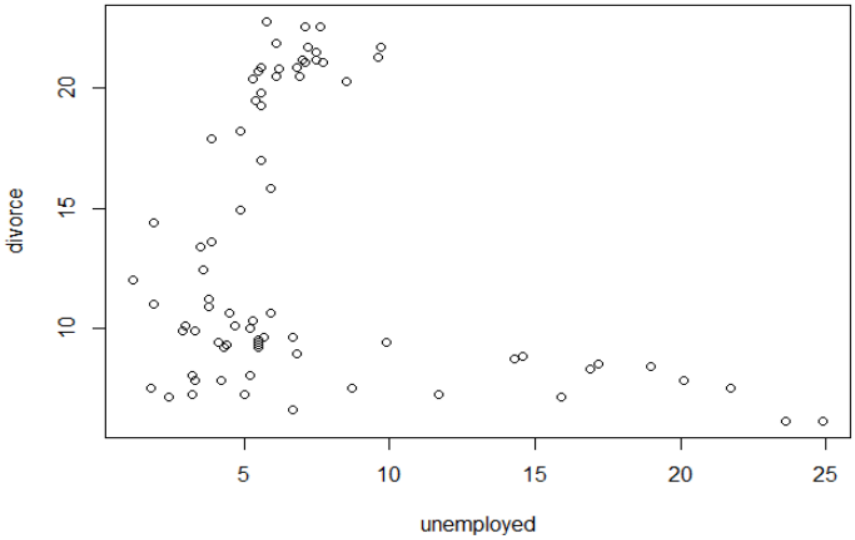
Jonothan Meyer
Linear Regression
02/26/21
Project 4

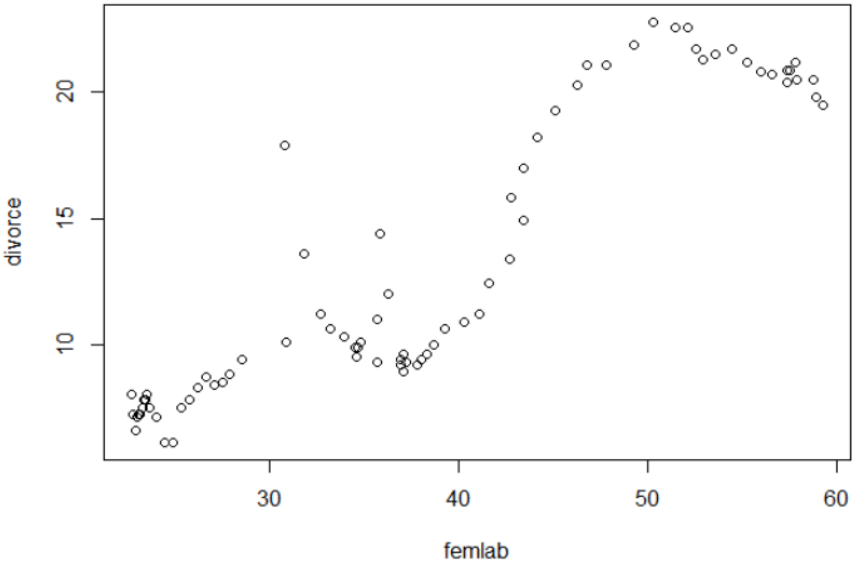
Exercise 1:
(1) Construct scatterplots of the response variable `divorce' against all other predictor variables except for `year.' Use these scatterplots to assess the criterion for a regression line in each case.

Scatterplot: Predictor: Unemployed, Response: Divorce



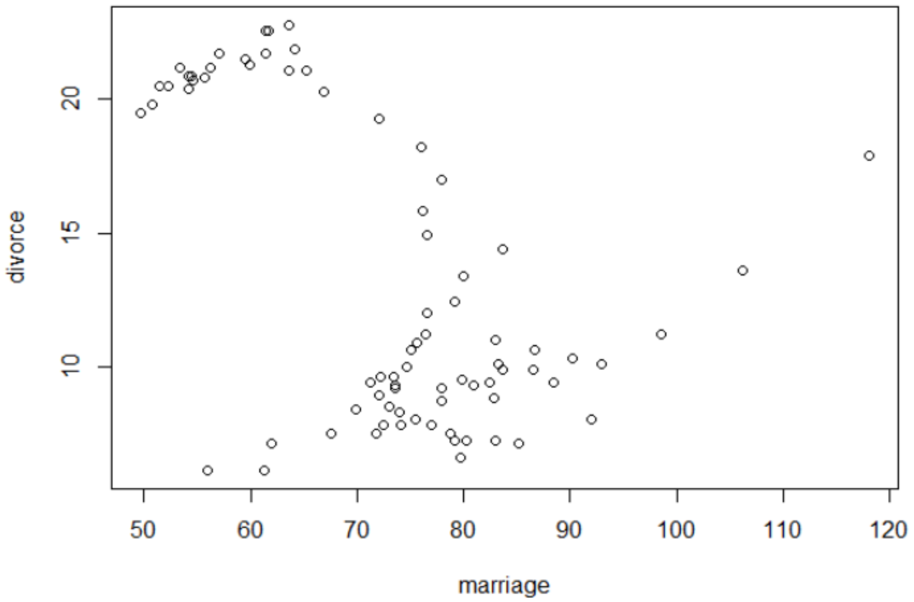
When plotting unemployment as the predictor against divorce it's clear just by looking at the scatterplot that a linear model would probably be inappropriate. However, if we were to examine further and deem it appropriate to filter out unemployed data greater than 10% then a linear model could work quite well.

Predictor: Female Participation in Labor Force Percentage (Femlab), Response: Divorce



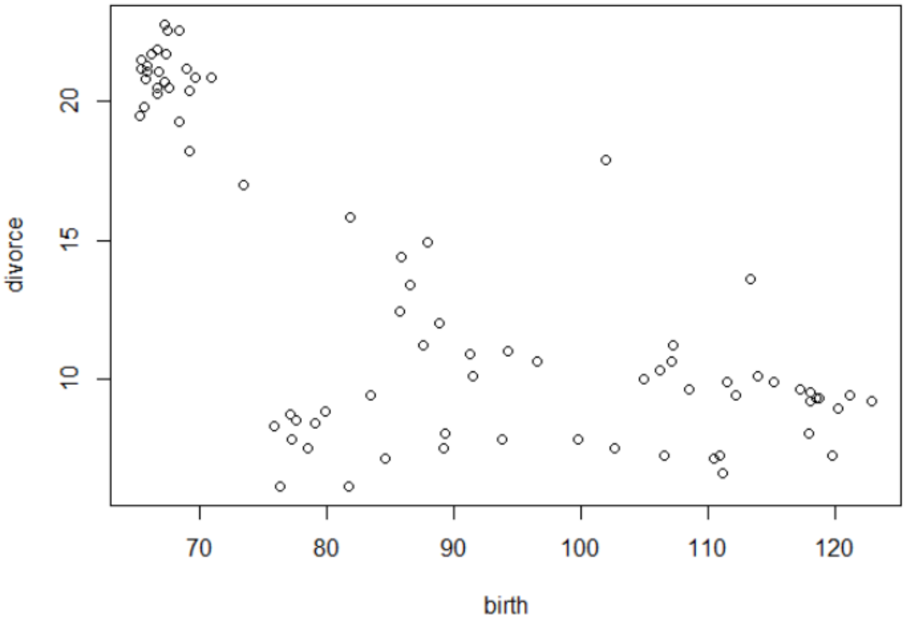
Initial impressions lead me to think that simple linear regression could be a reliable way to model this data. But further investigation would be needed on the distribution of residuals in order to make that determination since the data points look more like a polynomial function.

Predictor: Marriages per 1000 Unmarried Women, Response: Divorce



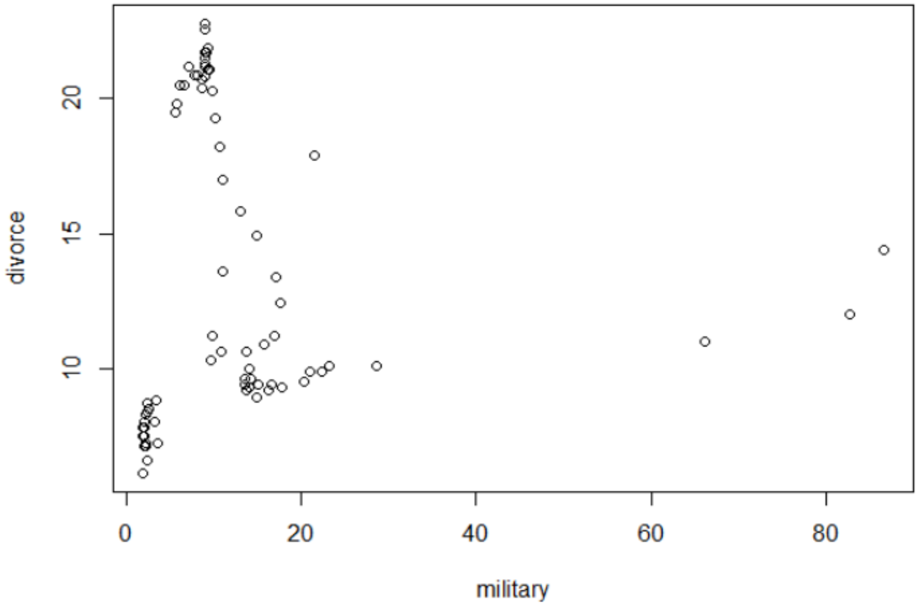
This scatterplot indicates a scenario similar to what was observed when plotting unemployment against divorce. The clusters of data located in the top left corner, and bottom middle don't give a clear indication that simple linear regression should be used. The least squares regression line would have a negative slope with the current data, despite appearing positive based on the eye test. As with the first scenario if we deemed it admissible to filter out marriage data less than 70 married women per 1000 then the regression line would be positive and seemingly appropriately model the data.

Predictor: Births per 1000 Women, Response: Divorce



Although this graph indicates that variance would be rather high, this plot looks like a good candidate to be modeled by simple linear regression. I would expect the least squares regression line to be negative and have a normal, wide, and equal distribution of residuals. More testing is needed to determine if simple linear regression would be the best route to go, but thus far, this predictor variable looks to be the best available in the data set to be modeled in this manner.

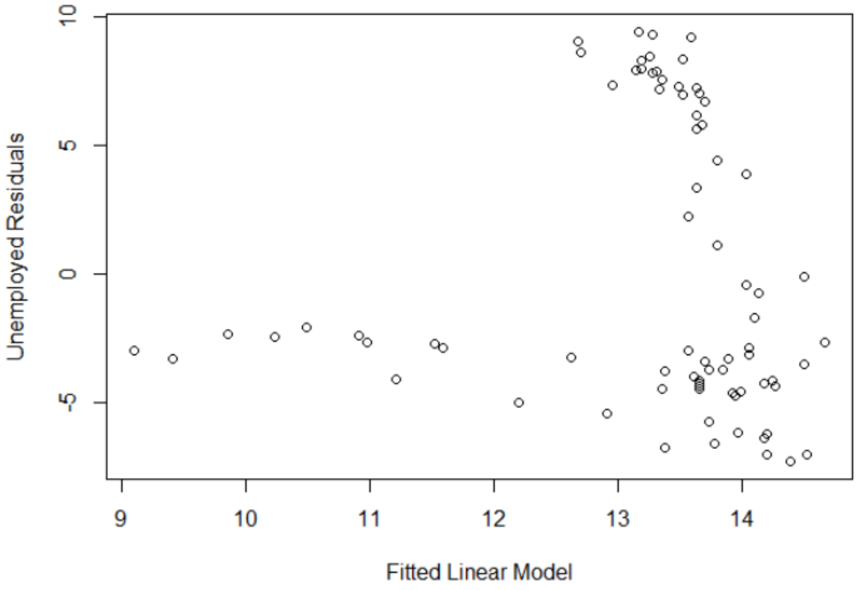
Predictor: Military Personal per 1000 Population, Response: Divorce



With this plot the slope of the least squares regression line could be nearly flat, or extremely steep if a few outliers were filtered out. There does appear to be dependence between data points, especially for a military value less than 30, and for this reason I don't think the linear regression model is appropriate. The residual data points would probably meet the criteria for this model, but because of the high variance in slope depending on which data points were included we should be skeptical about using this model.

