

6 Lecture 6:Feb 1

Last time

- SLR in Matrix Form

Today

- Simple correlation
- The statistical model of the SLR (JF chapter 6)

Simple correlation

Having calculated the least squares line, it is of interest to determine how closely the line fits the scatter of points. There are many ways of answering it. The standard deviation of the residuals, S_E , often called the *standard error of the regression* or the *residue standard error*, provides one sort of answer. Because of estimation considerations, the variance of the residuals is defined using *degrees of freedom* $n - 2$:

$$S_\epsilon^2 = \frac{\sum \epsilon_i^2}{n - 2}.$$

The residual standard error is,

$$S_\epsilon = \sqrt{\frac{\sum \epsilon_i^2}{n - 2}}$$

For the Davis's data, the sum of squared residuals is $\sum \epsilon_i^2 = 418.87$, and thus the standard error of the regression is

$$S_\epsilon = \sqrt{\frac{418.87}{101 - 2}} = 2.0569\text{kg}.$$

On average, using the least-squares regression line to predict measured weight from reported weight results in an error of about 2 kg.

Sum of squares:

- Total sum of squares (TSS) for Y: $\text{TSS} = \sum (y_i - \bar{y})^2$
- Residual sum of squares (RSS): $\text{RSS} = \sum (y_i - \hat{y}_i)^2$
- regression sum of squares (RegSS): $\text{RegSS} = \text{TSS} - \text{RSS} = \sum (\hat{y}_i - \bar{y})^2$
- $\text{RegSS} + \text{RSS} = \text{TSS}$

Sample correlation coefficient

Definition: The sample correlation coefficient r_{xy} of the paired data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ is defined by

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y}) / (n - 1)}{\sqrt{\sum (x_i - \bar{x})^2 / (n - 1) \times \sum (y_i - \bar{y})^2 / (n - 1)}} = \frac{s_{xy}}{s_x s_y}$$

s_{xy} is called the sample covariance of x and y :

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$s_x = \sqrt{\sum (x_i - \bar{x})^2 / (n - 1)}$ and $s_y = \sqrt{\sum (y_i - \bar{y})^2 / (n - 1)}$ are, respectively, the sample standard deviations of X and Y .

Some properties of r_{xy} :

- r_{xy} is a measure of the linear association between x and y in a dataset.
- correlation coefficients are always between -1 and 1 :

$$-1 \leq r_{xy} \leq 1$$

- The closer r_{xy} is to 1 , the stronger the positive linear association between x and y
- The closer r_{xy} is to -1 , the stronger the negative linear association between x and y
- The bigger $|r_{xy}|$, the stronger the linear association
- If $|r_{xy}| = 1$, then x and y are said to be perfectly correlated.
- $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x}$

R-square

The ratio of RegSS to TSS is called the *coefficient of determination*, or sometimes, simply “r-square”. it represents the proportion of variation observed in the response variable y which can be “explained” by its linear association with x .

- In simple linear regression, “r-square” is in fact equal to r_{xy}^2 . (But this isn’t the case in multiple regression.)
- It is also equal to the squared correlation between y_i and \hat{y}_i . (This is the case in multiple regression.)

For Davis’s regression of measured on reported weight:

$$\text{TSS} = 4753.8$$

$$\text{RSS} = 418.87$$

$$\text{RegSS} = 4334.9$$

Thus,

$$r^2 = \frac{4334.9}{4753.8} = 1 - \frac{418.87}{4753.8} = 0.9119$$