# Week 3 Assignment

By: Zach Adair

Regis University

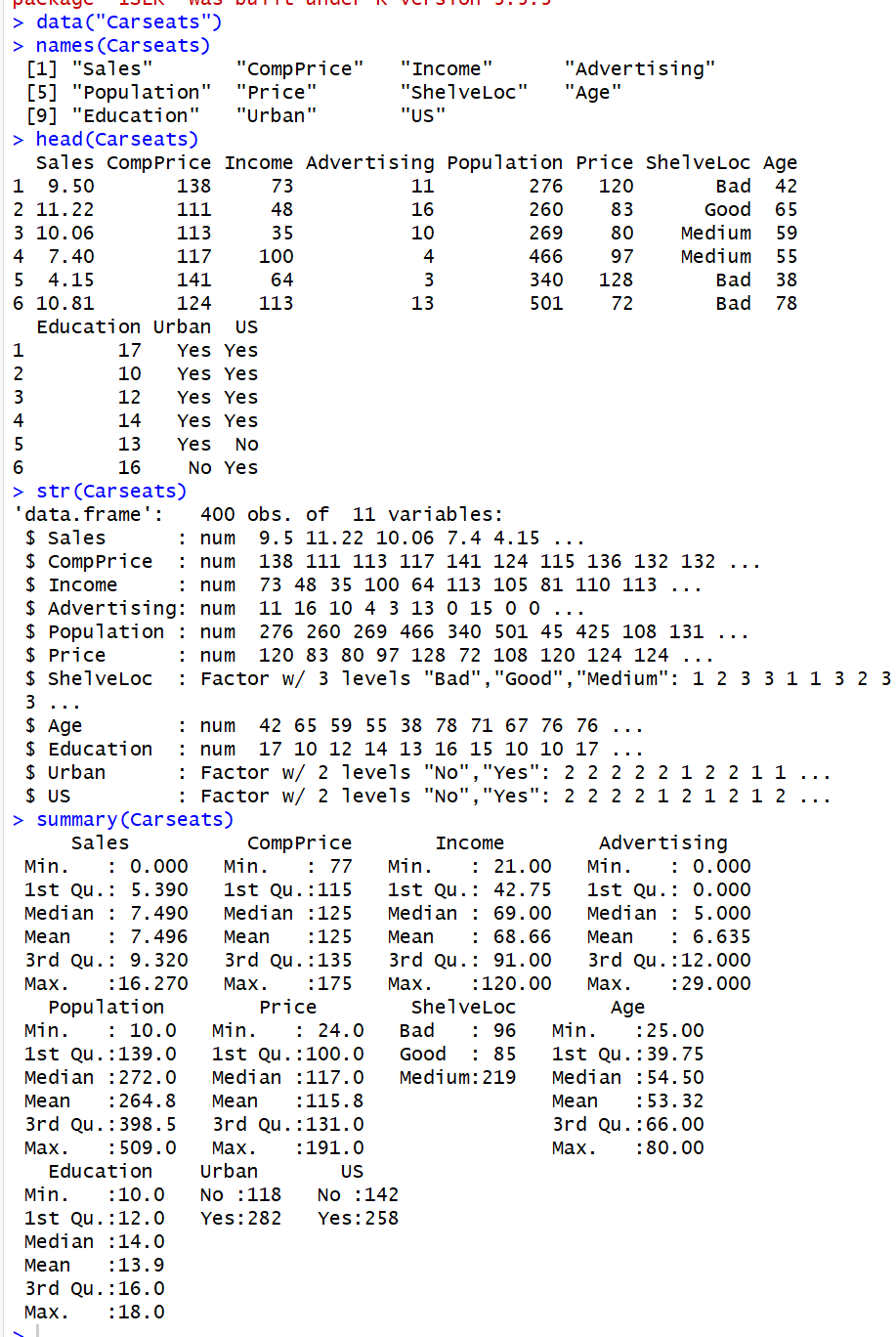
MSDS 660 – Statistical Methods & Experimental Design

## Introduction

In this assignment I will continue to work on the Carseats dataset just like from last week but this time we will use the dataset to run a Multiple Linear Regression at 95% confidence level. In this analysis I will be checking for multi-collinearity within variables I’ll be testing.

### Pick 2-3 predictors (independent variables) and one response (dependent variable). List them. Perform appropriate data explorations. State your research questions.