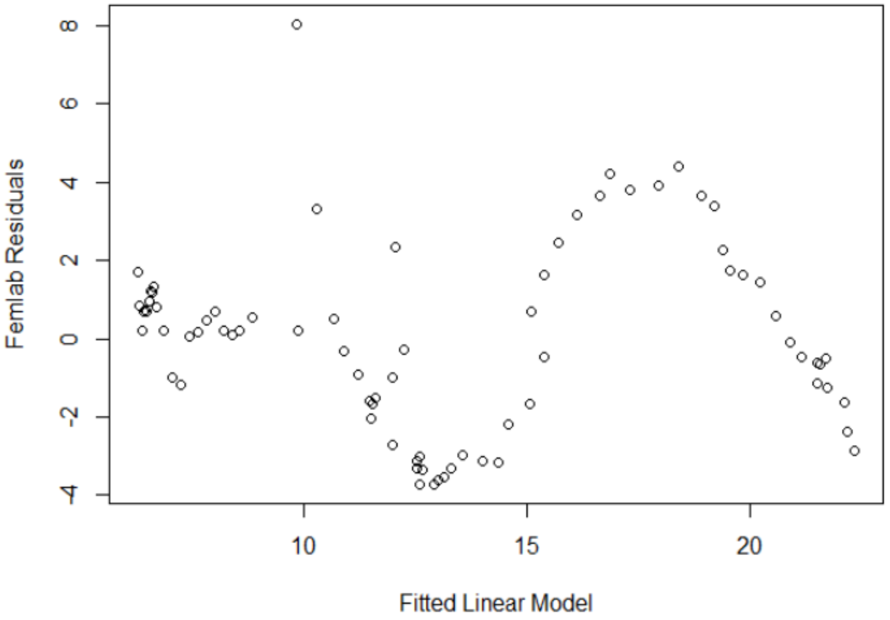
(2) Construct plots of the residual errors against the fitted value of linear models using response variable `divorce' and each predictor variable except for `year.' Use these plots to assess the assumption of independent error terms and the constant variance assumption.

Unemployment Rate Residuals



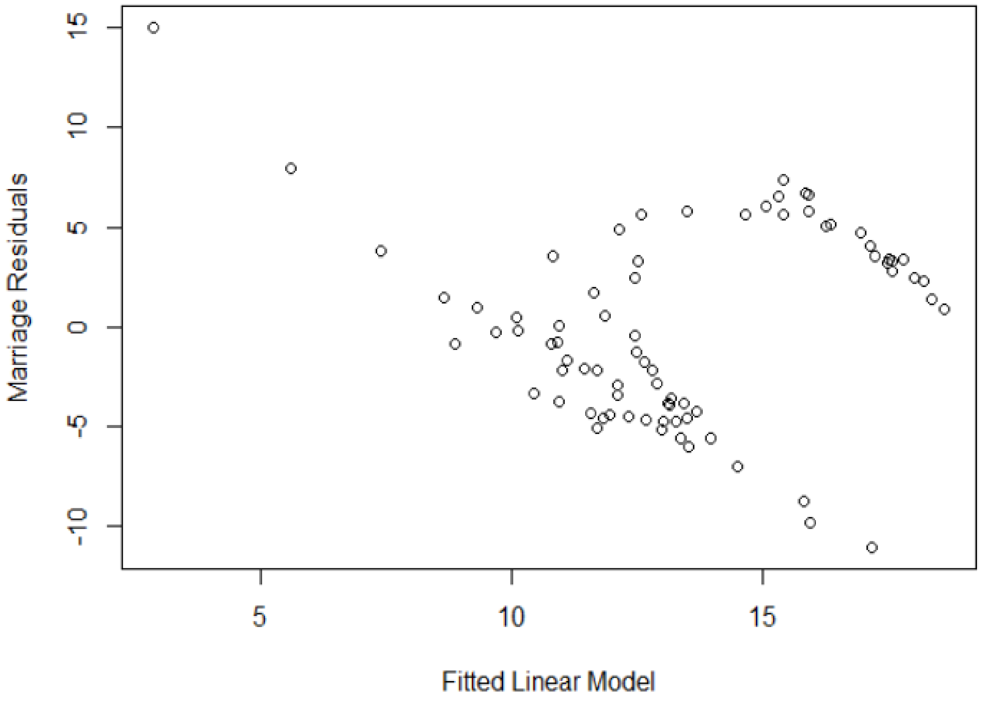
This plot demonstrates clearly that a linear model is probably inappropriate for the unemployment data against divorce. The data points do not appear at all independent, as each subsequent residual appears to be influenced by the previous. Without outlier constraints I would abandon this model after this point.

Female Participation in Labor Force (Femlab) Residuals



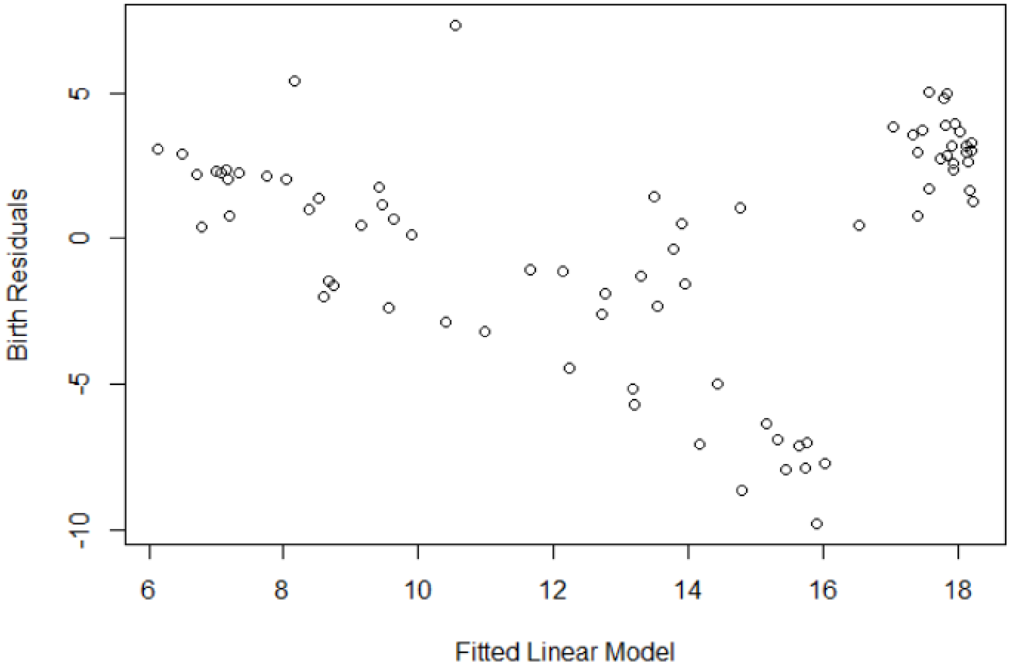
This graph also demonstrates violations in our necessary assumptions to use this model. These points look even more dependent than the previous graph. Connecting them would resemble smooth, low degree polynomial function. Additionally, they do not appear to be consistently distributed. After looking at this graph I would rule out this model to be used and explore other variables.

Marriages per 1000 Women Residuals

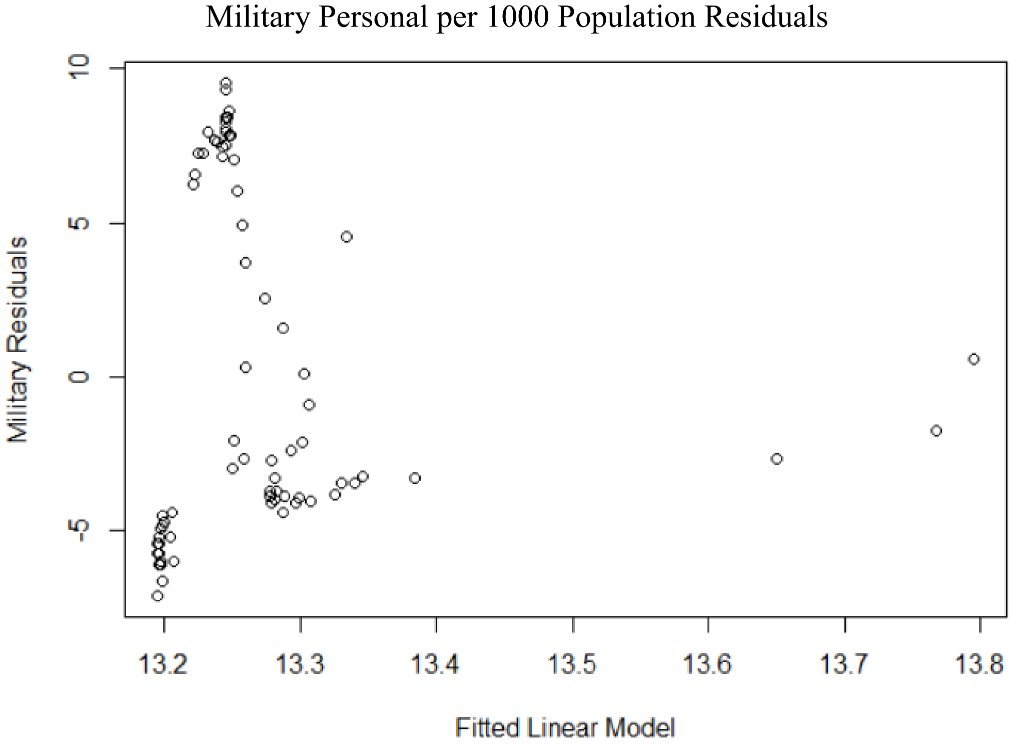


These data points do not appear to be independent, but do appear to be more consistently distributed. However, they are not consistently distributed to the point where I think linear regression is an appropriate model.

Births per 1000 Women Residuals



The birth residuals do appear to be somewhat independent and have somewhat consistent variance, other than a couple groupings of data. Based off this graph I think it would be appropriate to further consider these variables for constructing a mathematical model, and proceed towards further plotting to strengthen my confidence in that assessment.



Since the slope of the least squares regression line for the scatterplot of these variables is near zero the residual graph appears to be almost identical to it. The points do not appear independent, but do seem to have consistent variance. Linear regression might be an appropriate model if the assumptions were further investigated and confirmed with further testing.

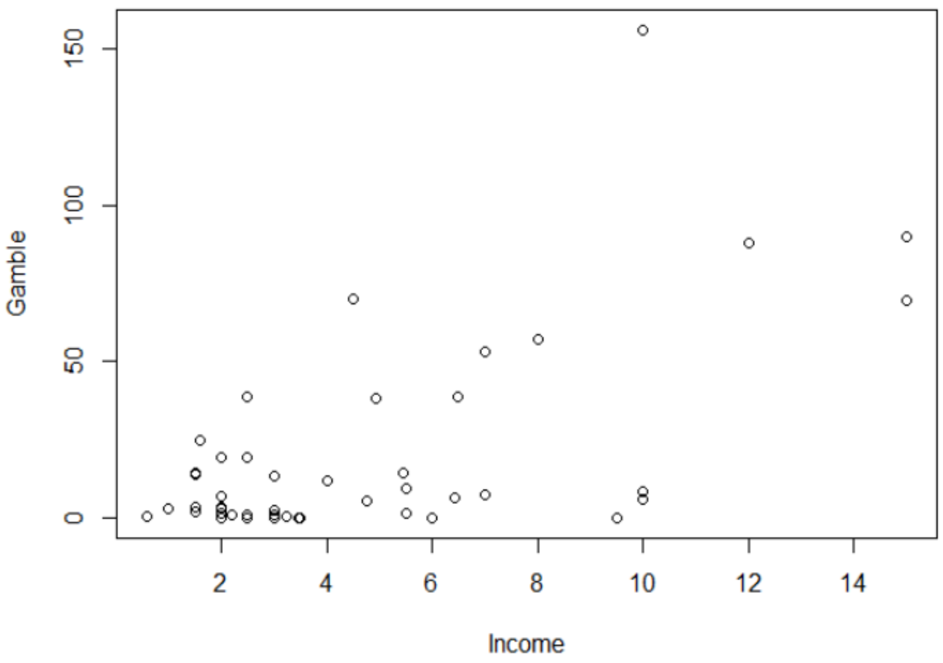
(3) Indicate which predictor variables are most suspect in terms of not satisfying the assumptions of the simple linear regression model linking that prediction variable to `divorce.'

