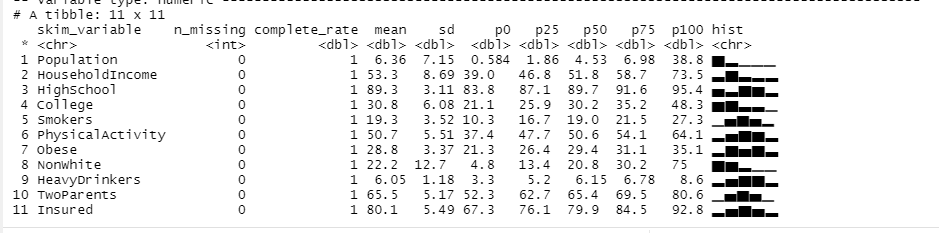
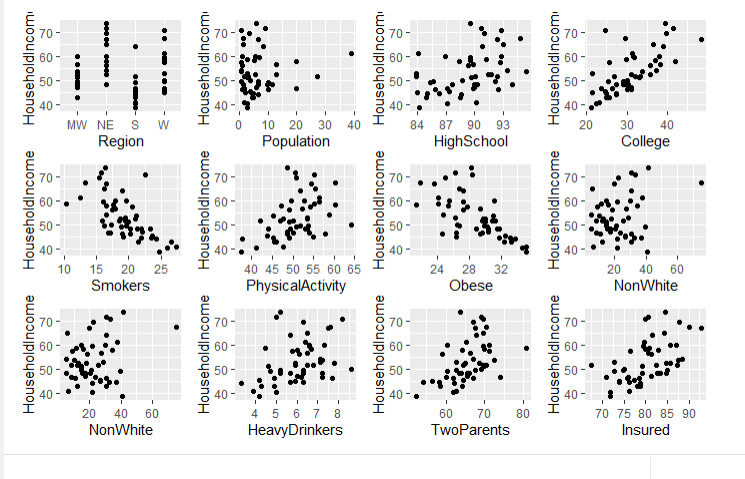
Assignment 2

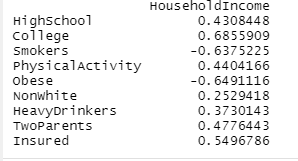
April 2021

1) In determining the response versus explanatory variables it is important to know what question is being asked of the data and what is trying to be solved. Response variables are those that could be predicted or the focus of a question. Explanatory variables explain changes. Here, response variables could be: Population, Household income, Insured, or Two Parents. Explanatory variables could be: Region, Highschool, college, smokers, physical activity, non white, heavy drinkers. Insure, two parents, and more could take on both roles. The population of interest is that of the United states broken down into states and regions.

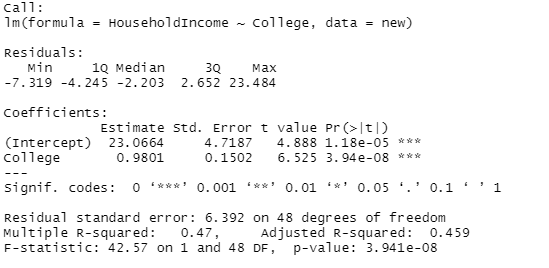
2) 



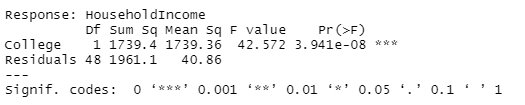
3)

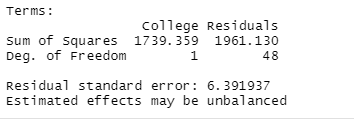


After reviewing the data of correlation and scatterplots, it does not seem that a linear regression is appropriate for analysis. All of the variables are too weak in relation against household income to create a strong linear regression model. However, a multiple regression model with the top variables might be more appropriate.

4) You would want to start with the college variable because it has the highest correlation to household income with .68 correlation. 

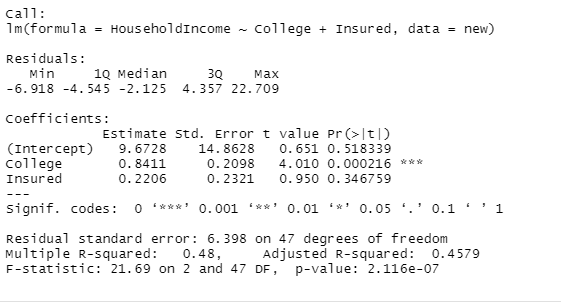
Equation: y\_hat = 23.0664 + 0.9801 \* X



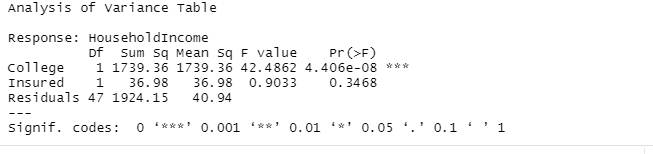


5) Sum of Squared Residuals is 1981.13, Sum of Squares Total is 3700.49, Sum of Squares due to regression is 1739.20, SSR/SST is 0.4700349. The anova table as put above matches these values.

