Assignment 5

Jennifer Benson

November 30, 2016

```
county.complete <- filter(county, !is.na(Romney_pct))
```

1. Research Question

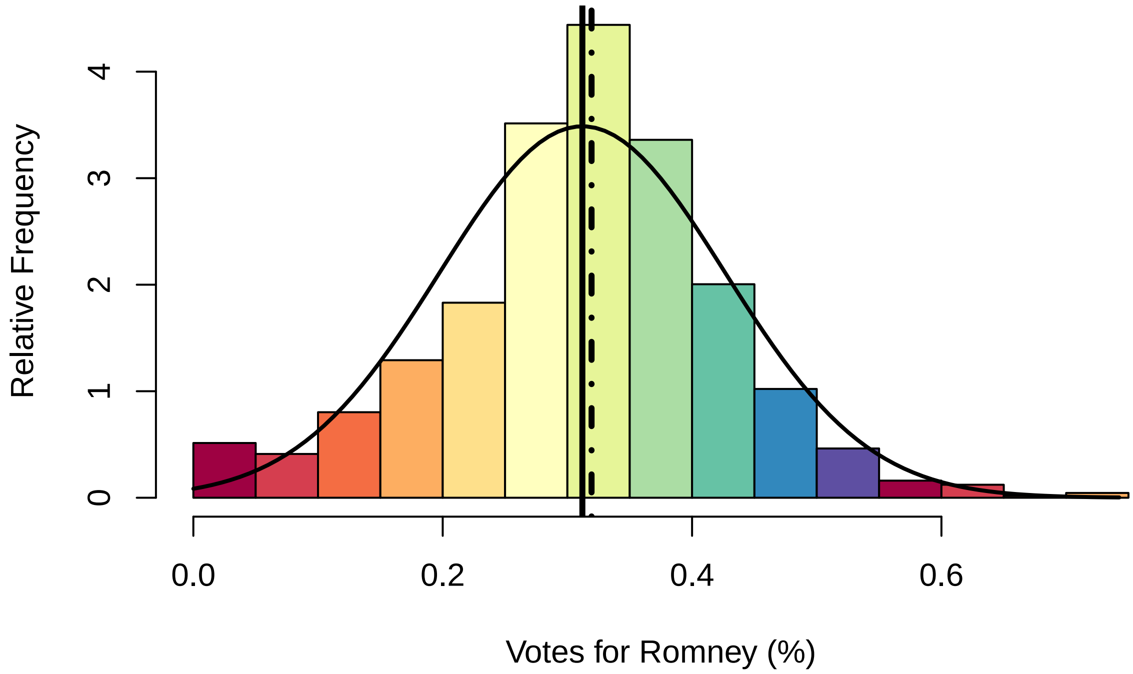
My research question is "Do more prosperous counties vote Republican in presidential elections?" Previous analyses has not established a relationship between median household income and the percent who voted for Romney. Most recently, this conclusion has been supported with a simple linear regression of the variables. Two of the four assumptions that justify the use of a linear regression model are not met. Firstly, there is a problem with linearity. Secondly, there is a lack of homoscedasticity since the variance of residuals is not the same for any value of median household income. Therefore, the conclusion that I can draw is that the data points are non-linear and seemingly random.

Education may be a confounding variable. Within the data-set there is a variable called *hs_grad*, which measures the percent who graduated from high school. Furthermore, it would be important to consider the variable *bachelors* in an independent regression. The variable measures the percent who graduated with a bachelors in each county. Another variable that might predict the dependent variable is the variable *foreign_spoken_at_home*, which measures the percent of people in each country that speak a language other than English at home.

2. Logged version of the dependent variable

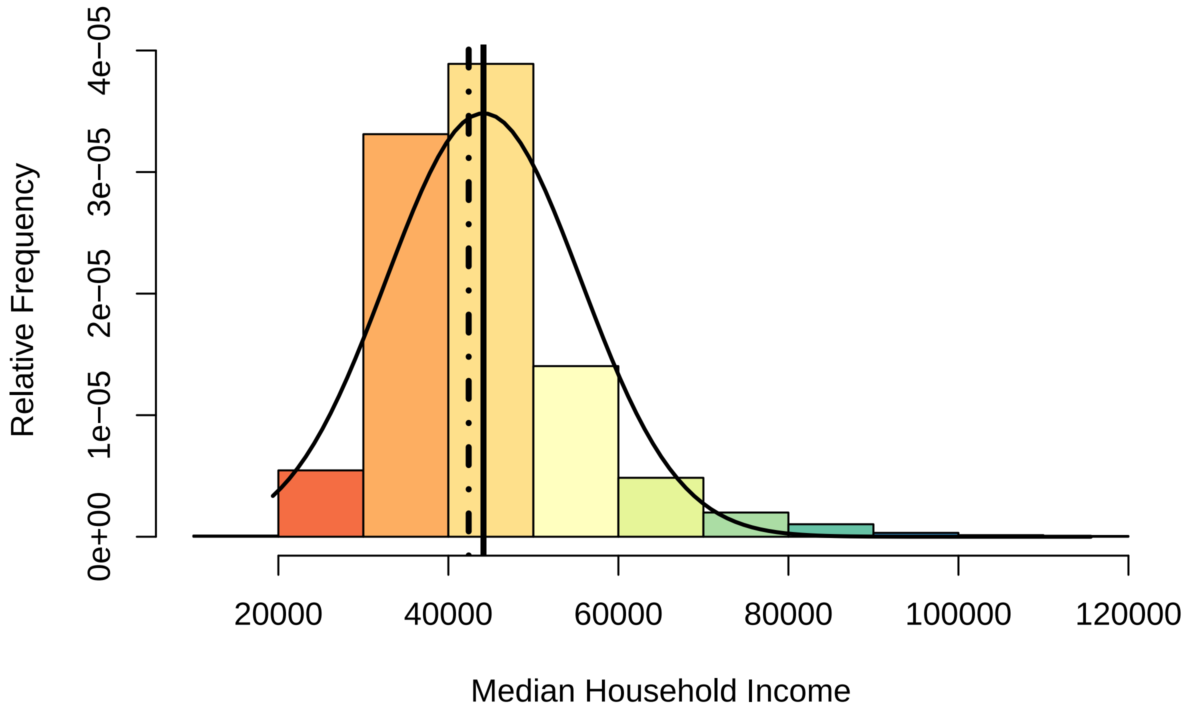
```
hist(county.complete$Romney_pct, prob=T, main="",
     ylab='Relative Frequency',
     xlab="Votes for Romney (%)",
     col=brewer.pal(11, "Spectral"))
#Add vertical lines at the mean and median
abline(v=mean(county.complete$Romney_pct, na.rm=TRUE), col="black", lwd=3)
abline(v=median(county.complete$Romney_pct, na.rm=TRUE), col="black", lty=4, lwd=3)

#create objects that will define the normal curve
xfit<-seq(min(county.complete$Romney_pct, na.rm=TRUE),
          max(county.complete$Romney_pct, na.rm=TRUE),length=100)
yfit<-dnorm(xfit,mean=mean(county.complete$Romney_pct, na.rm=TRUE),
            sd=sd(county.complete$Romney_pct, na.rm=TRUE))
#Now plot the normal curve over the histogram
lines(xfit, yfit, col="black", lwd=2)
```



```
hist(county.complete$median_household_income, prob=T, main = "" ,
     ylab='Relative Frequency',
     xlab="Median Household Income",
     col=brewer.pal(9, "Spectral"))
#Add vertical lines at the mean and median
abline(v=mean(county.complete$median_household_income, na.rm=TRUE),
       col="black", lwd=3)
abline(v=median(county.complete$median_household_income, na.rm=TRUE),
       col="black", lty=4, lwd=3)

#create objects that will define the normal curve
xfit<-seq(min(county.complete$median_household_income, na.rm=TRUE),
          max(county.complete$median_household_income, na.rm=TRUE),length=100)
yfit<-dnorm(xfit,mean=mean(county.complete$median_household_income, na.rm=TRUE),
            sd=sd(county.complete$median_household_income, na.rm=TRUE))
#Now plot the normal curve over the histogram
lines(xfit, yfit, col="black", lwd=2)
```

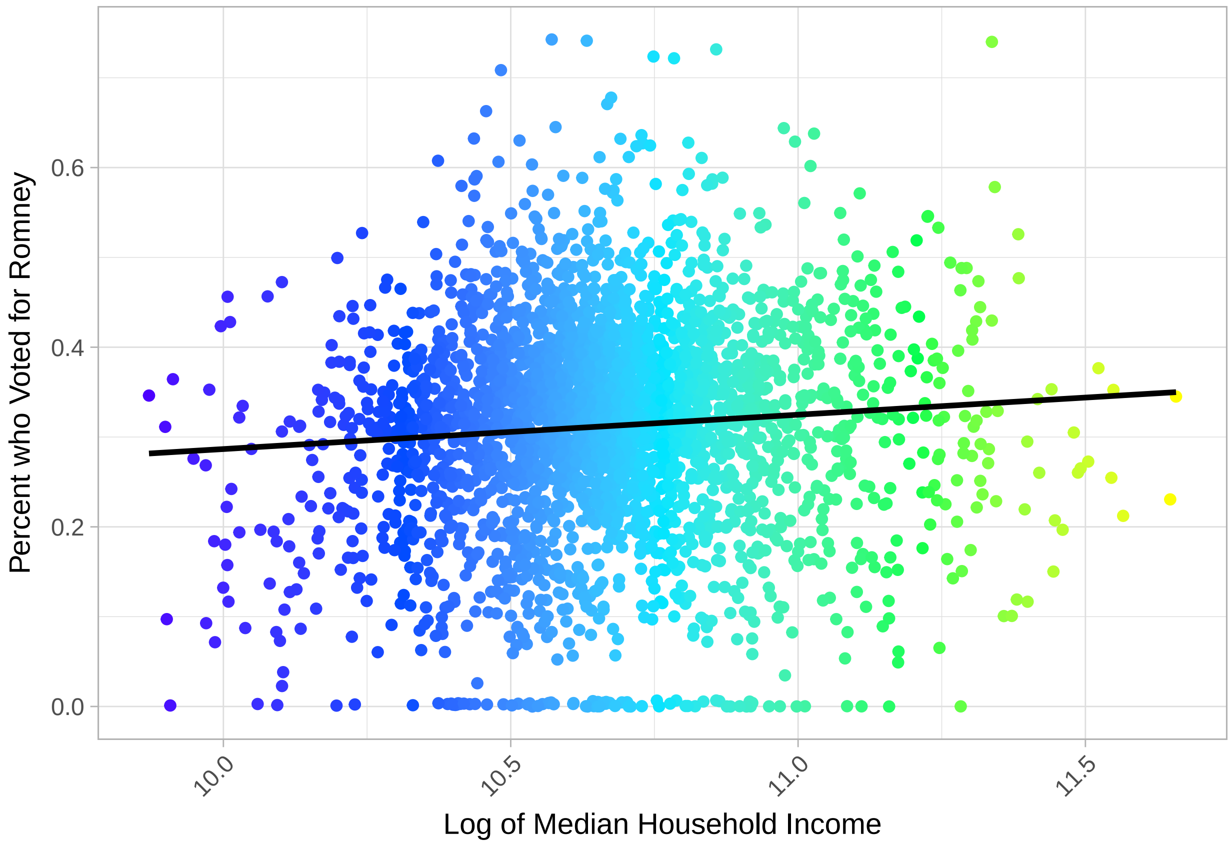


The histograms show that *median_household_income* is skewed. So, I am going to conduct a simple linear regression with a logged version of the variable.

```r
m1 <- lm(Romney_pct ~ log1p(median_household_income), data=county.complete)
summary(m1)
```

