Tutorial 7

1. A student stated: "Adding predictor variables to a regression model can never reduce R^2, so we should include all available predictor variables in the model." Comment.

   Bigger R^2, means the fitting is better. Better fitting does not imply better model.

2. For a model with X_1, X_2, X_3, X_4 predictors, we have $n = 30$ and

$$SSE(X_1) = 161.081, SSE(X_2) = 195.846, SSE(X_3) = 56.432, SSE(X_4) = 225.584$$

$$SSE(X_1, X_2) = 146.635, SSE(X_1, X_3) = 16.579, SSE(X_1, X_4) = 161.044,$$

$$SSE(X_2, X_3) = 45.660, SSE(X_2, X_4) = 195.403, SSE(X_3, X_4) = 56.431$$

$$SSE(X_1, X_2, X_3) = 12.436, SSE(X_1, X_2, X_4) = 146.604, SSE(X_1, X_3, X_4) = 16.383,$$

$$SSE(X_2, X_3, X_4) = 45.656, SSE(X_1, X_2, X_3, X_4) = 12.246, SST = 226.189$$

   (a) find $SSR(X_1, X_2|X_3, X_4)$

   (b) In model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, test $H_0 : \beta_1 = \beta_2 = 0$ with $\alpha = 0.05$.

   (c) Find the largest model in which every predictor variable is not significant at $\alpha = 0.05$.

(a)

$$\begin{aligned} SSR(X_1, X_2|X_3, X_4) &= SSR(X_1, X_2, X_3, X_4) - SSR(X_3, X_4) \\ &= SSE(X_3, X_4) - SSE(X_1, X_2X_3, X_4) \\ &= 56.431 - 12.246 = 44.1850 \end{aligned}$$

(b)

$$\begin{aligned} F^* &= \frac{SSR(X_1, X_2|X_3)/2}{SSE(X_1, X_2, X_3)/(n-4)} = \frac{(56.432 - 12.436)/2}{12.436/(30-4)} \\ &= 45.9913 > F(0.95, 2, 36) = 3.27 \end{aligned}$$

reject H_0.

(c) (1) We try to introduce X_3 because $SSE(X_3)$ is the smallest among $SSE(X_1), SSE(X_2), SSE(X_3)$. By F-test

$$F^* = \frac{SSR(X_1)/1}{SSE(X_1)/(n-2)} = \frac{(226.189 - 56.432)/1}{56.432/28} = 84.2287 > F(0.95, 1, 28) = 4.1960$$

(Thus, X_3 needs to be introduced)

(2) We try to introduce X_1 because $SSE(X_3, X_1)$ is the smallest among $SSE(X_2, X_3), SSE(X_3, X_4)$. By F-test

$$F^* = \frac{SSR(X_1|X_3)/1}{SSE(X_1, X_3)/(n-3)} = \frac{(56.432 - 16.579)/1}{16.579/27} = 64.9033 > F(0.95, 1, 27) = 4.22$$

(Thus, X_1 needs to be introduced)

(3) We try to introduce X_2 because $SSE(X_3, X_2, X_1)$ is the smaller than $SSE(X_1, X_3, X_4)$. By F-test

$$F^* = \frac{SSR(X_2|X_1, X_3)/1}{SSE(X_1, X_2, X_3)/(n-4)} = \frac{(16.579 - 12.436)/1}{12.436/26} = 8.6618 > F(0.95, 1, 26) = 4.23$$

(Thus, X_2 needs to be introduced)

(3) We try to introduce X_4. By F-test

$$F^* = \frac{SSR(X_4|X_1, X_2, X_3)/1}{SSE(X_1, X_2, X_3, X_4)/(n-5)} = \frac{(12.436 - 12.246)/1}{12.246/25} = 0.3879 < F(0.95, 1, 25) = 4.25$$

(Thus, X_4 should not be introduced)

