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

Minería de datos (*Data Mining*)
Regresión lineal (3ª Parte)

1

Profesora: *Dra. Fabiola Ocampo Botello*

El error estándar de la estimación

¿Cómo evaluar la confiabilidad de una ecuación de estimación de regresión encontrada?

Levin et al (2004) establecen que el **error estándar de la estimación (Se)** mide la variabilidad o dispersión de los valores observados alrededor de la recta de regresión.

2

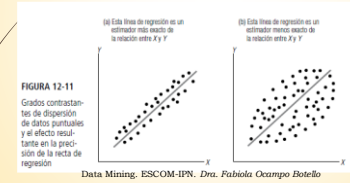

FIGURA 12-11
Grados contrastantes de dispersión de datos puntuales y el efecto resultante en la precisión de la recta de regresión

Imagen tomada de Levin, et. al (2004)

Levin et al (2004) establecen que una forma de calcular el error Є es mediante el error estándar de la estimación, *mide la variabilidad o dispersión de los valores observados alrededor de la recta de regresión.*
El cual tiene la siguiente fórmula:

Error estándar de la estimación

$$s_e = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}}$$

Ecuación 12-6.

3

donde,

- Y = valores de la variable dependiente
- $\hat{Y}$ = valores estimados con la ecuación de estimación que corresponden a cada valor de Y
- n = número de puntos utilizados para ajustar la línea de regresión

Imágenes tomadas de Levin, et. al (2004)

Imagen tomada de Levin, et. al (2004)

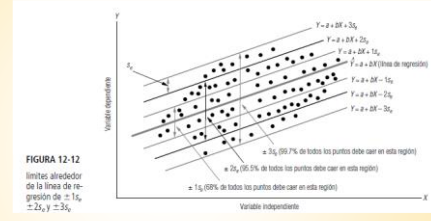

FIGURA 12-12
límites alrededor de la línea de regresión de $\pm 1s_e$, $\pm 2s_e$ y $\pm 3s_e$

4

Los valores observados para Y deben tener una distribución normal alrededor de cada valor estimado de Ŷ (Levin et al 2004:528, Anderson, Sweeney & Williams, 2008:567).

Los intervalos de confianza de la distribución normal expuesta en el tema de Normalidad vista en este curso (*vea el tema de la presentación de normalidad*) *la regla 68-95-99.7 significa el porcentaje de la cantidad de datos que se encuentran en 1, 2 ó 3 desviaciones estándar en la gráfica de la campana).*

Lo cual significa lo siguiente:

1. Alrededor del 68% (o de forma más precisa, el 68.3%), o un poco más de dos tercios de los datos se encuentran dentro de una desviación estándar de la media.
2. Alrededor del 95% (o de forma más precisa, el 95.4%) de los datos caen dentro de dos desviaciones estándar de la media.
3. Alrededor del 99.7% de los datos caen dentro de tres desviaciones estándar de la media.

5

Para analizar el cálculo del error estándar de medición, consideremos nuevamente el ejemplo de la directora del Departamento de Salubridad que deseaba conocer la relación entre la antigüedad de los camiones y la cantidad de reparaciones anuales presentado por Levin et al (2004).

$$\hat{Y} = 3.75 + 0.75X$$

El primer paso para conocer el error estándar de medición es calcular el numerador de la ecuación, esto es:

$$\sum (Y - \hat{Y})^2.$$

Tabla 12-11 Cálculo del numerador de la fracción en la ecuación 12-6

X (1)	Y (2)	$\hat{Y}$ (es decir, $3.75 + 0.75X$) (3)	Error individual $(Y - \hat{Y})$ (2) − (3)	$(Y - \hat{Y})^2$ $[(2) - (3)]^2$
5	7	3.75 + (0.75)(5)	7 − 7.5 = −0.5	0.25
3	7	3.75 + (0.75)(3)	7 − 6.0 = 1.0	1.00
3	6	3.75 + (0.75)(3)	6 − 6.0 = 0.0	0.00
1	4	3.75 + (0.75)(1)	4 − 4.5 = −0.5	0.25
				$\Sigma(Y - \hat{Y})^2 = 1.50$ ← Suma de los cuadrados de los errores

