Report

Team IG

12/09/2019

Introduction

Initially, this project is to explore if the racial diversity affects the negative relationship between union density and vocational education. Two questions are mainly analyzed. First, if the negative relationship actually exists between union density and vocational education. Second, if racial diversity mediates the relationship.

Research question

According to the VoC theory, the United States is defined as liberal market economies and one characteristic of this market type is that the union density is proportional to the vocational training level. While, the previous presented model contains 9 variables from dataset and defined as a linear model and the result shows below:

$$Highschool = 4.754\exp(5) - 1.692exp(4)Uniondensity + 6.629exp(-2)GDP - 1.045exp(3)Urbanity +$$

$$2.804exp(2)Population - 5.176exp(1)Export + 1.909exp(4)Unemployrate + 1.241exp(3)Costofliving - 9.598exp(5)Gini$$

The coefficient of high school which indicates there exists negative relationship between union density and high school. This apparently conflicts with properties of LME in VoC theory. Accordingly, a potential explaination comes that racial diversity in each state might leads to this negative relationship between vocational training and union density.

Literature review

Market-baised management: Market-Based Management (MBM) enables organizations to succeed in the long term by applying the principles that allow free societies to prosper. Just as upholding values such as free speech, property rights, and progress is important to a healthy, growing society, it is also pivotal in fostering a healthy, growing organization.

Nation-level/state level: US is called LME because its national governing framework does not has institutions that aggregate preference for public goods, such as high-skill development.

Sources:

- Stephen Amberg. Liberal Market Economy or Composite Regime? Institutional Legacies and Labor Market Policy in the United States. Retrieved from https://www.jstor.org/stable/40213467
- David V. Budescu/Mia Budescu. How to Measure Diversity When You Must. Retrieved from https://www.researchgate.net/publication/221810375_How_to_Measure_Diversity_When_You_Must

Dataset interpretation

Part I

Data Descriptions:
The initial dataset sent has 31 variables for the total of 51 states in the US. Not all the variables are used. Based on the content of the project, the dependent variable is the high school student that represents the vocational education, and the independent variables includes the predictor and other confounding variables. Predictor is the union density, and confounding variables are racial diversity, GDP, urbanity, population,

export, cost of living, unemployment rate and Gini. To better represent the racial diversity, a normalized ratio diversity variable has been added to the dataset rather than simply using the number of White, Black and Hispanic people.

Data Governance:

- Inconsistency. The source and the year of the data are inconsistent. Data are sourced separately from Bureau of Labor Statistics, Wikipedia, U.S. Census Bureau and Census Bureau's March Current Population Survey. There might be difference on each data from different source. Data for each variable is collected in different years, from 2016 to 2018.
- Misrepresentatives of vocational education. In the original dataset, the vocation education is represented by the number of high school students. However, based on the VoC theory, the government plays an important role in determine the level of vocational education. Therefore, it might be more helpful to use the supply side of vocational education rather than the demand side.
- Misrepresentatives of racial diversity. The racial diversity is initially represented by the number of White / Black / Hispanic people. However, based on the literature about racial diversity, the normalized racial diversity that is calculated by the proportion of different races is more reasonable. Therefore, a new variable named "normalized racial diversity" is added into the dataset.

Part Two

Data Descriptions:

Variable Name	Interpretation	Time Span
Urbanity	Quality of being urban in each state	2010
Unemployment	Unemployment rate in each state	2009-2017
GDP_Per_Capital	GDP/Population number	2005-2017
Union	Union degree for each state	2005-2017
Voe	Quantity of vocational training people	2007-2017
Racial(black/hispanic/white)	Racial proportion for each race	2005-2017
Totalpop	Total population	2009-2017
Studentpop	Student population	2009-2017

Data Governance:

- Missing data. Looking into the second dataset, values for some variables between 2005 and 2009 are missing. Therefore, data from 2009 to 2017 are used for model fitting.
- Because the urbanity only has data at the year of 2010, the same value is applied to every single year analysis.

