For this homework you are going to carry out a data analysis, with guidance. Because of the nature of the task, this homework may take a little longer than some of the others, but you have 9 days to do it instead of the usual 7 days. (This assignment differs from the Stat 110 assignment; see Nov 12 email for explanation.) Some useful R commands are given at the end of the assignment. Also, a tip when copying and pasting R output into a Word document is to change the font to Courier New, which looks like this. That will keep the alignment of the output.

The data used for this assignment is the StateSAT file, linked to the class website, and used as an example in class. It contains the 1982 SAT results by state, with average SAT score to be used as the response variable, and six possible explanatory variables. A description of the data set is linked to the class website in the same location as this assignment.

There are three goals for this analysis:

- To see what the relationship is between state expenditures per student ("Expend") and average SAT scores.
- To determine if any of the states should be justifiably excluded from the analysis because they are different from the others with regard to the goals of interest.
- To determine what combination of variables do a good job of explaining the average SAT scores.

Follow these steps for your analysis:

1. Create a correlation matrix for all of the variables. To do this, you will need to first create a data set without the column of state names, since correlations can only be found for numerical data. (See the end of this assignment for some handy R commands.)
   a. Show the correlation matrix. Based on the correlation matrix, which single variable would be the best predictor of SAT? Explain how you know.
   b. Look at the correlations among the six explanatory variables. Identify any pair(s) of variables that would result in a VIF value > 5, if those two variables were the only ones used as explanatory variables. Explain why it is logical that those two variables are highly correlated. (You need to read the descriptions of the variables to understand what they are.)

2. Fit the regression equation with SAT as the response and all of the other six variables as explanatory variables. Add columns to your data set that include the predicted values, hat values (leverage), Cook’s distance values, and studentized residuals. (See handout from lecture on Nov 9 to see how to do this.)
   a. Print and display in your homework the first 6 rows of the new data set, which includes the extra columns you were just asked to add.
   b. Identify which two of the 50 states have the highest leverage (hat values). For each one, give the name of the state and the hat value, then explain why that state has such a large hat value.
   c. For the two states identified in part (b), give the values of the studentized residual and Cook’s distance. Then explain whether you think there is justification to remove that state from additional analyses. (Remember the reasons for outliers and when they can be used to consider removing cases. Also remember our discussion of this data set in class.)
Remove the state with the largest hat value from the data set and use this new data set for the remainder of this assignment.

3. Redo the analysis using all of the predictor variables for the remaining 49 states. Use the “plot” command for this model, and include the plot of leverage versus standardized residuals in your homework. Choose one of the points identified by number on that plot, figure out which state is represents, and discuss why it is unusual. (You should have more than one choice, but you only need to choose one.) Do not remove any additional states from the analysis.

4. Run an “all subsets” regression. (Remember that you will need to install leaps and HH to do this. If you have trouble installing them using R Studio, use the install command in the R command line.)
   a. Print the summary using the HH version, which will give you all of the summary statistics you need to decide which model is best. Show the result of the summary command. (See lecture 12, Nov 4, for guidance.)
   b. Use a combination of the values of Cp, adjusted R$^2$ and the standard error of the residuals to identify the “best” model. Explain how you used that information to make your choice.
   c. For the model you chose in part (b), what are the values of Cp, adjusted R$^2$ and the standard error of the residuals?
   d. If you didn’t identify them already in part (b) specify which predictors are in your model.

5. Run the model you identified in #4. Show the part of the summary command output that includes the estimated coefficients and the t-test results that go with them. Do all of the predictor variables make a statistically significant contribution to the model? How do you know?

6. Run the model that has only Expend as a predictor.
   a. Show the part of the summary command output that includes the coefficients and the t-test results that go with them for this model.
   b. Is the coefficient for “Expend” positive or negative? Explain what that means about the relationship between state expenditure per student and average SAT scores.

7. Compare the contribution of “Expend” for the model in part 5 and the model in part 6. Explain any discrepancies between them.

8. Return to the beginning of this assignment, where three goals were stated. Write a short summary of the results for each of the three goals.

---

**Some useful R commands:**

Note that it’s best to retain the original data set when you remove a row or column so you can come back to it if you need to use it again later.

To create a new data set by removing column number c from the dataset called Data:

```r
> Newdata<-Data[, -c]
```

To create a new data set by removing row number r from the dataset called Data:

```r
> Newdata<-Data[-r, ]
```

To create a correlation matrix for the data set called Data (with numerical variables only):

```r
> cor(Data)
```

To print the first six rows of a dataset called Data:

```r
> head(Data)
```