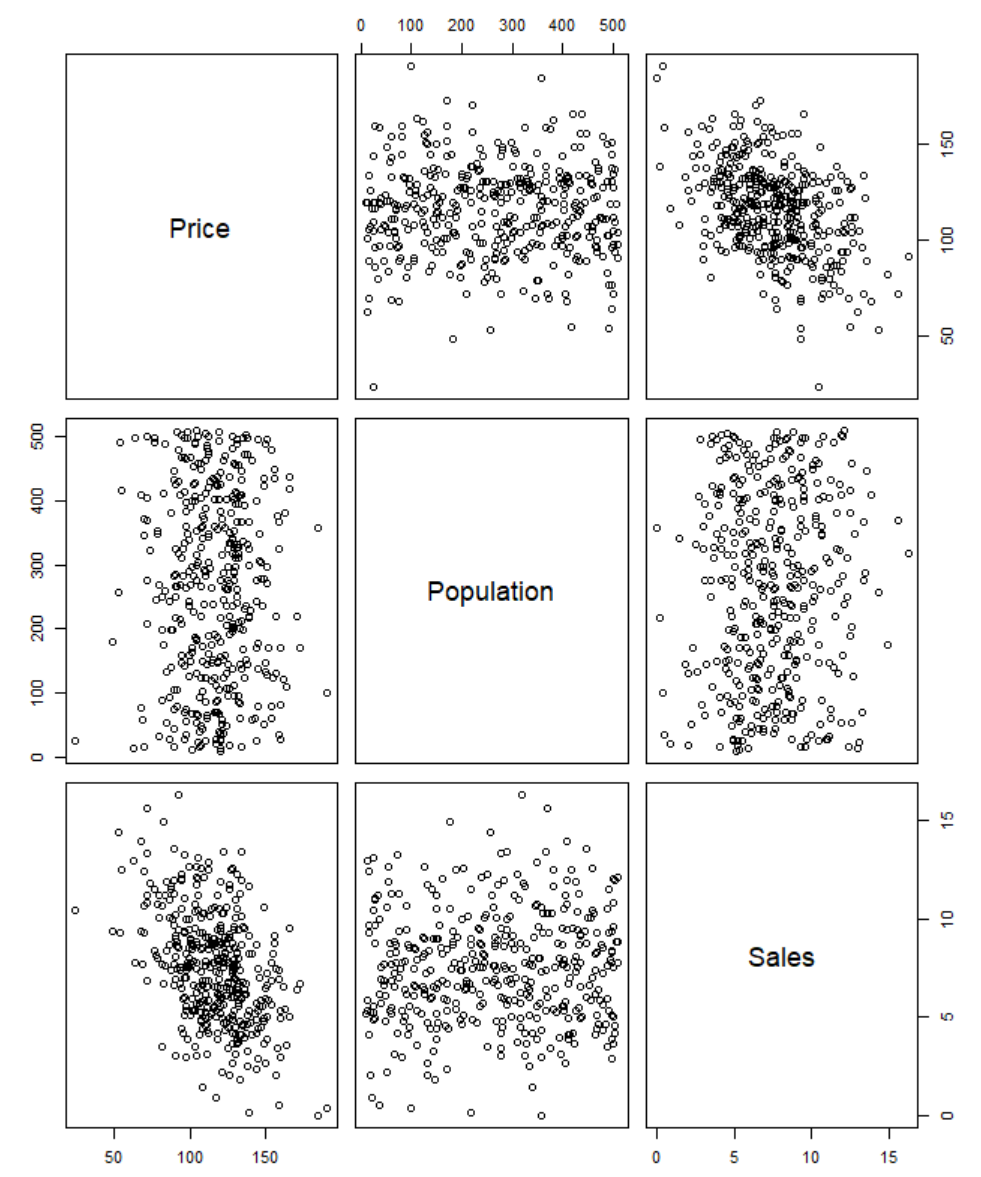
First I needed a little refresher about the data in Carseats so I wanted to look at the data a little bit before I figure out the predictors and dependent variable I’m going to select for analysis.