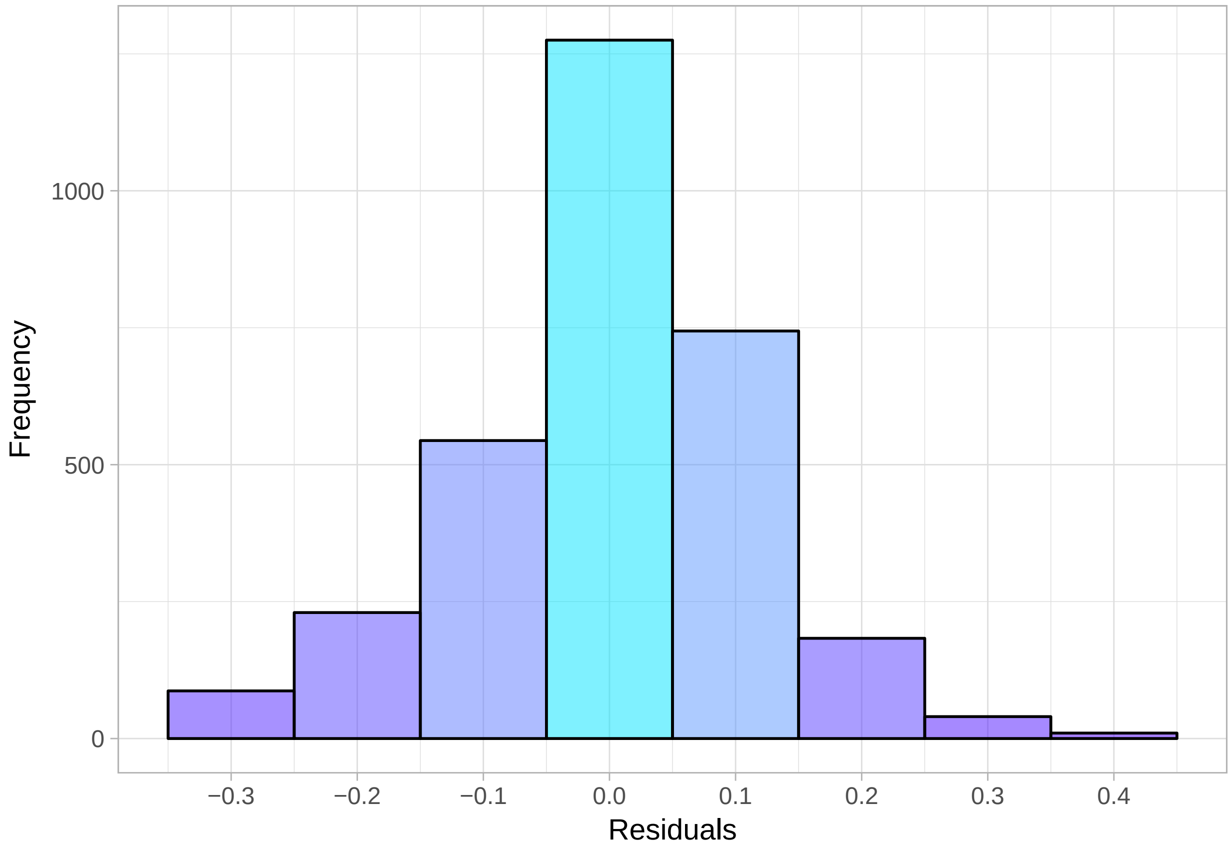
```
##
## Call:
## lm(formula = Romney_pct ~ log1p(median_household_income), data = county.complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33553 -0.05914  0.00873  0.06764  0.43417
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -0.095547   0.090121  -1.060    0.289
## log1p(median_household_income)  0.038215   0.008448   4.524 6.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.114 on 3111 degrees of freedom
## Multiple R-squared:  0.006534,   Adjusted R-squared:  0.006215
## F-statistic: 20.46 on 1 and 3111 DF,  p-value: 6.31e-06
```

```
#Check linearity with a scatterplot of x and y, with regression line added
breaks <- c(9, 9.5, 10, 10.5, 11, 11.5, 12)
a <- ggplot(county.complete, aes(log1p(median_household_income), Romney_pct, colour = log1p(median_hous
      geom_jitter() + geom_smooth(method=lm, colour="black", se=FALSE) +
     scale_colour_gradientn(colours = topo.colors(5),
                            breaks = breaks,labels = format(breaks))+
     scale_x_continuous(name="Log of Median Household Income", breaks=breaks)+
     ylab ("Percent who Voted for Romney")+
     theme_light()+
     theme(axis.text.x = element_text(angle = 45, hjust = 1),
           legend.position="none")
a
```

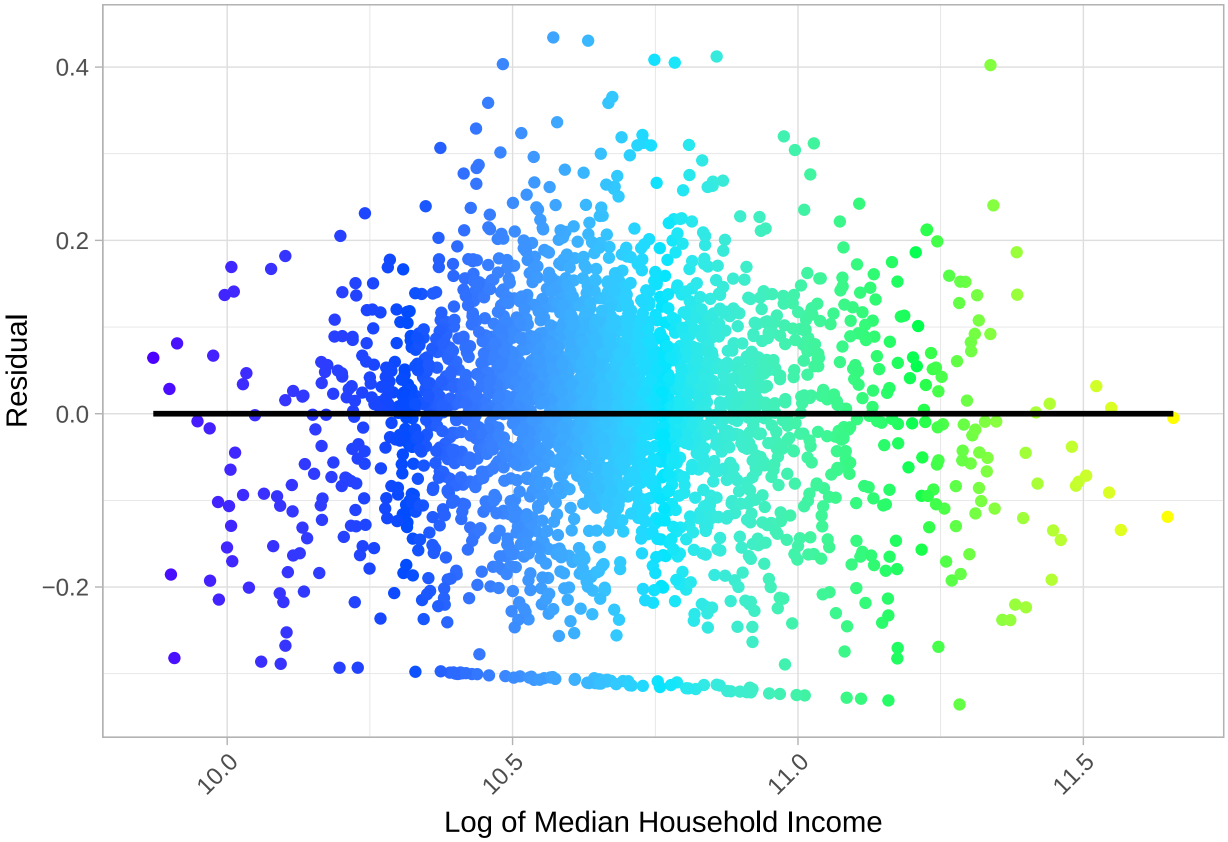


```
#Save residuals from m1 as a variable in dataname.complete
county.complete$m1residual <- m1$resid

#Assess normality of residuals with a histogram
 breaks=seq(-1, 1, by=0.1)
 m <- ggplot(county.complete, aes(x=m1residual)) +
      geom_histogram(binwidth=0.1,
                     colour="black", alpha=0.5,
                     aes(fill=..count..)) +
      scale_fill_gradientn(colours = topo.colors(2)) +
      scale_x_continuous(name="Residuals", breaks) +
      ylab ("Frequency") +
      theme_light() +
      theme(legend.position="none")
m
```



```
#Assess constant variability with a residual plot
breaks <- c(9, 9.5, 10, 10.5, 11, 11.5, 12)
b <- ggplot(county.complete, aes(log1p(median_household_income), m1residual, colour = log1p(median_hous
      geom_jitter() +  geom_smooth(method=lm, colour="black", se=FALSE) +
     scale_colour_gradientn(colours = topo.colors(5),
                            breaks = breaks,labels = format(breaks))+
     scale_x_continuous(name="Log of Median Household Income", breaks=breaks)+
     ylab ("Residual")+
     theme_light()+
     theme(axis.text.x = element_text(angle = 45, hjust = 1),
           legend.position="none")
b
```



The model does not meet all of the conditions for linear regression:
(a) Linearity: The data points appear to be distributed randomly with clustering.
(b) Homoscedasticity: The variance of the residuals are randomly distributed around the log of median household income.
(c) Independence: The observations are independent of each other.
(d) Normality: The distribution of the residuals is approximately normal.
Therefore, inference cannot be made based on the results of the regression.

