BSAN 450 Assignment 6

1) The purpose of this assignment is to relate the automobile accident rate in accidents per million miles (Rate) to 13 potential independent variables. The data include 39 sections of large highways in the state of Minnesota. The potential independent variables are:

Len: length of the segment in miles

Adt: average daily traffic count in thousands

Trks: truck volume as a percent of the total volume

Slim: speed limit in miles per hour

Lwid: lane width in feet

Shld: width in feet of outer shoulder on the roadway

Itg: number of freeway type interchanges per mile in the segment

Sigs: number of signalized interchanges per mile on the segment

Acpt: number of access points per mile in the segment

Lane: total number of lanes of traffic in both directions

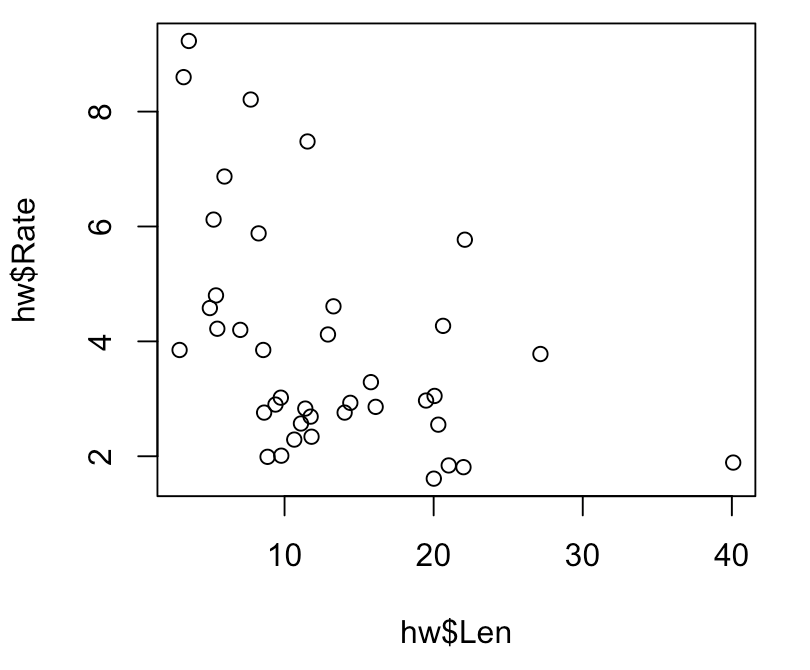
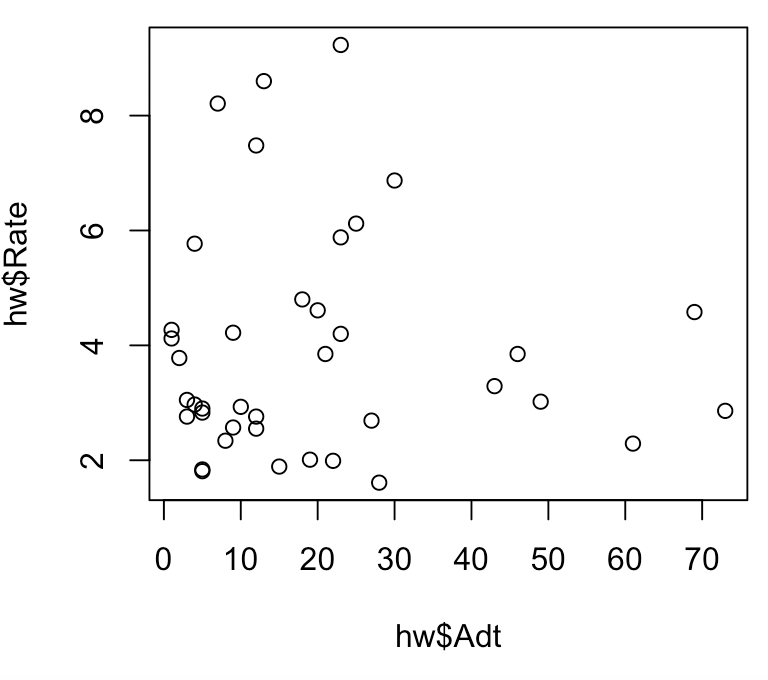
Fai: an indicator variable that is 1 if a federal interstate highway and otherwise 0

Pa: an indicator variable that is 1 if principle arterial highway and otherwise 0

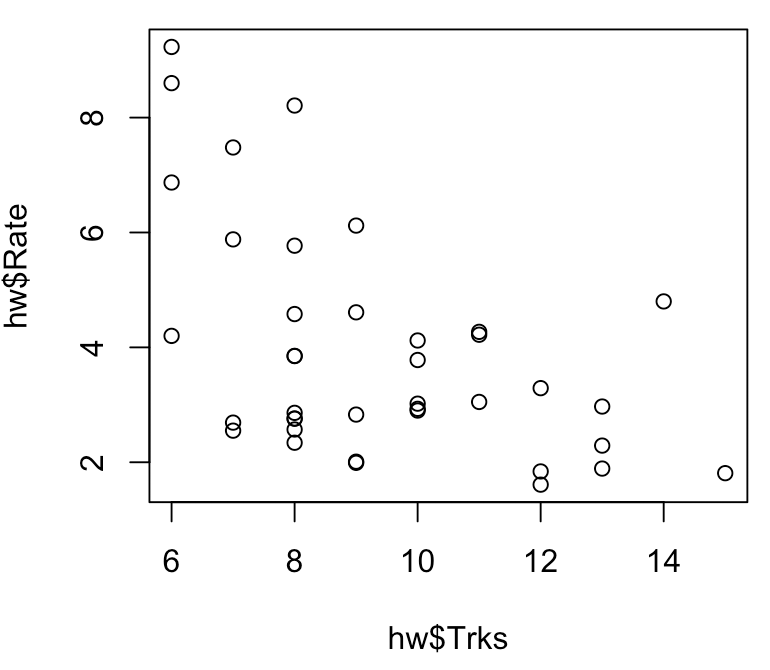
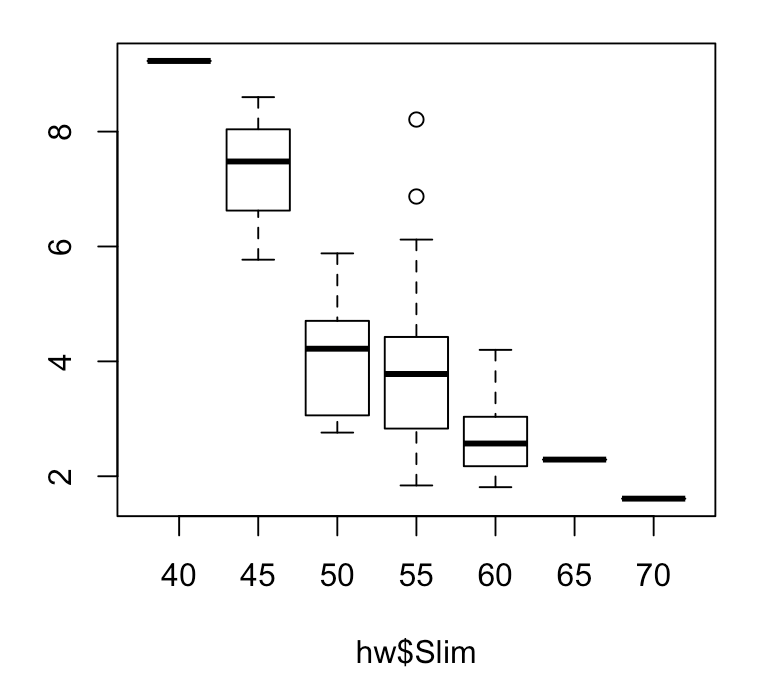
Ma: an indicator variable that is 1 if major arterial highway and otherwise 0

The data are in a file named highway.csv.