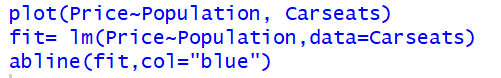


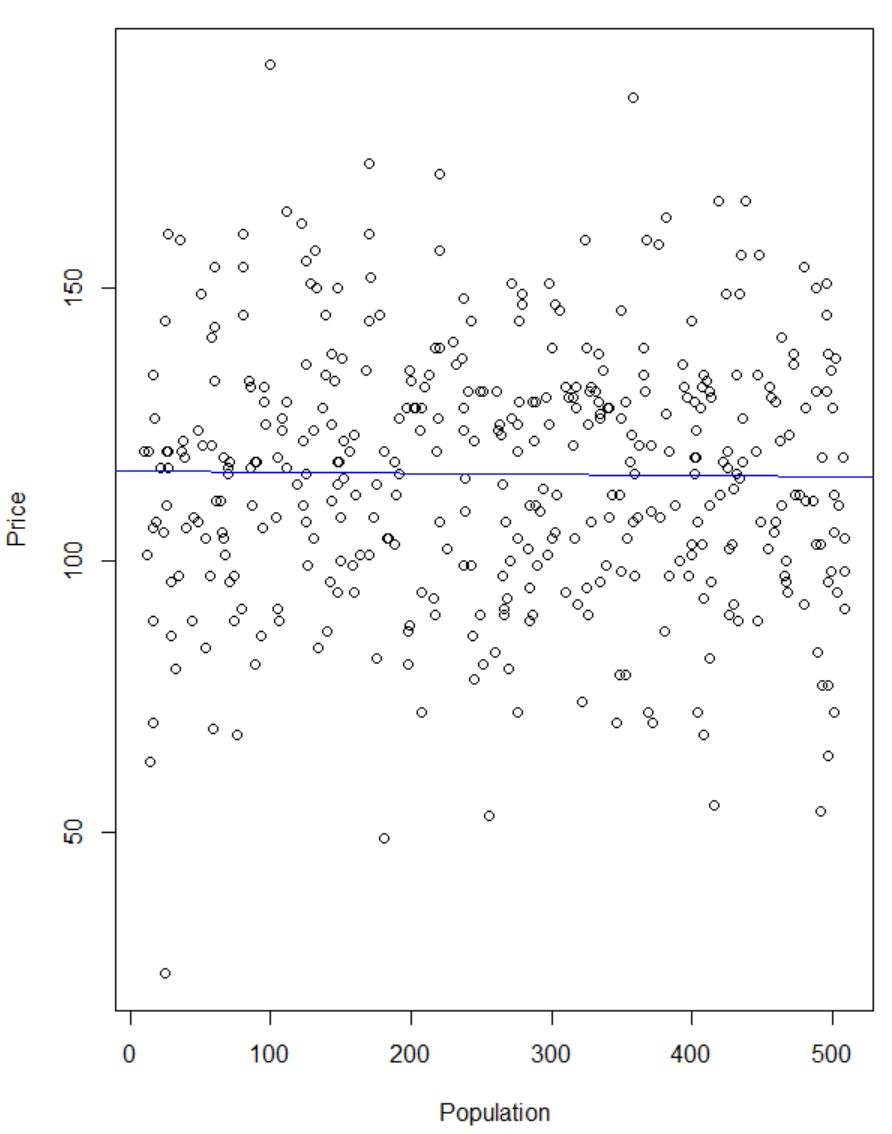
From my quick little summary analysis I found which variables are predictable variables and which are responses. The variables I chose to use as predictors are Price and Population, my response variable I chose to use is Sales, which tells us if the sell happened in America or not.

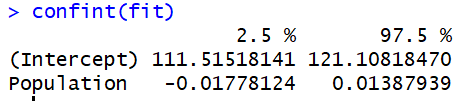
Before I continue on I’m going to do some more analysis on the data I have and in particular the 3 chosen variables.











From the plots and the measuring of the confidence interval for our variables we can see there might be a slight correlation between the two but not close to enough to think it was collinear in either a positive or negative connotation. Then the confidence interval at 95% for the coefficients is very slight with the difference between them being very small.

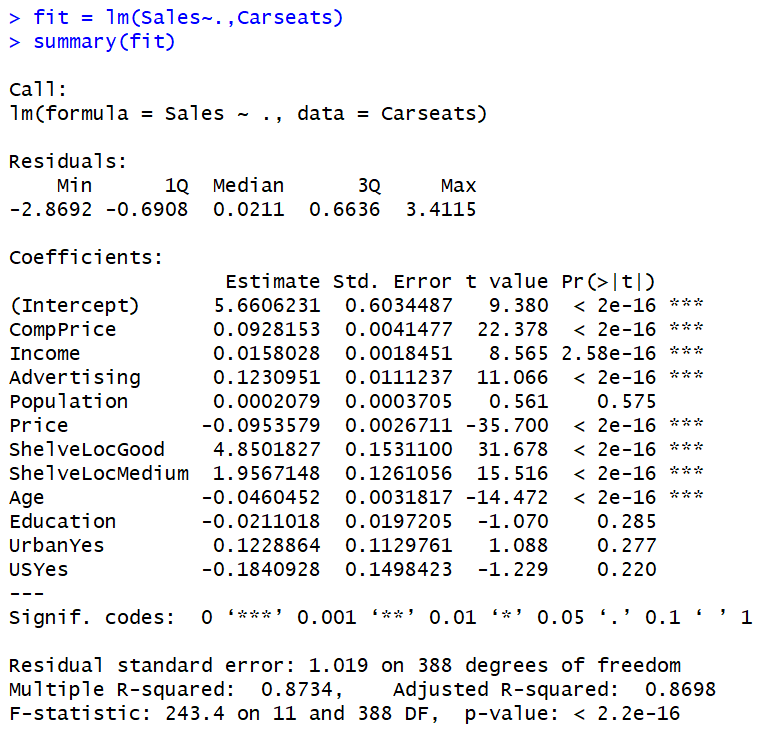
## Test the entire model (or significance of the model) using a global F-test. State both the null and alternative hypothesis.

I will be testing all the independent variables at one time and find out which variables appear to have a correlation with Sales and which ones do not.