Unemployed, Femlab, and Marriges are the most suspect predictor variables in terms of not satisfying the assumptions of the simple linear regression model that links that prediction variable to 'divorce'. Of those three I would say the 'Femlab' predictor was the greatest violator of the assumption of residual independence, and the 'Unemployed' predictor was the greatest violator of the assumption of constant variance of residuals. 'Births vs. Divorce' performed the best in terms of satisfying the necessary assumptions, and would be the best to build a linear regression model.

Exercise 2

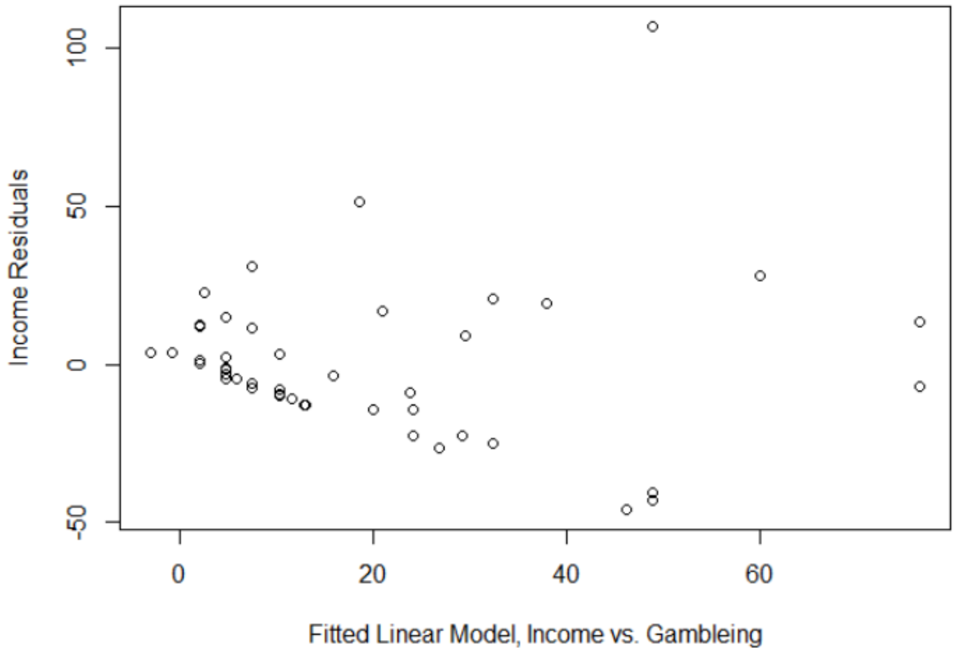
(1) Fit a model with `gamble' as the response and the `income ' as the predictor.

Scatterplot: Predictor: Income, Response: Gamble



(a) Check the constant variance assumption for the errors.

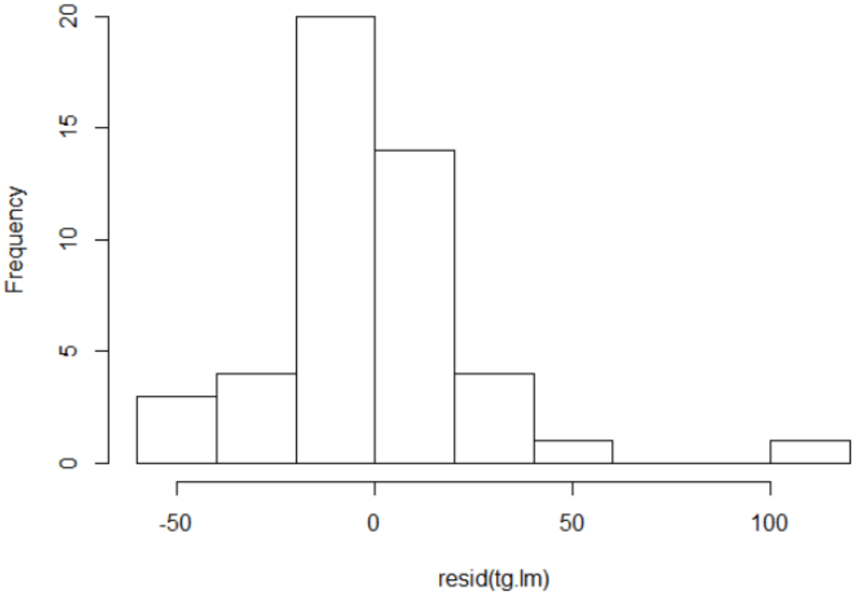
Income Residuals



It does not appear as though the constant variance assumption associated with the simple linear regression model is satisfied. The data points appear to have greater variance the higher the income level.

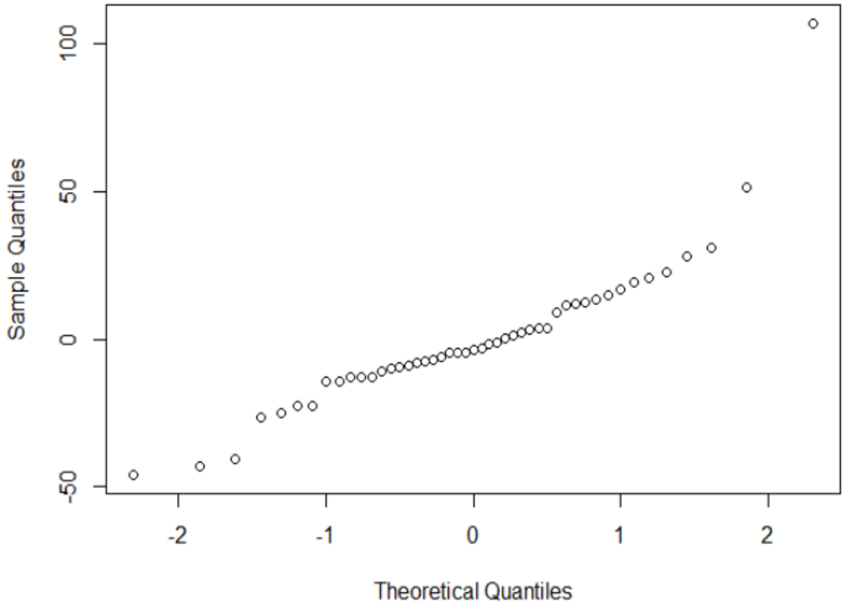
(b) Check the normality assumption.

Histogram of Residuals



Although we don't put too much stock in the histogram of residuals for a data set this small, it's still good practice to look at to identify a general normal distribution. The histogram of residuals appears to be somewhat normally distributed.

Normal QQ-Plot

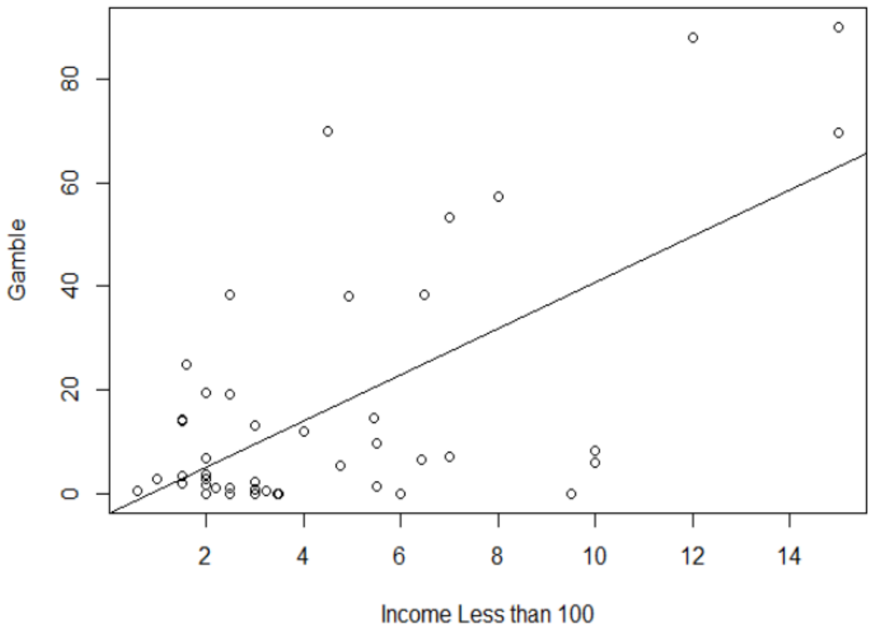


Based upon my current data acumen the normal QQ-Plot of the residuals of income vs. gambling from the teen gambling data does not appear to satisfy the linear model assumption of normality.

Taking both of these graphs into account my conclusion would be that the assumptions of normality necessary to conduct a linear model are not met when using income as the predictor against gambling for the teen gambling data set.

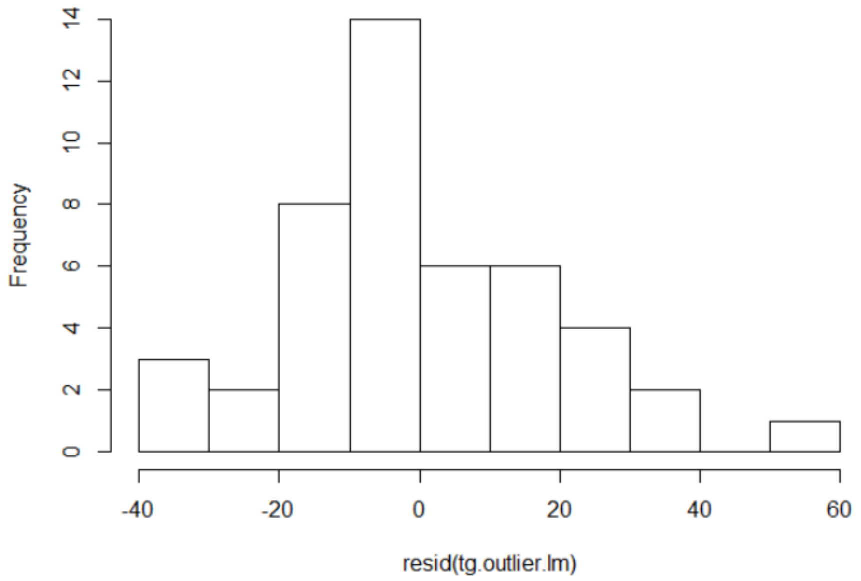
(c) Check for outliers.
An outlier located roughly at point ~(10, 150) was removed by filtering the income data to only include data points of less than 100.

