STAT 501 – Homework 6 Solutions

1. (**8x5=40 points**) Suppose that X_1, X_2, and X_3 are potential predictor variables in a model used to predict Y. Use the given SSE values in the following table to answer the questions below.

$SSE(X_1) = 510$	$SSE(X_1, X_2) = 450$	$SSE(X_1, X_2, X_3) = 330$
$SSE(X_2) = 905$	$SSE(X_1, X_3) = 400$	
$SSE(X_3) = 720$	$SSE(X_2, X_3) = 640$	

The notation $SSE(X_1, X_2)$ means the sum of squares for a multiple regression model that includes X_1 and X_2 as predictors (and does not include X_3).

a) Calculate $SSR(X_3|X_1)$. Show your work.

$SSR(X_3|X_1) = SSE(X_1) - SSE(X_1, X_3) = 510 - 400 = 110.$

b) Explain in words what is measured by the quantity calculated in the previous part.

This is the reduction in SSE that would come from adding X_3 to a model that already included X_1.

c) Calculate $SSR(X_1|X_2, X_3)$. Show your work.

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

d) Explain in words what is measured by the quantity calculated in the previous part.

This is the reduction in SSE that would come from adding X_1 to a model that already included X_2 and X_3.

e) Consider the "full" model, $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon_i$. What is the "reduced" model associated with the null hypothesis H_0: $\beta_2 = \beta_3 = 0$?

The reduced model is $Y_i = \beta_0 + \beta_1 X_{i,1} + \varepsilon_i$.

f) Suppose that n = 70. Calculate the value of an F-statistic for testing H_0: $\beta_2 = \beta_3 = 0$ for the model $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon_i$. It is not necessary to carry out the test – just calculate the value of F. Show your work. (Hint: Use the general linear F-test.)

$$\mathbf{F} = \frac{\frac{SSE(X1)-SSE(X1,X2,X3)}{(n-2)-(n-4)}}{\frac{SSE(X1,X2,X3)}{n-4}} = \frac{\frac{510-330}{2}}{\frac{330}{66}} = \mathbf{18}.$$

g) Calculate the value of the coefficient of partial determination $R^2_{Y,2|1}$.

$$\mathbf{R^2_{Y,2|1}} = \frac{SSR(X2|X1)}{SSE(X1)} = \frac{SSE(X1)-SSE(X1,X2)}{SSE(X1)} = \frac{510-450}{510} = \mathbf{0.118}.$$

h) Write a sentence that interprets the value calculated in the previous part.

X_2 explains 11.8% of the variation in Y that cannot be explained by X_1.

2. **(10+6x5=40 points)**

a) Fill in the blanks in the following tables. The column labeled "Seq SS" represents "sequential sums of squares" (measures the reduction in the SS when a term is added to a model that contains only the terms before it), while the column labeled "Adj SS" represents "adjusted sums of squares" (measures the reduction in the SS for each term relative to a model that contains all of the remaining terms). *[Hint: The t-statistics in the Coefficients table assume all other predictors are included in the model, so if we square these we get the F-statistics in the Anova table based on Adjusted Sums of Squares.]*

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
X1	1	67.444	33.031	34.52	0.000
X2	1	3.883	***0.163 =0.17x88.976/93***	***0.17 = –0.41²***	***0.681***
X3	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
X1	0.34	0.058	5.88	0.000
X2	-0.01	0.0245	***–0.41 = –0.01/0.0245***	***0.681***
X3	0.06	0.0103	5.56	0.000

b) Calculate $SSR(X_3|X_1)$, that is the sequential sum of squares obtained by adding X_3 to a model already containing only the predictor X_1. Show your work.

$SSR(X_3|X_1) = SSR(X_1,X_3) - SSR(X_1) = SSR(X_1,X_2,X_3) - SSR(X_2|X_1,X_3) - SSR(X_1)$
$=100.866 - 0.163 - 67.444 = 33.259.$

c) Explain in words what is measured by the quantity calculated in the previous part.

$SSR(X_3|X_1)$ is the reduction in the error sum of squares when X_3 is added to the model in which X_1 is the only predictor. Alternatively, $SSR(X_3|X_1)$ is the increase in the regression sum of squares when X_3 is added to the model in which X_1 is the only predictor.

d) Discuss the conceptual difference between the sequential sum of squares (Seq SS) and adjusted sum of squares (Adj SS) in terms of the predictor X_2. For this data, what are the numerical values of these sums of squares for the predictor X_2?