The regression equation for the model is:
Y = -0.095547 + 0.038215*log(X)
β_0, the y-intercept, is -0.095547 and β_1, the slope, is 0.038215. This slope means that a one-unit change in log(X) will produce an expected increase in Y of 0.038215 units.

The null hypothesis is that the slope or β_1 is 0, which means that there would be no relationship. The alternative hypothesis is that β_1 is not 0, which would indicate a relationship between the two variables. The t-value for my β_1 coefficient is 4.524 and the p-value is 6.31e-06. At $\alpha = 0.001$ the coefficient is statistically significant. Since the p-value, 6.31e-06, is less than α, 0.001, we can reject the null hypothesis and conclude that the coefficient is statistically significant.

This model further emphasizes that there is no linear relationship between *median_household_income* and *Romney_pct*. In the context of the research question, this means that more prosperous counties, at least as defined now, are not correlated with the percent who voted Republican.

Log transforming variables is preferable when there appears to be a non-linear relationship between the independent and dependent variables. In my case, the relationship between *median_household_income* and *Romney_pct* is seemingly random. Therefore, neither a logged or unlogged version of the variable(s) is preferable.

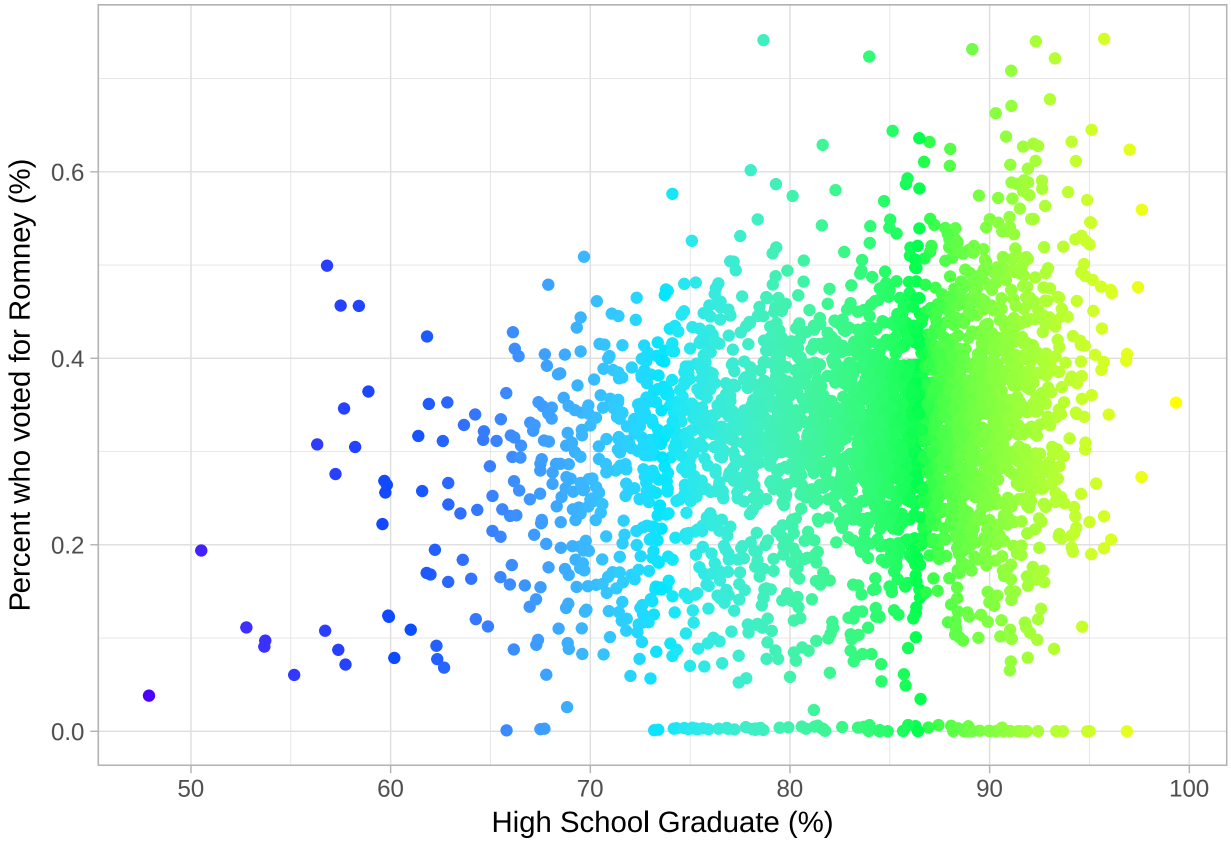
3. Multiple Regression

The variable I am using is *hs_grad*, which measures the percent in each county that have at least a bachelors degree. I believe this is a confounding variable because level of education impacts median household income which is hypothesized to be related to levels of Republican votes per county. I expect that *hs_grad* will be negatively correlated with *Romney_pct*.

```
summary(county.complete$hs_grad)
```

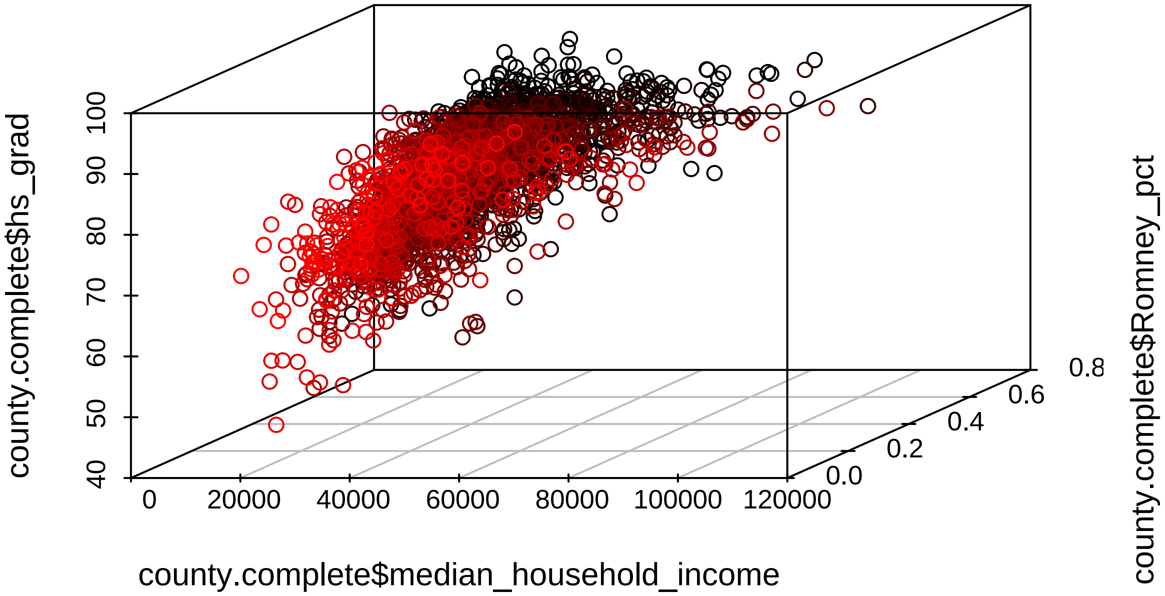
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   47.90   78.40   84.60   83.08   88.50   99.30
```

```
breaks <- c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
e <- ggplot(county.complete, aes(hs_grad, Romney_pct, colour = hs_grad)) +
     geom_jitter() +
     scale_colour_gradientn(colours = topo.colors(5),
                            breaks = breaks,labels = format(breaks))+
     scale_x_continuous(name="High School Graduate (%)", breaks=breaks)+
     ylab("Percent who voted for Romney (%)")+
     theme_light()+
     theme(legend.position="none")
e
```



The above scatter plot shows that I was wrong and *hs_grad* and *Romney_pct* are negatively correlated, although this relationship is weak.

```
scatterplot3d(county.complete$median_household_income,county.complete$Romney_pct,county.complete$hs_gra
```



```
#Re-estimate your bivariate model (if you felt you needed to log a variable, you should continue using
m1 <- lm(Romney_pct ~ median_household_income, data=county.complete)
summary(m1)
```

```
##
## Call:
## lm(formula = Romney_pct ~ median_household_income, data = county.complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33248 -0.05955  0.00786  0.06823  0.43359
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.863e-01  8.151e-03  35.125  < 2e-16 ***
## median_household_income 5.826e-07  1.788e-07   3.259  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1142 on 3111 degrees of freedom
## Multiple R-squared:  0.003401,   Adjusted R-squared:  0.003081
## F-statistic: 10.62 on 1 and 3111 DF,  p-value: 0.001132
```

```r
#Now estimate a multiple regression model that includes your original independent variable and your add
m2 <- lm(Romney_pct ~ median_household_income + hs_grad, data=county.complete)
summary(m2)
```

