STA 101 Spring 2018
Homework 3 - Due Wednesday, April 25^{th}

"Hand Written" Homework

These problems may be completed without the use of R (except for calculation of p-values).

1. Suppose we were trying to build a linear model for Y = selling price of a car, based on X_1 = current total distance of car (in Kilometers), and X_2 = type of fuel (CNG (Combustible Natural Gas), Diesel, Gasoline). Price is given in thousands of dollars, and X_1 is given in thousands of kilometers.
   A table with various models and their SSE's follow:

Y	SSE	Fit
1	18877.24	10.73
X1	12744.88	14.51+(-0.06)X_1
X2	18795.47	9.42+(1.87)$X_{2,D}$+(1.26)$X_{2,G}$
X1+X2	10819.38	17.64+(-0.07)X_1+(1.46)$X_{2,D}$+(-2.6)$X_{2,G}$

   There are $n = 1436$ rows total in this dataset. Use this table to answer the following questions:

   (a) State the null and alternative for testing if X_2 can be dropped from the full model (with both X_1, X_2).
   (b) Find the test-statistic for the hypothesis in (a), and the corresponding p-value.
   (c) State your conclusion for the hypothesis in (a) if $\alpha = 0.05$.
   (d) State the null and alternative for testing if X_1 can be dropped from the full model (with both X_1, X_2).
   (e) Find the test-statistic for the hypothesis in (d), and the corresponding p-value.
   (f) State your conclusion for the hypothesis in (d) if $\alpha = 0.05$.

2. Continue with problem 1.

   (a) Find what the proportion of reduction in error was when adding X_1 to the empty model (the model with no X's).
   (b) Find what the proportion of reduction in error was when adding X_2 to the empty model (the model with no X's).
   (c) Find and interpret $R^2\{X_1|X_2\}$
   (d) Find and interpret $R^2\{X_2|X_1\}$
   (e) Based on all the information from problem 1 and this problem, what would you chose to be your final model and why?

3. A hospital administrator wished to study the relation between patient satisfaction (Y) and X_1 = age in years of the patient, X_2 = severity of illness (an index score), and X_3 = anxiety level of the patient (an index score). For Y, X_2, X_3, higher values are associated with more satisfaction, more severity, and more anxiety respectively. A table with various models and their SSE's follow:

Model	SSE
Y˜1	13369.30
Y˜X1	5093.92
Y˜X2	8509.04
Y˜X3	7814.39
Y˜X1+X2	4613.00
Y˜X1+X3	4330.50
Y˜X2+X3	7106.39
Y˜X1+X2+X3	4248.84

   There are $n = 46$ total rows in this dataset. *Data Source: "Applied Linear Statistical Models", Kutner, Nachtsheim, Neter, & Li.*

   (a) State the null and alternative for testing if X_2 and X_3 can be dropped from the full model (with X_1, X_2, X_3).
   (b) Find the test-statistic for the hypothesis in (a), and the corresponding p-value.
   (c) Interpret your p-value from (b) in terms of the problem.
   (d) State your conclusion for the hypothesis in (a) if $\alpha = 0.01$.
   (e) Find and interpret $R^2\{X_2, X_3|X_1\}$.
   (f) Give a reason why we would not choose the model with the lowest value of SSE by default.

4. Continue with problem 3. Various model selection criteria are shown below:

	AIC	BIC	PRESS	R2adj
Y˜1	393.458	395.286	13970.098	0.000
Y˜X1	351.072	354.729	5569.562	0.610
Y˜X2	374.674	378.331	9254.489	0.349
Y˜X3	370.756	374.413	8451.432	0.402
Y˜X1+X2	348.510	353.996	5235.192	0.639
Y˜X1+X3	345.603	351.089	4902.751	0.661
Y˜X2+X3	368.387	373.873	8115.912	0.444
Y˜X1+X2+X3	346.727	354.042	5057.886	0.659

   (a) Pick the "best" model, using AIC as the criteria.
   (b) Pick the "best" model, using BIC as the criteria.
   (c) Pick the "best" model, using PRESS as the criteria.
   (d) Which model would you pick as your final model, and why?

5. Answer the following questions with TRUE or FALSE. It is also good practice to explain your answers, and the only way to get partial credit should your answer be incorrect.

   (a) We can compare a model to before using a transformation on X or Y to a model after using a transformation by using the corresponding values of SSE.
   (b) An "underfit" model has included unimportant X values.

(c) The "best" model depends on your overall goal for the model.

(d) As the number of predictor variables (X's) increase, $R^2 = \frac{SSTO-SSE}{SSTO}$ always increases.

R Problems

Note: You do not have to use R Markdown to turn in the homework, but the homework must be turned in in a reasonable format. The answers to the questions should be in the body of the homework, and the code used to obtain those answers should be in an appendix. There should be no code in the body of the homework. You can accomplish this in R, Word, LaTex, Google Docs, etc.

I. Online you will find the file `salary3.csv`. Among its columns, we are interested in (note, I have rearranged the columns):

Column 1: `sl`: The three month salary of the subject in dollars (Y).

Column 2: `yd`: The number of years since the subject earned their highest degree (X_1) (i.e., years of experience).

Column 3: `dg`: The highest degree earned (doctorate, masters) of the subject (X_2).

Column 4: `sx`: The gender of the subject (male, female) (X_3)

Column 5: `rk`: The rank of the subject (assistant, associate, full) (X_4)

Data Source: S. Weisberg (1985). Applied Linear Regression, Second Edition. New York: John Wiley and Sons.

(a) Test to see if X_4 can be dropped from the model, comparing to the full model with X_1, X_2, X_3, X_4. Specify the null and alternative in terms of β's, the value of F_S, the corresponding p-value.

(b) What is your conclusion in terms of the problem based on your information from (a) if $\alpha = 0.01$?

(c) Test to see if both X_2 and X_3 can be dropped from the model, comparing to the full model with X_1, X_2, X_3, X_4. Specify the null and alternative in terms of β's, the value of F_S, the corresponding p-value.

(d) What is your conclusion in terms of the problem based on your information from (c) if $\alpha = 0.01$?

(e) Based on your observations from (b) and (d), fit the "best" model and write down its estimated linear equation.

II. Continue with the data from problem I.

(a) What is the additional reduction in error we expect to see when we add X_4 to a model with only X_1 in it already?

(b) Find and interpret the value $R^2\{X_1|X_4\}$.

(c) What is the additional reduction in error we expect to see when we add X_2, X_3 to a model with X_1, X_4 in it already?

(d) Find and interpret the value $R^2\{X_1, X_4|X_2, X_3\}$

(e) Do the above values agree with your "best model" from II(e)? Explain.

III. Continue with the data from problem I. Consider the "full list" of models in this case to be:

```
all.models = c("Y ~ X1", "Y ~ X2", "Y ~ X3", "Y ~ X4"
"Y ~ X1 + X3", "Y ~ X1 + X2", "Y ~ X1 + X4",
"Y ~ X1 + X2 + X4", "Y ~ X1 + X3 + X4",
"Y ~ X1 + X2 + X3 + X4")
```

(a) Using the function `All.Criteria` in the R handout from week 7, find the model fit critera (except CP Mallows) for all of the above models. List your results.

(b) What is the best model according to BIC?

(c) What is the best model according to AIC?

(d) What is the best model according to PRESS?

(e) What is the best model according to R^2_{adj}?

(f) If you were trying to build a predictive model, which model would you use? Explain.

(g) If you were trying to build a "correct" model, which model would you use? Explain.