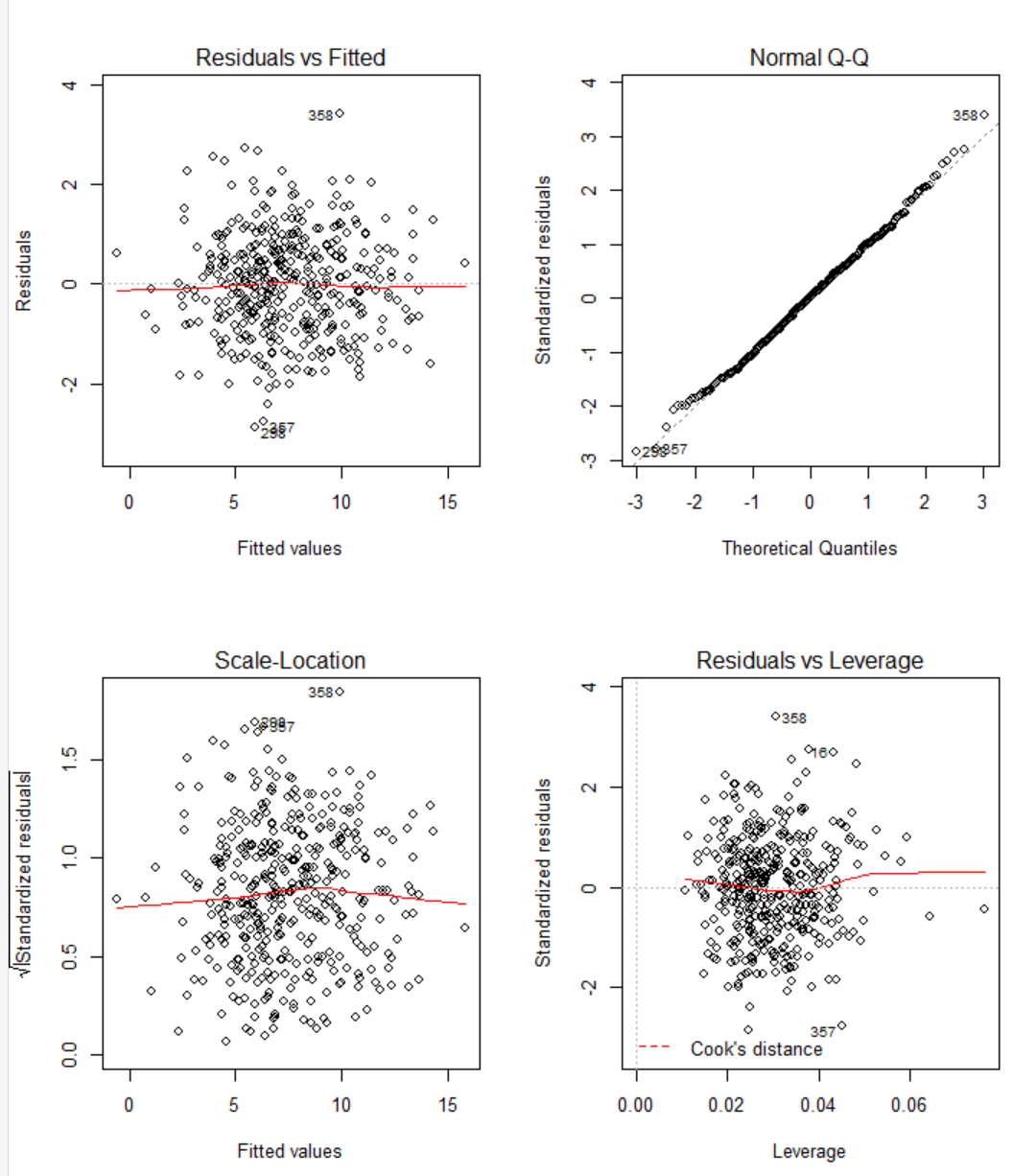
Null Hypothesis: There is no relationship between the dependent variable Sales and all the independent variables.

Alternative Hypothesis: There is a relationship with at least one of the independent variables and the dependent variable of Sales.

I will test this by running a linear model of all individual coefficients upon Sales.



From the linear model it is apparent that the variables population, education, Urban and US all don’t have any effect on the result of Sales but Comp Price, Income, Advertising, Price, SelveLocGood, ShelveLocMedium and Age all have an effect on the dependent variable. When examining the R-squared numbers those are both very pretty high so that is good because that is the percentage of variance explained. The F-statistic is what our result if we dropped out all the predictor variables from the model, since we can’t drop all of those variables and have a credible linear regression model it isn’t as valuable of information as it will be later on when less variables are being modeled. Next I plotted the model just to get more insights on the linear model.



The plot of the model would appear there is no correlation in the data but the model appears to fit fairly well from the Normal Q-Q plot except on the two ends where the data is skewed.

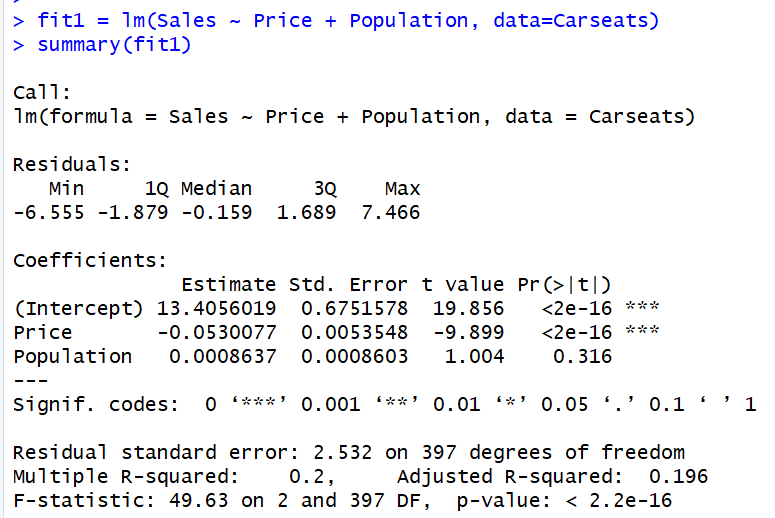
## Test significance of each explanatory variable (X). State both null and alternative hypothesis of each explanatory variable (list each in pairs)