Scatterplot (w/least squares regression line): Predictor: Income <100, Response: Gamble



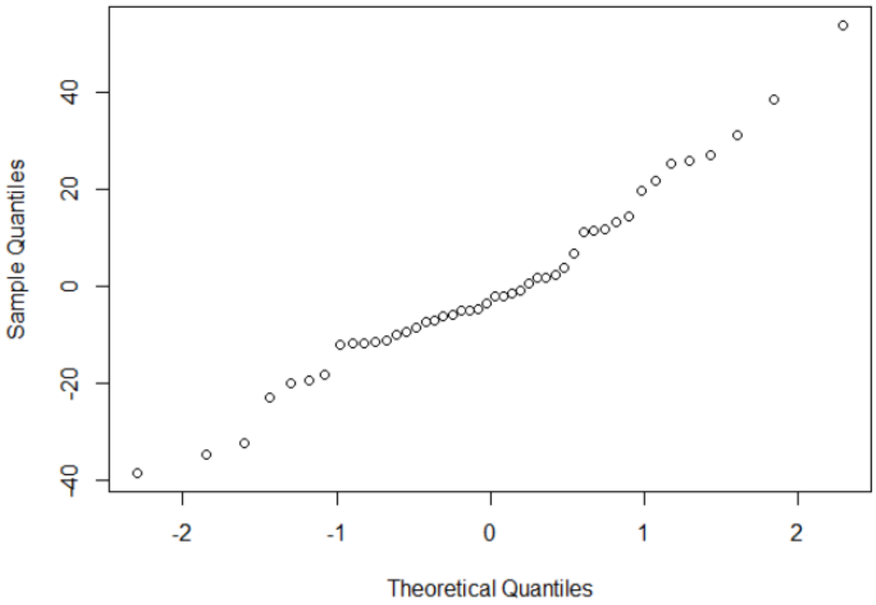
(d) Check the structure of the relationship between the predictors and the response.

Histogram of Residuals (with data that excludes outliers)



The residuals from the least square regression line of the teen gambling data that excludes outliers appears to be basically normally distributed.

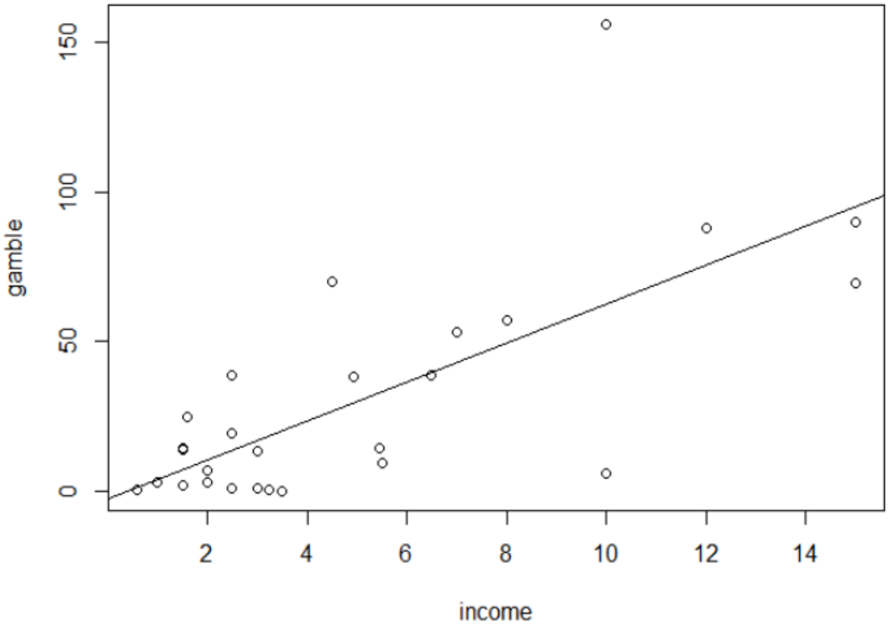
Normal QQ-Plot of Residuals (with data that excludes outliers)



The normal QQ-plot of residuals does not appear to fully meet the assumptions of normality for linear regression. This is primarily because the variance does not appear to be consistent.

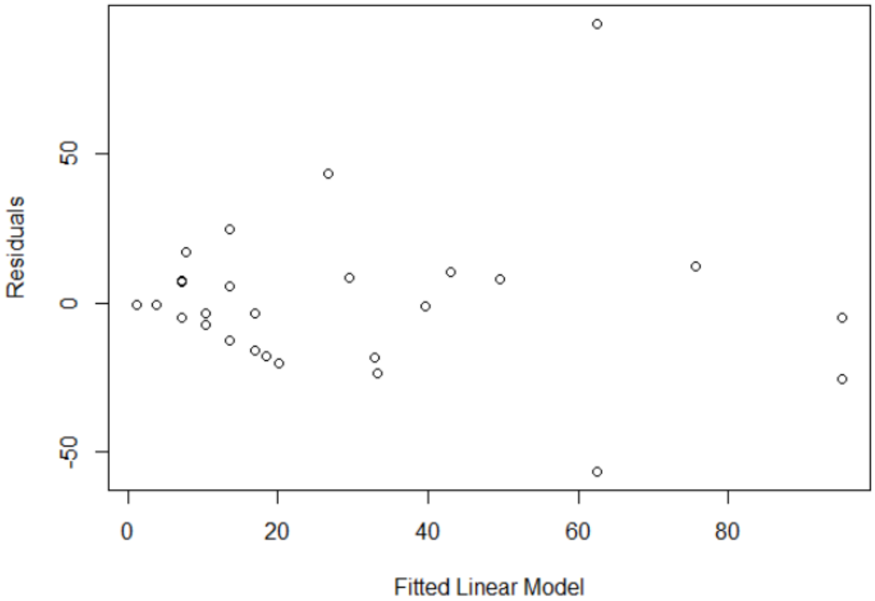
(2) Complete the previous part, but only for males in the original data set.

Scatterplot (w/Least Squares Line): Predictor: Male Income, Response: Gamble



(a) Check the constant variance assumption for the errors.

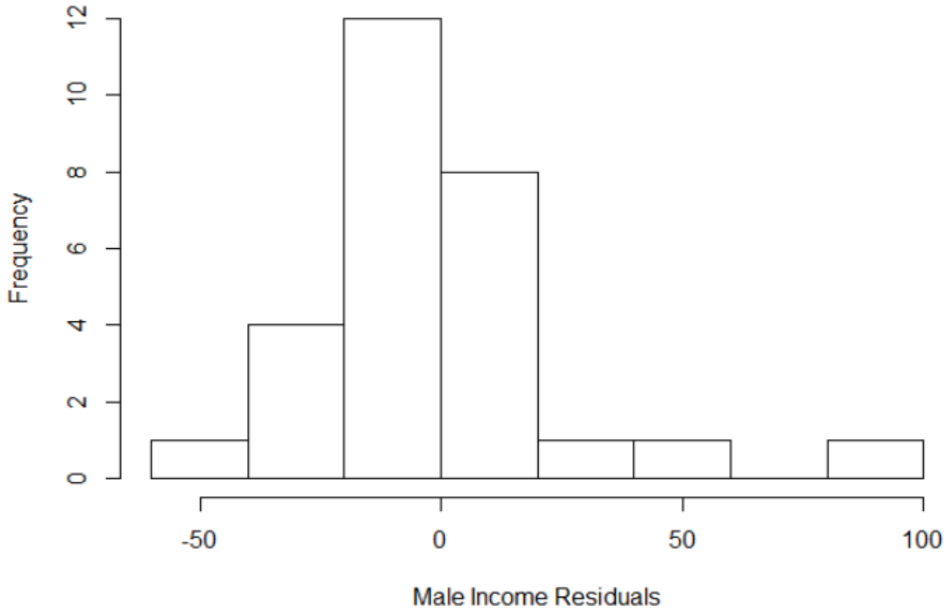
Income Residuals for Male Teen Gamblers



The income residuals for the linear model of male teen gamblers appears to not have consistent variance. This is mainly because a couple data points are outliers that give the impression that the varience expands as income increases. Otherwise the residuals appear basically normally consistent and are independent.

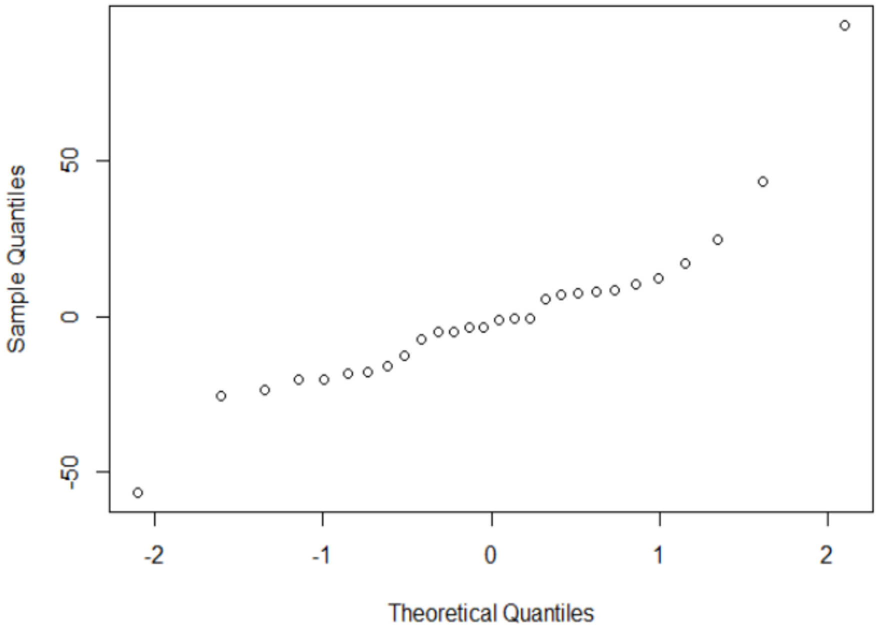
(b) Check the normality assumption.

Histogram of Male Income Residuals



The histogram of income residuals for male teen gamblers appears to be approximately normally distributed, especially considering the size of the same.

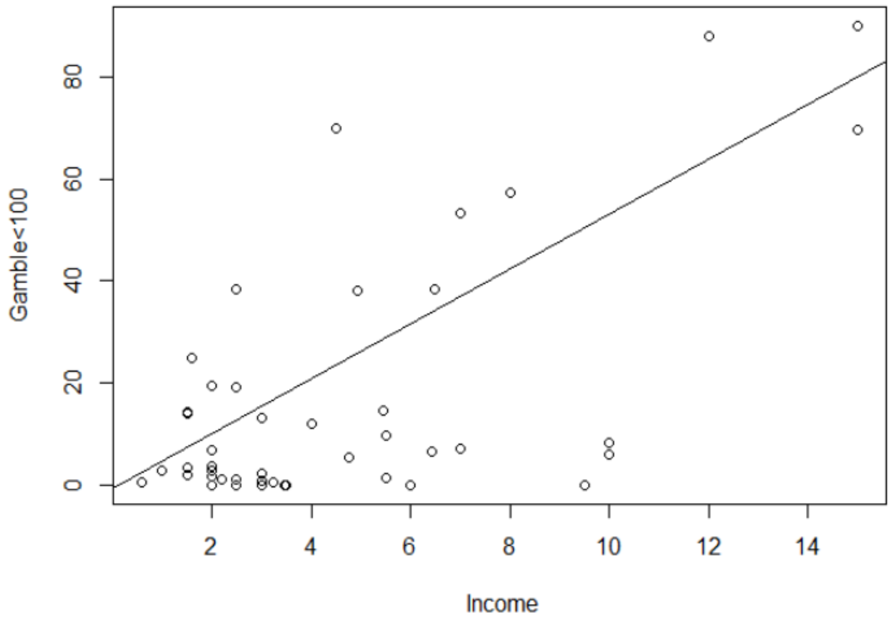
Normal QQ-Plot of Male Income Residuals



Based on the QQ-Plot, the income residuals of teen male gamblers appears to be approximately normally distributed.