(4) our final model is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

3. State the number of degrees of freedom that are associated with each of the following extra sums of squares: (1) $SSR(X_1 \mid X_2)$; (2) $SSR(X_2 \mid X_1, X_3)$ (3) $SSR(X_1, X_2 \mid X_3, X_4)$; (4) $SSR(X_1, X_2, X_3 \mid X_4, X_5)$

(1): 1, (2): 1, (2): 2, (4): 3

4. Refer to **Patient satisfaction** date (see the **Example** of Chapter 2)

a. Obtain the analysis of variance table that decomposes the regression sum of squares into extra sums of squares associated with X_2; with X_1, given X_2; and with X_3, given X_2 and X_1.

b. Test whether X_3 can be dropped from the regression model given that X_1 and X_2 are retained. Use the F^* test statistic and level of significance 0.025. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

a.

$$SSR(X_2) = 4860.3$$

$$SSR(X_1|X_2) = 3896.0$$

$$SSR(X_3|X_1, X_2) = 364.2$$

b.

$$F = (364.2/1)/(4248.8/42) = 3.600169$$

Since $3.600169 < F(0.975, 1, 42) = 5.4239$, X_3 can be dropped. The p-value is 0.06468

See **R-code**

5. Refer to **Patient satisfaction** data above, test whether $\beta_1 = -1.0$ and $\beta_2 = 0$; use $\alpha = 0.025$. State the alternatives, full and reduced models, decision rule and conclusion.

Full model

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Reduced model

$$y = \beta_0 - X_1 + \beta_3 X_3 + \varepsilon$$

SSE(F) = 4248.8, SSE(R) = 4427.7

$$F^* = \frac{(SSE(R) - SSE(F))/2}{SSE(F)/42} = \frac{(4427.7 - 4248.8)/2}{4248.8/42} = 0.8842 < F(0.975, 2, 42) = 4.0510$$

Thus, $\beta_1 = -1.0$ and $\beta_2 = 0$ can be accepted.

See **R-code**

6. Show that:

a. $SSR(X_1, X_2, X_3, X_4) = SSR(X_1) + SSR(X_2, X_3 \mid X_1) + SSR(X_4 \mid X_1, X_2, X_3)$.

b. $SSR(X_1, X_2, X_3, X_4) = SSR(X_2, X_3) + SSR(X_1 \mid X_2, X_3) + SSR(X_4 \mid X_1, X_2, X_3)$.

a.

$$\begin{aligned} & SSR(X_1, X_2, X_3, X_4) \\ = \;& SSR(X_1) + [SSR(X_2, X_3, X_1) - SSR(X_2, X_3)] \\ & \quad + [SSR(X_4, X_2, X_3, X_1) - SSR(X_2, X_3, X_1)] \\ = \;& SSR(X_1) + SSR(X_2, X_3 \mid X_1) + SSR(X_4 \mid X_1, X_2, X_3). \end{aligned}$$

b.

$$\begin{aligned} & SSR(X_1, X_2, X_3, X_4) \\ = \;& SSR(X_2, X_3) + [SSR(X_1, X_2, X_3) - SSR(X_2, X_3)] \\ & \quad + [SSR(X_4, X_1, X_2, X_3) - SSR(X_1, X_2, X_3)] \\ = \;& SSR(X_2, X_3) + SSR(X_1 \mid X_2, X_3) + SSR(X_4 \mid X_1, X_2, X_3). \end{aligned}$$

7. The following regression model is being considered in a water resources study: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \beta_4 \sqrt{X_{i3}} + \varepsilon_i$. State the reduced model for testing whether or not: (1) $\beta_3 = \beta_4 = 0$, (2) $\beta_3 = 0$, (3) $\beta_1 = \beta_2 = 5$, (4) $\beta_4 = 7$.

(1)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

(2)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_4 \sqrt{X_{i3}} + \varepsilon_i$$

(3)

$$Y_i = \beta_0 + 5X_{i1} + 5X_{i2} + \beta_3 X_{i1} X_{i2} + \beta_4 \sqrt{X_{i3}} + \varepsilon_i$$

(4)

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + 7\sqrt{X_{i3}} + \varepsilon_i$$