STAT 420

Homework #10

(due Thursday, May 6, by 9:00 p.m.)

Please include your name, your NetID, and your section number at the top of the first page.

No credit will be given without supporting work.

Include a printout of the relevant code and output or plot.

1. Using the `sat` dataset from `faraway` library, fit a model with the total SAT score as the response and expend, salary, ratio, and takers as predictors. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant. Do not provide any plots about which you have nothing to say.

```
> library(faraway)
> data(sat)
> head(sat)
           expend ratio salary takers verbal math total
Alabama     4.405  17.2 31.144      8    491  538  1029
Alaska      8.963  17.6 47.951     47    445  489   934
Arizona     4.778  19.3 32.175     27    448  496   944
Arkansas    4.459  17.1 28.934      6    482  523  1005
California  4.992  24.0 41.078     45    417  485   902
Colorado    5.443  18.4 34.571     29    462  518   980
```

See http://cran.r-project.org/web/packages/faraway/faraway.pdf (p. 86) for more info if desired.

a) Check the constant variance assumption for the errors.

b) Check the normality assumption.

c) Check for large leverage points.

d) Check for outliers.

e) Check for influential points.

1.(continued) Using the `sat` dataset from `faraway` library, fit a model with the total SAT score as the response and expend, salary, ratio, and takers as predictors.

```
> library(faraway)
> data(sat)
> fit = lm(total ~ expend + ratio + salary + takers, data=sat)
```

f) Implement the Backward Elimination variable selection method to determine the "best" model. Use $\alpha_{crit} = 0.10$.

g) Implement the AIC Backward Elimination variable selection method to determine the "best" model.

h) Implement the AIC Forward Selection variable selection method to determine the "best" model.

i) Surprisingly, part (g) and part (h) produce different "best" models. If we use AIC, which one of these two models, part (g) or part (h), is preferred?

j) If we use Adjusted R^2, which one of these two models, part (g) or part (h), is preferred?

2. Recall Problem **3** from Homework #6:

Suppose a model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

was fit to $n = 34$ data points. The following results were obtained:

```
> summary( lm( y ~ x1 + x2 + x3 + x4 + x5 ) )$r.squared
[1]  0.40107
> summary( lm( y ~ x1 + x2 + x3 + x4 + x5 ) )$sigma
[1]  4
> summary( lm( y ~ x1 + x3 + x5 ) )$sigma^2
[1]  17.6
```

In part (b), we tested $H_0 : \beta_2 = \beta_4 = 0$ at a 5% level of significance.

SYY = 748. RSS_{Full} = 448. RSS_{Null} = 528.

c) Find the values of AIC for the Null and the Full models. Which model is preferred, Full or Null? *Justify your answer*.

d) Find the values of BIC for the Null and the Full models. Which model is preferred, Full or Null? *Justify your answer*.

e) Find the values of Adjusted R^2 for the Null and the Full models. Which model is preferred, Full or Null? *Justify your answer*.