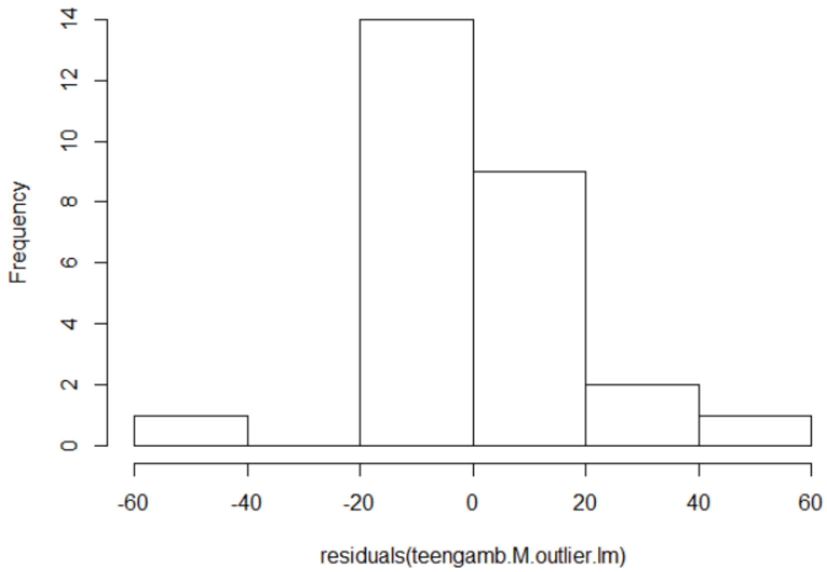
(c) Check for outliers.
An outlier was identified located at ~(10, 150). The linear model produced an r^2 of 0.50 with the outlier contained in the data. After it was removed the r^2 value increased to 0.58. This gives one indication that the linear regression model is a better fit with the outlier removed.

Scatter Plot (with outlier removed and least square regression line): Predictor: Male Income, Response: Gamble



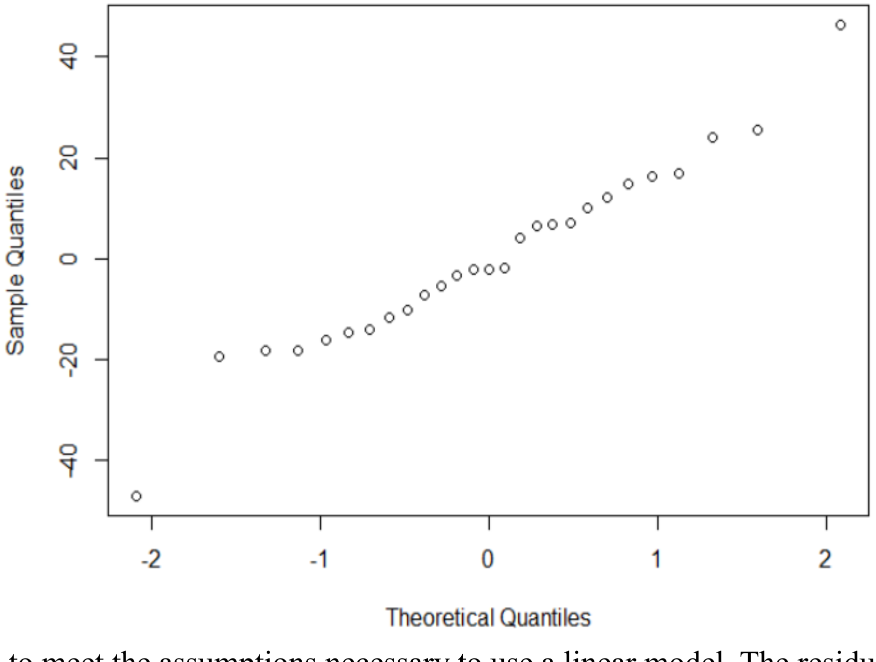
(d) Check the structure of the relationship between the predictors and the response.

Histogram of Residuals (with data that excludes outliers)



The residuals with the outlier data point removed appear to be approximately normally distributed.

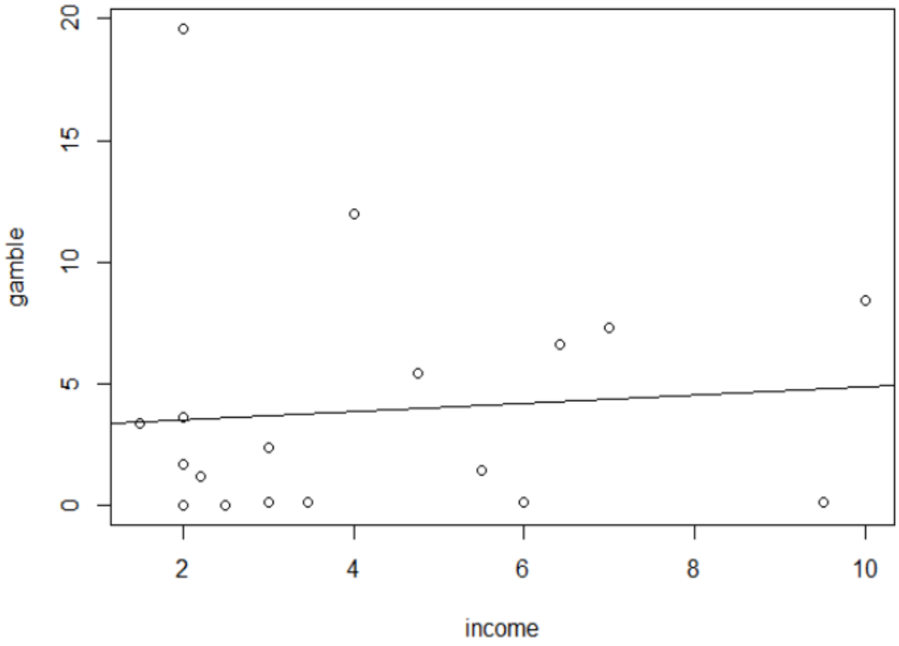
Normal QQ-Plot of Residuals without Outlier Data Point



The data appears to meet the assumptions necessary to use a linear model. The residuals appear to be normally distributed, independent, and maintain consistent variance.

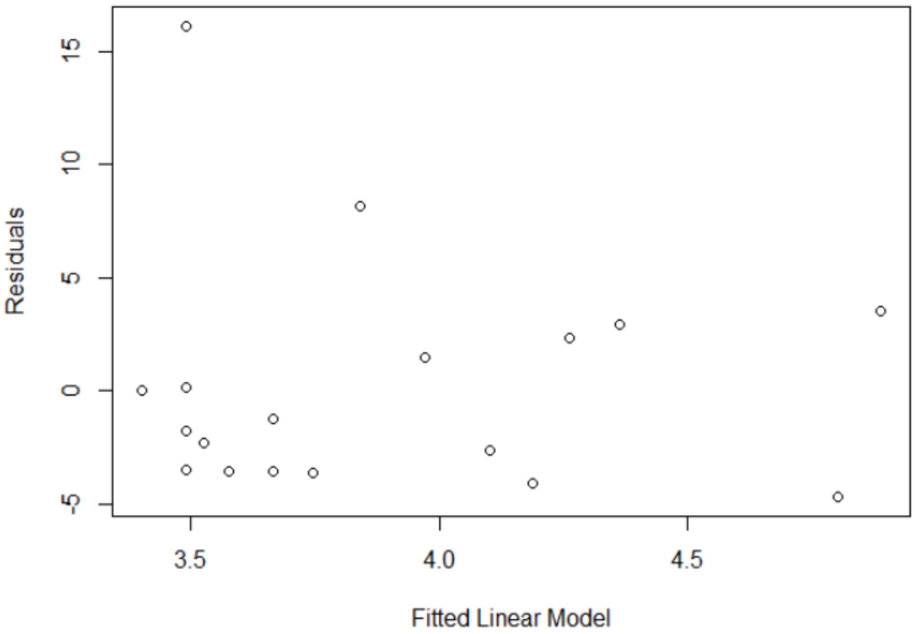
(3) Complete the previous part, but only for females in the original data set.

Scatterplot (w/Least Squares Line): Predictor: Female Income, Response: Gamble



(a) Check the constant variance assumption for the errors.

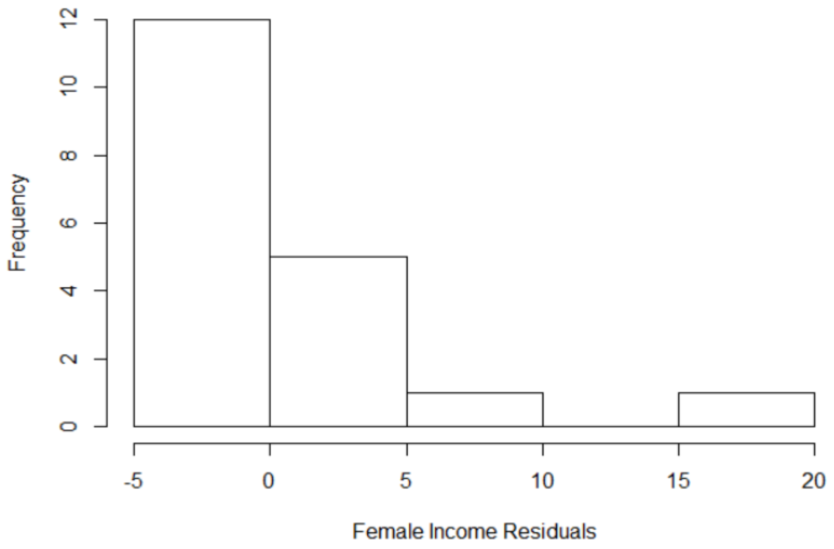
Income Residuals for Female Teen Gamblers



Variance does appear to be constant for the income residual errors for female teen gamblers. One outlier skews the data a bit, but otherwise it looks as though the assumption is met.

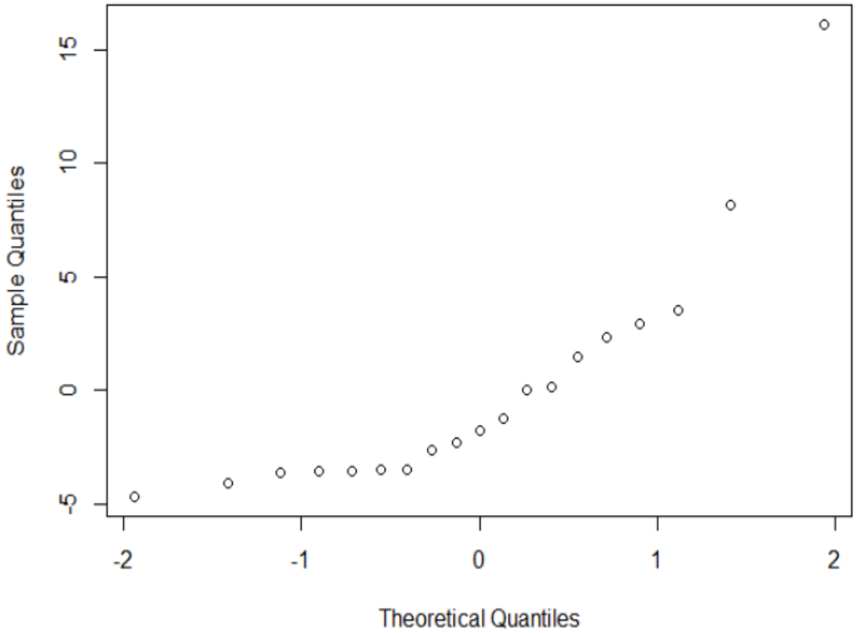
(b) Check the normality assumption.

Histogram of Female Income Residuals



This histogram does not appear normal at all. However, since the data set is so small we don't have to read too deeply into this graph, but should investigate a bit further into a normal QQ graph.