6

Imágenes tomadas de Levin, et. al (2004)

Una vez calculado $\sum (Y - \hat{Y})^2$ se procede a aplicar la ecuación 12-6.

$$s_e = \sqrt{\frac{\Sigma(Y - \hat{Y}^2)}{n - 2}}$$
$$= \sqrt{\frac{1.50}{4 - 2}}$$
$$= \sqrt{0.75}$$
$$= 0.866 \leftarrow \text{Error estándar de la estimación de \$86.60}$$

$$\hat{Y} = 3.75 + 0.75X$$

7

Imagen tomada de Levin, et. al (2004)

Intervalos de confianza para la estimación (o el valor esperado)

Los intervalos de confianza para la estimación se refieren a *la posibilidad de realizar afirmaciones de probabilidad acerca del intervalo alrededor del valor estimado de Ŷ.*

Regresando al ejemplo de la directora del Departamento de Salubridad, la ecuación encontrada fue:

$$\hat{Y} = 3.75 + 0.75X$$

Si se considera un camión con cuatro años de antigüedad, el gasto calculado es:

$$\hat{Y} = 3.75 + 0.75(4)$$
$$= 3.75 + 3.00$$
$$= 6.75 \leftarrow \text{Gasto anual de reparaciones esperado de \$675}$$

8

El error estándar calculado fue de 0.866 ($86.60).

Imágenes tomadas de Levin, et. al (2004)

Suponga que en el caso de la directora del Departamento de Salubridad desea tener una confianza del 68% de que el gasto real de reparaciones está dentro de ±1 desviación estándar de la desviación de Ŷ. Los intervalos de confianza son:

$$\hat{Y} + 1s_{\hat{Y}} = \$675 + (1)(\$86.60)$$
$$= \$761.40 \leftarrow \text{Límite superior del intervalo de predicción}$$
$$\hat{Y} - 1s_{\hat{Y}} = \$675 - (1)(\$86.60)$$
$$= \$588.40 \leftarrow \text{Límite inferior del intervalo de predicción}$$

Si deseara tener una confianza del 95.5%, lo cual representa el valor de ±2 desviaciones estándar de la desviación de Ŷ. Se tiene:

$$\hat{Y} + 2s_{\hat{Y}} = \$675 + (2)(\$86.60)$$
$$= \$848.20 \leftarrow \text{Límite superior}$$
$$\hat{Y} - 2s_{\hat{Y}} = \$675 - (2)(\$86.60)$$
$$= \$501.80 \leftarrow \text{Límite inferior}$$

Imágenes tomadas de Levin, et. al (2004)

Levin, et. al (2004:528) establecen que los estadísticos aplicados para los intervalos de confianza se basan en la normalidad de los datos sólo para muestras grandes (n > 30).

Para evitar el cálculo de valores inexactos, es necesario aplicar la distribución t, ya que es adecuada para muestra de tamaño n < 30.

Debido a que en el ejemplo de la directora del Departamento de Salubridad la muestra es de tamaño n = 4.

Mason, Lind & Marshal (2000:286) indican que cuando el tamaño de la muestra n, es al menos igual a 30 se acepta que el teorema de límite central asegurará una distribución normal de las medias muestrales.

Ahora suponga que la directora del Departamento de Salubridad desea tener una seguridad aproximada del 95% de que los gastos anuales de reparación caerán en el intervalo de la estimación.

Grados de libertad	0.75	0.9	0.95	0.975	0.99	0.995	0.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.886	2.920	4.303	6.965	9.925	31.598
3	0.765	1.638	2.353	3.182	4.541	5.841	12.941
4	0.741	1.533	2.132	2.776	3.747	4.604	8.610

Tabla t

$$\hat{Y} + t(s_{\hat{Y}}) = \$675 + (2.920)(\$86.60)$$
$$= \$675 + \$252.87$$
$$= \$927.87 \leftarrow \text{Límite superior}$$

$$\hat{Y} - t(s_{\hat{Y}}) = \$675 - (2.920)(\$86.60)$$
$$= \$675 - \$252.87$$
$$= \$422.13 \leftarrow \text{Límite inferior}$$

La directora puede estar 95% segura de que los gastos anuales de reparación de un camión de cuatro años de antigüedad estarán entre $422.13 y $927.87.

