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

Profesora: *Dra. Fabiola Ocampo Botello*

1

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

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).

3

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 $\boldsymbol{\beta_1 = 0}$, 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

Si $\boldsymbol{\beta_1 <> 0}$, indica que x y y están relacionadas linealmente. (Hipótesis alterna)

4

Para probar si una ecuación de regresión es significativa, se debe realizar una prueba de hipótesis para determinar si el valor de β_1 es distinto de cero.

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

5

Estimación de σ2

σ^2, la varianza de ϵ, representa también la varianza de los valores de y respecto a la recta de regresión

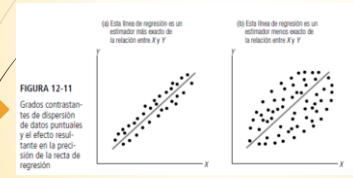

FIGURA 12-11
Grados contrastantes de dispersión de datos puntuales y el efecto resultante en la precisión de la recta de regresión

Imagen tomada de Levin, et. al (2004)

6

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 linea de regresión estimada.

El error cuadrado medio (ECM) proporciona una estimación de σ^2; esta estimación es SCE dividida entre sus grados de libertad.

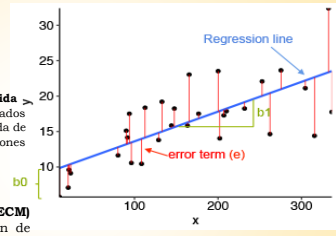

Imagen Creative Commons
En: http://www.sthda.com/english/articles/40-regression-analysis/167-simple-linear-regression-in-r/

7

ERROR CUADRADO MEDIO (ESTIMACIÓN DE σ^2)

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

ERROR ESTÁNDAR DE ESTIMACIÓN

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

8

Se calcula el error estándar de la estimación:

Para el caso de la Pizzería "Polito", se tiene:

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

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

9

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 β_1

H_0: $\beta_1 = 0$

Ha: $\beta_1 \neq 0$

10

ESTADÍSTICO DE PRUEBA

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

REGLA DE RECHAZO

Método de valor-p: Rechazar H_0 si valor-p <= α

Método de valor crítico: Rechazar H_0 si t <= $-t_{\alpha/2}$ o si t >= $t_{\alpha/2}$

Donde se toma de la distribución $t_{\alpha/2}$ con n–2 grados de libertad

11

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

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

12

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.

Referencias bibliográficas

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