$

  y

  $$E(\frac{\sum \hat{\mu}_i}{n-2}) = E(\hat{\sigma}^2) = \sigma^2$$

- Si $\beta_2 = 0$, entonces ambas expresiones proporcionan el mismo valor para σ^2. Una distribución T (con k gl) al cuadrado es una distribución F $(1 : k)$ gl.

- Las pruebas T y F, proporcionan, dos formas alternas pero complementarias de testear $H_0 : \beta_2 = 0$.

- Cuando el análisis es de dos variables la prueba F es innecesaria, con más variables se pueden testear hipótesis globales.

Aplicaciones del análisis de regresión: problema de predicción

- obteniedo:

$$\hat{Y}_i = 24{,}4545 + 0{,}5091X_i$$

- Qué podemos hacer con esta regresión:
- 1 $\Longrightarrow$ Predicción media: valor de la media condicional de Y correspondiente al valor de X (ejemplo X_0).
- 2 $\Longrightarrow$ Predicción individual: predicción de un valor individual Y, correspondiente a X_0.

Predicción media

- Suponga que $X_0 = 100$, y deseamos saber $E(Y \mid X_0 = 100)$. Obteniedo:

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

$$= 24{,}4545 + 0{,}5091(100)$$

$$= 75{,}364$$

- Dado que $\hat{Y}_0$ es un estimador, puede ser diferente del verdadero valor $\Longrightarrow$ Error de predicción $\Longrightarrow$ debemos encontrar la distribución muestral de $\hat{Y}_0$.

- Y_0 se distribuye con media $= (\beta_1 + \beta_2 X_0)$, y

- $Var(\hat{Y}_0) = \sigma^2(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2})$

- $t = \frac{\hat{Y}_0 - (\beta_1 + \beta_2 X_0)}{ee(\hat{Y}_0)}$

Predicción media

- Intervalo de cofianza para el verdadero $E(Y_0 \mid X_0)$

$$Pr(\beta_1+\beta_2X_0-t_{\frac{\alpha}{2}}ee(\hat{Y}_0) \leq \beta_1+\beta_2X_0 \leq \beta_1+\beta_2X_0+t_{\frac{\alpha}{2}}ee(\hat{Y}_0)) = 1-\alpha$$

Predición media: Ejemplo...tabla 3.3 (Guajarati)

TABLA 3.3 DATOS PRIMARIOS BASADOS EN LA TABLA 3.2

	Y_i (1)	X_i (2)	Y_iX_i (3)	X_i^2 (4)	$x_i = X_i - \bar{X}$ (5)	$y_i = Y_i - \bar{Y}$ (6)	x_i^2 (7)	x_iy_i (8)	$\hat{Y}_i$ (9)	$\hat{u}_i = Y_i - \hat{Y}_i$ (10)	$\hat{Y}_i\hat{u}_i$ (11)
	70	80	5 600	6 400	−90	−41	8 100	3 690	65.1818	4.8181	314.0524
	65	100	6 500	10 000	−70	−46	4 900	3 220	75.3636	−10.3636	−781.0382
	90	120	10 800	14 400	−50	−21	2 500	1 050	85.5454	4.4545	381.0620
	95	140	13 300	19 600	−30	−16	900	480	95.7272	−0.7272	−69.6128
	110	160	17 600	25 600	−10	−1	100	10	105.9090	4.0909	433.2631
	115	180	20 700	32 400	10	4	100	40	116.0909	−1.0909	−126.6434
	120	200	24 000	40 000	30	9	900	270	125.2727	−6.2727	−792.0708
	140	220	30 800	48 400	50	29	2 500	1 450	136.4545	3.5454	483.7858
	155	240	37 200	57 600	70	44	4 900	3 080	145.6363	8.3636	1 226.4073
	150	260	39 000	67 600	90	39	8 100	3 510	156.8181	−6.8181	−1 069.2014
Suma	1 110	1 700	205 500	322 000	0	0	33 000	16800	1 109.9995 ≈1 110.0	0	0.0040 ≈ 0.0
Media	111	170	nc	nc	0	0	nc	nc	110	0	0

$$\hat{\beta}_2 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = 16\,800/33\,000 = 0.5091$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 111 - 0.5091(170) = 24.4545$$

Cálculos ejemplo...

- obteniedo:

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$
$$= 24{,}4545 + 0{,}5091(100)$$
$$= 75{,}364$$

- Con media $= (\beta_1 + \beta_2 X_0)$.
- $Var(\hat{Y}_0) = 42{,}159(\frac{1}{10} + \frac{(100-170)^2}{33000}) = 10{,}4759$

Cálculos ejemplo...

- Intervalo de confianza al 95 % de confianza para el verdadero $E(Y \mid X_0) = \beta_1 + \beta_2 X_0$:

$$75{,}3645 - 2{,}306(3{,}2366) \leq E(Y_0 \mid X = 100) \leq 75{,}3645 + 2{,}306(3{,}2366)$$

$$(67{,}9010 \leq E(Y_0 \mid X = 100) \leq 82{,}8381) = 1 - \alpha$$

- Interpretación: En muestreo repetido dado $X_0 = 100$, en 95 de cada 100 intervalos como el anterior estará incluido el verdadero valor medio de Y_0.