**First pair of explanatory variables I’m going to choose are the ones I selected originally which are Price and Population:**

Null Hypothesis: There is no relationship between the Price and Population variables and the variable Sales.

Alternative Hypothesis: There is a relationship between the Price and Population variables and the variable Sales.

To test our hypothesis, first we must set up a linear model for our three variables in question.

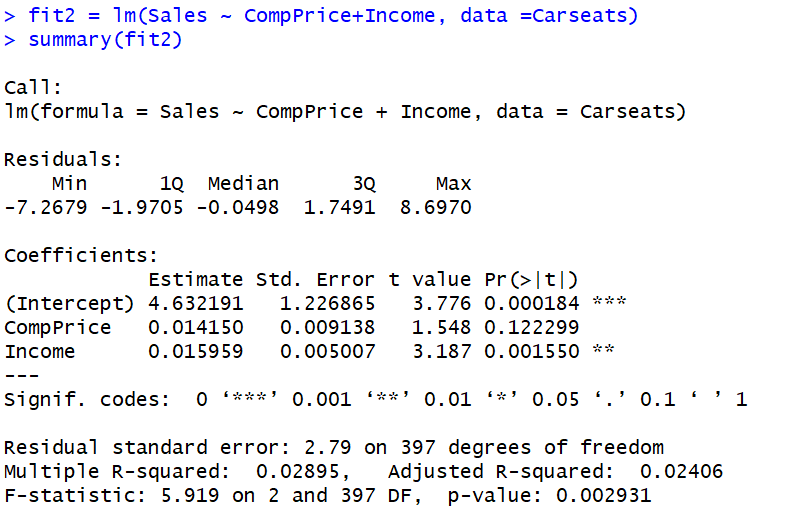


From our linear model we would conclude that there is no correlation between Sales and Population but it appears with the p-value for coefficient Price being less than 0.05 that we would reject the null hypothesis. When looking at the R-Squared figures and the F-Statistic and the overall p-value of the model we would again reject the null hypothesis because the R-squared values are so low, so is the F-statistic and the p-value once again is below the alpha level and would lead us to reject the null hypothesis.

**CompPrice and Income are the new independent variables:**

Null Hypothesis: There is no relationship between CompPrice and Income variables and the Sales variable.

Alternative Hypothesis: There is a relationship between the CompPrice and Income variables and the Sales variable.

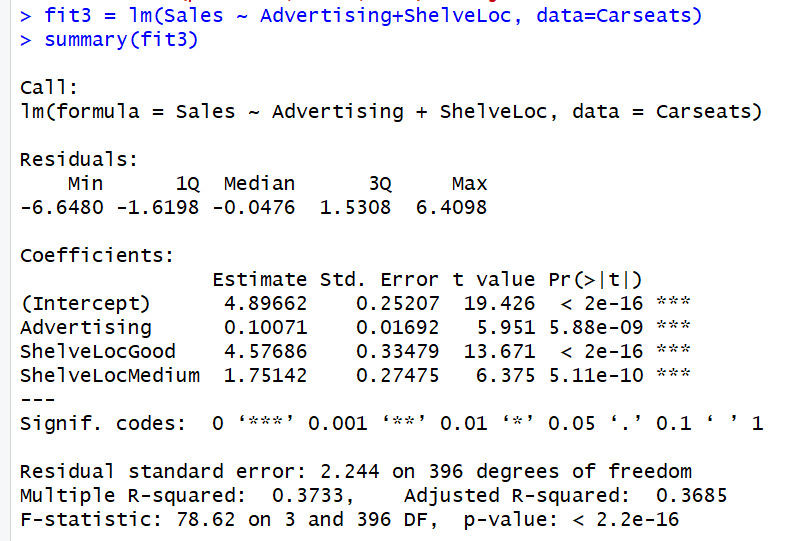


From our two new independent variables we would conclude CompPrice doesn’t have a relationship with the dependent variable of Sales but Income does. But from our first linear model with all the variables CompPrice and Income were flagged as variables which effected the dependent Sales variable. In this model our R-squared figures are both very small as well as our F-statistic which would mean it would be smart here to reject the null hypothesis for the alternative, that CompPrice and Income have a relationship with the Sales variable.

**Advertising and ShelveLoc are the new independent variables:**

Null Hypothesis: There is no relationship between the variables Advertising and ShelveLoc and the dependent variable of Sales.

Alternative Hypothesis: There is a relationship between the variables Advertising and ShelveLoc and the dependent variable of Sales.