```
##
## Call:
## lm(formula = Romney_pct ~ median_household_income + hs_grad,
##     data = county.complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35488 -0.06240  0.00702  0.06796  0.44649
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -2.370e-02  2.349e-02  -1.009    0.313
## median_household_income -1.136e-06  2.124e-07  -5.350 9.45e-08 ***
## hs_grad                  4.645e-03  3.314e-04  14.013  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1108 on 3110 degrees of freedom
## Multiple R-squared:  0.06259,    Adjusted R-squared:  0.06199
## F-statistic: 103.8 on 2 and 3110 DF,  p-value: < 2.2e-16
```

The regression equation for the multiple regression model is:
Y = -2.370e-02 - 1.136e-06X_1 + 4.645e-03X_2

For β_1 the t-value is -5.350 and the p-value is 9.45e-08. At $\alpha = 0.05$ the coefficient is statistically significant and the null hypothesis can be rejected since the p-value, 9.45e-08, is less than α, 0.05. In this case the hypotheses are:
H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$

For β_2 the t-value is 3.259 and the p-value is 0.00113. At $\alpha = 0.05$ the coefficient is statistically significant and the null hypothesis can be rejected since the p-value, 0.00113, is less than α, 0.05. In this case the hypotheses are:
H_0: $\beta_2 = 0$
H_A: $\beta_2 \neq 0$

The estimate for β_1 in the initial regression is 5.826e-07 whilst for the multiple regression β_1 is -1.136e-06. Both values are extremely close to 0 and their statistical significance has not changed.

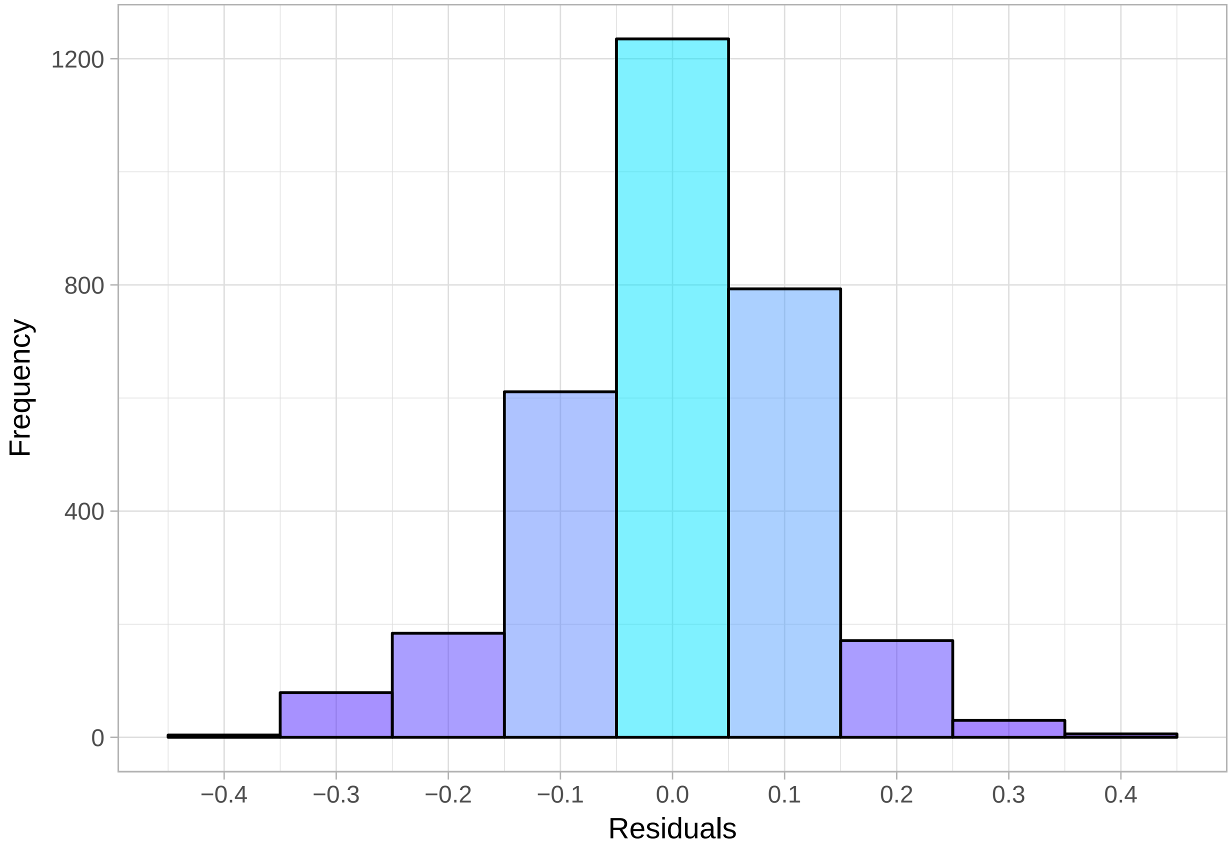
The value of multiple R^2 is 0.06259 and the value of the adjusted R^2 is 0.06199. Multiple R^2 is an indicator of how much of the variance in the dependent variable can be explained by the independent variables whilst adjusted R^2 provides the same information but adjusts for the number of terms in the model. Since the R^2 values are extremely low, this indicates high variability that is unexplained by the independent variables in the model. However, the interpretation of the p-value and coe???cient does not change,

Holding all other variables constant, for a one-unit change in median household income and high school grad, we would expect on average a (- 1.136e-06 + 4.645e-03) change in the percent who voted for Romney.

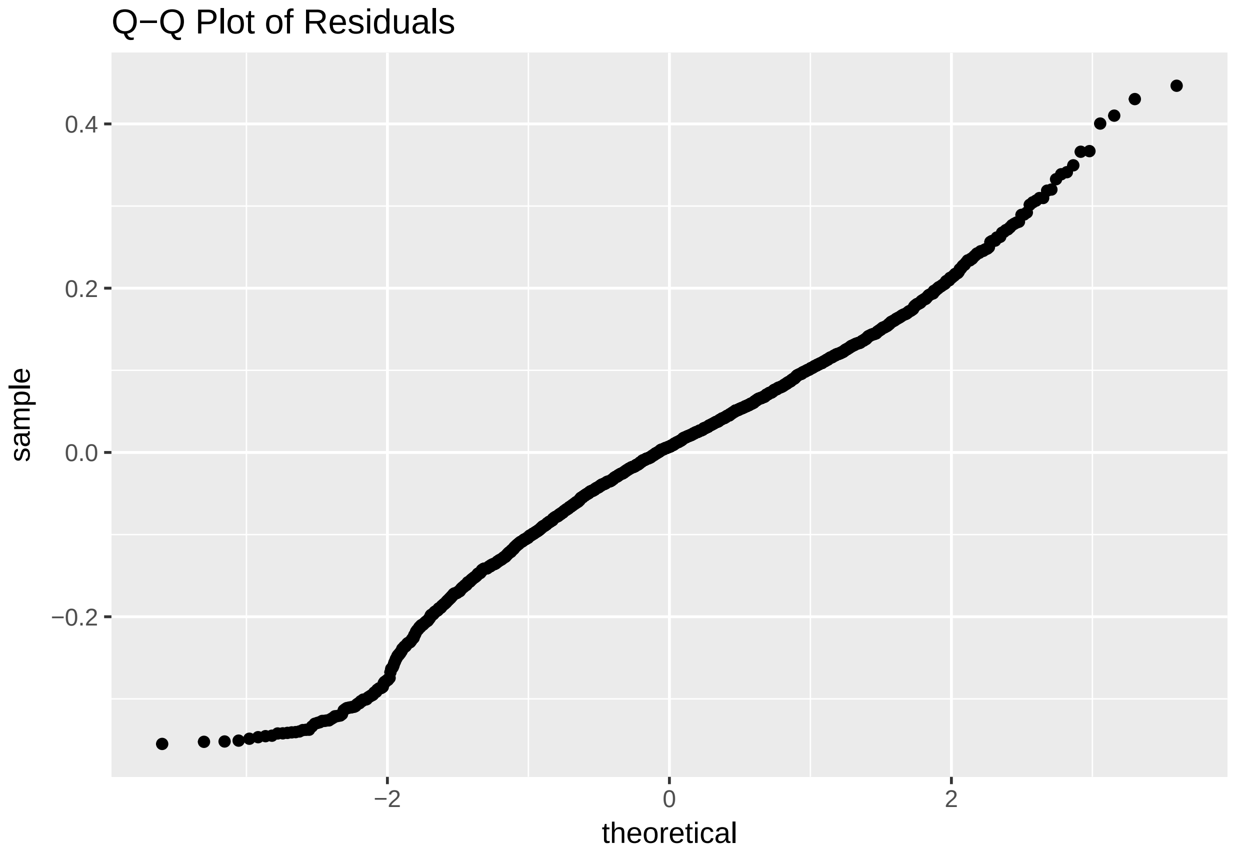
Nearly Normal Residuals

