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.

   ```r
   > Newdata<-StateSAT[,,-1]
   > cor(Newdata)
   SAT     Takers     Income       Years      Public      Expend        Rank
   SAT   1.00000000 -0.8578100  0.5844666  0.33096886 -0.08035688 -0.06287764  0.87990910
   Takers -0.85780996  1.0000000 -0.6619351 -0.10154350  0.12355625  0.28363041 -0.94283311
   Income  0.58446657 -0.6619351  1.0000000  0.13476231 -0.30656703  0.13151942  0.53269989
   Years   0.33096886 -0.1015435  0.1347623  1.00000000 -0.41711822  0.05982861  0.07022360
   Public -0.08035688  0.1235563 -0.3065670 -0.41711822  1.00000000  0.28459116  0.05062355
   Expend -0.06287764  0.2836304  0.1315194 -0.26496897  0.28459116  1.00000000 -0.26496897
   Rank   0.87990910  0.9428331  0.5326999  0.07022360 -0.26496897  1.00000000
   Rank is the best predictor. It’s the variable with the highest absolute value of correlation with SAT.

   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.)

   **Solution:** As explained in the lecture on the board on November 2, VIF > 5 corresponds to $R^2 > 80\%$ or 0.80, which means $|correlation| \geq \sqrt{0.8} = 0.894$. (You could also figure out that VIF > 5 corresponds to $R^2 > 80\%$ using the formula for VIF.) The only pair of variables meeting this condition is *Rank* and *Takers*, which have a correlation of -0.94. It is logical that they have a strong
negative correlation because “Rank” is the average high school rank of the students in that state who took the test, and “Takers” is the percent of eligible students in the state who took it. In the states with a low percent of students taking it, only the best students were likely to take it. They are the ones with high ranks in their class.

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.

   **Solution:** Here is the R code and results:
   ```r
   > Full<lm(SAT ~ Takers + Income + Years + Public + Expend + Rank, data = StateSAT)
   > StateSAT$predict <- predict(Full)
   > StateSAT$hat <- hatvalues(Full)
   > StateSAT$cook <- cooks.distance(Full)
   > StateSAT$rstud <- rstudent(Full)
   > head(StateSAT)
   
   State   SAT  Takers  Income Years Public  Expend  Rank  predict        hat         cook
   1 Iowa 1088     3   326 16.79  87.8 25.60  89.7 1057.044 0.11609974 0.0293128015
   2 SouthDakota 1075 2   264 16.07  86.2 19.95  90.6 1037.626 0.16926150 0.0705184879
   3 NorthDakota 1068 3   317 16.57  88.3 20.62  89.8 1041.743 0.11000956 0.0197098894
   4 Kansas  1045  5   338 16.30  83.9 27.14  86.3 1021.304 0.06036139 0.0079018502
   5 Nebraska  1045  5   293 17.25  83.6 21.05  88.5 1048.468 0.12261873 0.0003943522
   6 Montana  1033  8   263 15.91  93.7 29.48  86.4 1013.135 0.21547157 0.0284361670
   rstud
   1 1.2583163
   2 1.5835686
   3 1.0579621
   4 0.9263967
   5 0.1389303
   6 0.8485461
   
   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.

   **Solution:** High hat values indicate that there is something about one or more of the explanatory variables that sets that case apart from the others. The two states with the highest values are Alaska with \( h_i = 0.582 \), and Louisiana with \( h_i = 0.363 \). Alaska has the highest values of any state for Income and Expend (by far), and the 2nd highest for Public. Louisiana has by far the lowest value of Public at 44.8%, meaning that only 44.8% of the test takers attended public schools. It also has the 2nd largest value for Income, which is the median income of the families of the test takers.

   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.)

   **Solution:** Alaska: studentized residual = -3.3512, Cook’s distance = 1.805. Louisiana: studentized residual = -1.506, Cook’s distance = 0.179. It makes sense to remove Alaska. It is different from the other states with regard to expenditures, which are of particular interest in this analysis. The very high expenditures per student are mostly likely because schools are located so far apart and in remote areas, in addition to the fact that incomes are high and
therefore salaries are high. The unusually large values of the studentized residual and Cook’s distance provide additional evidence that Alaska is very different from the other states.

Remove the state with the largest hat value from the data set and use this new data set for the remainder of this assignment. *Note that this means removing Alaska.*

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, 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. (You will need to install one or two packages to do this.)
   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.)