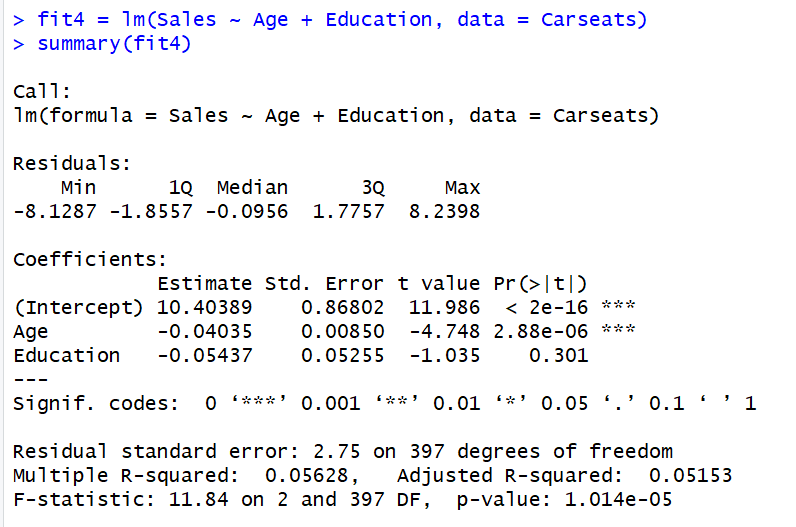


With these two new variables it is evident from our summary that both Advertising and ShelveLoc have an effect on our dependent variable, Sales. The coefficients individual p-values are both far below the alpha level of 0.05 so the null would be rejected for both of coefficients. The R-squared numbers are both small as well which lean toward rejecting our null and then looking at the F-statistic and the overall p-value it really solidifies that Advertising and ShelveLoc do appear to have some correlation.

**Age and Education are the new independent variables:**

Null Hypothesis: There is no relationship between the variables Age and Education and the dependent variable of Sales.

Alternative Hypothesis: There is a relationship between the two independent variables, Age and Education, and the dependent variable, Sales.

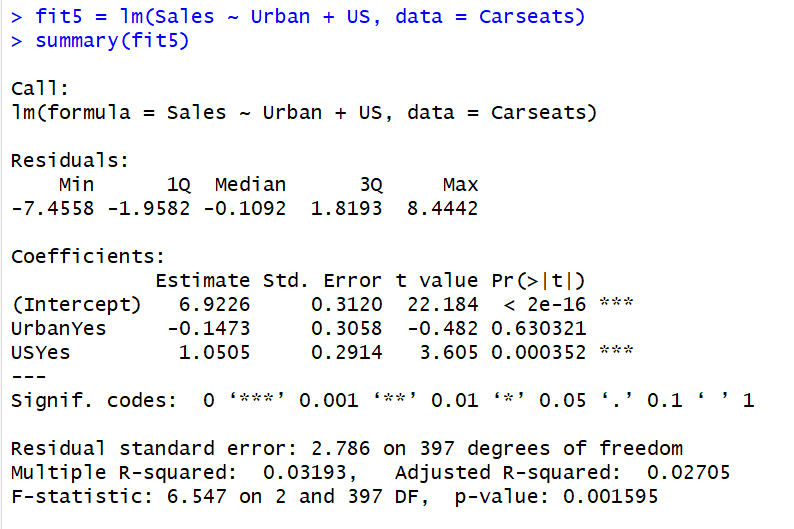


From our analysis with just the two independent variables chosen we got a result which was expected from our original linear model with all the independent variables. Age again looked like it actually does have a relationship with our dependent variable of Sales while Education looks like it doesn’t have a relationship with Sales. In this situation we would reject the null hypothesis for the Age variable but we would not do so for the Education variable. For the overall model, we would reject the null hypothesis even with both R-squared figures above 0.5 because of a fairly low F-statistic and a very small p-value.

**Last two independent variables: Urban and US**

Null Hypothesis: There is no relationship between the two independent variables, Urban and US, and the dependent variable, Sales.

Alternative Hypothesis: There is a relationship between the two independent variables, Urban and US, and the dependent variable, Sales.



Comparing our newest linear model to our original model with all the independent variables I see a difference in the result for the independent variable of US. In the original model from earlier it’s p-value would indicate there is no relationship between it and Sales while in our new model there does seem to be a relationship between the two. Urban still appears to be a variable that has no influence on the Sales variable though so with this result we would reject the null hypothesis for US but for Urban our null hypothesis would hold true. For the overall model it appears there the variables have a relationship with Sales thus the overall linear model would state that we reject the null hypothesis.

After testing each coefficient all together and then in pairs it was apparent which ones indeed had influence and which ones did not.

CompPrice – Had influence on Sales because in the two models it was in the p-value was below 0.05 and it makes sense that it would because CompPrice and Sales are something that usually have a relationship in any business venture.

Income – Like CompPrice, Income also had an influence on Sales in both models it was in and it makes sense because the higher the Income the higher Sales the Carseats had.