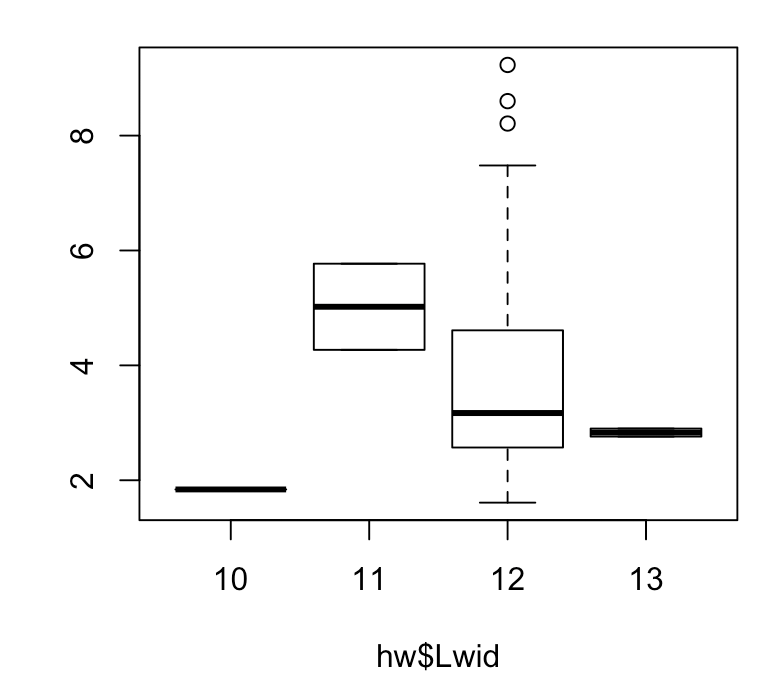
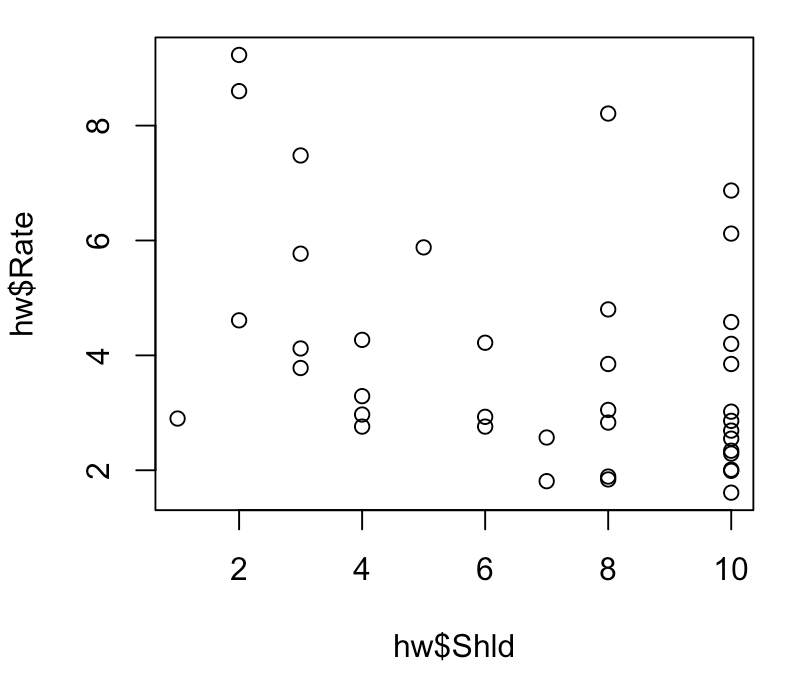
1. Produce appropriate plots of Rate versus all of the possible independent variables. What do these plots suggest about which independent variables should be included in a regression model?

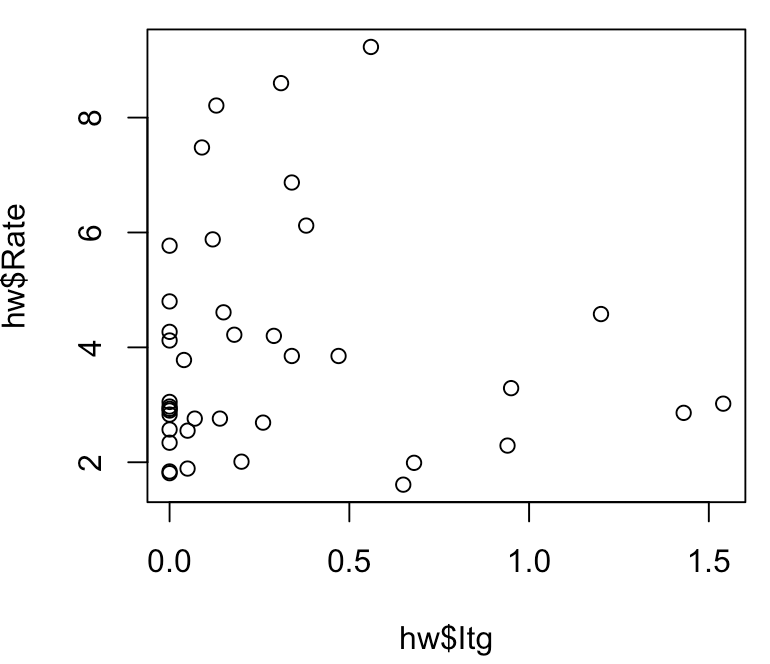
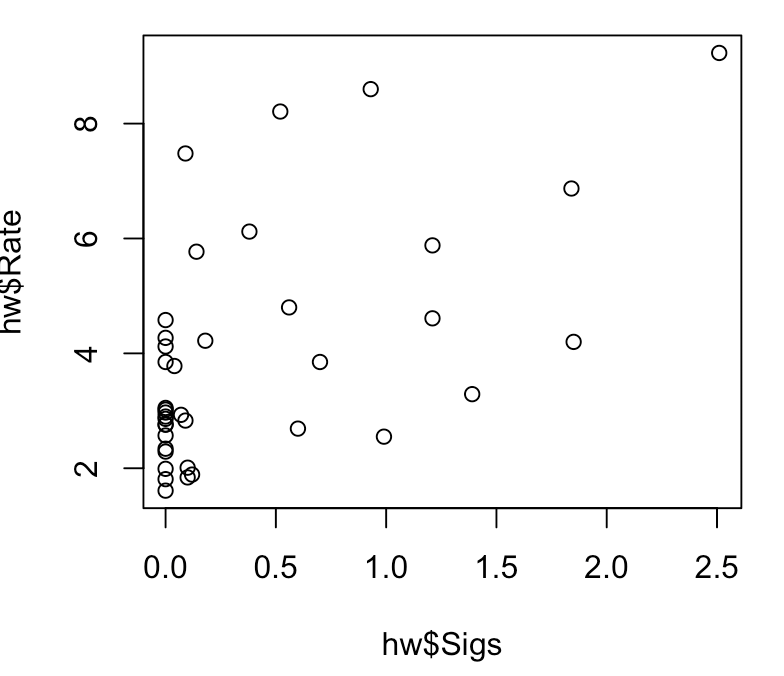
Both Len and Adt graphs look like there is a probable negative linear relationship