Normal QQ-Plot of Female Residuals

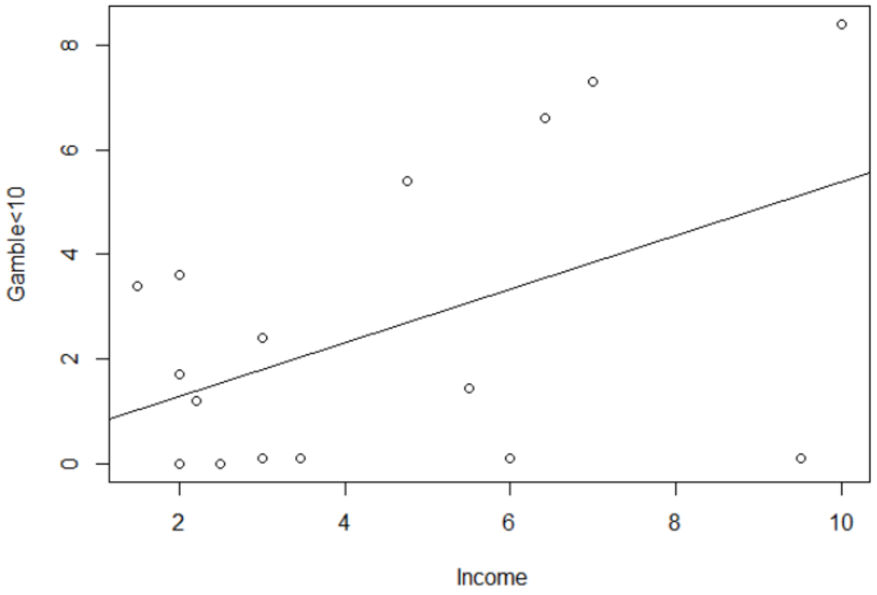


The normal QQ-Plot of female residuals shows that this data does not meet he assumptions of normality and constant variance. Without removing an outlier I don't believe linear regression is a good model to represent this data.

(c) Check for outliers.

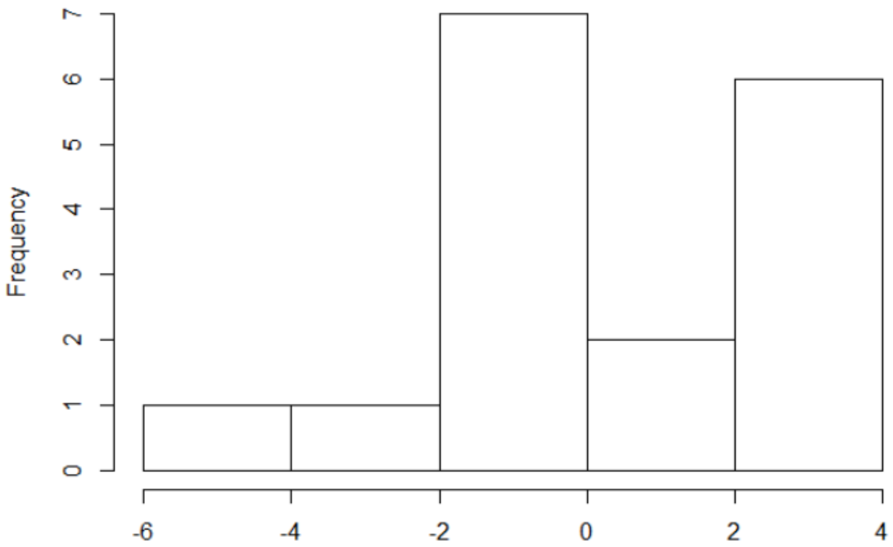
An outlier was identified located at ~(2, 20) and another at ~(4,12). The linear model produced an r^2 of 0.008 with the outlier contained in the data. After it was removed the r^2 value increased to 0.23. This gives one indication that the linear regression model is a better fit with the outliers removed.

Scatter Plot with Least Squares Line: Predictor: Female Income, Response: Gambling in Pounds/Year with Outliers Removed



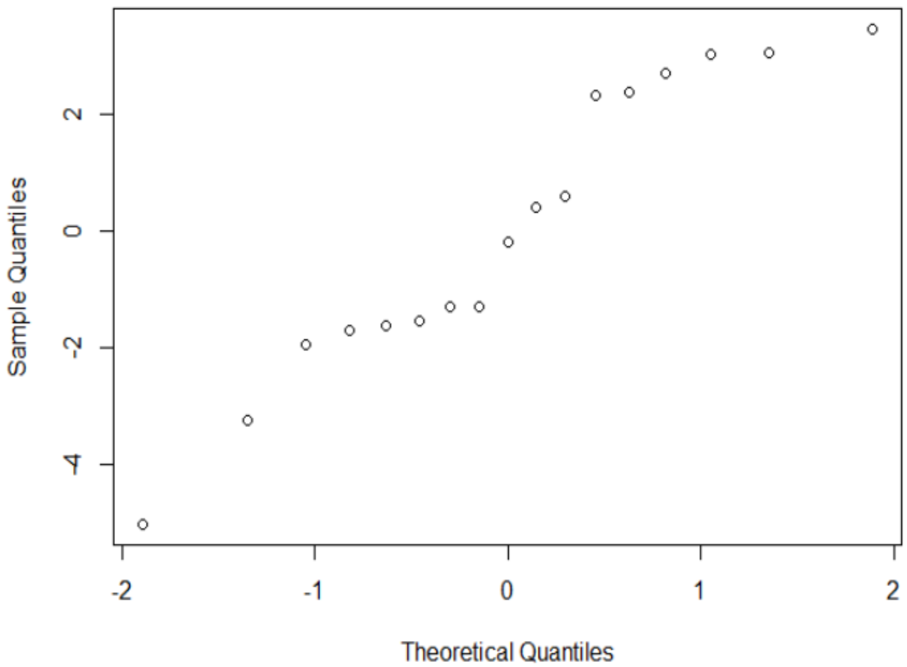
(d) Check the structure of the relationship between the predictors and the response.

Histogram of Residuals from Data that Excludes Outliers



The income residual histogram for females with the outliers removed still does not look to be normal.

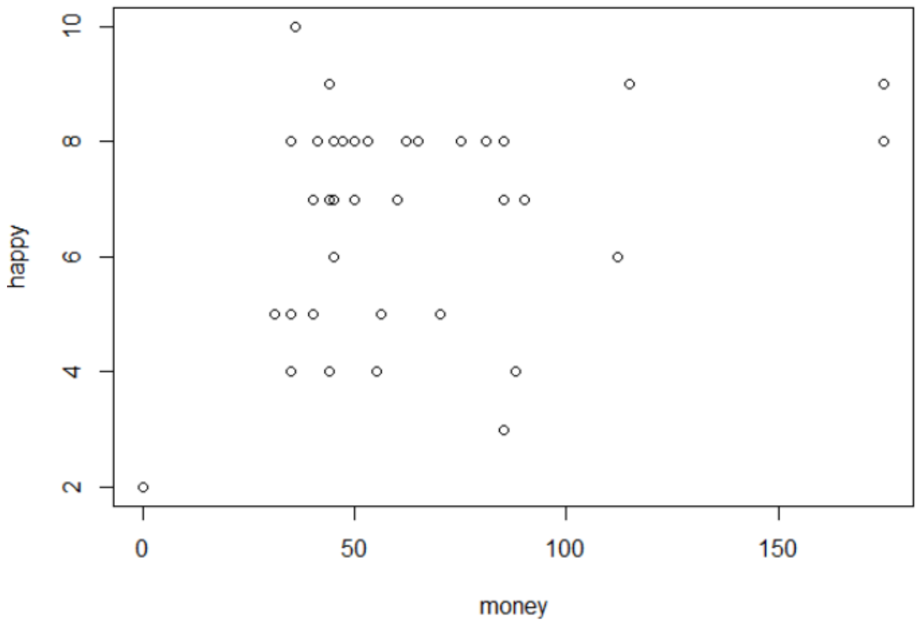
Normal QQ-Plot of Residuals without Outlier Data Points



This graph demonstrates that with the removal of the outliers that a linear model is appropriate because the assumptions of residual normality, independence, and consistent variance are suitable, especially considering how small the sample size is.

Exercise 3: Fit a model with `happy' as the response and the `money ' as the predictor for the 'happy' data set.

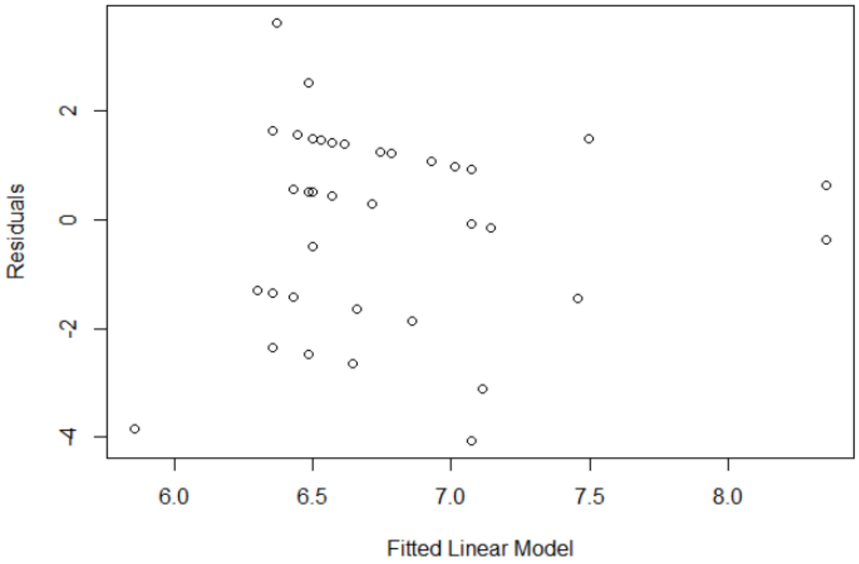
Scatterplot: Predictor: Family Income in Thousands of Dollars (Money)
Response: Happiness on a 10 Point Scale where 10 is the Max (Happy)



'Money' was choosen as the predictor variable because based on the scatterplot it looks like a good candidate for a linear regression model.

(a) Check the constant variance assumption for the errors.

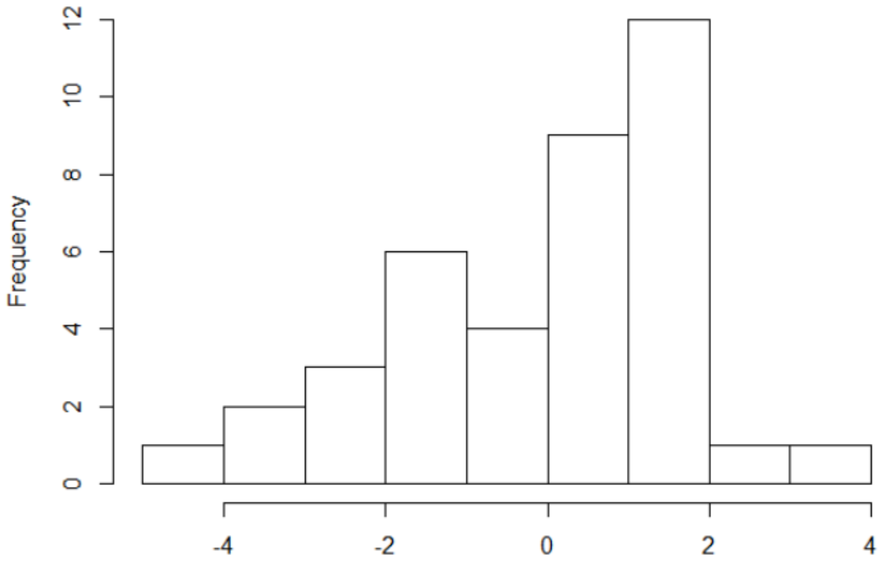
Money Residuals



The constant variance assumption for the errors appears to be satisfied from looking at this plot. There seems to be independence from the data points, and they appear to be have consistent varience.

