Variación e Intervalos de Predicción

Utilizando datos apareados (x, y) describiremos la variación que puede explicarse por la correlación lineal entre *x* y *y* y la variación que no puede explicarse. Luego, procederemos a considerar un método para construir un intervalo de predicción, que es un estimado del intervalo de un valor predicho de *y* (los estimados de intervalos de parámetros se conocen como intervalos de confianza, en tanto que los estimados de intervalos de variables suelen denominarse intervalos de predicción).

Variación

Suponga que tenemos un conjunto de datos apareados que contienen el punto muestral (x, y), que $\hat{y}$ es el valor predicho de *y* (obtenido por medio de la ecuación de regresión), y que la media de los valores y muestrales es $\bar{y}$.

La **desviación total** de (x, y) es la distancia vertical $y - \bar{y}$, que es la distancia entre el punto (x, y) y la recta horizontal que pasa por la media muestral $\bar{y}$.

La **desviación explicada** es la distancia vertical $\hat{y} - \bar{y}$ que es la distancia entre el valor predicho *y* y la recta horizontal que pasa por la media muestral $\bar{y}$.

La **desviación sin explicar** es la distancia vertical $y - \hat{y}$ que es la distancia vertical entre el punto (x, y) y la recta de regresión. (La distancia también se conoce como residual).

$$Desviación\ total\ =\ desviación\ explicada\ +\ desviación\ sin\ explica\square$$

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

Variación e Intervalos de Predicción

Esta última expresión implica desviaciones a partir de la media y se aplica a cualquier punto (x, y) particular. Si sumamos los cuadrados de las desviaciones utilizando todos los puntos (x, y), obtenemos cantidades de *variación*. La *variación total* se expresa como la suma de los cuadrados de los valores de desviación totales, la *variación explicada* es la suma de los cuadrados de los valores de desviación explicados, y la *variación sin explicar* es la suma de los cuadrados de los valores de desviación sin explicar.

$$Variación\ total = variacion\ explicada + variación\ \text{sin}\ explicar$$

O bien

$$\sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2$$

El **coeficiente de determinación** es la cantidad de variación en *y* que está explicada por la recta de regresión. Se calcula como

$$r^2 = \frac{variación\ explicada}{variación\ total}$$

Intervalos de predicción

La creación de un intervalo de predicción requiere una medida de la dispersión de los puntos muestrales alrededor de la recta de regresión.

El **error estándar** del estimado es una medida colectiva de la dispersión de los puntos muestrales alrededor de la recta de regresión, y se define de manera formal como sigue:

Variación e Intervalos de Predicción

Definición

El error estándar del estimado, denotado por s_e, es una medida de las diferencias (o distancias) entre los valores muestrales observados de *y* y los valores predichos que se obtienen por medio de la ecuación de regresión. Está dado por

$$s_e = \sqrt{\frac{\sum(y - \hat{y})^2}{n - 2}}$$

Donde $\hat{y}$ es el valor predicho de y.

O por medio de la siguiente formula equivalente:

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n - 2}}$$

Así como la desviación estándar es una medida de la desviación de los valores a partir de su media, el error estándar del estimado s_e es una medida de la desviación de los puntos de los datos muestrales a partir de su recta de regresión. La lógica que subyace en la división entre n-2 es similar a la lógica que condujo a la división entre n-1 para la desviación estándar ordinaria. Es importante señalar que valores relativamente pequeños reflejan puntos que están cercanos a la recta de regresión y los valores relativamente grandes se presentan cuando hay puntos que se alejan de la recta de regresión.

Variación e Intervalos de Predicción

Intervalo de predicción para una *y* individual

Dado el valor fijo x_0, el intervalo de predicción para una *y* individual es

$$\hat{y} - E < y < \hat{y} + E$$

donde el margen de error E es

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}$$

Y x_0 representa el valor dado de x, $t_{\alpha/2}$ tiene n-2 grados de libertad, y s_e se calcula a partir de la fórmula anterior

$$s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n - 2}}$$

Referencia:
Triola, M., (2013). Estadística. Decimoprimera edición. Pearson educación. México
Apuntes de clase Estadística 2 FCFM Rivera Rosales Elsa Edith