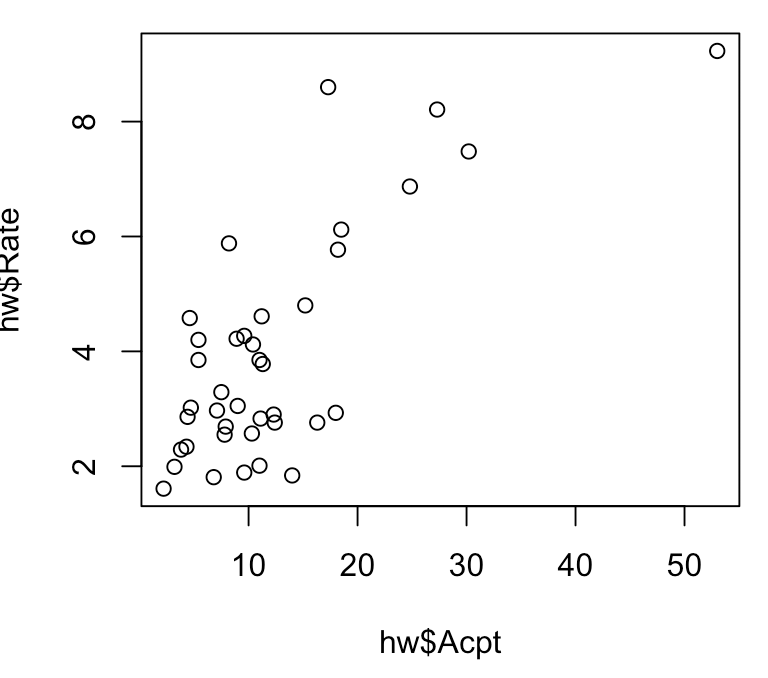
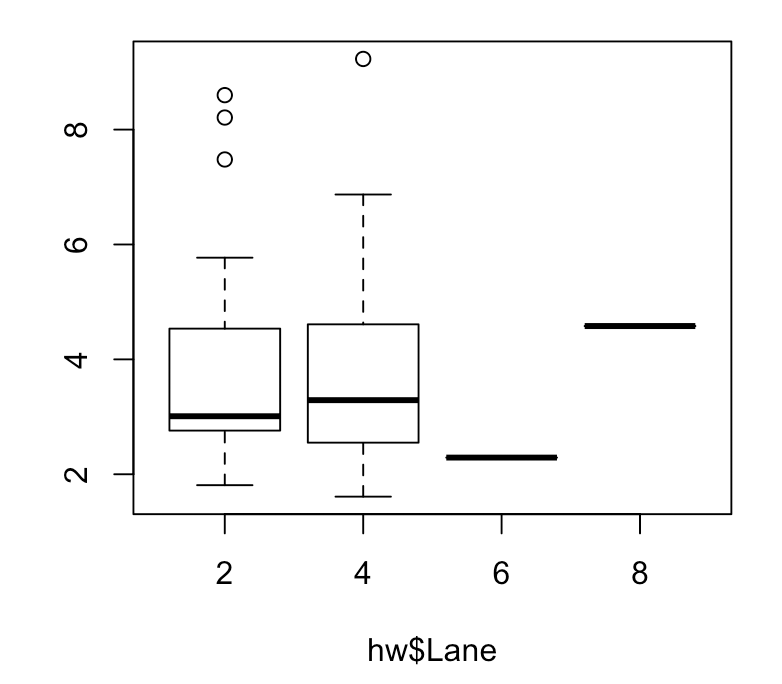
Both Trks and Slim graphs suggest a negative linear relationship.

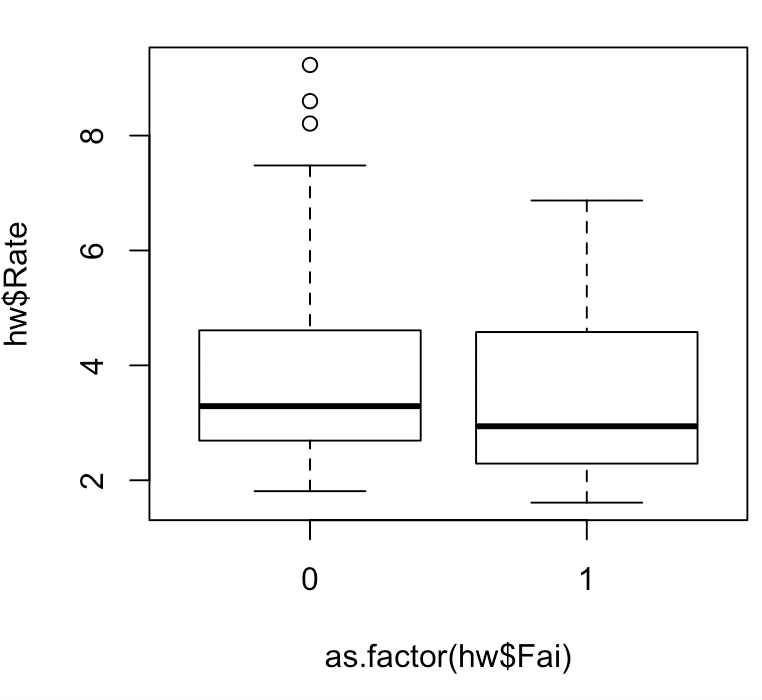
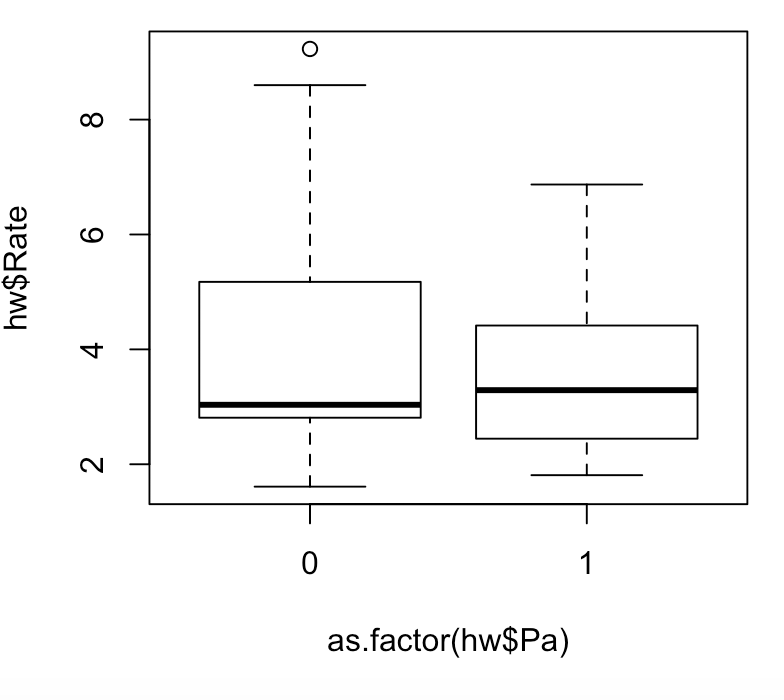
Lwid appears to have a potential linear relationship. Shld appears to have a polynomial relationship with Rate.

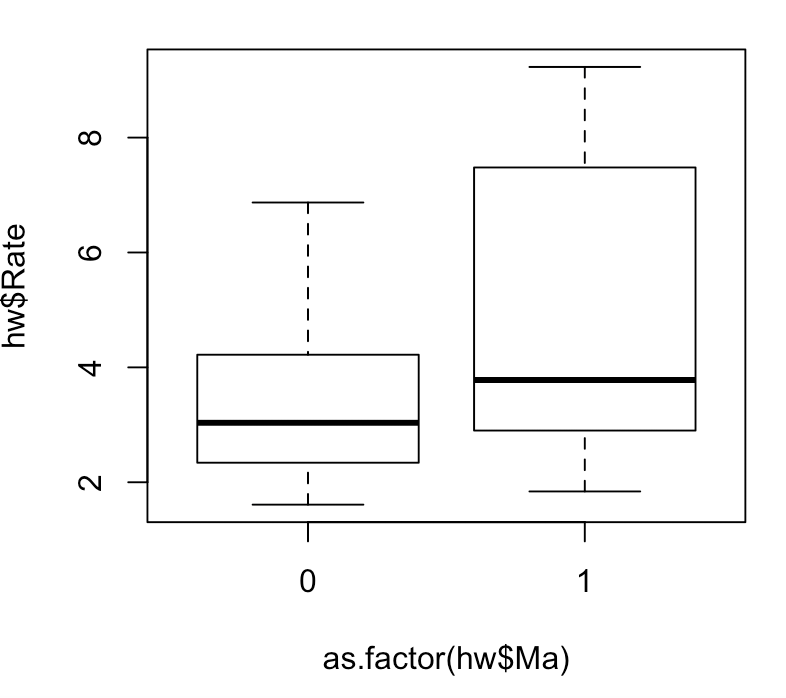
Both Itg and Sigs graphs do not appear to have a strong relationship to Rate.

Acpt appears to have a linear relationship with Rate. Lane might have a slight linear relationship with rate.