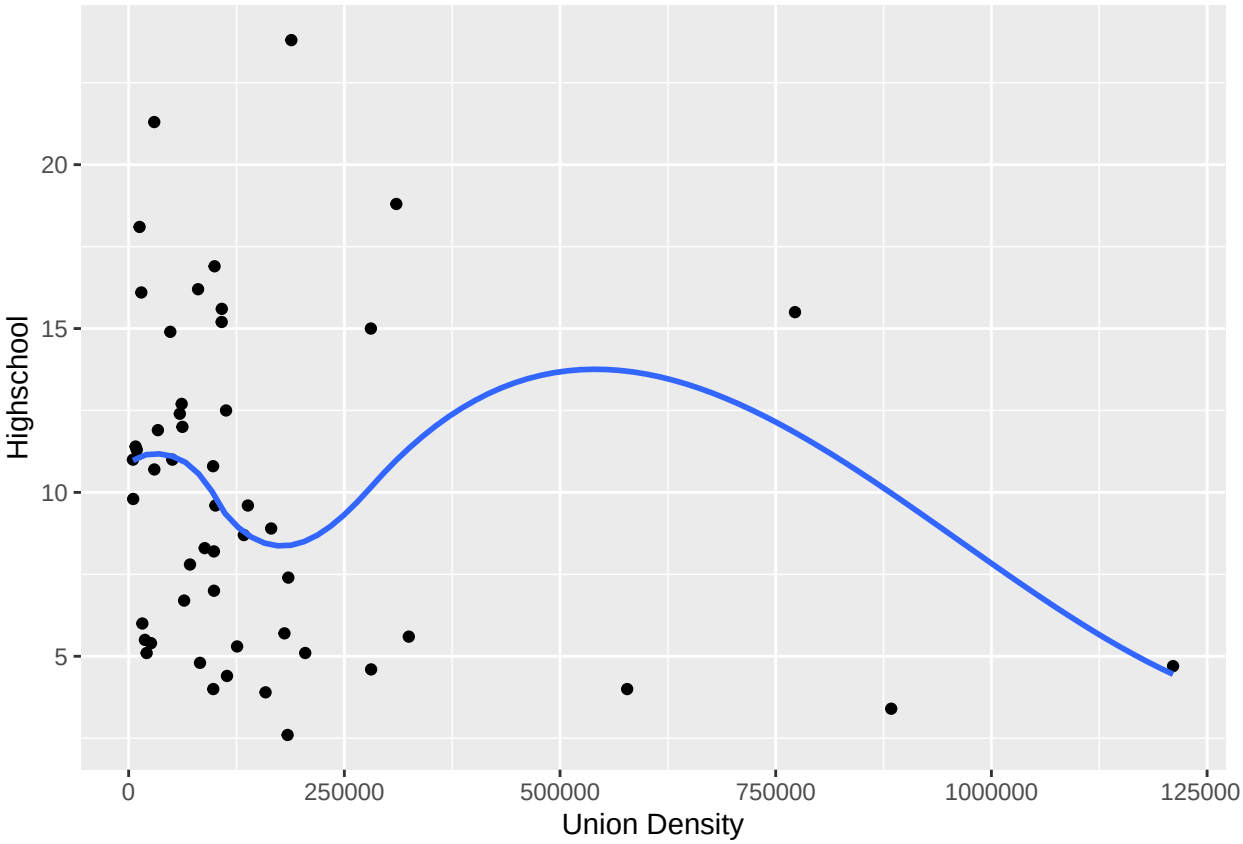
Methodology

According to the research question, the following things needs to be done: Data check, Conducting the initial EDA(exploratory data analysis), model check.