Advertising – Also ended up having a relationship with Sales, it appeared the more Advertising the company did the more Sales they made.

Population – In all models Population was in there didn’t appear to be any relationship between it and Sales.

Price – Price is a variable which appeared to have a relationship with Sales. It appeared to be a negative relationship meaning the higher the Sales the lower the Price but still a correlation nonetheless.

ShelveLoc – ShelveLoc at status Good and Medium appeared to have a relationship with Sales. The ShelveLoc has only 3 options, 1 – Bad, 2 – Medium, and 3 – Good, for the options it has so it can be an easy a bit harder to validate the relationship based on how few of options it has but from our analysis it would appear when the ShelveLoc is either Good or Medium then you would get a result of higher Sales.

Age – There appears to be a relationship between Age and Sales. From our two models with Age it would appear that Age has a negative correlation with Sales.

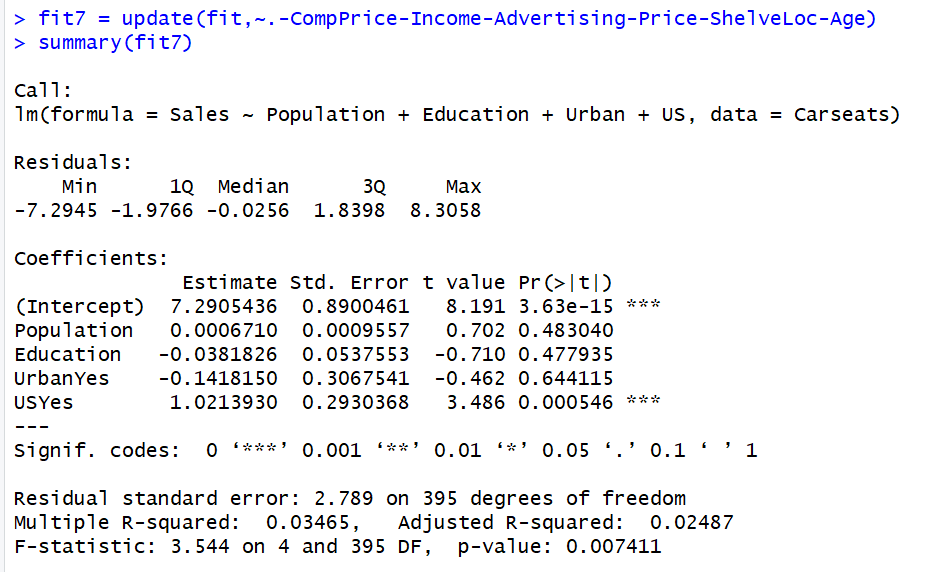
Education – Education appears to not have any correlation between it and Sales and is based numerically by the number of years someone was in school for.

Urban – There doesn’t appear to be a relationship between it Urban and Sales. The variable Urban is either a 1 or 2 based on if the buyer is in an urban population or is not.

US – When all independent variables are being analyzed together there doesn’t appear to be a relationship between US and Sales, but when coupled with one other variable like in our last model when US is and independent variable along with Urban it appears to have a relationship with the Sales variable. US is measured like Urban is, by either a 1 or 2 whether the customer is in America or is not.