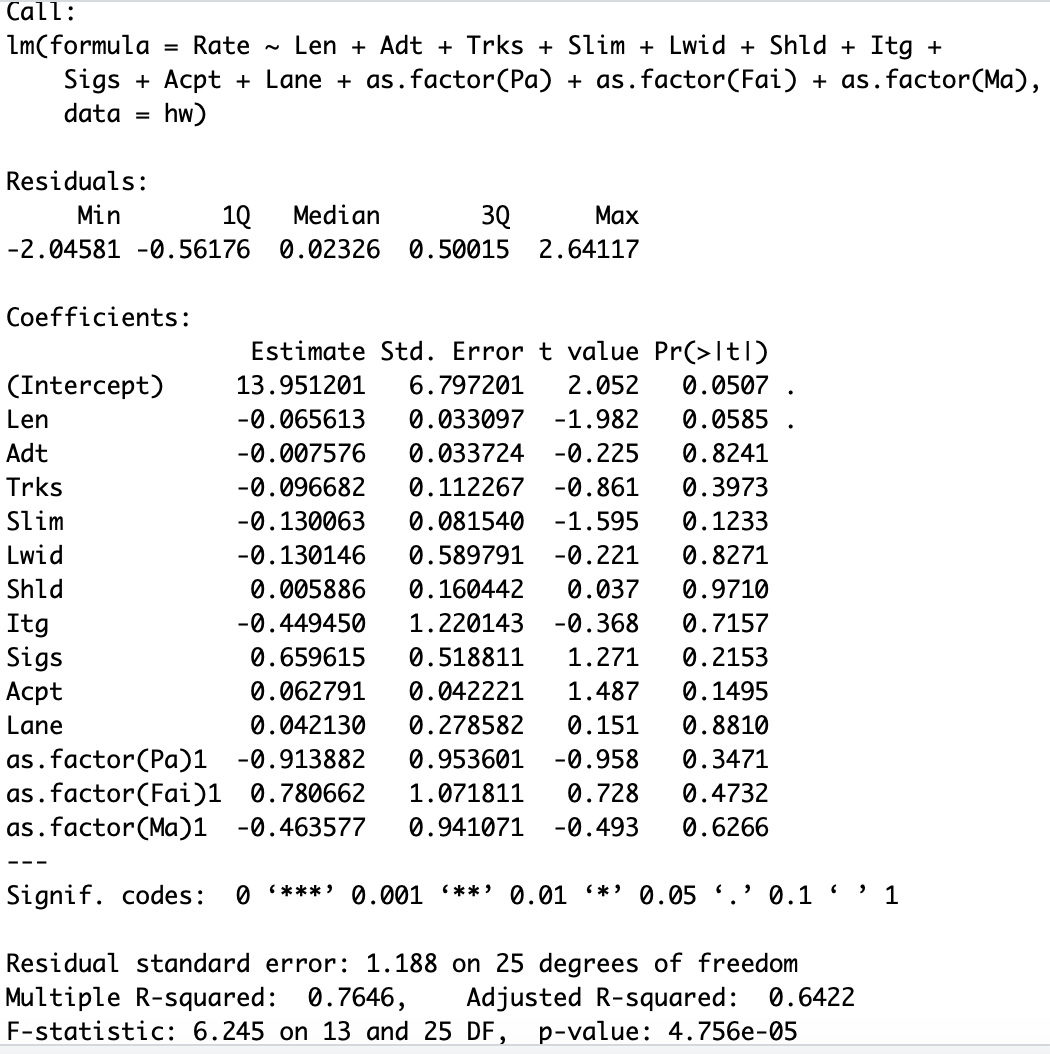
 

Neither Fai nor Pa appear to have a strong relationship with Rate.

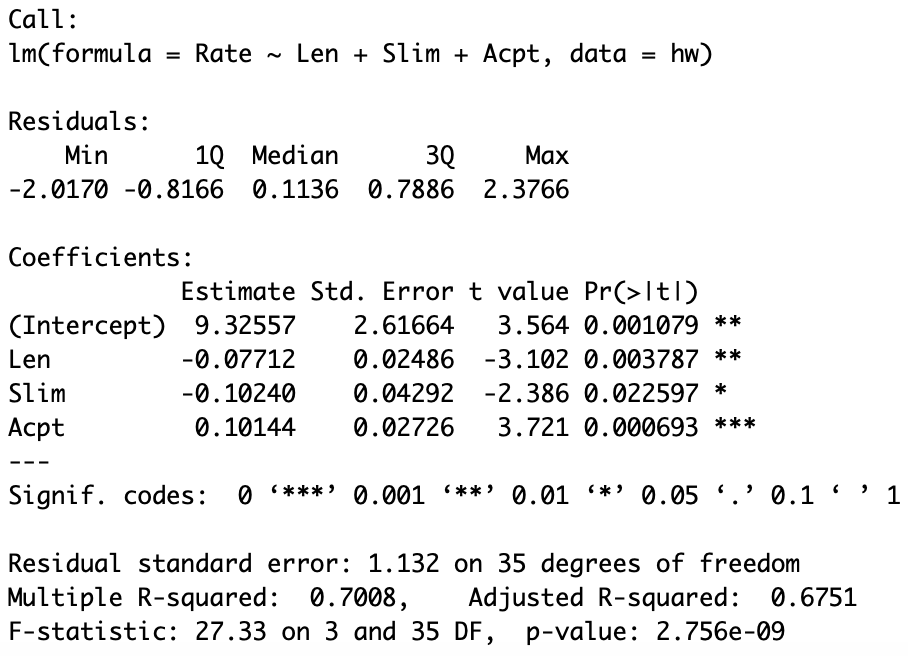


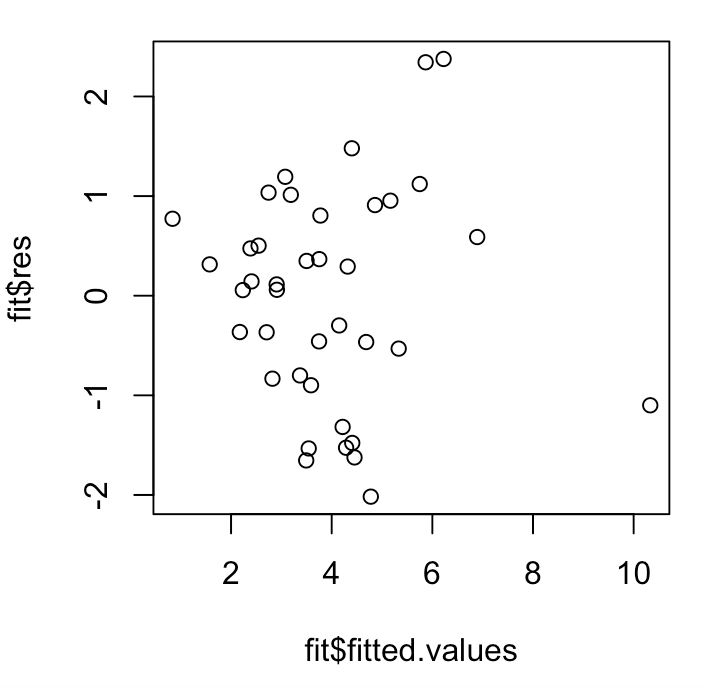
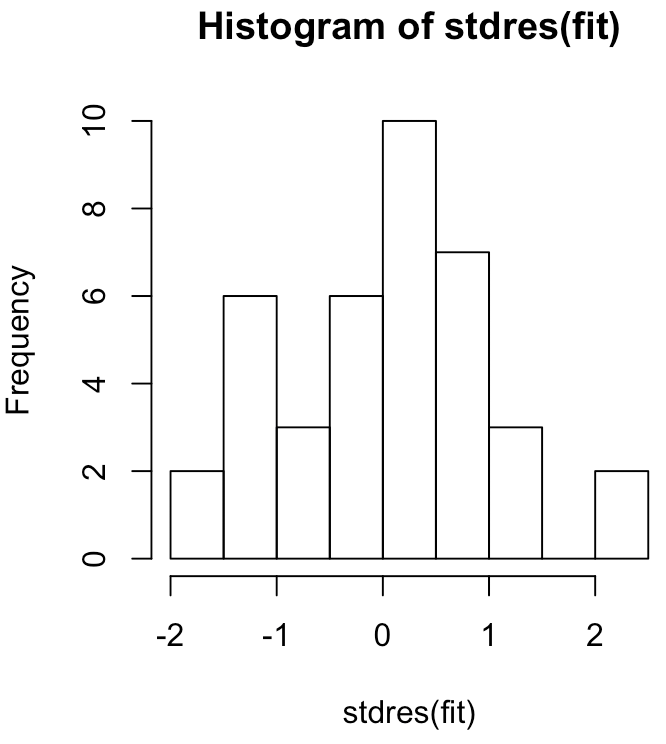
Ma does not appear to have a strong relationship with the rate.