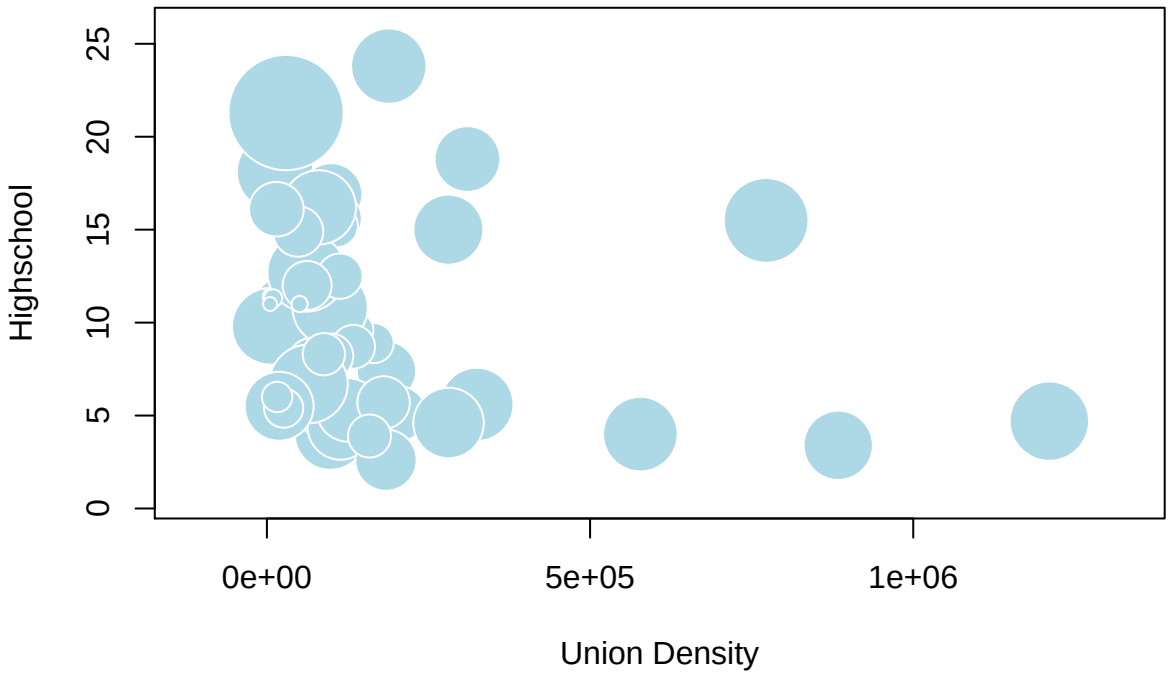
- To better measure the racial diversity, a new variable based on White/Black/Hispanic is created.
- In the initial EDA part, scatterplots and bubble plot are presented to show the rough corelationship between union density and vocational training to get the general idea of the question.
- Make transformation for linear regression model to see if the result turns to be more reasonable.
- Drawing the component residual plot(CR plot) to check every correlationship between independent variables and response variables.
- Partial corelation test has been used to check if there still exists significant negative relationship between two variables when control the other 7 confounding variables.
- Mediation analysis has been conducted to check if racial diverisity could be regarded as a intermediate influencing factor of union density and vocational training.

Part ONE

Exploratory Data Analysis



This plot visualizes the potential relationship with x-axis set as union density and y-axis set as vocational education. From this plot, no obvious negative relationship can be observed.



Moreover, a bubble plot is presented. Same as former plot, setting union density as independent variable and high school as dependent variable, but this time replacing every spot with a single circle. Each circle represents for one state and the radius of circle stands for racial diversity level, which means, the higher racial diversity level, the larger circle is. Same as previous result, by observing this plot we could not confirm these two variables have a negative relationship.