   **Solution:** Here is the R code and results:

   ```R
   > Best <- regsubsets(SAT ~ Takers + Income + Years + Public + Expend + Rank, data = NoAlaska)
   > summaryHH(Best)
   
   model p  rsq  rss  adjr2   cp   bic  stderr
   1     R 2 0.776 55079 0.771 53.25 -65.4   34.2
   2   E-R 3 0.850 36815 0.843 22.67 -81.3   28.3
   3  Y-E-R 4 0.890 27016 0.883  7.19 -92.5   24.5
   4 I-Y-E-R 5 0.902 23968 0.893  3.75 -94.5   23.3
   5 I-Y-P-E-R 6 0.904 23547 0.893  5.00 -91.5   23.4
   6 T-I-Y-P-E-R 7 0.904 23546 0.890  7.00 -87.6   23.7
   
   Model variables with abbreviations
   
   model
   R     Rank
   E-R   Expend-Rank
   Y-E-R    Years-Expend-Rank
   I-Y-E-R Income-Years-Expend-Rank
   I-Y-P-E-R Income-Years-Public-Expend-Rank
   T-I-Y-P-E-R Takers-Income-Years-Public-Expend-Rank
   model with largest adjr2
   4
   ```

   The points specifically identified are 16, 22 and 31*, which are Mississippi, Louisiana and Oregon. The reasons Louisiana has such high leverage were given in part 2b, and apply even with Alaska removed. Oregon doesn’t have any single variable that stands out, but has a unique combination, especially with a low value of “Years” and high value of “Public.” Mississippi has a very low percent of Takers and very high rank.

   *Depending on your code, this point might be labeled 30 instead of 31. But it’s still Oregon.
b. Use a combination of the values of Cp, adjusted R² and the standard error of the residuals to identify the “best” model. Explain how you used that information to make your choice.

**Solution:** The models with acceptable Cp values are the ones with 4, 5 and 6 variable. The smaller models have Cp values much larger than \( p + 1 \). Of the models with acceptable Cp values, the ones with 4 and 5 variables show the same adjusted R² to 3 decimal places (0.893 each). But of those two models, the one with 4 variables has the slightly smaller “stderr” value of 23.3, where “stderr” is the standard error of the residuals.

c. For the model you chose in part (b), what are the values of Cp, adjusted R² and the standard error of the residuals?

**Solution:** Cp = 3.75, adjusted R² = 0.893, standard error of the residuals = 23.3.

d. If you didn’t identify them already in part (b) specify which predictors are in your model.

**Solution:** Income, Years, Expend, Rank

5. Run the model you identified in #4. Show the part of the 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?

**Solution:**
```r
> BestModel <- lm(SAT ~ Income + Years + Expend + Rank, data = NoAlaska)
> summary(BestModel)
```

```
Coefficients:              Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  -255.2894     88.6074  -2.881 0.006103 **
Income         0.2412     0.1020   2.365 0.022479 *
Years         18.9447     5.1563   3.674 0.000644 ***
Expend         3.3851     0.7775   4.354 7.87e-05 ***
Rank           9.3764     0.6589  14.229  < 2e-16 ***
```

All four predictors make a statistically significant contribution to the model. The \( p \)-values for the tests of whether the corresponding population coefficient is 0 are all small (below 0.05).

6. Run the model that has only Expend as a predictor.
   a. Show the part of the output that includes the coefficients and the t-test results that go with them for this model.

**Solution:**
```r
> ExpendOnly<-lm(SAT ~ Expend, data = NoAlaska)
> summary(ExpendOnly)
```

```
Coefficients:              Estimate  Std. Error t value  Pr(>|t|)
(Intercept)    961.7244     49.8885  19.277 < 2e-16 ***
Expend         -0.5923     2.1779  -0.272    0.787
```

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.
**Solution:** The coefficient is negative. This means that in the absence of any of the other explanatory variables, the more states spend on average per student, the lower the average SAT score.

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

**Solution:** In part 5 the coefficient is positive and statistically significantly different from 0. In part 6 the coefficient is negative and has a high p-value of 0.787, which means the population coefficient could be 0. The difference is that in part 5 the model includes Income, Years and Rank as well. So the coefficient and test in that case represent the additional contribution of Expend, given that the other variables are in the equation. In other words, it’s the contribution of Expend after accounting for the difference in the quality of the students taking the test. It is misleading to look at Expend by itself, because the quality of students taking the test was so different across states.

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.

**Solution:**
- To see what the relationship is between state expenditures per student (“Expend”) and average SAT scores.

If other variables are not taken into account, it appears that there is a very weak, negative relationship. But that’s probably because the quality of the students who took the exam differ from state to state. For states with similar quality of students taking the exam (Rank), family income (Income) and years of specific coursework (Years), additional expenditure has a significantly positive relationship with SAT scores, with each additional hundred dollars of expenditure predicted to add 3.385 points to the average SAT score.

- 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.

Alaska could be justifiably excluded because it appears to come from a different set of circumstances (population) than the other states, with almost no students who were not in public schools (Public = 96.5), very high expenditures (Expend = 50.1), and very high family incomes (Income = 401).

- To determine what combination of variables do a good job of explaining the average SAT scores.

The variables that did the best job are Income, Years, Expend and Rank. (You are not expected to have included what follows in your homework, but it is useful for understanding the results.) As noted in Question 1b, there is a very strong (negative) correlation between Takers and Rank. If Rank had been excluded as an option, the best model would have included Takers.