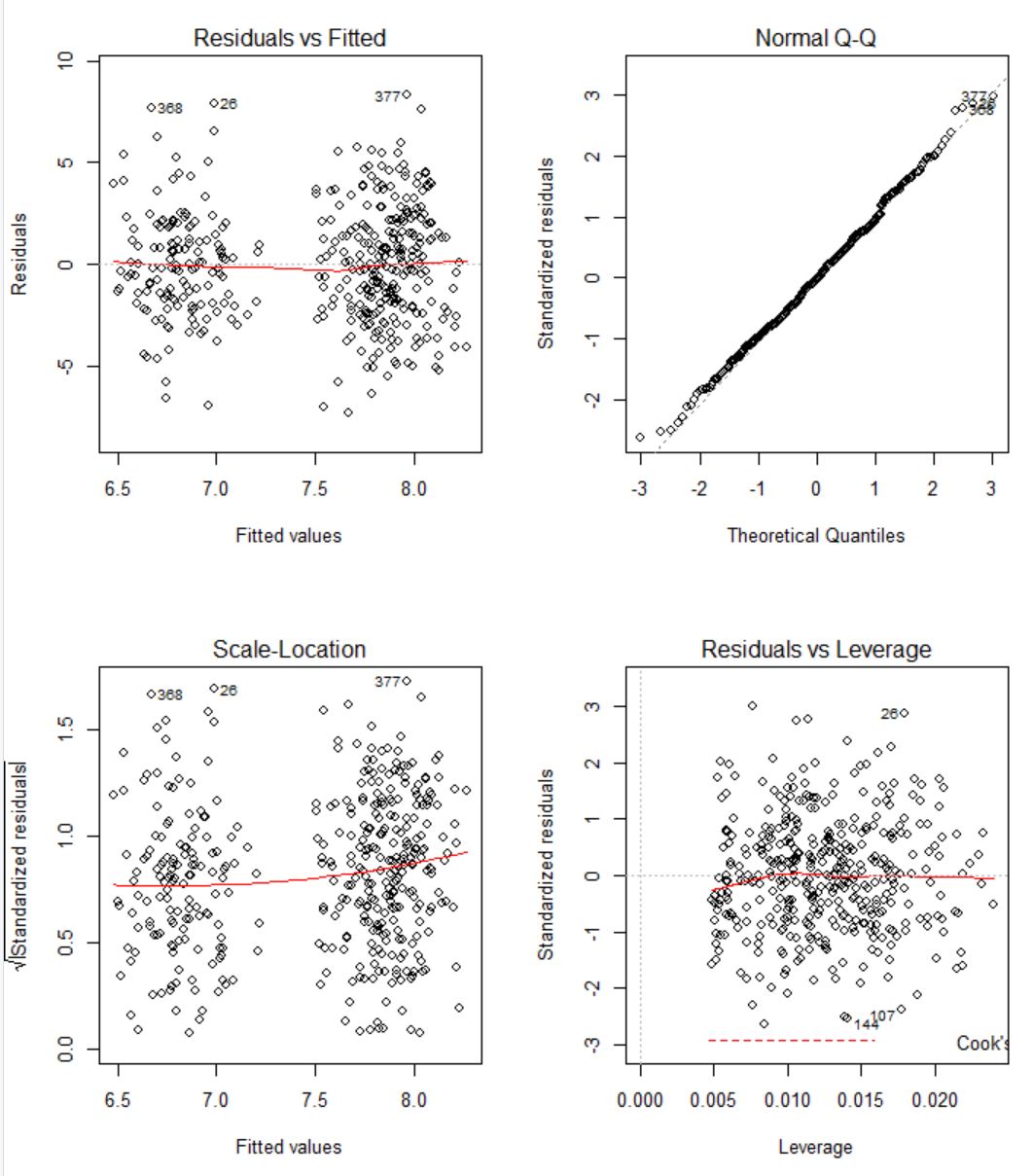
## Try a different model. Which model fits the data better? How do you select the model? Explain your answers.

For my new model I’m going to only select the independent variables which appeared to have no relationship with the dependent variable of Sales, I’m going to be looking for the variables with their correlation possibly changing either to be less influential on Sales or more. My expectation is that some of the variables will stay as non-correlated independent variables but I believe there will be at least two which become correlated with Sales. The independent variables I will be using are Population, Education, Urban and US.



From the look of our quick summary of our new model it appears I was wrong about at least two changing and becoming correlated with Sales, only one US had it’s p-value fall below the alpha of 0.05. Besides US all of the p-values actually grew, solidifying the null hypothesis of there being no correlation between the independent variables and the dependent variable. Because US did however have a correlation with it and Sales we can see that the overall model itself did have to reject the null hypothesis. From here I’m going to plot this model.



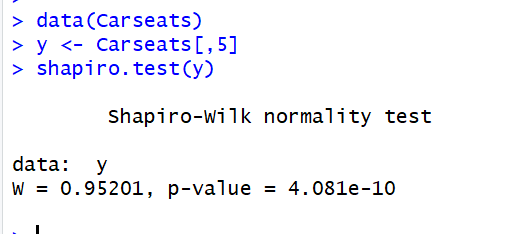


From our plots we see that our data is quite sporadic and there seems to be no mappable correlation positively or negatively. But in the Normal Q-Q plot we see that our data lines up very well with our line which makes it appear that the model fits very well. The fit of this model appears to fit better than the original model because there appears to be less data skewing away from the regression line from the Normal Q-Q.

## Are there any other tests should you include in this analysis?

The tests that I should include is a Shapiro-Wilk test to see if the data has normality or not. Doing this will help an analyst figure out what kind of tests will be best to use, either a parametric or non-parametric statistical test could work for your data but it takes knowing the normality of your data to know which of these tests fit best. Also testing collinearity between some of the variables so that way I wouldn’t use them in the same linear model because in a linear model we don’t want two independent variables with a very high level of collinearity positively or negatively.

If I wanted to test the normality of the data I could easily do that in RStudio with the Shapiro-Wilk test:



With the p-value being less than 0.05 we know the Carseats data is not normal, so it would be best to run non-parametric statistical tests like, 1-sample Wilcoxon signed rank test, Friedman test, Kruskal-Wallis test or Mann-Whitney test to run on the data for further analysis.

## Address any other concerns you might have.

The concerns I might have from what I’ve analyzed is I never figured out which variable has the most influence on my dependent variable of Sales so I would want to distinguish which one that is. Also which variables have a collinear relationship so I can make sure I don’t use them in the same model when trying to find a relationship between my dependent variable and the different independent variables.

## Summary

From this assignment I got to use look deeper into linear regression with multiple linear regression. Within multiple linear regression we are able to measure multiple variables and their effect on a dependent variable which I find to be a very powerful tool of analysis. Doing this in RStudio as well and learning the different things I’m able to do is also very unique because there are so many other things I saw from research which I researched but didn’t use. Overall, it was a great week of learning about multiple linear regression and how it can be done in RStudio.