Methodology and Result

Replicate Original Model

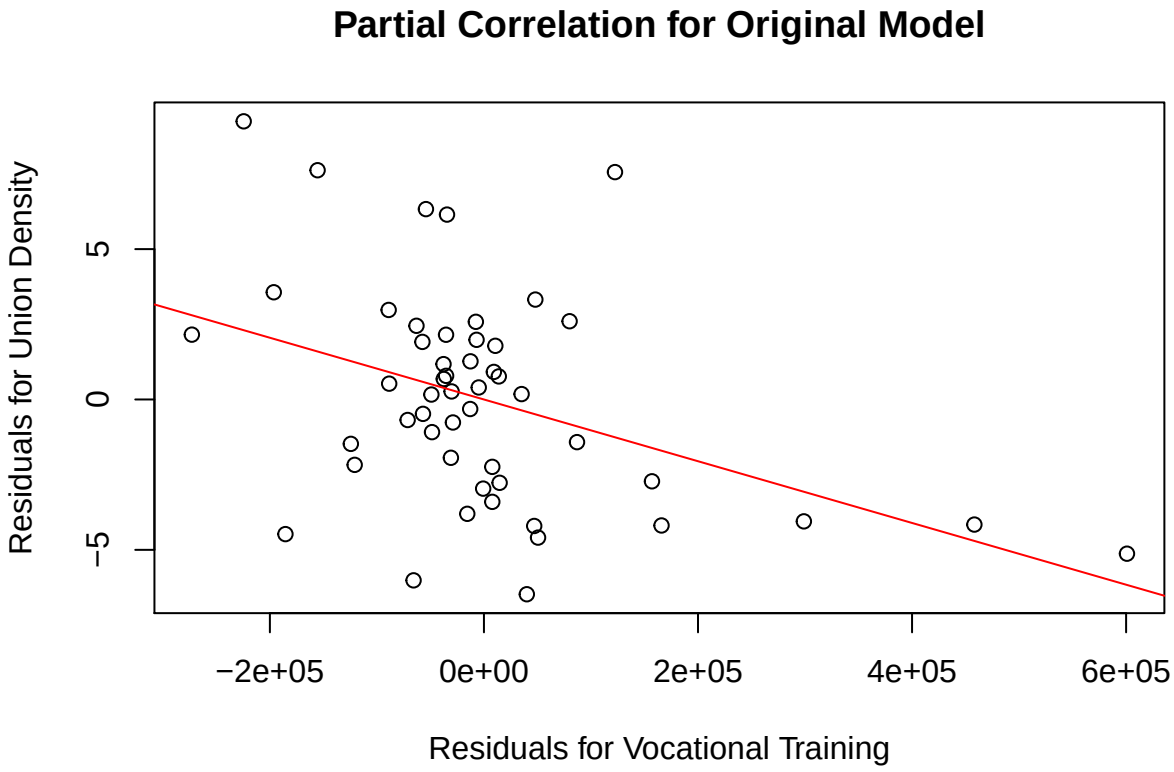
Since the initial EDA does not show a distinctive pattern of the negative relationship between vocational training and unionization, which is different from the result provided by the model in the second-year paper, then a replication of the original model conducted to confirm the model result.

```
##
## Call:
## lm(formula = Highschool ~ Union_density + GDP + Urbanity + Population +
##     Export + Unemployment_rate + Cost_of_living + Gini, data = olddata3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -261208  -64360  -25121   37190  514055
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        4.754e+05  5.014e+05   0.948  0.34852
## Union_density     -1.692e+04  5.694e+03  -2.971  0.00489 **
## GDP                6.629e-02  1.358e+00   0.049  0.96129
## Urbanity          -1.045e+03  1.826e+03  -0.572  0.57028
## Population         2.804e-02  3.834e-03   7.313  5.2e-09 ***
## Export            -5.176e-01  5.463e-01  -0.947  0.34882
## Unemployment_rate  1.909e+04  2.453e+04   0.778  0.44086
## Cost_of_living     1.241e+03  1.669e+03   0.743  0.46136
```

```
## Gini              -9.598e+05  1.142e+06  -0.840  0.40546
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 144200 on 42 degrees of freedom
## Multiple R-squared:   0.67,  Adjusted R-squared:  0.6072
## F-statistic: 10.66 on 8 and 42 DF,  p-value: 5.037e-08
```

Partial Correlation for Original Model

Because there are other factors that will impact the outcome, vocational education, partial correlation analysis is implemented to better investigate the relationship between union density, specifically and vocational education. Partial correlation is a measure of the strength and direction of a linear relationship between two continuous variables while controlling for the effect of other continuous variables. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and −1. A value of ś1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker.



When conducting partial correlation analysis, a method for computing the partial correlation coefficients has to be specified. Two methods are used in this project. One is the Pearson method, which evaluates the linear relationship between two continuous variables. A relationship is linear when a change in one variable is associated with a proportional change in the other variable. The following formula is used to calculate the Pearson r correlation:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum {x_i}^2 - (\sum x_i)^2}\sqrt{n\sum {y_i}^2 - (\sum y_i)^2}}$$

r_{xy} = Pearson r correlation coefficient between x and y
n = number of observations
x_i = value of x (for ith observation)
y_i = value of y (for ith observation)

Another is the Spearman method, which evaluates the monotonic relationship between two continuous or ordinal variables. In a monotonic relationship, the variables tend to change together, but not necessarily at a constant rate. The Spearman rank correlation test does not carry any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal. The following formula is used to calculate the Spearman rank correlation:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ρ = Spearman rank correlation
d_i = the difference between the ranks of corresponding variables
n = number of observations