For this data, the sequential sum of squares for $X_2 = SSR(X_2|X_1) = 3.883$ = increase in the regression sum of squares when X_2 is added to the model which has only X_1 as a predictor. Sequential sum of squares depends on the order the predictors are entered into the model. It is the portion of the regression sum of squares explained by a predictor, after accounting for the previously entered predictors. The adjusted sum of squares for $X_2 = SSR(X_2|X_1, X_3) = 0.163$ = increase in the regression sum of squares when X_2 is added to the model which already contains all other predictors, X_1 and X_3 in this case. Adjusted sum of squares does not depend on the order the predictors are added to the model.

e) Calculate the value of an F-statistic for testing $H_0: \beta_2 = \beta_3 = 0$ within the model $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \varepsilon_i$. It is not necessary to carry out the test – just calculate the value of F. Show your work.

$$\mathbf{F = \frac{SSR(X2,X3|X1)/2}{MSE} = \frac{(3.883+29.539)/2}{88.976/93} = \frac{16.711}{0.957} = 17.467.}$$

f) Calculate the value of the coefficient of partial determination $R^2_{Y,2|1}$.

$$\boldsymbol{R^2_{Y,2|1} = \frac{SSR(X2|X1)}{SSE(X1)} = \frac{3.883}{3.883+29.539+88.976} = 0.032.}$$

g) Write a sentence that interprets the value calculated in the previous part.

X_2 explains 3.2% of the variation in Y that cannot be explained by X_1.

3. **(6+7+5+2 = 20 points)** Consider the "GroceryRetailer" dataset. A large, national grocery retailer tracks productivity and costs of its facilities closely. Data were obtained from a single distribution center for a one-year period. Each data point for each variable represents one week of activity. The variables included are the number of cases shipped (X_1), the indirect costs of the total labor hours as a percentage (X_2), a qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X_3), and the total labor hours (Y).

   a) Obtain the ANOVA table that decomposes the regression sum of squares into extra sums of squares associated with X_1; X_3 given X_1; and with X_2 given X_1 and X_3. Give their values along with the associated dfs.

```
Regression Analysis: Y versus X1, X3, X2

Analysis of Variance
Source       DF   Seq SS   Seq MS  F-Value  P-Value
Regression    3  2176606   725535    35.34    0.000
  X1          1   136366   136366     6.64    0.013
  X3          1  2033565  2033565    99.04    0.000
  X2          1     6675     6675     0.33    0.571
Error        48   985530    20532
Total        51  3162136
```

$SSR(X_1) = 136{,}366$; $SSR(X_3 \mid X_1) = 2{,}033{,}565$; $SSR(X_2 \mid X_1, X_3) = 6{,}675$ with corresponding dfs: 1, 1, 1.

   b) Test whether X_2 can be dropped from the regression model given that X_1, and X_3 are retained. Use the F test statistic and $\alpha = 0.05$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

$H_0: \beta_2 = 0$, vs. $H_a: \beta_2 \neq 0$.
$SSR(X_2 \mid X_1, X_3) = 6{,}675$, $SSE(X_1, X_2, X_3) = 985{,}530$,

Test Statistic: $F = \frac{SSR(X2 \mid X1, X3)/1}{SSE/48} = \frac{6675}{985530/48} = 0.3251$.

Critical Value: $F(0.95; 1, 48) = 4.04265$.
Conclusion: Since $0.3251 < 4.04265$, we do not reject H0.
P-value = $P(F > 0.3251) = 0.571$.

c) Now test H_0: $\beta_2 = 0$ vs. H_a: $\beta_2 \neq 0$ in the model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_2 X_2$ using a t-test. Give the value of the test statistic, the p-value and conclusion.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	4150	196	21.22	0.000	
X1	0.000787	0.000365	2.16	0.036	1.01
X3	623.6	62.6	9.95	0.000	1.01
X2	-13.2	23.1	-0.57	0.571	1.02

Test statistic: t = - 0.57 with p-value = 0.571. So, we fail to reject H_0 in favor of H_a and conclude X_2 is not useful in this model.

d) From part (a), use sequential sum of squares to test H_0: $\beta_2 = \beta_3 = 0$ in the model $E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_2 X_2$. Give the test statistic, P-value and conclusion.

From part (a) we get,
SSR(X2,X3|X1) = SSR(X3|X1) + SSR(X2|X1,X3) = 2033565 + 6675 = 2040240**. So,**

Test Statistic: F = $\frac{\text{SSR(X1,X2|X3)/2}}{\text{MSE}} = \frac{2040240/2}{20532} = 49.68439509$.

P-value = P(F > 49.68439509) ≈ 0.000.
Conclusion: Since P-value is very small, we reject H_0.