Q. Suppose we have data on (Y_i, X_{2i}, X_{3i}), $i = 1, 2, \cdots, n$.
Consider the following models on Y:

$$M_1 : Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

$$M_2 : Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

How can we assess which model is a better fit to the given data

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum e_i^2 \Rightarrow TSS = ESS + RSS$$

↳ ESS [Sum of squared terms].

Approach 1:

Fit M_1 ⇒ Obtain R_1^2
Fit M_2 ⇒ Obtain R_2^2
} → Compare them, then the one with higher R^2 is a better fit.

$R^2 = \frac{ESS}{TSS}$. and as M_2 has 2 explanatory variables vs M_1 having 1 explanatory variable, $ESS_2 \geq ESS_1$.

∴ In General, $R_2^2 \geq R_1^2$ irrespective of the consideration whether adding X_3 as an explanatory variable is significant or not.

∴ R^2 cannot be used to compare Goodness of Fit of Models M_1 and M_2.

$$M_1' : Y_i = \beta_1 + \beta_2 X_{2i} + u_i$$

$$M_2' : Y_i = \gamma_1 + \gamma_2 X_{3i} + u_i$$

(*)(*)

Note:

(i) For comparing the Goodness of Fit of 2 Models with the same no. of explanatory variables use R^2 [Eg: compare

(i) For comparing the Goodness of Fit of 2 Models with the same no. of explanatory variables use R^2 [Eg: compare M_1' and M_2']

(ii) For comparing the Goodness of Fit of 2 Models with different no. of explanatory variables use "Adjusted R^2".

Define: Adjusted R^2 $(\bar{R}^2) = 1 - \left[\frac{(RSS/(n-k))}{TSS/(n-1)}\right]$ [n = Sample size, k = No. of parameters]

Note: $k \uparrow \Rightarrow ESS \uparrow \Rightarrow RSS \downarrow \Rightarrow (n-k) \downarrow$

$\therefore$ $RSS \downarrow \Rightarrow \left(\frac{RSS}{n-k}\right) \downarrow$ and $(n-k) \downarrow \Rightarrow \left(\frac{RSS}{n-k}\right) \uparrow$

$\hookrightarrow \bar{R}^2 \uparrow$ $\hookrightarrow \bar{R}^2 \downarrow$

$\therefore$ Increase in 'k' will not always lead to increase in $\bar{R}^2$. Hence $\bar{R}^2$ addresses the problem present in R^2.

$\therefore$ $\bar{R}^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)}$

$\Rightarrow$ $\bar{R}^2 = 1 - \frac{RSS}{TSS} \cdot \left(\frac{n-1}{n-k}\right)$

$\Rightarrow$ $(1-\bar{R}^2) = \frac{RSS}{TSS}\left(\frac{n-1}{n-k}\right)$

We know $R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$ $\Rightarrow$ $\frac{RSS}{TSS} = (1-R^2)$

$\Rightarrow$ $\boxed{(1-\bar{R}^2) = (1-R^2)\left(\frac{n-1}{n-k}\right)}$ $\Rightarrow$ Relationship b/w R^2 & $\bar{R}^2$

$\Rightarrow$ $(1-\bar{R}^2) = \frac{(1-R^2)}{\left(\frac{n-k}{n-1}\right)}$ > 0 > 0.

If $(1-R^2) > \left(\frac{n-k}{n-1}\right) \Rightarrow (1-\bar{R}^2) > 1 \Rightarrow \boxed{\bar{R}^2 < 0}$

$\therefore$ Adj R^2 can be -ve as well, whereas $0 \leq R^2 \leq 1$

and for comparing models for Goodness of Fit using Adj R^2, the model with higher Adj R^2 is a better fit to the data.

Criteria of "Good" Estimates

Recall: True Model: $Y_i = \alpha + \beta X_i + u_i$ → Random disturbance term

Estimated Model: $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$

$\hat{\alpha}$ = estimate of α, $\hat{\beta}$ = estimate of β.

(i) $\hat{\alpha}$ and $\hat{\beta}$ should be unbiased estimates of α, β, i.e $E(\hat{\alpha}) = \alpha$ and $E(\hat{\beta}) = \beta$.

(ii) $\hat{\alpha}$ and $\hat{\beta}$ should have low variances. (imply high reliability of the estimates)

(iii) $\hat{\alpha}$ and $\hat{\beta}$ should be consistent estimators of α, β.

$\therefore$ as $n \to \infty$, $\hat{\alpha} \xrightarrow{P} \alpha$ and $\hat{\beta} \xrightarrow{P} \beta$.

Assumptions on u_i :-

i) $E(u_i) = 0$ [Zero-mean assumption].

ii) $Var(u_i) = \sigma^2$ [Homoskedasticity].

iii) $Cov(u_i, u_j) = 0 \quad \forall\, i \neq j$ [No autocorrelation].

iv) $Cov(X, u) = 0$ [explanatory variable is uncorrelated with the random disturbance term].

(ii) $\text{Var}(u_i) = \sigma^2$

By defn: $E(u_i^2) - \left[\overbrace{E(u_i)}^{=0}\right]^2 = \sigma^2 \Rightarrow E(u_i^2) = \sigma^2$

(iii) $\text{cov}(u_i, u_j) = 0$.

By defn: $E(u_i u_j) - \overbrace{E(u_i)}^{=0}\,\overbrace{E(u_j)}^{=0} = 0 \Rightarrow E(u_i u_j) = 0$.

Summarizing:

True Model: $Y_i = \alpha + \beta X_i + u_i$

$E(u_i) = 0$, $E(u_i^2) = 0 \;\forall\, i$, $E(u_i u_j) = 0 \;\forall\, i \neq j$

Estimated Model (OLS): $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$

where $\hat{\beta} = \dfrac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$, $\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$