Homework 6

Roly Vicaría
STAT501 Fall 2015

October 4, 2015

0.0.1 Question 1

a) $SSR(X_3|X_1) = SSE(X_1) - SSE(X_1, X_3) = 510 - 500 = 10$

b) The quantity in part (a) represents the reduction in the error sum of squares that resulted when adding X_3 to the model when X_1 was already in the model.

c) $SSR(X_1|X_2, X_3) = SSE(X_2, X_3) - SSE(X_1, X_2, X_3) = 640 - 330 = 310$

d) The quantity in part (c) represents the reduction in the error sum of squares that resulted when adding X_1 to the model when X_2 and X_3 were already in the model.

e) $Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i$

f) $F^* = \frac{SSE(X_1) - SSE(X_1, X_2, X_3)}{(n-2)-(n-4)} \div \frac{SSE(X_1, X_2, X_3)}{n-4} = \frac{510-330}{(70-2)-(70-4)} \div \frac{330}{70-4} = 18$

g) $R^2_{Y,2|1} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)} = \frac{510-450}{510} = 0.1176$

h) The value in part (g) represents the marginal contribution of X_2 given that X_1 is already in the model. It shows that about 11.7% of the variation in Y is reduced after adding X_2 to the model.

0.0.2 Question 2

a)

Source	df	Seq SS	Adj SS	F-statistic based on Adj SS	p-value based on Adj SS
Regression	3	100.866	100.866	35.14	0.000
X_1	1	67.444	33.031	34.52	0.000
X_2	1	3.883	0.160	0.167	0.684
X_3	1	29.539	29.539	30.88	0.000
Error	93	88.976	88.976	—	—
Total	96	189.842	189.842	—	—

Coefficients

Term	Coef	SE coef	t-statistic	p-value
Constant	0.58	1.24	0.45	0.652
X_1	0.34	0.058	5.88	0.000
X_2	-0.01	0.0245	-0.408	0.342
X_3	0.06	0.0103	5.56	0.000

b) Calculate $SSR(X_3|X_1)$:

$$\begin{aligned} SSR(X_1, X_2, X_3) &= SSR(X_1) + SSR(X_3|X_1) + SSR(X_2|X_1, X_3) \\ SSR(X_3|X_1) &= SSR(X_1, X_2, X_3) - SSR(X_1) - SSR(X_2|X_1, X_3) \\ &= 100.866 - 67.444 - 0.160 \\ &= 33.262 \end{aligned}$$

c) One way to think of the difference between the sequential sum of squares and the adjusted sum of squares is that in the sequential sum of squares, the order matters, whereas in the adjusted sum of squares, order doesn't matter. In other words, when looking at the ANOVA table, the sequential sum of squares can be thought of as a snowball that grows as you go down the list of predictors. Each line computes the sum of square of that given predictor assuming the previous predictors are already in the model, but not the ones that follow. In the case of X_2, the sequential sum of square only assumed the presence of X_1. The adjusted sum of square looks at each predictor assuming every other predictor is already in the model. In the case of X_2, we assumed that X_1 and X_3 were already present.

d)

$$\begin{aligned} F^* &= \frac{SSR(X_2, X_3|X_1)}{p-q} \div \frac{SSE(X_1, X_2, X_3)}{n-p} \\ &= \frac{SSR(X_1, X_2, X_3) - SSR(X_1)}{4-2} \div \frac{SSE(X_1, X_2, X_3)}{97-4} \\ &= \frac{100.866 - 67.444}{2} \div \frac{88.976}{93} \\ &= 17.467 \end{aligned}$$

e) $R^2_{Y,2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{3.883}{122.398} = 0.0317$

f) The value in part (e) represents the marginal contribution of X_2 given that X_1 is already in the model. It shows that about 3.2% of the variation in Y is reduced after adding X_2 to the model.

0.0.3 Question 3

a) Analysis of Variance

Source	DF	Seq SS
Regression	3	2176606
X1	1	136366
X3	1	2033565
X2	1	6675
Error	48	985530
Total	51	3162136

b) $H_0 : \beta_2 = 0$
$H_a : \beta_2 \neq 0$

Full model SSE: 985530
Reduced model SSE: 992204

$$\begin{aligned} F^* &= \frac{SSR(X_2|X_1, X_3)}{1} \div \frac{SSE(X_1, X_2, X_2)}{52 - 4} \\ &= \frac{6675}{1} \div \frac{985530}{48} \\ &= 0.3251 \end{aligned}$$

$p - value = P(X > 0.3251) = 0.571$

Because the p-value is greater than $\alpha = 0.05$, we conclude that we can drop X_2 from the model.

c) This is the same hypothesis as part (b). The t-statistic is $-13.2/23.1 \approx -0.57$ with a p-value of 0.571. The conclusion is the same as above, that we fail to reject the null hypothesis.

d)

$$\begin{aligned} F^* &= \frac{SSR(X_2, X_3|X_1)}{2} \div \frac{SSE(X_1, X_2, X_3)}{48} \\ &= \frac{SSR(X_1, X_2, X_3) - SSR(X_1)}{2} \div \frac{SSE(X_1, X_2, X_3)}{48} \\ &= \frac{2176606 - 136366}{2} \div \frac{985530}{48} \\ &= 49.685 \end{aligned}$$

p-value is 0.000. Therefore we can reject the null hypothesis and conclude that at least one β_i (for $i = 2, 3$) is not equal to 0.