```
county.complete$residual <- m2$resid
```

```
 breaks=seq(-1, 1, by=0.1)
 m <- ggplot(county.complete, aes(x=residual)) +
      geom_histogram(binwidth=0.1,
                     colour="black", alpha=0.5,
                     aes(fill=..count..)) +
      scale_fill_gradientn(colours = topo.colors(2)) +
      scale_x_continuous(name="Residuals", breaks) +
      ylab ("Frequency") +
      theme_light() +
      theme(legend.position="none")
m
```



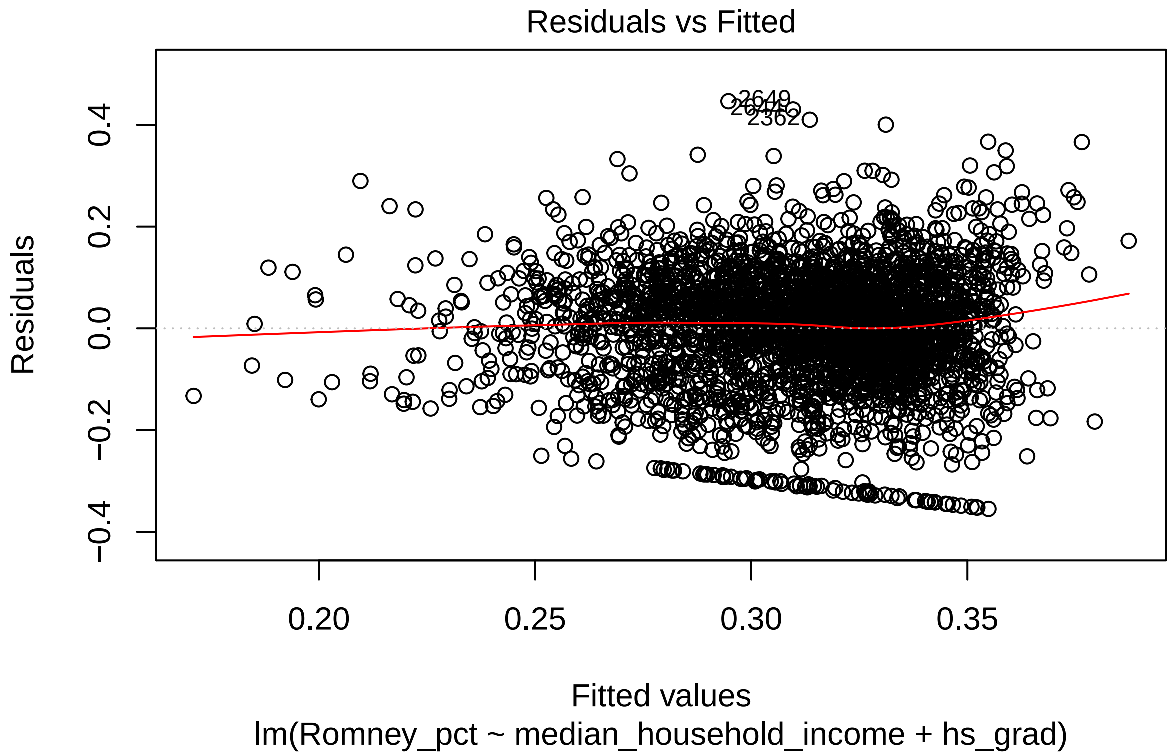
```
ggplot(county.complete, aes(sample = residual)) +
  ggtitle("Q-Q Plot of Residuals") +
  stat_qq()
```



The residuals are approximately normal so this condition is met.

Constant Variability

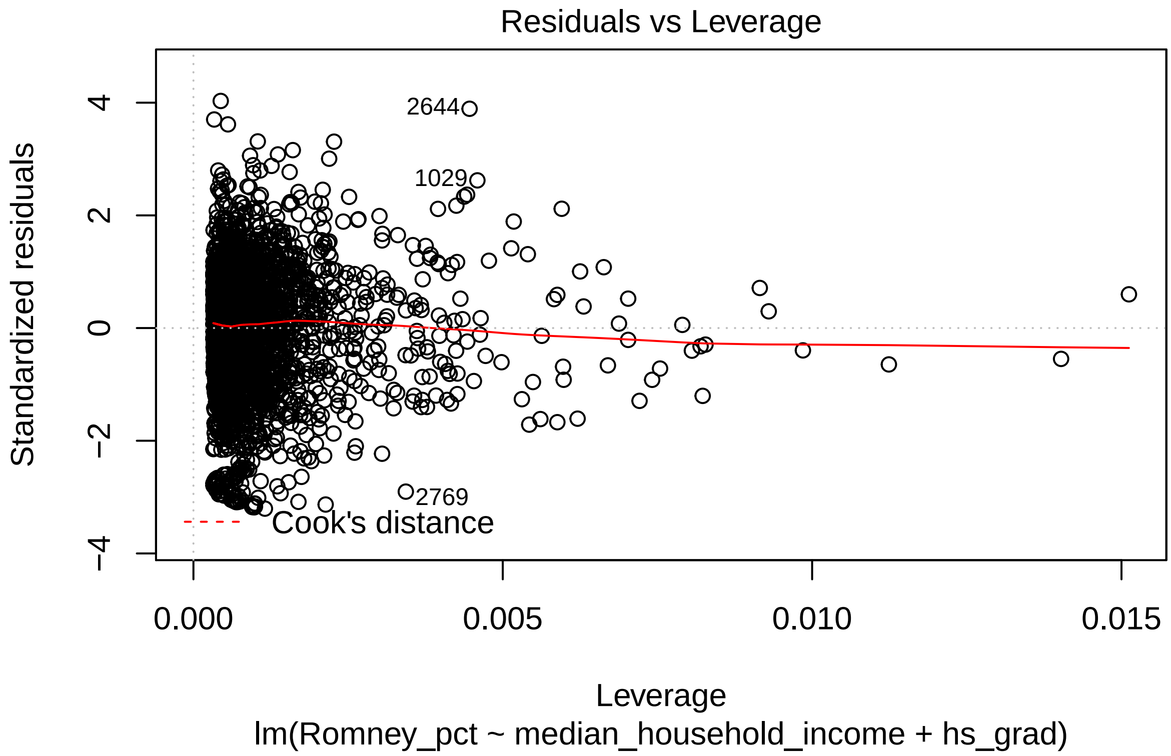
```
plot(m2,1)
```



The residuals appear to be approximately normal (especially when you exclude the outlier).

No multi-collinearity between predictors

```
plot(m2,5)
```



```
vif(m2)
```

```
## median_household_income                  hs_grad
##                1.500325                 1.500325
```

4. Play around with some other models

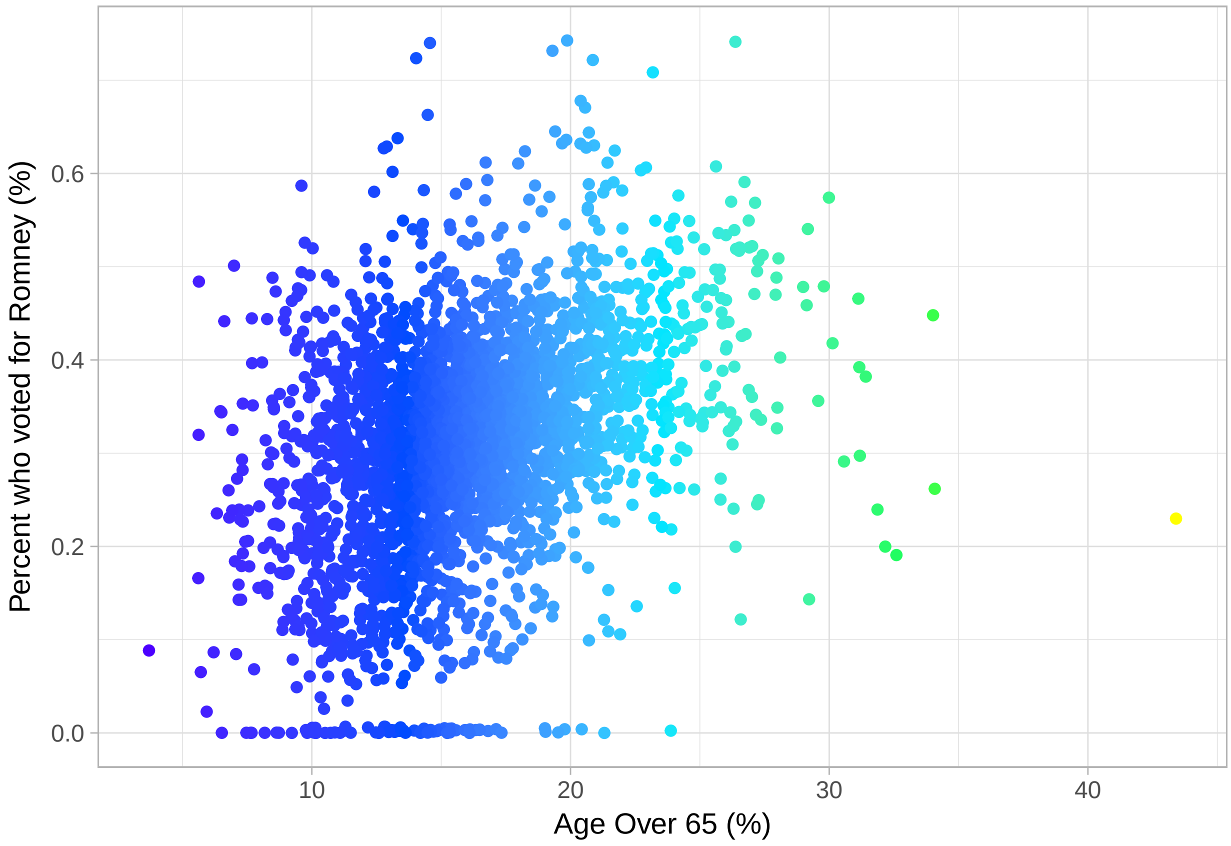
Variable: ***age_over_65***

The variable I am adding is *age_over_65* , which measures the percent in each county that are older than 65 years old. I believe this variable also influences the percent who voted for Romney because older people may have more 'traditional' and 'conservative' ideals. Therefore, I expect that *age_over_65* will be positively correlated with *Romney_pct*.

```
summary(county.complete$age_over_65)
```

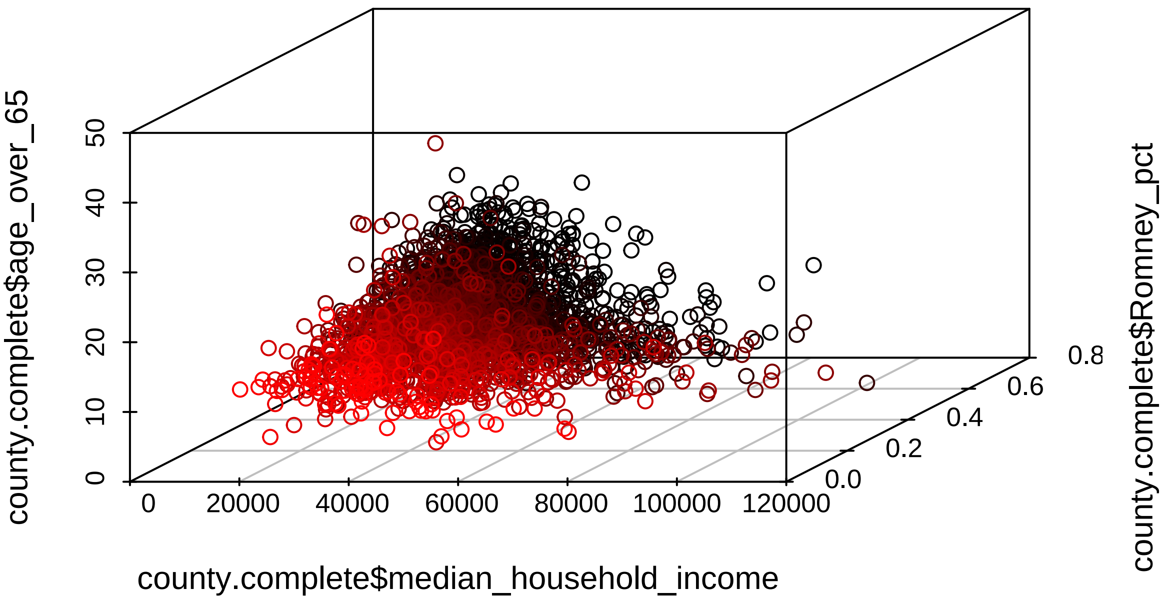
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.70   13.20   15.60   15.95   18.20   43.40
```

```
breaks <- c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
e <- ggplot(county.complete, aes(age_over_65, Romney_pct, colour = age_over_65)) +
     geom_jitter() +
     scale_colour_gradientn(colours = topo.colors(5),
                            breaks = breaks,labels = format(breaks))+
     scale_x_continuous(name="Age Over 65 (%)", breaks=breaks)+
     ylab("Percent who voted for Romney (%)")+
     theme_light()+
     theme(legend.position="none")
e
```