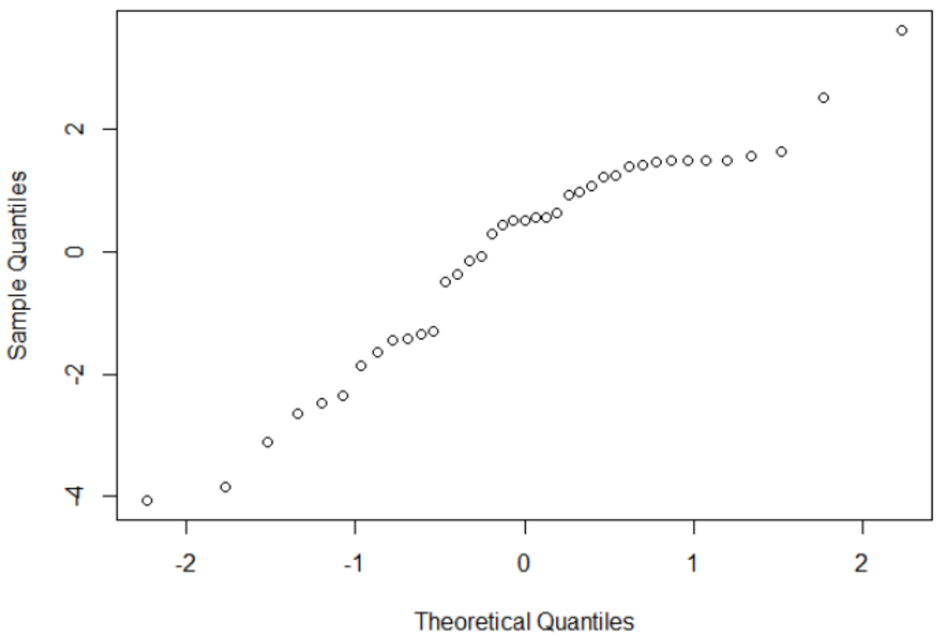
(b) Check the normality assumption.

Histogram of Residuals



The histogram of the residuals does not look to be normally distributed. It looks more like it is increasing at a linear rate, but the sample size is small, so a linear model should not be abandoned just yet.

QQ-Normal Plot

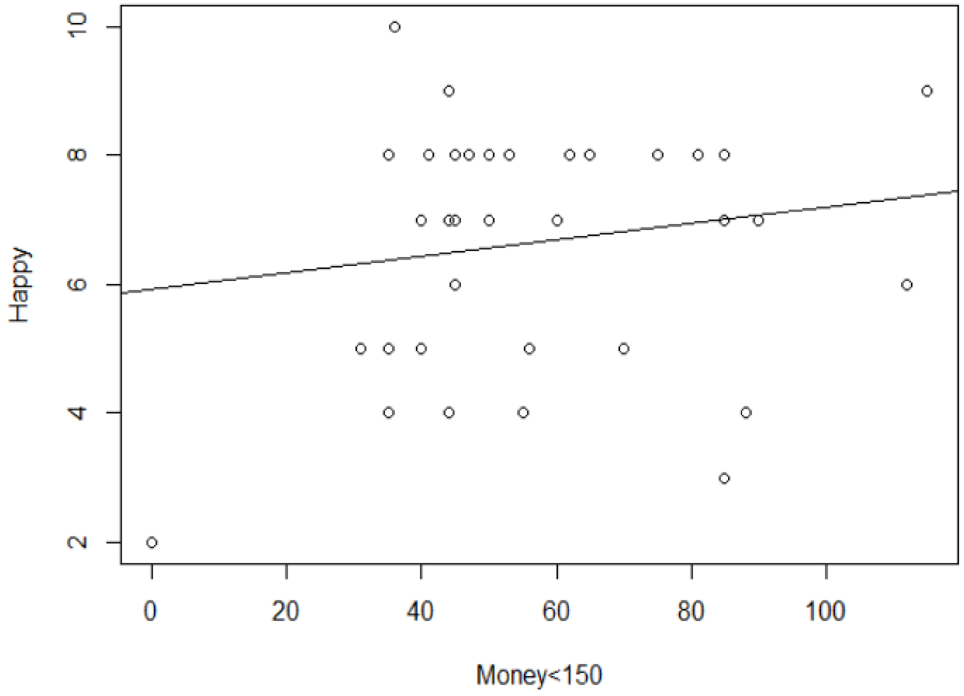


Based on the QQ-Normal plot the residuals appear to be approximately normally distributed.

(c) Check for outliers.

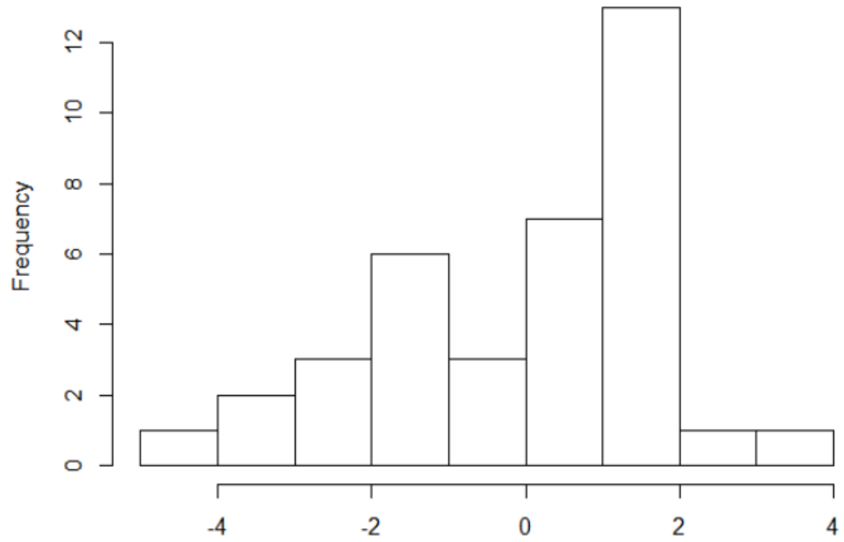
Two outliers were removed from the data set by constricting the data to only include money less than 150. By doing this the data appeared to have more consistent variance and it is expected the residuals will be more equally distributed.

Scatterplot w/Least Squares Line: Predictor: Family Income (Money) without Outliers
Response: Happiness on 10 Point Scale where 10 is Max



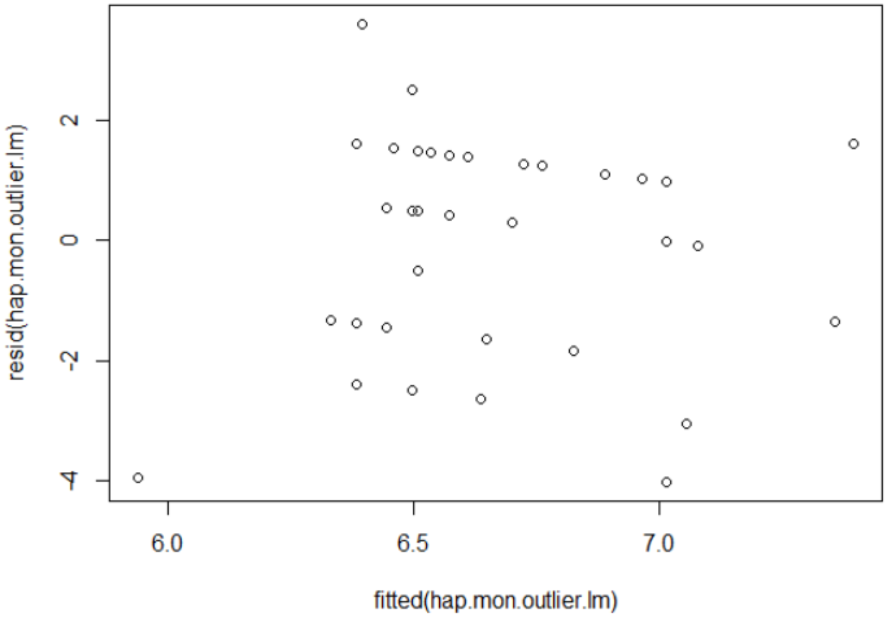
Removing the ouliers surprisingly didn't change the slope or intercept much at all, which indicates to me that it might not be appropriate to remove them. This thought process was reinforced more after looking at the histogram of residuals and fitted linear model plotted against the residuals (found below).

Histogram of Residuals



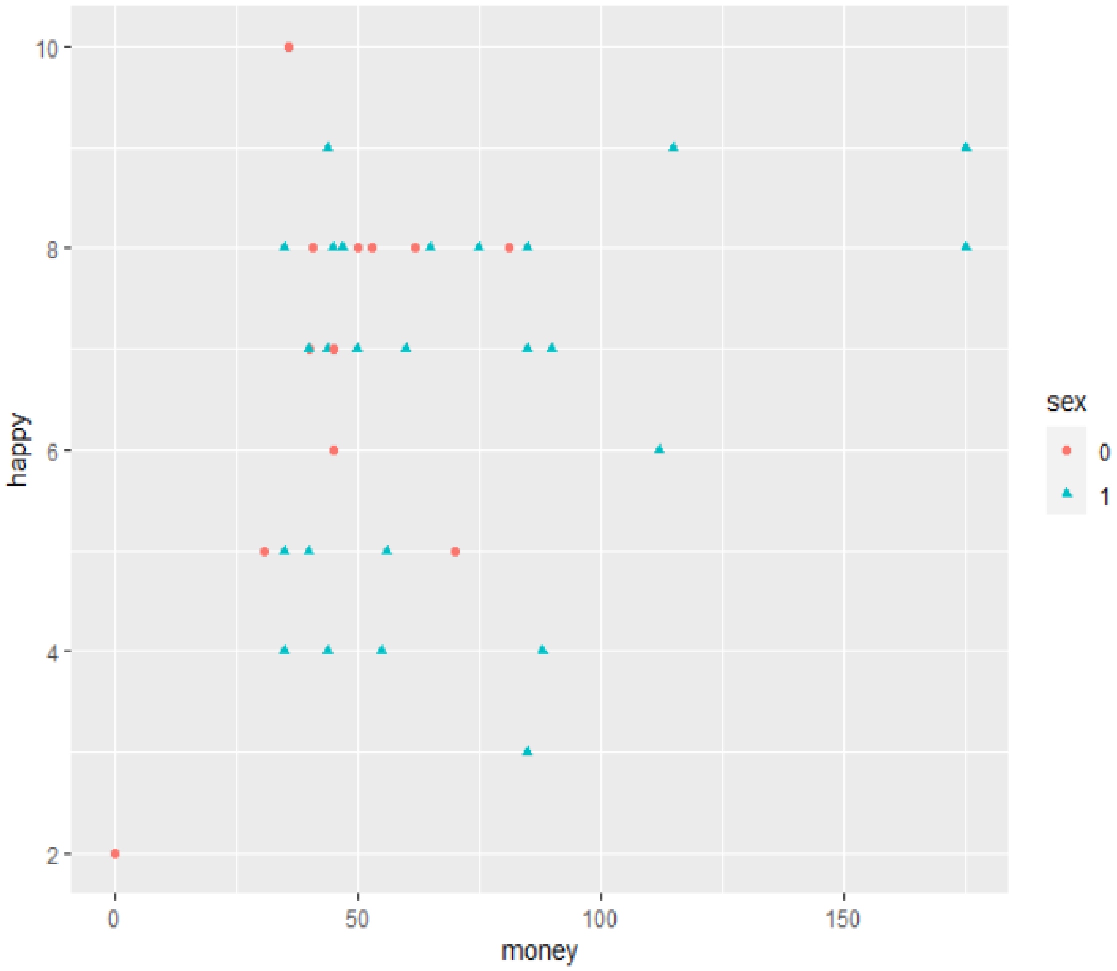
Appears to be slightly less normally distributed when the outliers were removed from the original data.

