Homework - 4. A20545853 MATH 564.

4. Given, simple linear regression.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, \ldots, n.$$

To show that $R^2 = r^2$, where r is the correlation between x and y

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Let us consider, a estimate linear regression

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon_i \qquad i = 1, \ldots, n$$

we know that, $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$$R^2 = \frac{SS_{Reg}}{SS_T} = 1 - \frac{RSS}{SS_T}.$$

where,

- Total sum of squares, $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- Regression sum of squares, $SS_{Reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Residual sum of squares, $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Let's compute R^2;

$$R^2 = \frac{SS_{Reg}}{SS_T}$$

$$= \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$= \frac{\sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$= \frac{\sum_{i=1}^{n} (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (\because \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x})$$

$$= \frac{\hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

we know that, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$R^2 = \hat{\beta}_1^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$= \left[\frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]^2 \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$R^2 = \frac{\left(\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}$$

we can see that, $r = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$

Therfore, we can show that,

$$R^2 = r^2$$

where r is the correlation between x and y

given by:

5. To show when moving from a simple linear regression model with one predictor to a Multilinear regression with two predictors the R^2 (coefficient of determination) with either increase or stay the same. ~~Explain~~ :

And explain your resoning using the definition of R2 and the effect of adding a predictor on the Residual sum of squares. (RSS).

(i) Let us consider two models :

Model 1 : Simiple Linear regression model with one predictor

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Model 2 : Multi linear regression model with ~~and~~ two predictors

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

Here, $$R_1^2 = 1 - \frac{RSS_1}{TSS}$$ (unexplained variance)

R_1^2 is for model 1.

$$R_2^2 = 1 - \frac{RSS_2}{TSS_2}$$

R_2^2 for model 2.

→ In model 1, RSS_1 represents the unexplained variance in y using only one predictor (x_i)

→ In model 2, here we have one more predictor, So this model has more flexibility. This results, RSS_2 can either decrease or remain the same. . But cannot increase.

$$RSS_1 \geq RSS_2.$$

~~adding~~ Dividing both sides by TSS.

$$\frac{RSS_1}{TSS} \geq \frac{RSS_2}{TSS}$$

by substracting, the above with 1, we get,

$$1 - \frac{RSS_1}{TSS} \leq 1 - \frac{RSS_2}{TSS}$$

$$R_1^2 \leq R_2^2$$

There fore, by adding a new predictor to the model will remain the same or will increase but will not decrease.