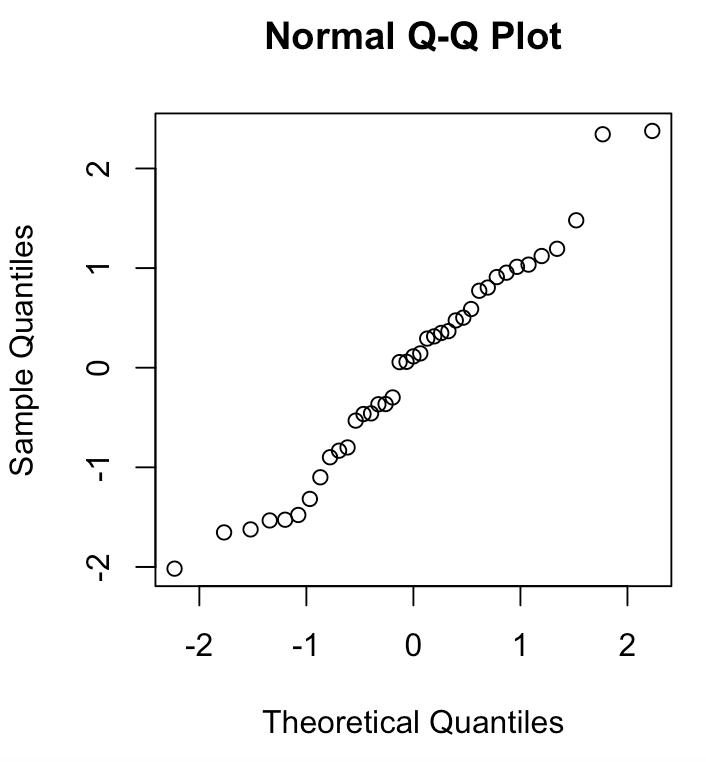
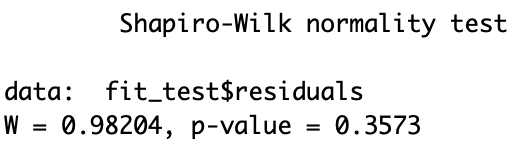
1. Propose a multiple regression model to fit to this data. If necessary make any appropriate changes to your model and perform the diagnostic checks for the model you believe best fits the data.



This linear model indicates that multiple of the variables are not statistically significant. I removed the variables that had a p-value above .05.



Adjusting the model to only include the four statistically significant variable results in the above model which the diagnostics indicate no issue.

2) The data are the construction costs of 32 light water reactor power plants, together with characteristics of the plants and details of their construction. The objective of the modeling is to learn which of the variables influence capital costs so that we can predict the cost of constructing a new plant with certain specified characteristics. The data is in a file named powerplant.csv. The variables are:

C = Cost in 1 million dollars, adjusted to 1976 base

D = Date construction permit issued

T1 = Time between application for permit and issue of permit

T2 = Time between issue of operating license and construction permit

S = Power plant net capacity (MWe)

PR = prior existence of a light water reactor on the same site (=1)

NE = plant constructed in northeast region of US (=1)

CT = use of cooling tower (=1)

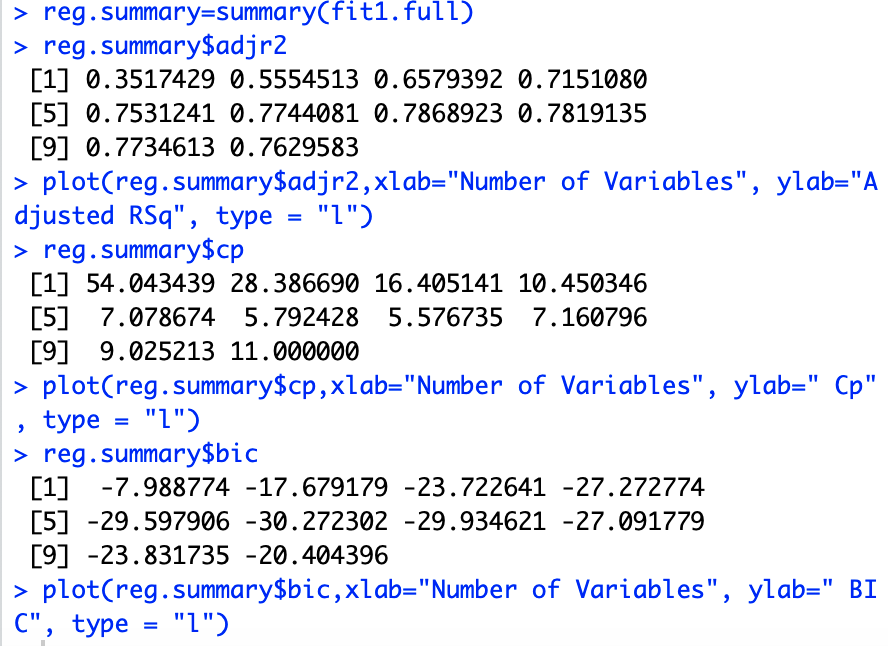
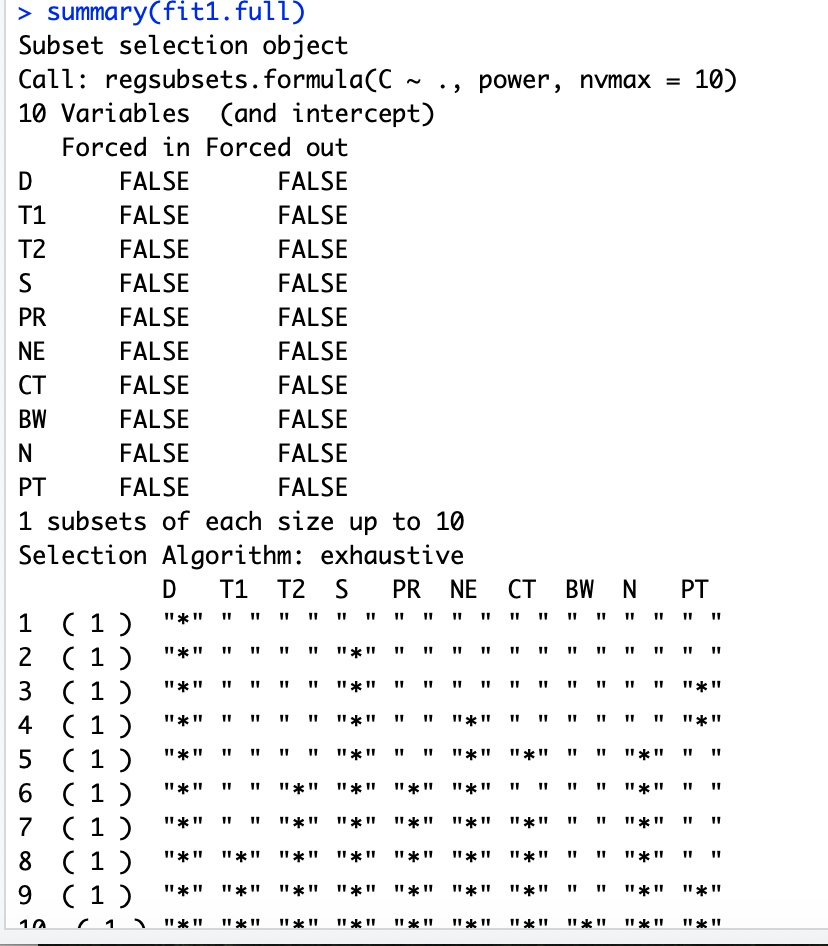
BW = nuclear steam supply system manufactured b Babcock-Wilcox (=1)