Fitted Linear Model vs. Residuals



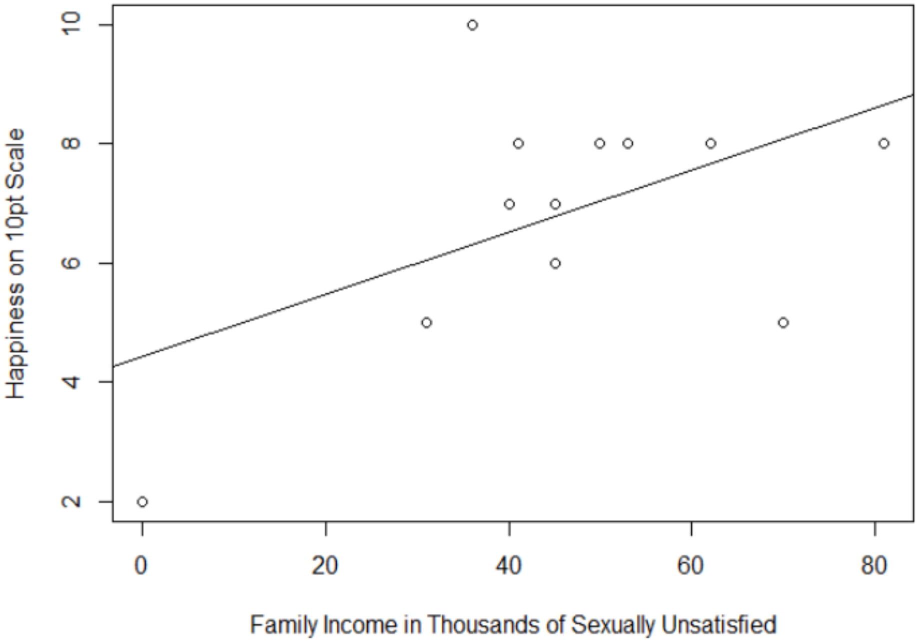
It appears that removing the outliers did not increase our confidence in the assumption of consistent variance within the residuals. In conclusion, it is not appropriate to remove the outliers from this data for the purpose of applying a better fit simple linear regression model.

Scatterplot: Predictors: Family Income in Thousands (Money), Sexual Satisfaction where 1=Satisfied), Response: Happiness on 10 Point Scale



This graph is interesting because although there doesn't seem to be a significant interaction between sexual satisfaction and happiness among the students, however there does seem to be a relationship between sexual satisfaction and how much money someone makes.

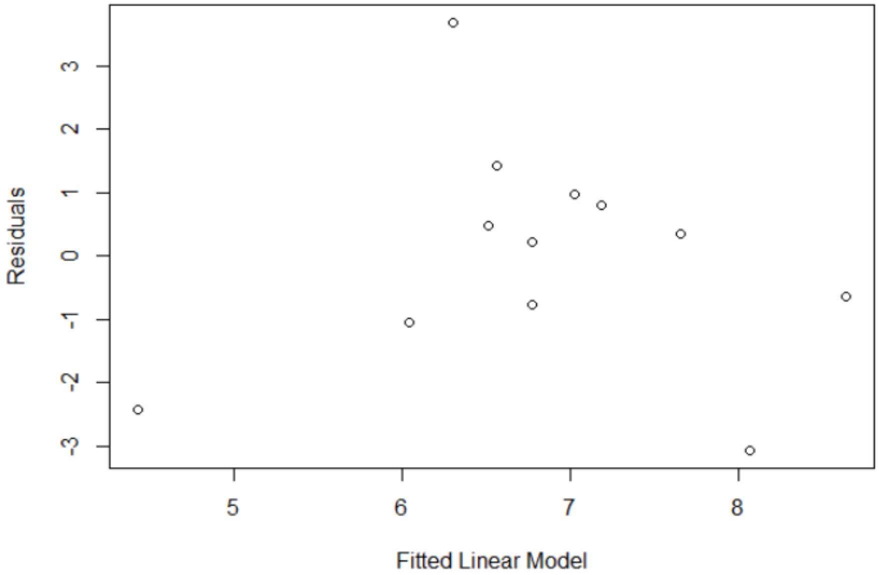
Scatterplot: Predictor: Money of only those who are sexually unsatisfied, Response: Happy



There appears to be a bigger relationship between happiness and money among those who are sexually unsatisfied. The slope is greater for this data group when fit into a linear model. However, the data points themselves do not appear based off the "eye" test to have a super strong linear relationship, but then again the data set is too small to really make and real conclusions from this information.

(a) Check the constant variance assumption for the errors.

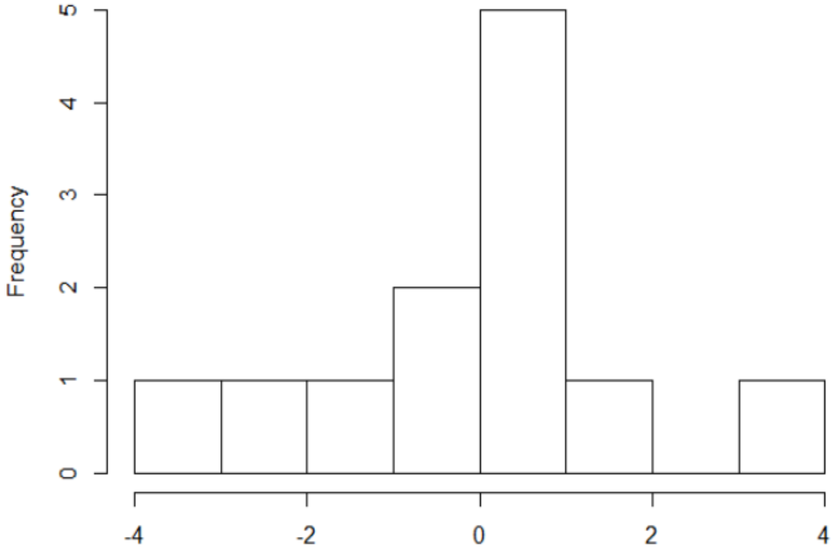
Fitted Linear Model vs. Residuals of those who are sexually unsatisfied



All the residuals appear to be independent, but not consistently normally distributed. Overall I would say this raises a medium orange flag in terms of concerns for violating the linear regression model assumptions.

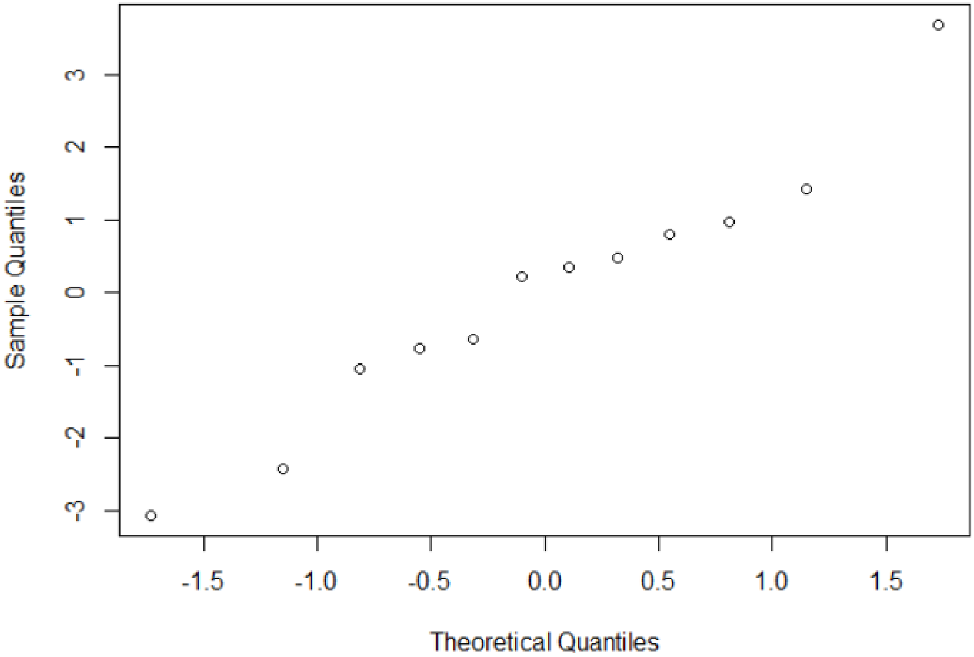
(b) Check the normality assumption.

Histogram of Residuals for those who are sexually unsatisfied



However, the histogram of the residuals shows that they are indeed normally distributed.

Normal QQ-Plot of those who are sexually unsatisfied



The normal QQ-Plot, in addition with the conclusions drawn from the histogram, and fitted model vs. residuals, lead us to think that the assumptions of normality are satisfied to use a linear model for this data. All points appears to have consistent variance, are independent, and normally distributed.

(c) Check for outliers.

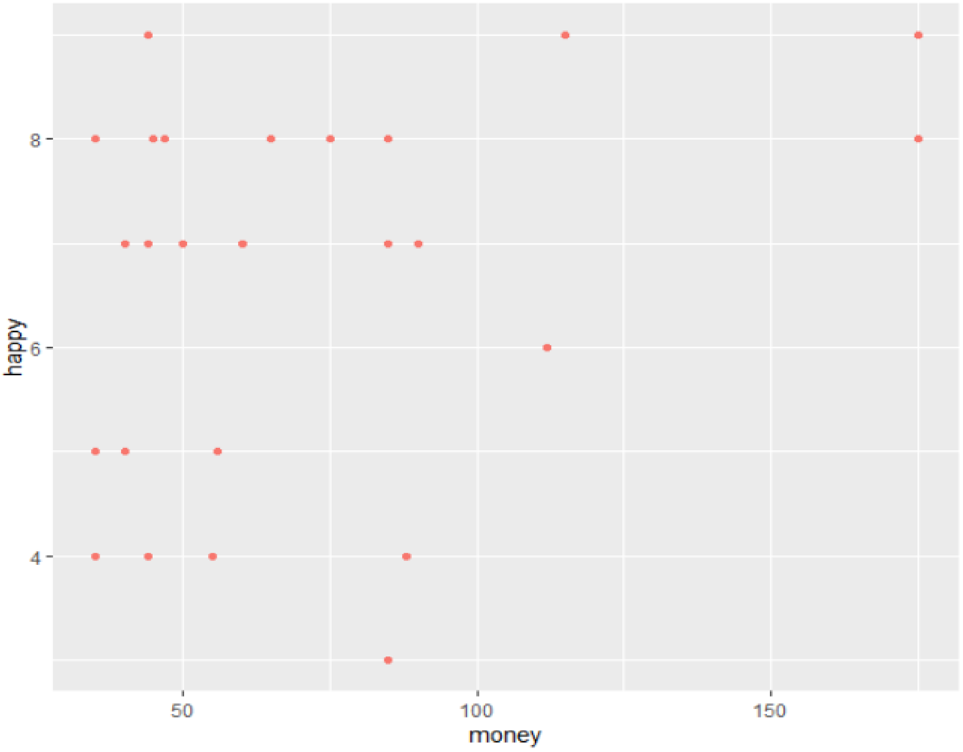
There doesn't seem to be any outliers when only looking at the students who are sexually unsatisfied in the data set. Therefore, I think for this group it would be inappropriate to remove any data points.

(d) Check the structure of the relationship between the predictors and the response.

When looking at the sexually unsatisfied data partition the estimated Beta0=4.43, with a estimated Beta1=0.05. Although the estimated R^2=0.26 it does seem like the linear regression model does a fair job at representing the relationship between how wealthy a sexually unsuccessful students family is, and how happy they are. The assumptions of normality, independence and consistent varience seem to be satisfied to use this model. So, for each thousand dollar increase in wealth we can expect a 1/20 increase to their happiness. If this linear model is an accurate predictor for the response then a sexually unsatisfied student's family would need to hold 111,400 pounds for them to report a happiness level of 10.

(1) Repeated exercise 2 again but with partitioned data of only the sexually satisfied students.

Scatterplot: Predictor: Wealth of Sexually Satisfied Students in Thousands (Money), Response: Happiness on 10 Point Scale (Happy)



There does not seem to be much of a linear relationship between family wealth and happiness of a student when they are sexually satisfied. Since this is the case with the 'eye' test I don't think it is worthwhile to explore this relationship further. The previous linear model using the sexually unsatisfied students seems to act as a predictor far better.