Predicción individual.....

- Si el interés está en predecir un valor individual Y, Y_0 correspondiente a un valor dado de X (X_0), el mejor estimador lineal insesgado de Y_0 esta dado por:

$$= (\beta_1 + \beta_2 X_0)$$

- $$Var(Y_0 - \hat{Y}_0) = E(Y_0 - \hat{Y}_0) = \sigma^2 (1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2})$$

- $$t = \frac{Y_0 - \hat{Y}_0}{ee(Y_0 - \hat{Y}_0)}$$

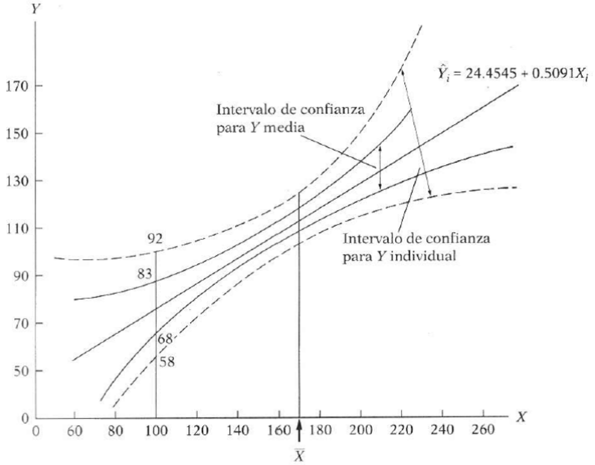
- Con el mismo ejemplo, hallar el intervalo de confianza para la predicción individual, debería darles:

$$(58{,}6345 \leq Y_0 \mid X_0 = 100 \leq 92{,}0945)$$

Predicción individual.....

- Evaluar lo siguiente: ¿Es posible decir con una confianza del 95 % que el consumo de una familia será de 110 (miles) si el ingreso es 180 (miles)?

Comparación predicción media vs predicción individual



Intervalos (bandas) de confianza para Y media y para valores individuales de Y.

Volvamosa a una salida de regresión

$$
\begin{array}{llll}
\hat{Y}_i = 24.4545 & + \quad 0.5091X_i & & \\
\text{ee} = (6.4138) & (0.0357) & & r^2 = 0.9621 \\
t = (3.8128) & (14.2605) & & \text{g de l} = 8 \\
p = (0.002571) & (0.000000289) & & F_{1,8} = 202.87
\end{array}
$$

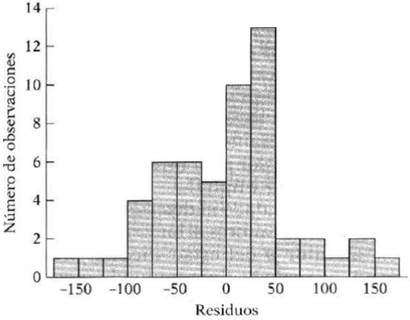
Qué hacer después de obtener la salida

- Signo de los coeficientes
- Significancia de los coeficientes
- Bonda de ajuste
- Estudiar el cumplimiento de los supuestos

Prueba de normalidad

- Histograma de los residuos
- Gráfica de probabilidad normal
- Prueba Jarque-Bera

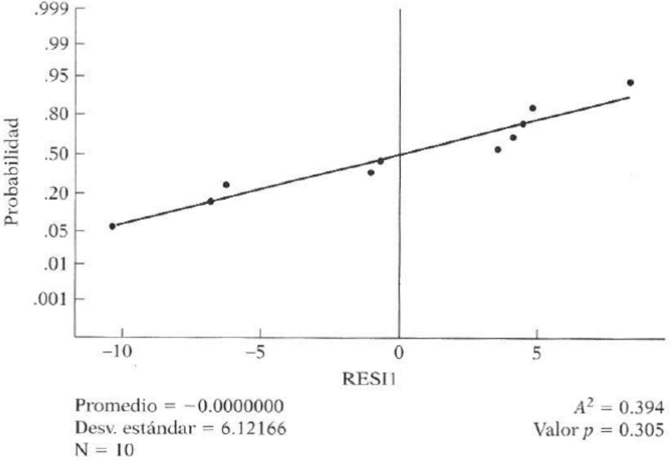
Histograma de residuos



Series: residuos
Muestra 1 55
Observaciones 55

Media	-1.19×10^{-14}
Mediana	7.747849
Máximo	171.5859
Mínimo	-153.7664
Desviación estándar	66.23382
Asimetría	0.119816
Curtosis	3.234473
Jarque-Bera	0.257585
Probabilidad	0.879156

Gráfica de probabilidad normal



Prueba Jarque-Bera (JB)

- Prueba asintótica o de grandes muestras.
- Calcula la asimetría y la curtosis de los residuos.
- $$JB = n(\frac{S^2}{6} + frac(K-3)^2 24)$$

  n=tamaño de la muestra; S=coeficiente de asimetría; K=coeficiente de curtosis.