Pearson Method

```
## [1] -0.4167175
```

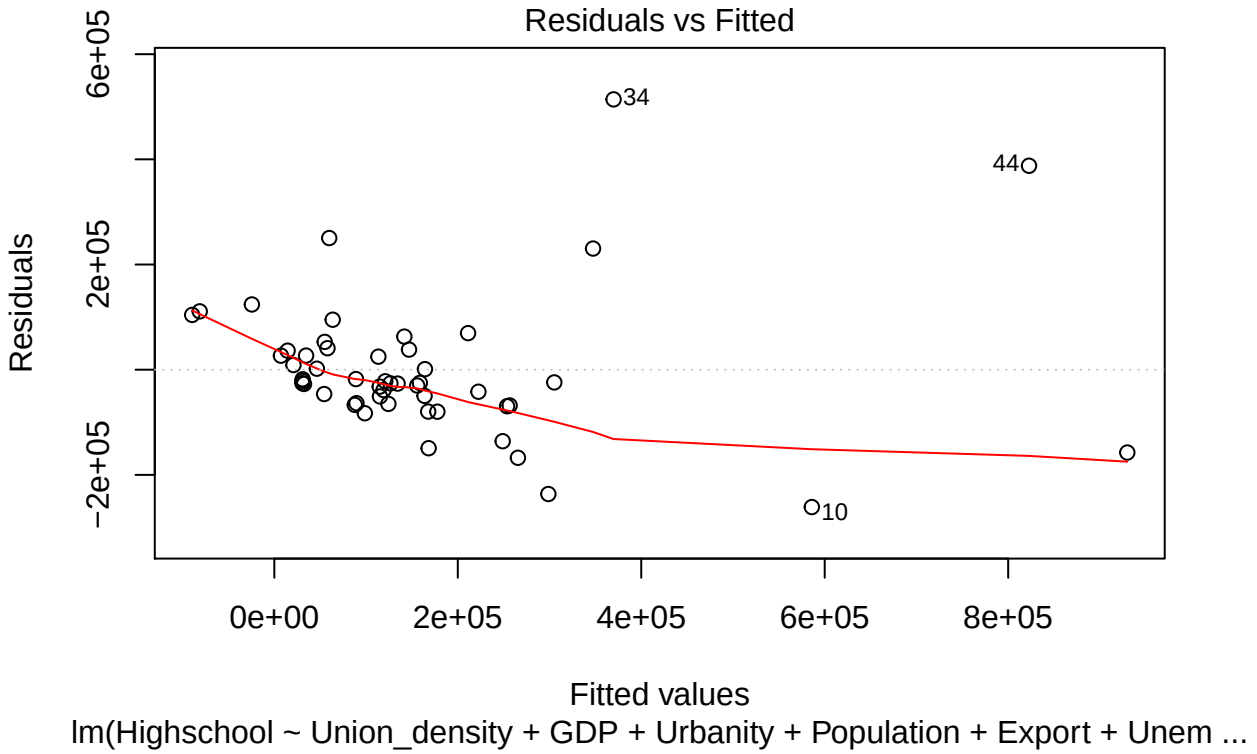
```
## [1] 0.002351165
```

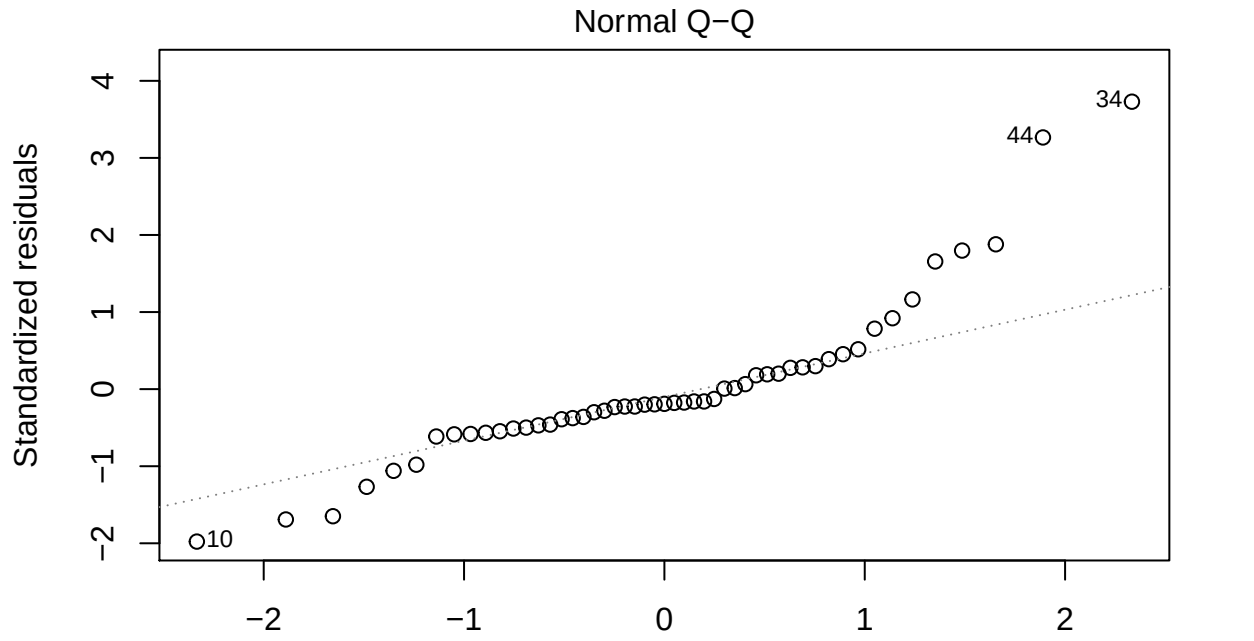
Spearman Method

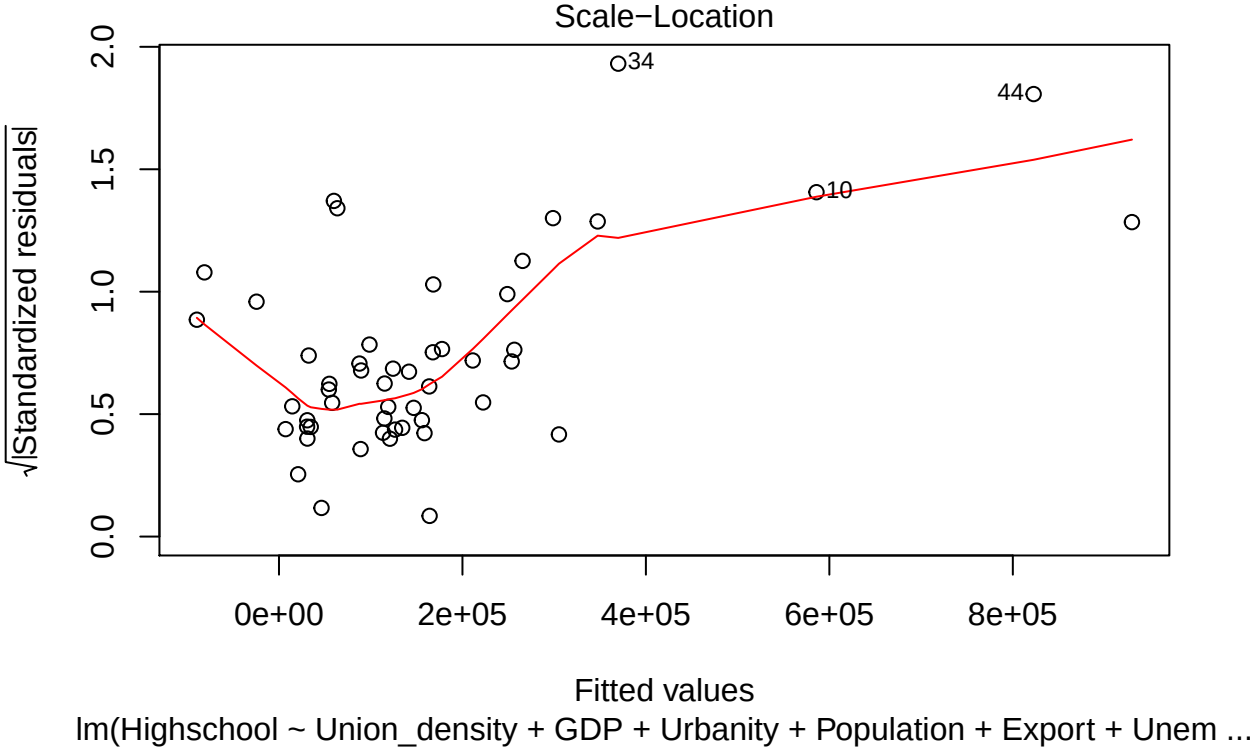
```
## [1] -2.7575
```

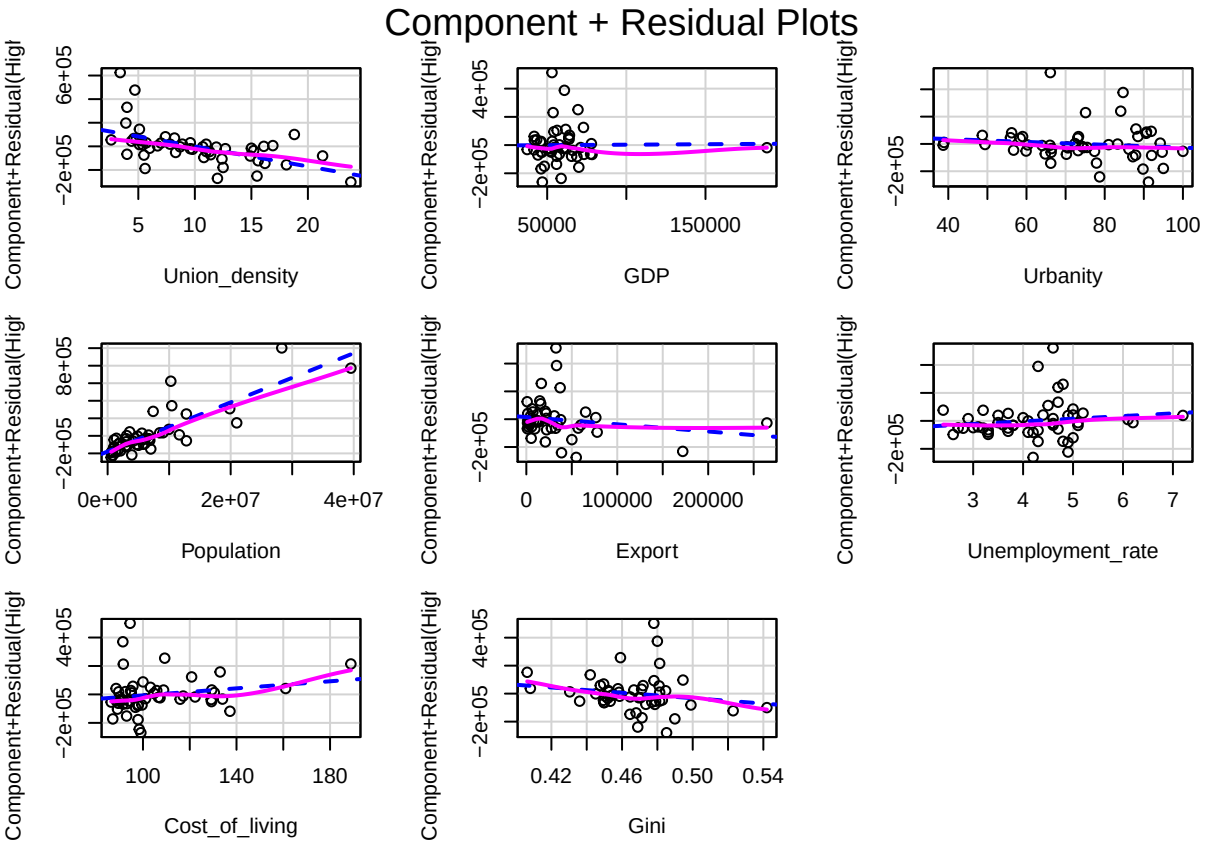
```
## [1] 0.008161668
```

From the output of two methods, correlations between union density and vocational education are negative. The small P-value (<0.05) indicates that the negative coefficient of Union density is statistically significant in the primary model. From the plot on residual of Union density and residual of Highschool, the red line, which represents the relationship between union density and vocational education, has an apparent downward trend. However, by looking at the points on the graph, there is a leverage point in the lower right corner, indicating some data transformation might be needed.