6)

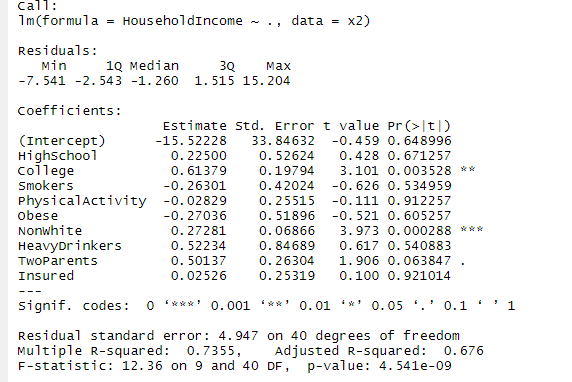


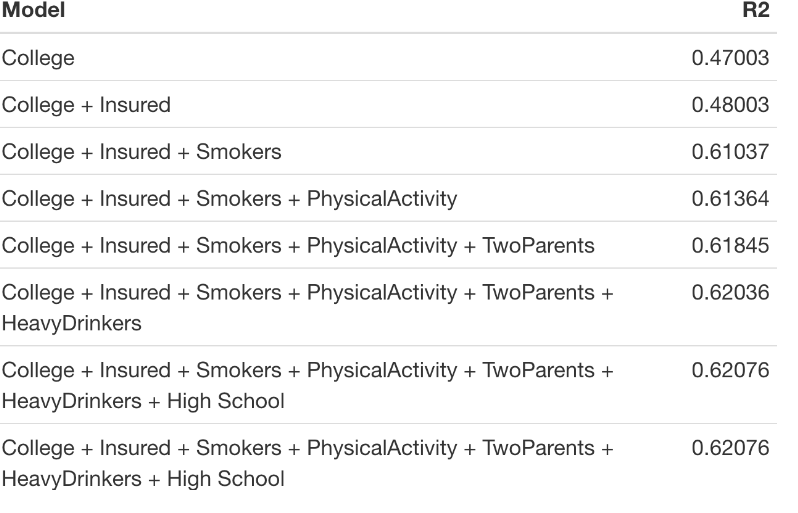
Equation = y\_hat = 9.6728 + .8411 \* X1 = .2206 \* X2



R-Squared value is .48. The coefficient decreases from model 1 to model 2 for college. The r squared value only increased by .01. There is not enough evidence to reject the null hypothesis of zero and Insured doesn’t/t explain much of the variability in the models so it can be rejected.

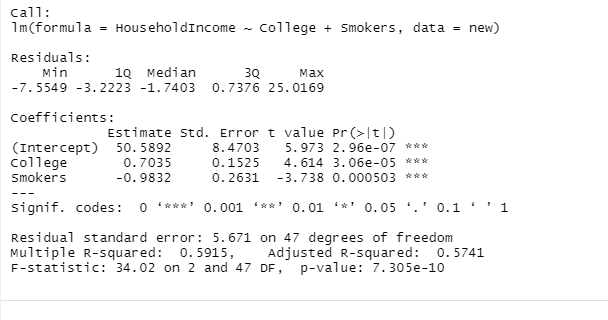
7)

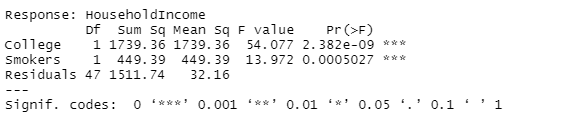




It seems that the best variables for a predictive model are College and Smokers. The r2s seem to have very similar values the more variables added to it with twoparents and PhysicalActivty being more minimal in terms of effect. The criteria that seems appropriate are those which would have a significant impact on the model and since most of the variables do not they can be discarded. The interpretations do not become more counterintuitive but they do show which variables have minimal or no impact. This means that it would be important to understand all the r squared or other coeffecients regarding impact otherwise you could end up with a bloated model which includes variables that have little to no impact.

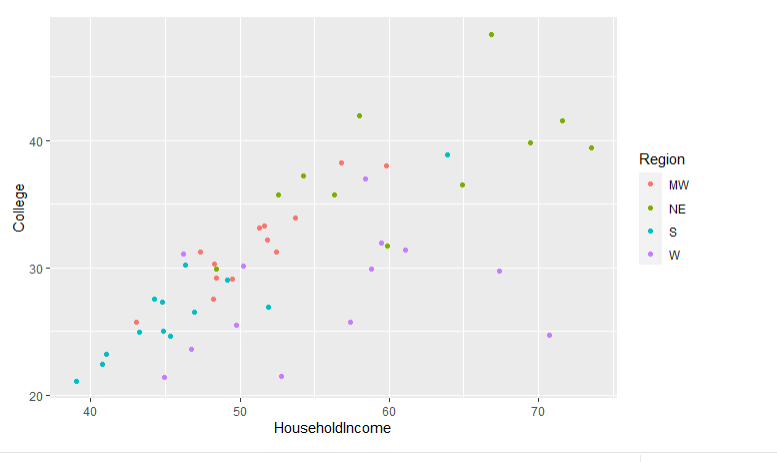
8)





This model which is household income against college and smokers has an r squared value of .59, much higher than our previous models. It was necessary to refit this model because those are the best fitting variables which show the most direct impact on household income.

9)



I wanted to see if the variables that we didn’t look at like region would have an effect on the data. From the chart it looks like there is as those in the West with low college percentages are still making quite a lot of income, whilst in New England it is those who are highly college educated who make the most income.

10)

The conclusions I can draw are that very few factors affect the r squared or fit for regression models. It seems only College or smokers affects household income in a great amount whilst the other variables are marginal. It was interesting to look at how region affects household income and college, with those in the west not being very educated but still making quite a lot of money. The story in this data is that region, college education and smoking have a great impact upon the income of the household. Despite their low affects on r squared it would still be a good idea to keep the other variables in the model if one wants to get a more nuanced view of the model. This could especially help with region and state as there might be different regions or states that have higher or lower affects which would be interesting to note.