The above scatterplot shows that I was wrong and *age_over_65* and *Romney_pct* are positively correlated, although this relationship is weak.

```
scatterplot3d(county.complete$median_household_income,county.complete$Romney_pct,county.complete$age_ove
```



```
#Re-estimate your bivariate model (if you felt you needed to log a variable, you should continue using
m1 <- lm(Romney_pct ~ median_household_income, data=county.complete)
summary(m1)
```

```
##
## Call:
## lm(formula = Romney_pct ~ median_household_income, data = county.complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33248 -0.05955  0.00786  0.06823  0.43359
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.863e-01  8.151e-03  35.125  < 2e-16 ***
## median_household_income 5.826e-07  1.788e-07   3.259  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1142 on 3111 degrees of freedom
## Multiple R-squared:  0.003401,   Adjusted R-squared:  0.003081
## F-statistic: 10.62 on 1 and 3111 DF,  p-value: 0.001132
```

```
#Now estimate a multiple regression model that includes your original independent variable and your add
m3 <- lm(Romney_pct ~ median_household_income + hs_grad + age_over_65, data=county.complete)
summary(m3)
```

```
## 
## Call:
## lm(formula = Romney_pct ~ median_household_income + hs_grad + 
##     age_over_65, data = county.complete)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.39116 -0.05699  0.00728  0.06372  0.42869 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            -8.656e-02  2.203e-02  -3.929 8.73e-05 ***
## median_household_income  1.174e-06  2.236e-07   5.253 1.60e-07 ***
## hs_grad                  2.025e-03  3.303e-04   6.129 9.94e-10 ***
## age_over_65              1.120e-02  5.074e-04  22.063  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.103 on 3109 degrees of freedom
## Multiple R-squared:  0.1895, Adjusted R-squared:  0.1887 
## F-statistic: 242.3 on 3 and 3109 DF,  p-value: < 2.2e-16
```

The regression equation for the multiple regression model is:
Y = -8.656e-02 + 1.174e-06X_1 + 2.025e-03X_2 + 1.120e-02X_3

For β_1 the t-value is 5.253 and the p-value is 1.60e-07. At $\alpha = 0.05$ the coefficient is statistically significant and the null hypothesis can be rejected since the p-value, 1.60e-07, is less than α, 0.05. In this case the hypotheses are:
H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$

For β_2 the t-value is 6.129 and the p-value is 9.94e-10. At $\alpha = 0.05$ the coefficient is statistically significant and the null hypothesis can be rejected since the p-value, 9.94e-10, is less than α, 0.05. In this case the hypotheses are:
H_0: $\beta_2 = 0$
H_A: $\beta_2 \neq 0$

For β_3 the t-value is 22.063 and the p-value is < 2e-16. At $\alpha = 0.05$ the coefficient is statistically significant and the null hypothesis can be rejected since the p-value, < 2e-16, is less than α, 0.05. In this case the hypotheses are:
H_0: $\beta_3 = 0$
H_A: $\beta_3 \neq 0$

The estimate for β_1 in the initial regression is 5.826e-07 whilst for the multiple regression β_1 is 1.174e-06. Both values are extremely close to 0 and their statistical significance has not changed.

The value of multiple R^2 is 0.1895 and the value of the adjusted R^2 is 0.1887. Multiple R^2 is an indicator of how much of the variance in the dependent variable can be explained by the independent variables whilst adjusted R^2 provides the same information but adjusts for the number of terms in the model. Since the R^2 values is low, this indicates high variability that is unexplained by the independent variables in the model. It is exciting to note that in this case 18.95% of the variability can be explained by the independent variables whilst previously only 6.259% of the variability could be explained by the independent variables. As a side note, the interpretation of the p-value and coe???cient does not change, Holding all other variables constant, for a one-unit change in median household income, high school grad, and those ages over 65, we would expect on increases of 1.174e-06, 2.025e-03, and 1.120e-02 respectively in the independent variable, *Romney_pct*.

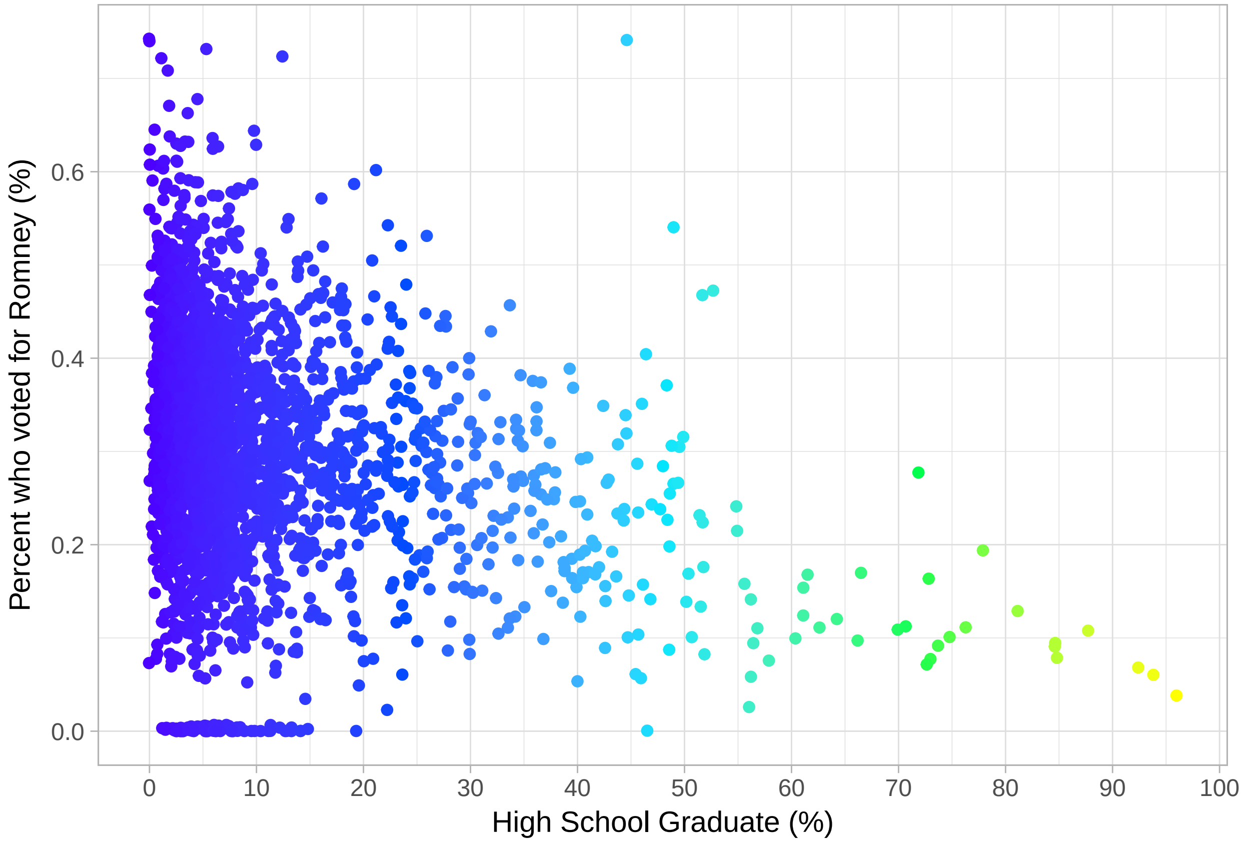
Variable: *foreign_spoken_at_home*

The variable I am adding is *foreign_spoken_at_home*, which measures the percent in each county that speak a foreign language at home. I believe this variable also influences the percent who voted for Romney as data I have seen for the most recent election indicated that voting eligible people in non-white households voted predominantly voted more liberal. Therefore, I expect that *foreign_spoken_at_home* will be negatively correlated with *Romney_pct*.

```
summary(county.complete$foreign_spoken_at_home)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.800   4.800   8.927   9.800  96.000
```

```
breaks <- c(0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
e <- ggplot(county.complete, aes(foreign_spoken_at_home, Romney_pct, colour = foreign_spoken_at_home)) +
     geom_jitter() +
     scale_colour_gradientn(colours = topo.colors(5),
                            breaks = breaks,labels = format(breaks))+
     scale_x_continuous(name="High School Graduate (%)", breaks=breaks)+
     ylab("Percent who voted for Romney (%)")+
     theme_light()+
     theme(legend.position="none")
e
```



The scatterplot indicates that there is a weak negative correlation between the two variables.

```
#Re-estimate your bivariate model (if you felt you needed to log a variable, you should continue using
m1 <- lm(Romney_pct ~ median_household_income, data=county.complete)
summary(m1)
```

```
##
## Call:
## lm(formula = Romney_pct ~ median_household_income, data = county.complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.33248 -0.05955  0.00786  0.06823  0.43359
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.863e-01  8.151e-03  35.125  < 2e-16 ***
## median_household_income 5.826e-07  1.788e-07   3.259  0.00113 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1142 on 3111 degrees of freedom
## Multiple R-squared:  0.003401,   Adjusted R-squared:  0.003081
## F-statistic: 10.62 on 1 and 3111 DF,  p-value: 0.001132
```

```r
#Now estimate a multiple regression model that includes your original independent variable and your add
m4 <- lm(Romney_pct ~ median_household_income + hs_grad + age_over_65 + foreign_spoken_at_home, data=cou
summary(m4)
```

```
##
## Call:
## lm(formula = Romney_pct ~ median_household_income + hs_grad +
##     age_over_65 + foreign_spoken_at_home, data = county.complete)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38847 -0.05452  0.00608  0.06164  0.43365
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2.538e-02  2.433e-02   1.043    0.297
## median_household_income  1.937e-06  2.325e-07   8.329   <2e-16 ***
## hs_grad                 5.690e-04  3.554e-04   1.601    0.109
## age_over_65             1.069e-02  5.018e-04  21.304   <2e-16 ***
## foreign_spoken_at_home -1.859e-03  1.834e-04 -10.132   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1014 on 3108 degrees of freedom
## Multiple R-squared:  0.2154, Adjusted R-squared:  0.2144
## F-statistic: 213.3 on 4 and 3108 DF,  p-value: < 2.2e-16
```

The regression equation for the multiple regression model is:
Y = 2.538e-02 + 1.937e-06X_1 + 5.690e-04X_2 + 1.069e-02X_3 - 1.859e-03X_4

For β_1 the t-value is 8.329 and the p-value is $<$ 2e-16. At $\alpha = 0.05$ the coefficient is statistically significant and the null hypothesis can be rejected since the p-value, <2e-16, is less than α, 0.05. In this case the hypotheses are:
H_0: $\beta_1 = 0$
H_A: $\beta_1 \neq 0$

For β_2 the t-value is 1.601 and the p-value is 0.109. At $\alpha = 0.05$ the coefficient is not statistically significant and the null hypothesis can be rejected since the p-value, 0.109, is greater than α, 0.05. In this case the hypotheses are:
H_0: $\beta_2 = 0$
H_A: $\beta_2 \neq 0$

For β_3 the t-value is 21.304 and the p-value is $<$ 2e-16. At $\alpha = 0.05$ the coefficient is statistically significant and the null hypothesis can be rejected since the p-value, $<$ 2e-16, is less than α, 0.05. In this case the hypotheses are:
H_0: $\beta_3 = 0$
H_A: $\beta_3 \neq 0$

For β_4 the t-value is -10.132 and the p-value is $<$ 2e-16. At $\alpha = 0.05$ the coefficient is statistically significant and the null hypothesis can be rejected since the p-value, $<$ 2e-16, is less than α, 0.05. In this case the hypotheses are:
H_0: $\beta_4 = 0$
H_A: $\beta_4 \neq 0$

The only coefficient estimate that had a change in statistical significance is β_2. Initially, the coeffiencent was 4.645e-03 whilst for currently β_2 is 5.690e-04. Both values are extremely close to 0 so did not change much and their statistical significance has changed. With the addition of a new independent variable *hs_grad* has lost its statistical significance.