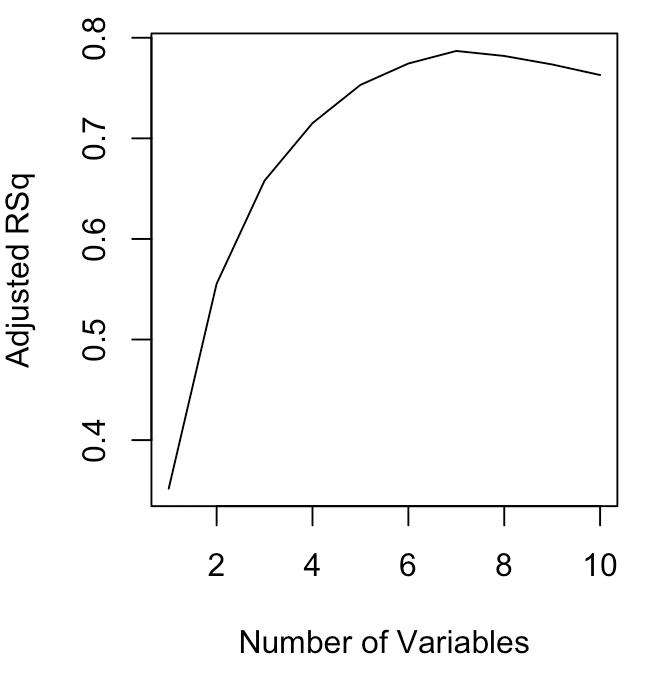
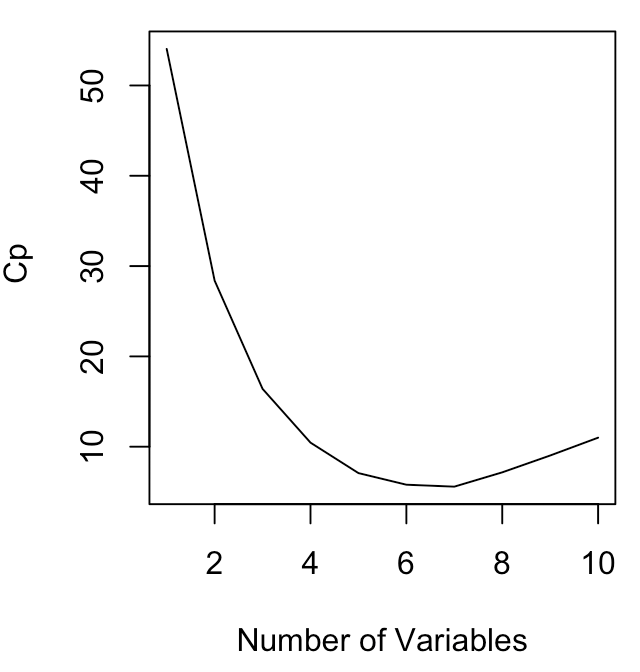
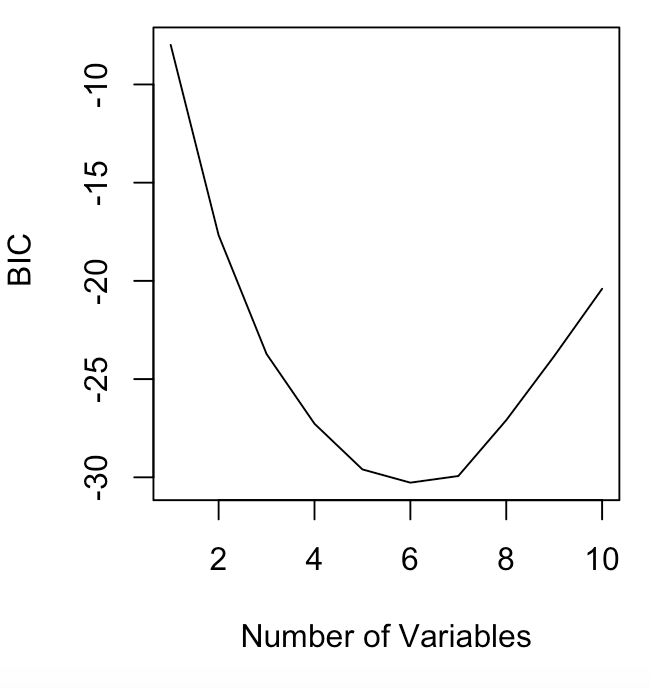
N = Cumulative number of power plants constructed by each architect-engineer

PT = partial turnkey plant (=1); a special feature in the contract that may affect capital cost.

a) Use the best subsets approach to help determine how many of the independent variables are needed in a multiple regression model that has the variable C as the dependent variable. Print out and plot the values of the adjusted R squared values, the Cp values and the BIC values for models with different numbers of variables. How many variables should be included in the multiple regression model? Justify your answer.

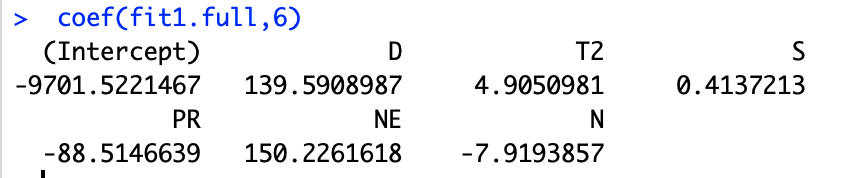
The R commands to do this are:

Six variables should be used, because having six variables results in the lowest BIC and provides a close second for minimizing Cp and maximizing Adjusted RSq.  

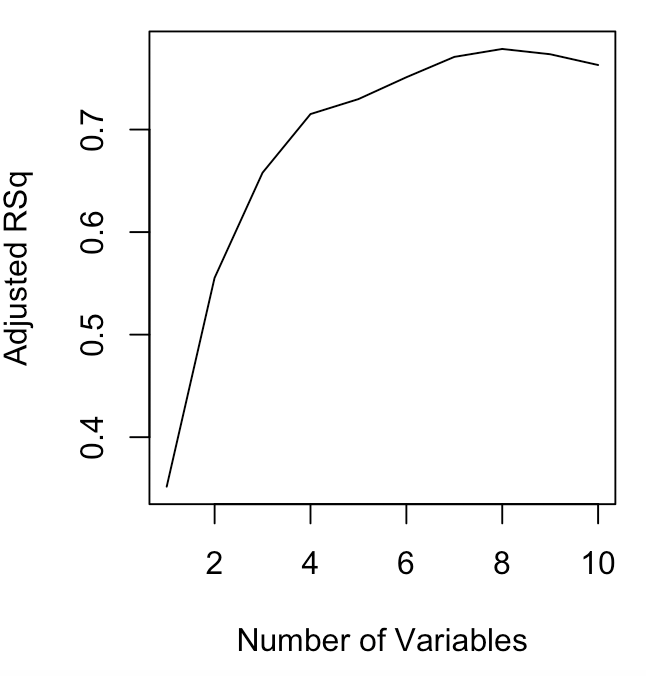
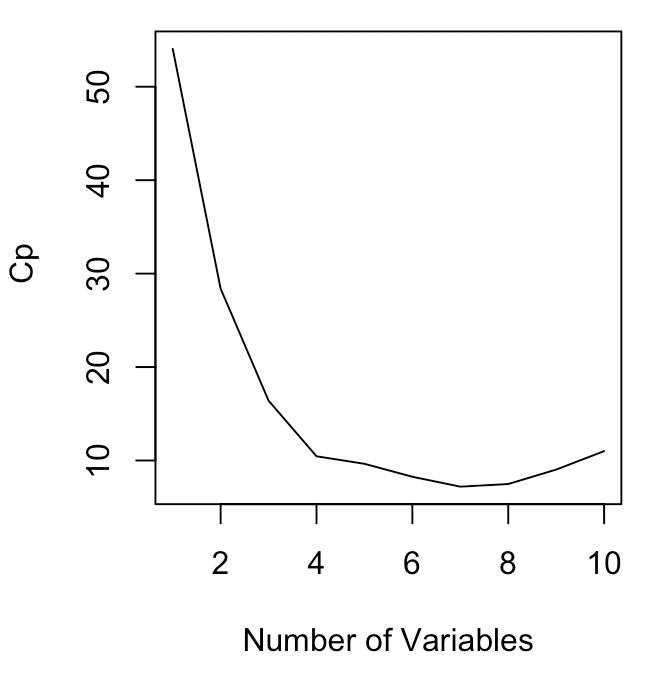
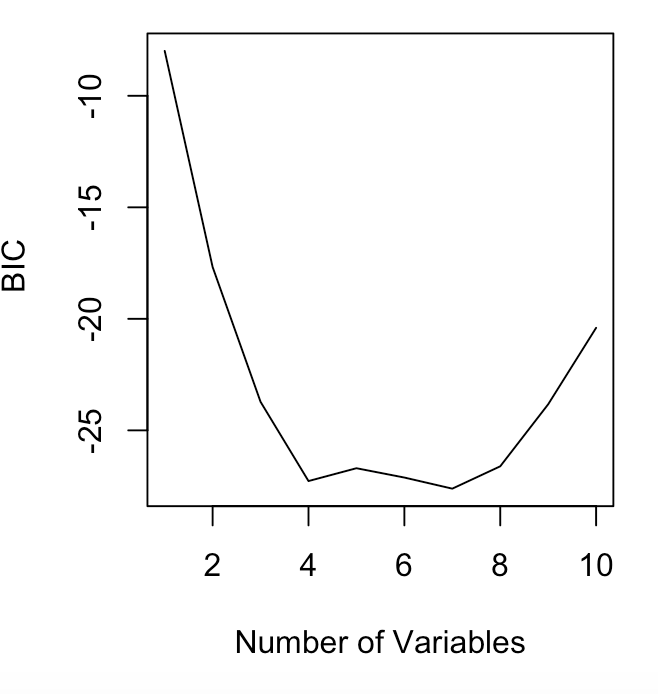
b) Print out the variables that are included in the model or models that you believe are the most appropriate for this data based on the adjusted R square values, the Cp values, and the BIC values.

