Chapter 3, Problem 7

Samvel Khalatyan

Feb 27, 2014

Proposition 1. *Prove that in case of simple linear regression:*

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

the R^2 is equal to correlation between X and Y squared, e.g.:

$$R^2 = corr^2(x, y) \tag{2}$$

We'll be using following definitions to prove above proposition.

Definition 1.

$$R^2 = \frac{TSS - RSS}{TSS} \tag{3}$$

Definition 2.

$$TSS = \sum (y_i - \bar{y})^2 \tag{4}$$

Definition 3.

$$RSS = \sum (y_i - \hat{y}_i)^2 \tag{5}$$

Definition 4.

$$corr(x, y) = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \tag{6a}$$

$$\sigma_x^2 = \sum (x_i - \bar{x})^2 \tag{6b}$$

$$\sigma_y^2 = \sum (y_i - \bar{y})^2 \tag{6c}$$

Proof. Substitute 4 and 5 into 3:

$$R^2 = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{7}$$

Lets work on the numerator:

$$\begin{aligned} A &= \sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2 && \text{(8a)} \\ &= \sum \left[(\cancel{y_i} - \bar{y}) - (\cancel{y_i} - \hat{y}_i)\right]\left[(y_i - \bar{y}) + (y_i - \hat{y}_i)\right] && \text{(8b)} \\ &= \sum (\hat{y}_i - \bar{y})(2y_i - \bar{y} - \hat{y}_i) && \text{(8c)} \end{aligned}$$

Recall that:

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} && \text{(9a)} \\ \hat{\beta}_1 &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_j - \bar{x})^2} && \text{(9b)} \end{aligned}$$

Substitue 9a into the expression for $\hat{y}_i$:

$$\begin{aligned} \hat{y}_i &= \hat{\beta}_0 + \hat{\beta}_1 x_i && \text{(10)} \\ &= \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i && \text{(11)} \\ &= \bar{y} + \hat{\beta}_1 (x_i - \bar{x}) && \text{(12)} \end{aligned}$$

Lets analyze two terms from 8c:

$$\begin{aligned} \hat{y}_i - \bar{y} &= \hat{\beta}_1 (x_i - \bar{x}) && \text{(13a)} \\ 2y_i - \bar{y} - \hat{y}_i &= 2y_i - \bar{y} - \bar{y} - \hat{\beta}_1 (x_i - \bar{x}) && \text{(13b)} \\ &= 2(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) && \text{(13c)} \end{aligned}$$

Substitue these expreesions back into 8c:

$$\begin{aligned} A &= \sum \hat{\beta}_1 (x_i - \hat{x}) \left[2(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})\right] && \text{(14)} \\ &= \hat{\beta}_1 \sum (x_i - \bar{x}) \left[2(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x})\right] && \text{(15)} \\ &= \hat{\beta}_1 \left[2 \sum (x_i - \bar{x})(y_i - \bar{y}) - \hat{\beta}_1 \sum (x_i - \bar{x})^2\right] && \text{(16)} \end{aligned}$$

Recall formula for $\hat{\beta}_1$. Obviously the very last term in 16 is nothing but:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) \tag{17}$$

Therefore we get:

$$\begin{aligned} A &= \hat{\beta}_1 \sum (x_i - \bar{x})(y_i - \bar{y}) && \text{(18)} \\ &= \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum (x_j - \bar{x})^2} && \text{(19)} \end{aligned}$$

Recall that A was a denominator of R^2. Thus:

$$R^2 = \frac{\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum (x_j - \bar{x})^2 \sum (y_k - \bar{y})^2} \tag{20}$$

Substitue eqs. (6a) to (6c) into 20:

$$R^2 = corr^2(x, y) \tag{21}$$

$\square$