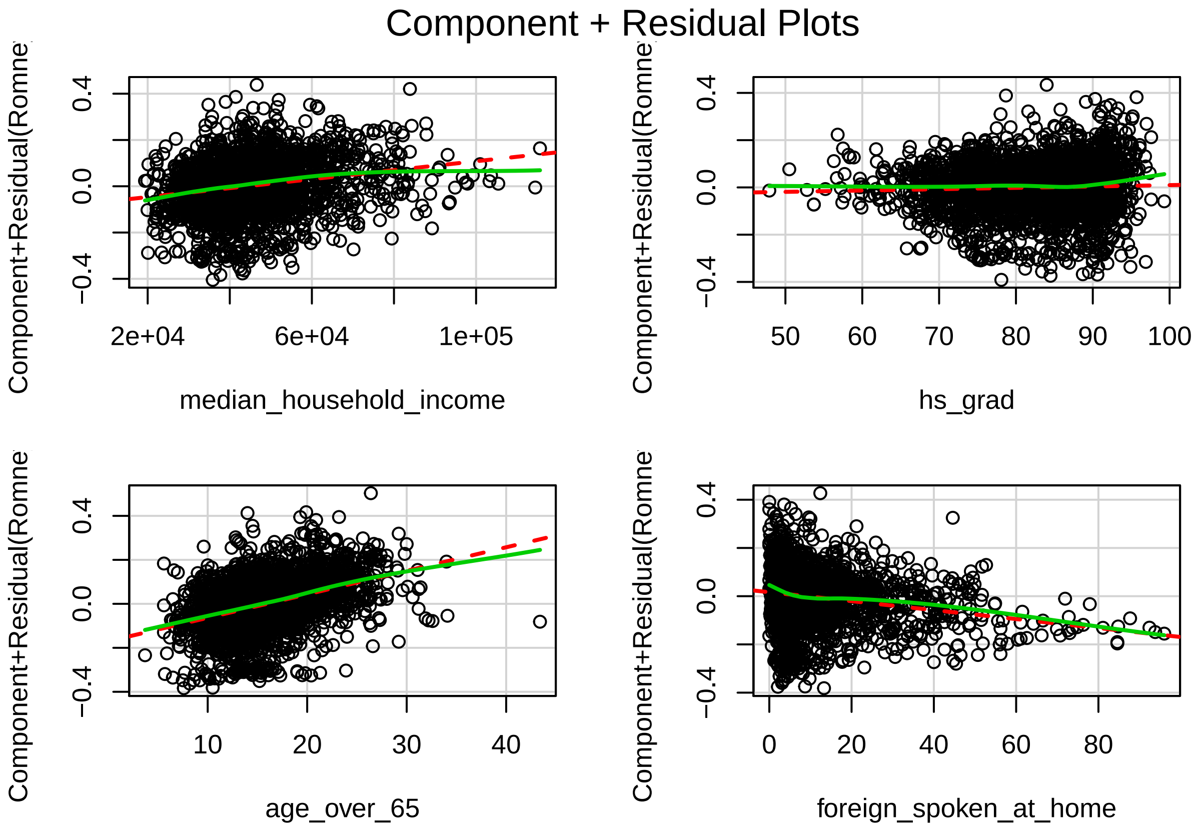
The value of multiple R^2 is 0.2154 and the value of the adjusted R^2 is 0.2144. Multiple R^2 is an indicator of how much of the variance in the dependent variable can be explained by the independent variables whilst adjusted R^2 provides the same information but adjusts for the number of terms in the model. Since the R^2 values is low, this indicates high variability that is unexplained by the independent variables in the model. It is exciting to note that in this case 21.44% of the variability can be explained by the independent variables whilst previously only 6.259% of the variability could be explained by the independent variables. However, *hs_grad* had lost its statistical significance; this means that adding more variables may explain more of the variability but is not an indication of a good model. As a side note, the interpretation of the p-value and coe???cient does not change,

Holding all other variables constant, for a one-unit change in median household income, high school grad, those aged over 65, and those who speak a foreign language at home we would expect on increases of 1.937e-06, 5.690e-04, 1.069e-02 and -1.859e-03 respectively in the independent variable, *Romney_pct*.

4A. Checking the assumptions of the multiple regression models

For this section I am going to check the assumptions for my last model which incorporates the most variables. Therefore, this can also be applied to checking the assumptions for previous models. The conditions that need to be met are:
### Linearity

```r
crPlots(m4 <- lm(Romney_pct ~ median_household_income + hs_grad + age_over_65 + foreign_spoken_at_home,
```

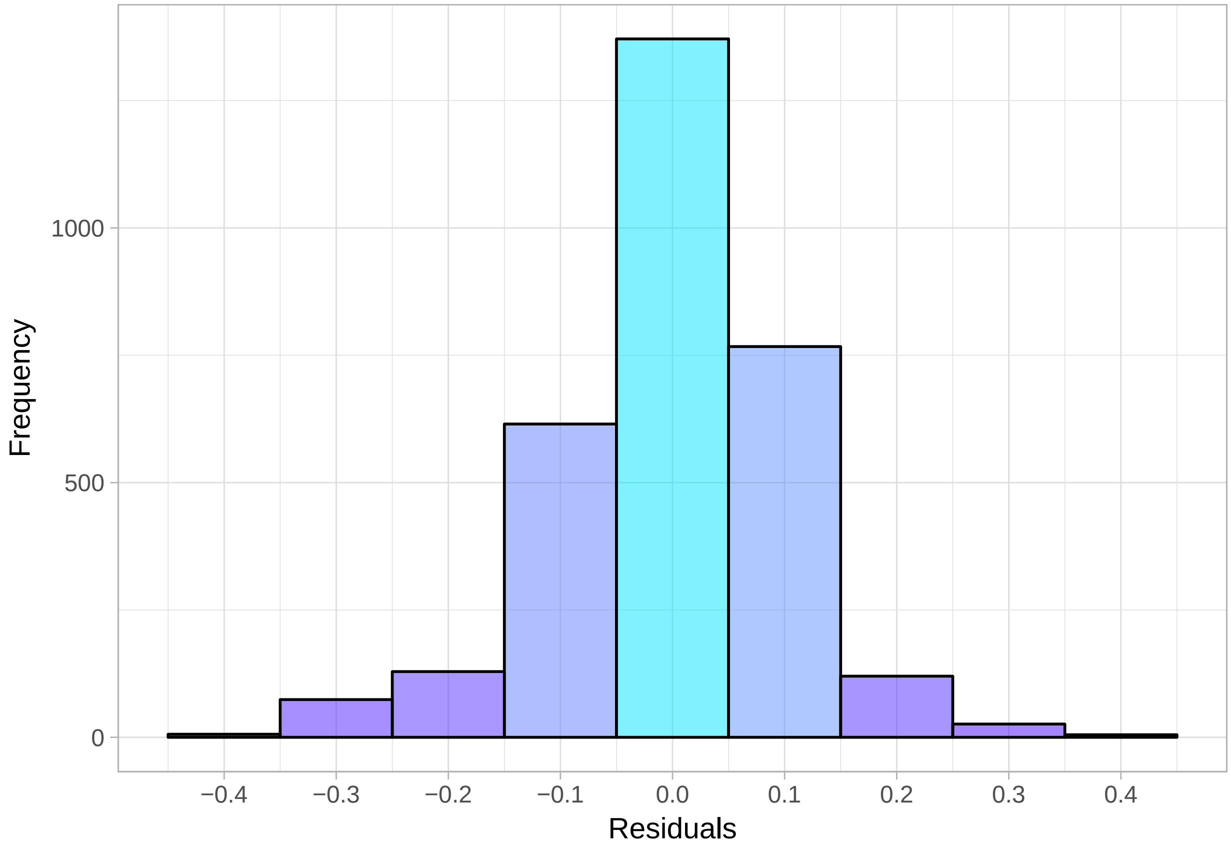


This condition is NOT met.