The R command to do this for a model with 5 variables is:



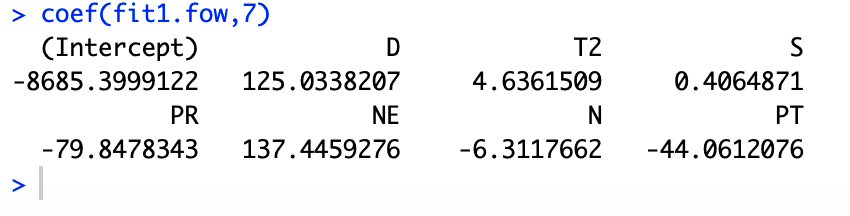
c) Use the forward stepwise regression approach to help determine how many of the independent variables are needed in a multiple regression model that has the variable C as the dependent variable. Print out and plot the values of the adjusted R squared values, the Cp values and the BIC values for models with different numbers of variables. How many variables should be included in the multiple regression model? Justify your answer.

The R commands to do this are:

    
There should be seven variable in the model beceause in the graphs seven variables is approximately the max/min on the graphs above.

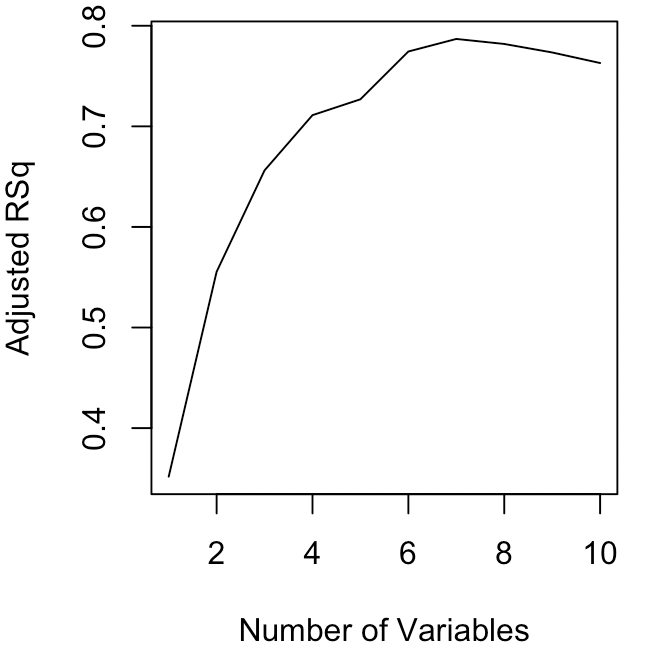
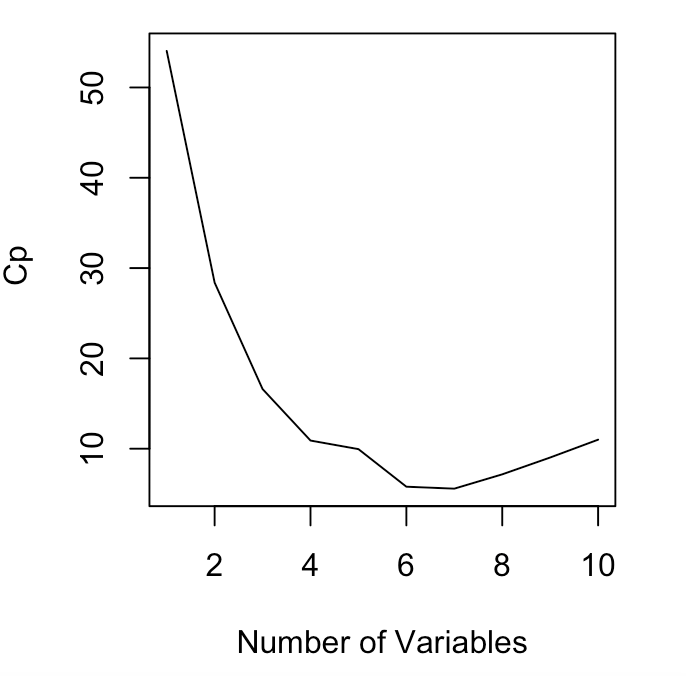
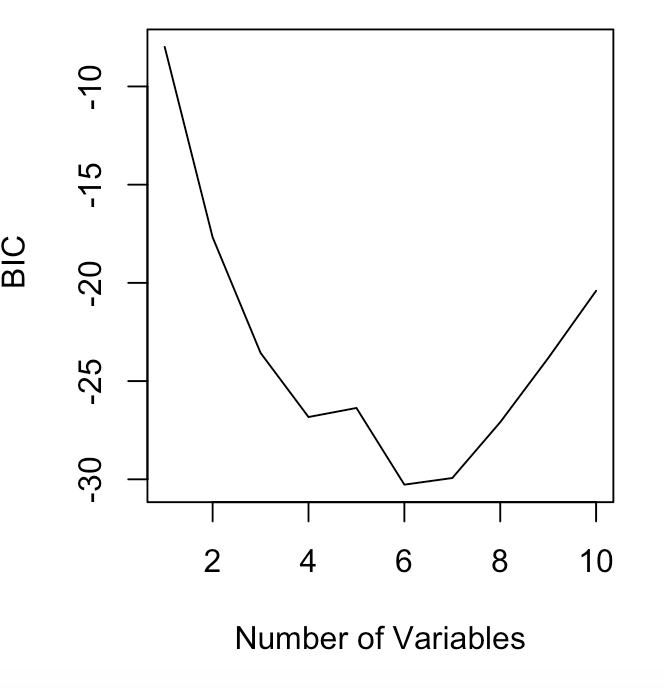
d) Print out the variables that are included in the model or models that you believe are the most appropriate for this data based on the adjusted R square values, the Cp values, and the BIC values.

The R commands to do this for a model with 5 variables is:



e) Use the backward stepwise regression approach to help determine how many of the independent variables are needed in a multiple regression model that has the variable C as the dependent variable. Print out and plot the values of the adjusted R squared values, the Cp values and the BIC values for models with different numbers of variables. How many variables should be included in the multiple regression model? Justify your answer.