Imágenes y ejemplo tomados de Levin, et. al (2004)

¿Qué son los grados de libertad?

Levin et al (2004:297) define los grados de libertad como el número de valores que se pueden escoger libremente.

Ejemplos:

Suponga que se tienen dos valores de muestra a y b y tienen una media de 18. Suponga que el valor de a = 10.

$$\frac{a+b}{2} = 18$$

Si $a = 10$

entonces $\frac{10+b}{2} = 18$

de modo que $10 + b = 36$

por tanto $b = 26$

Como se tiene el valor de la media muestral, entonces se tiene un grado de libertad.

Suponga que se tienen siete valores de muestra y la media muestral es 16.

$$\frac{a+b+c+d+e+f+g}{7} = 16$$

Los valores que se pueden especificar libremente es: 7 – 1 = 6

Imágenes y ejemplo tomados de Levin, et. al (2004)

Consideraciones de la aplicación de la regresión lineal

Levin, et. al (2004) mencionan algunos errores que es común cometer cuando se utilizan los métodos de correlación y regresión.

- **Extrapolación más allá del rango de los datos observados**

  Un error común es suponer que la línea de estimación puede aplicarse en cualquier intervalo de valores. Una ecuación de estimación es válida sólo para el mismo rango dentro del cual se tomó la muestra inicialmente.

- **Causa y efecto**

  Los análisis de regresión y correlación no pueden, de ninguna manera, determinar la causa y el efecto.
  Pueden existir más variables que produzcan la variación de los datos.

- **Uso de tendencias anteriores para estimar tendencias futuras**

  Se debe reevaluar los datos históricos que se usarán para estimar la ecuación de regresión. Las condiciones pueden cambiar y violar una o más de las suposiciones de las cuales depende el análisis de regresión.

- **Interpretación errónea de los coeficientes de correlación y determinación**

  Si $r = 0.6$ y $r^2 = 0.6 \times 0.6 = 0.36$, entonces significa que el 36% de la variación total se explica por la recta de regresión, r^2 es una medida sólo de qué tan bien una variable describe a la otra, no de qué tanto cambio en una variable es originado por la otra variable.

- **Descubrimiento de relaciones cuando no existen**

  Al aplicar el análisis de regresión, algunas veces se encuentra una relación entre dos variables que no están vinculadas.

GRÁFICAS DE LOS RESIDUALES CORRESPONDIENTES A TRES ESTUDIOS DE REGRESIÓN

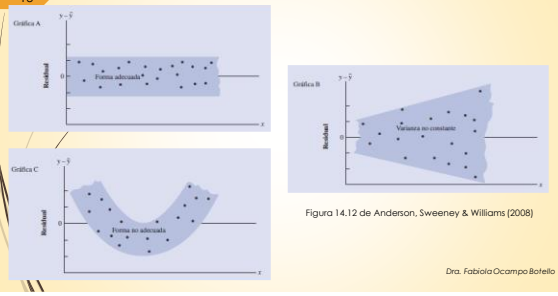

Figura 14.12 de Anderson, Sweeney & Williams (2008)

Referencias bibliográficas

Anderson, Sweeney & Williams. (2008). Estadística para administración y economía, 10ª edición. Cengage Learning.
Bennet, Briggs & Triola (2011). Razonamiento estadístico. Pearson. México.
Carollo Limeres, M. Carmen. (2012). Regresión lineal simple. Apuntes del departamento de estadística e investigación operativa . Disponible en: http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_50140116_Regr_%20simple_2011_12.pdf
Kerlinger, F. N. & Lee, H. B. (2002). Investigación del comportamiento. Métodos de investigación en ciencias sociales. 4ª ed. México: Mc. Graw Hill.
Levin, Rubín, Balderas, Del Valle y Gómez. (2004). Estadística para administración y economía. Séptima Edición. Prentice-Hall.
Mason, Lind & Marshal. (2000). Estadística para administración y economía. Alfaomega. 10ª edición.