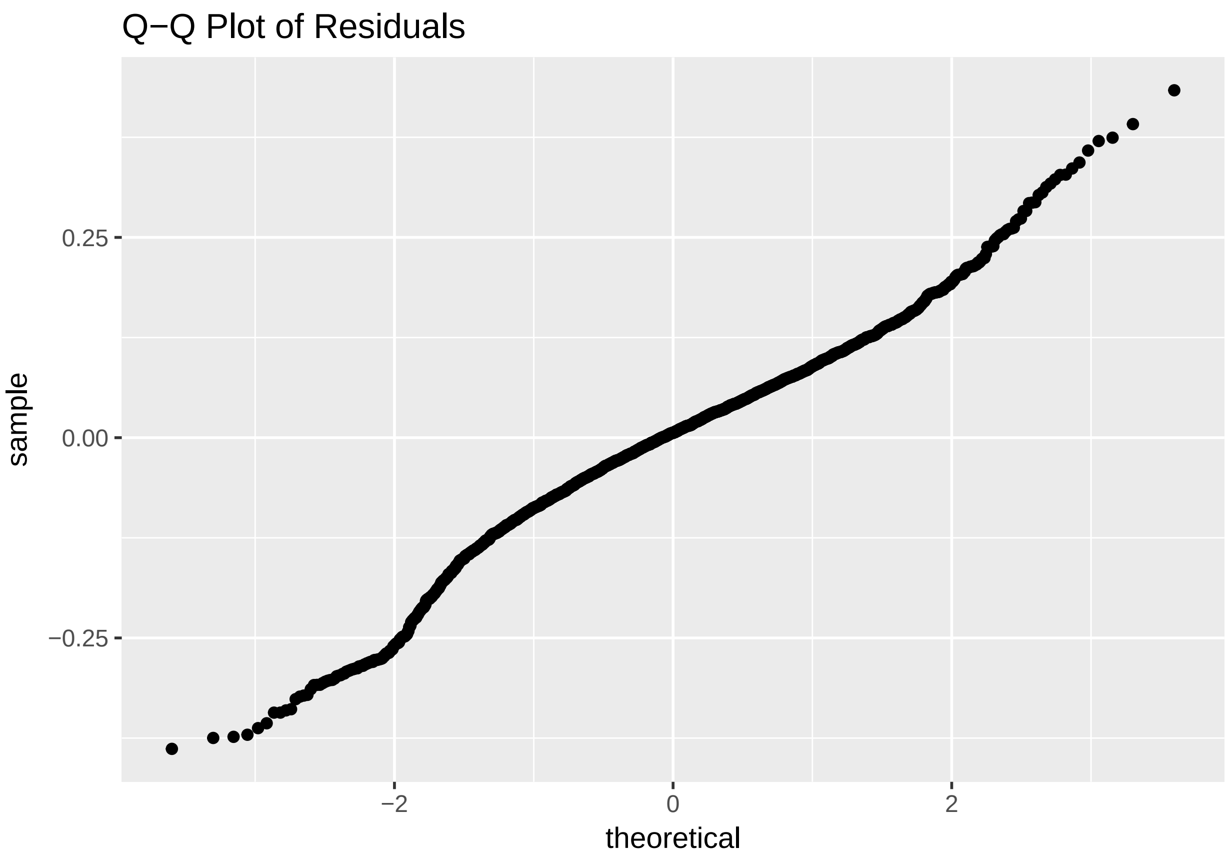
Nearly Normal Residuals

```
county.complete$residual <- m4$resid
```

```
 breaks=seq(-1, 1, by=0.1)
 m <- ggplot(county.complete, aes(x=residual)) +
      geom_histogram(binwidth=0.1,
                     colour="black", alpha=0.5,
                     aes(fill=..count..)) +
      scale_fill_gradientn(colours = topo.colors(2)) +
      scale_x_continuous(name="Residuals", breaks) +
      ylab ("Frequency") +
      theme_light() +
      theme(legend.position="none")
m
```



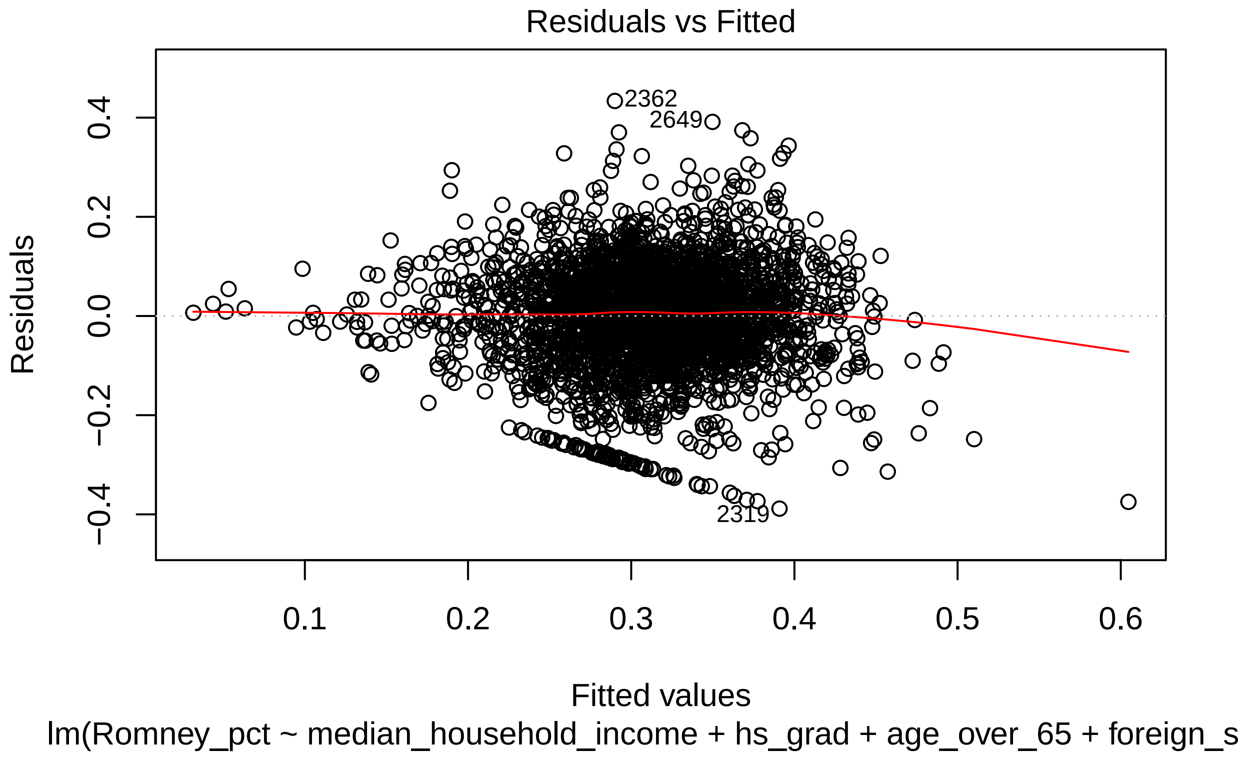
```
ggplot(county.complete, aes(sample = residual)) +
  ggtitle("Q-Q Plot of Residuals") +
  stat_qq()
```



The residuals are approximately normal so this condition is met.

Constant Variability

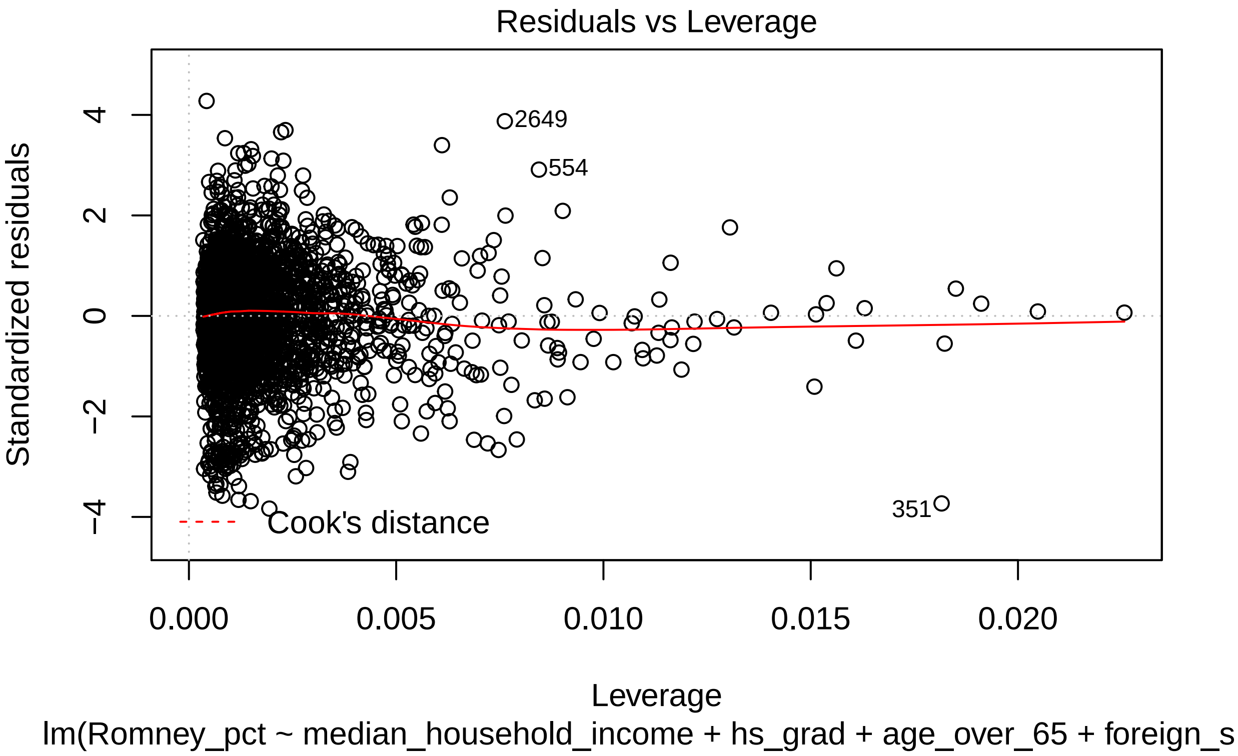
```r
plot(m4,1)
```



The residuals appear to be approximately normal (especially when you exclude the outlier).

No multi-collinearity between predictors

```r
plot(m4,5)
```



```r
vif(m4)
```

```
## median_household_income                 hs_grad             age_over_65
##                2.146760                2.059580                1.303347
##  foreign_spoken_at_home
##                1.303411
```

Since none of the VIF values are greater than 10 there is no indication of multi-collinearity between the predictors.

5. Return to your research question

Write a short paragraph that relates the results of your favored multiple regression model to your research question. This should include a formal interpretation of your model, as well as a narrative discussion of what the model tells you about the empirical relationship you set out to investigate. What statistical issues have you encountered? How do these impact the conculsions you can draw?
My favored multiple regression model is the latest one in which I used the variables *median_household_income*, *age_over_65*, *hs_grad* and *foreign_spoken_at_home* along with the dependent variable *Romney_pct*. The formal interpretation of the regression model is:
Holding all other variables constant, for a one-unit change in median household income, high school grad, those aged over 65, and those who speak a foreign language at home we would expect on increases of 1.937e-06, 5.690e-04, 1.069e-02 and -1.859e-03 respectively in the independent variable, *Romney_pct*.
However, due to the fact that the first assumption of a multiple regression model is not met no inference can be made from the model. Therefore, I have run into statistal issues that impact the conclusions that I can make. The relationship between the variables ultimately appears to be random.

6. Reflection

Statistics wise, I learned about multiple regression models and the underlying assumptions. However, the most striking thing I have taken away from this assignment is how to deal with dissapointment. Going in to the assignment I was excited to be able to improve my linear regression and thought that I would be able to make a meaningful, effective regression model. When I was not able to get the results that I wanted I was frustrated at first, but then I realized that proving my underlying assumptions going into the assignments at the beginning of the quarter wrong is meaningful itself.