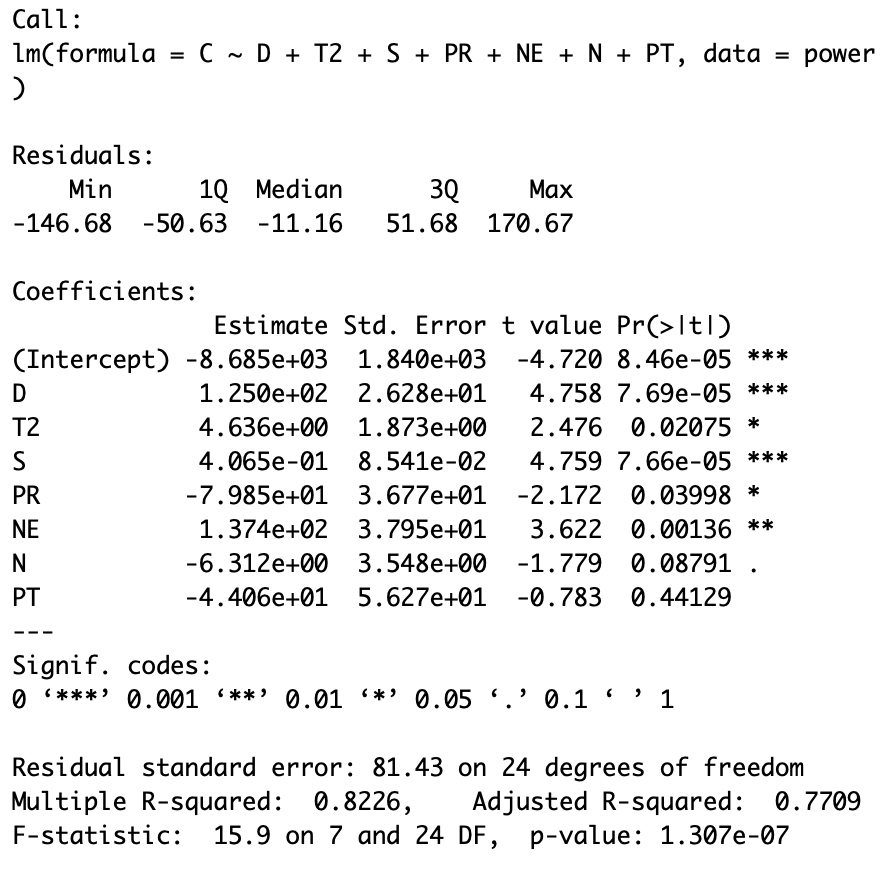
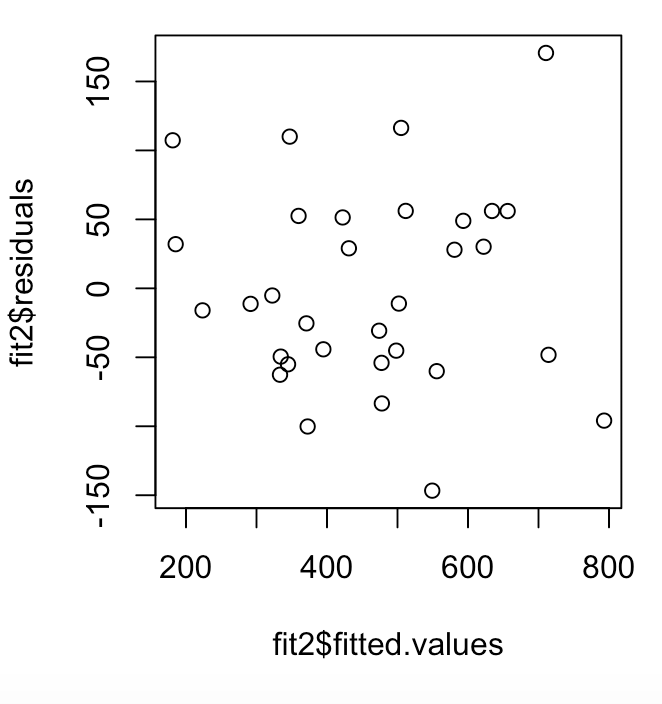
The R commands to do this are:

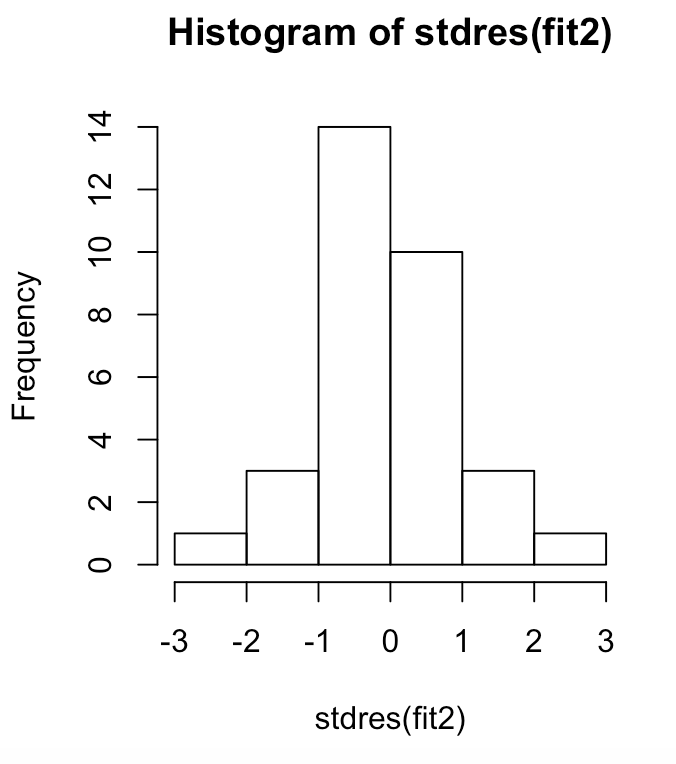
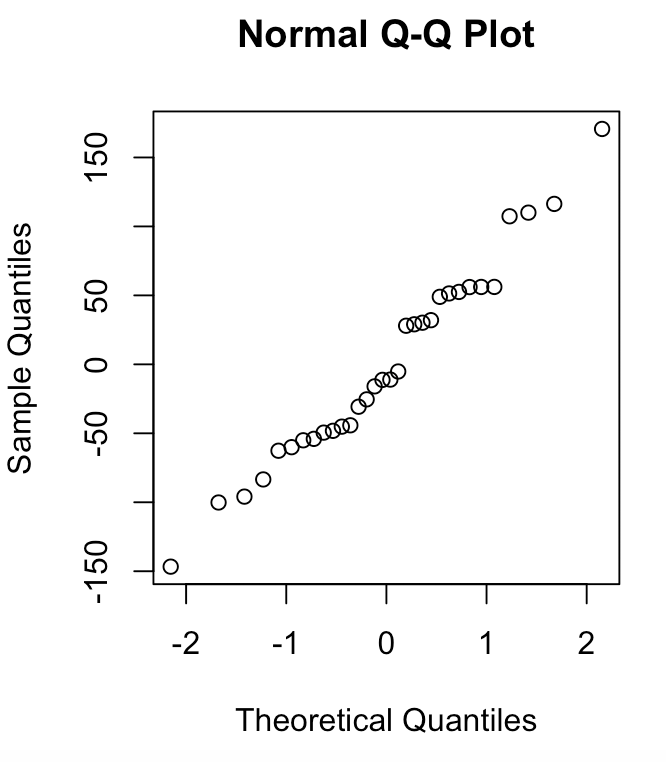
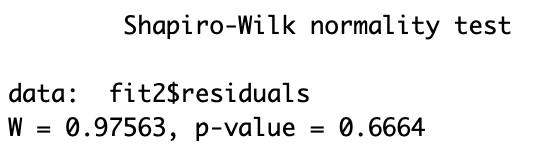
  Seven variables should be included. Seven is the value where most of the graphs are at mins/maxs.

f) Based on the results for the best subsets regression, the forward stepwise regression, and the backward stepwise regression which model would you suggest? Justify your answer.

I would suggest a model including seven variables which should be D, T2, S, PR, NE, N, PT. This is because when running both the forward and backwards stepwise regression, the graphs indicated seven would be ideal. Also in both cases the impacting variables where the variable mentioned above.

g) For the model you picked in part f, use the lm() function to estimate the parameters in the model and print out the summary of the results. Plot the residuals versus the fitted values, a histogram of the residuals and a Normal probability plot of the residuals. Is there any indication of a problem with this model?

Everything looks good. The qqnorm graph is a little shakey, but overall fine.