

Instituto Politécnico Nacional  
Escuela Superior de Cómputo  
Secretaría Académica  
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)  
*Regresión lineal (4ª Parte)*  
*Significancia estadística*

1

Profesora: Dra. Fabiola Ocampo Botello

## Verificación de la ecuación de estimación

La ecuación de regresión estimada no debe ser usada hasta que se realice un análisis para determinar si el modelo empleado es adecuado, es decir, conocer su significancia estadística.

2



Esta foto de Autor desconocido está bajo licencia [CC BY-SA-NC](#)

Data Mining, ESCOM-IPN. Dra. Fabiola Ocampo Botello

### Definición de significancia estadística

Un conjunto de medidas u observaciones en una investigación estadística es estadísticamente significativa si es poco probable que haya ocurrido por el azar (Bennet, Briggs & Triola, 2011:234).



Esta foto de Autor desconocido está bajo licencia [CC BY](#)

3

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

Anderson, Sweeney & Williams (2008) establecen lo siguiente:

En una ecuación de regresión lineal simple, la media o valor esperado de  $y$  es la función lineal de  $x$ , de la forma:

$$E(y) = \beta_0 + \beta_1 x$$

Si  $\beta_1 = 0$ , indica que no existe relación lineal entre  $x$  y  $y$ .  
(Hipótesis nula)

$$E(y) = \beta_0 + \beta_1(0) = \beta_0$$

*$x$  y  $y$  no están relacionados linealmente*

4

Si  $\beta_1 \neq 0$ , indica que  $x$  y  $y$  están relacionadas linealmente.  
(Hipótesis alterna)

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

Para probar si una ecuación de regresión es significativa, se debe realizar una prueba de hipótesis para determinar si el valor de  $\beta_1$  es distinto de cero.

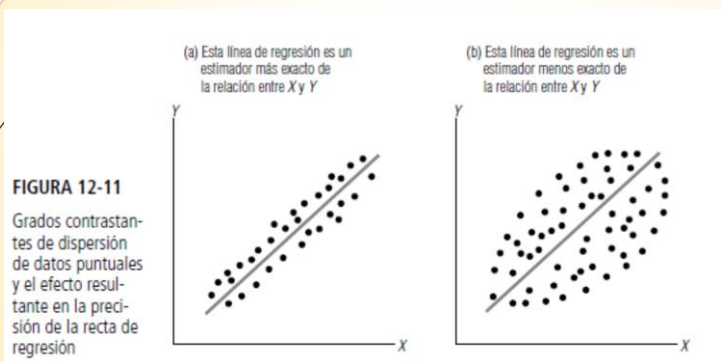
Hay dos pruebas que son las más usadas. En ambas, se requiere una estimación de  $\sigma^2$ , la varianza de  $\epsilon$  en el modelo de regresión.

5

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

## Estimación de $\sigma^2$

$\sigma^2$ , la varianza de  $\epsilon$ , representa también la varianza de los valores de  $y$  respecto a la recta de regresión



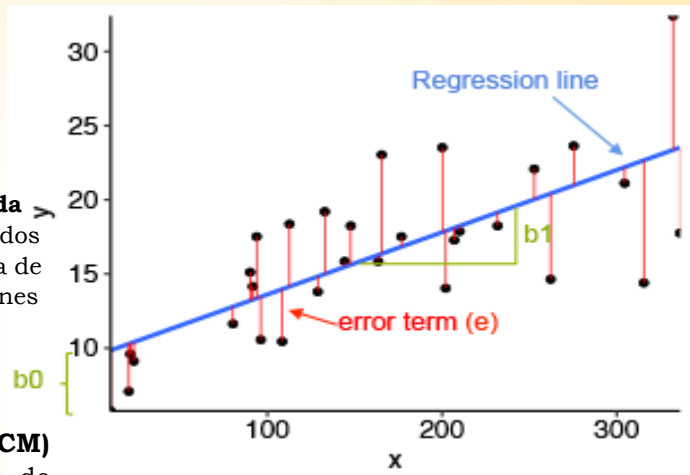
6

Imagen tomada de Levin, et. al (2004)

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

A las desviaciones de los valores de  $y$  de la recta de regresión estimada se les conoce como **residuales**.

**SCE (Suma de cuadrados debida al error)**, la suma de los cuadrados de los residuales, es una medida de la variabilidad de las observaciones reales respecto a la línea de regresión estimada.



- 7 **El error cuadrado medio (ECM)** proporciona una estimación de  $\sigma^2$ ; esta estimación es SCE dividida entre sus grados de libertad.

Imagen Creative Commons

En: <http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/>

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

## ERROR CUADRADO MEDIO (ESTIMACIÓN DE $\sigma^2$ )

$$s^2 = ECM = \frac{SCE}{n-2}$$

## ERROR ESTÁNDAR DE ESTIMACIÓN

$$s = \sqrt{ECM} = \sqrt{\frac{SCE}{n-2}}$$

8

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

Se calcula el error estándar de la estimación:

Para el caso de la Pizzería “Polito”, se tiene:

$$s^2 = \text{ECM} = \frac{1530}{8} = 191.25$$

$$s = \sqrt{191.25} = 13.82$$

9

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

## Prueba t

En el modelo de regresión lineal  $E(y) = \beta_0 + \beta_1 x + \epsilon$

Si  $x$  y  $y$  están relacionados linealmente, entonces  $\beta_1 \neq 0$

### PRUEBA DE t DE SIGNIFICANCIA PARA LA REGRESIÓN LINEAL

Se generan las hipótesis considerando el parámetro  $\beta_1$

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

10

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

**ESTADÍSTICO DE PRUEBA**

$$t = \frac{b_1}{s_{b1}}$$

**REGLA DE RECHAZO**

Método de valor-p:

Rechazar  $H_0$  si valor-p  $\leq \alpha$ 

Método de valor crítico:

Rechazar  $H_0$  si  $t \leq -t_{\alpha/2}$  o si  $t \geq t_{\alpha/2}$ Donde se toma de la distribución  $t_{\alpha/2}$  con  $n-2$  grados de libertad

11

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

Considerando el ejemplo de la pizzería Polito, Se toma el p-valor del resultado que proporcionó el Knime.

S	Variable	D	Coeff.	D	Std. Err.	D	t-value	D	P> t
	NoEstud	5		0.58		8.617		0	
	Intercept	60		9.226		6.503		0	

Si  $\alpha = 0.01$  como nivel de significancia, p-valor  $< \alpha$  se rechaza  $H_0$  y se concluye que hay una relación estadísticamente significativa entre el tamaño del campus y las ganancias de la pizzería.

12

La probabilidad de que los valores observados se deban al azar es 0.

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

## Prueba F

Se utiliza para probar que existe una relación estadísticamente significativa cuando hay más de una variable independiente.

Cuando sólo se tiene una variable independiente, la prueba F lleva a la misma conclusión que la prueba t; es decir, si la prueba t indica que  $\beta_1 \neq 0$  y por lo tanto que existe una relación significativa, la prueba F también indicará que existe una relación significativa.

13

Data Mining. ESCOM-IPN. Dra. Fabiola Ocampo Botello

14

### Referencias bibliográficas

Anderson, Sweeney & Williams. (2008). Estadística para administración y economía, 10ª edición. Cengage Learning.  
Bennet, Briggs & Triola (2011). Razonamiento estadístico. Pearson. México.

Dra. Fabiola Ocampo Botello