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Highschool ~ Union_density + GDP + Urbanity + Population + Export + Unem ...
34
44
10


Scale−Location
√|Standardized residuals|
Fitted values
lm(Highschool ~ Union_density + GDP + Urbanity + Population + Export + Unem ...
34
44
10



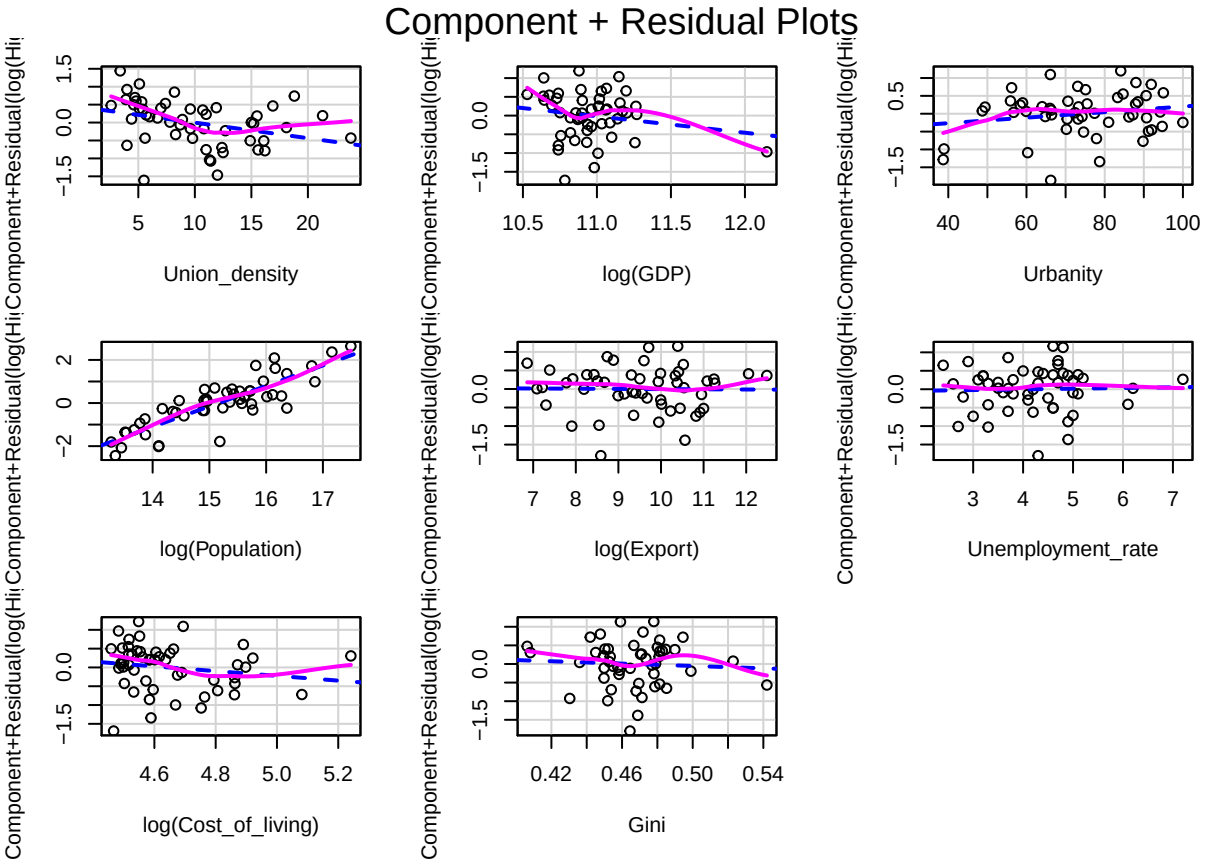
The plots above are used to check the assumptions of Linear Regression. These assumptions are linearity assumption, normality assumption, and constant variance assumption. The first plot shown here is Residuals vs Fitted plot which is used to check linearity assumption. Ans ideal plot should show the equally spread residuals around the horizontal line without any distinct patterns. The plot above indicates the violation of the linearity assumption. The normality assumption can be checked by the Q-Q plot. The ideal plot of residuals should approximately follow a straight line. Some points are skewed at the top right of the plot which indicates the violation of normality assumption. The third plot, which is a scale-location plot can be used to detect the assumption of equal variance. An ideal plot will show a horizontal line with randomly spread points. In our case, the constant variance assumption is violated. If the assumptions of Linear Regression are not met, it indicates that a transformation in the linear regression is required.

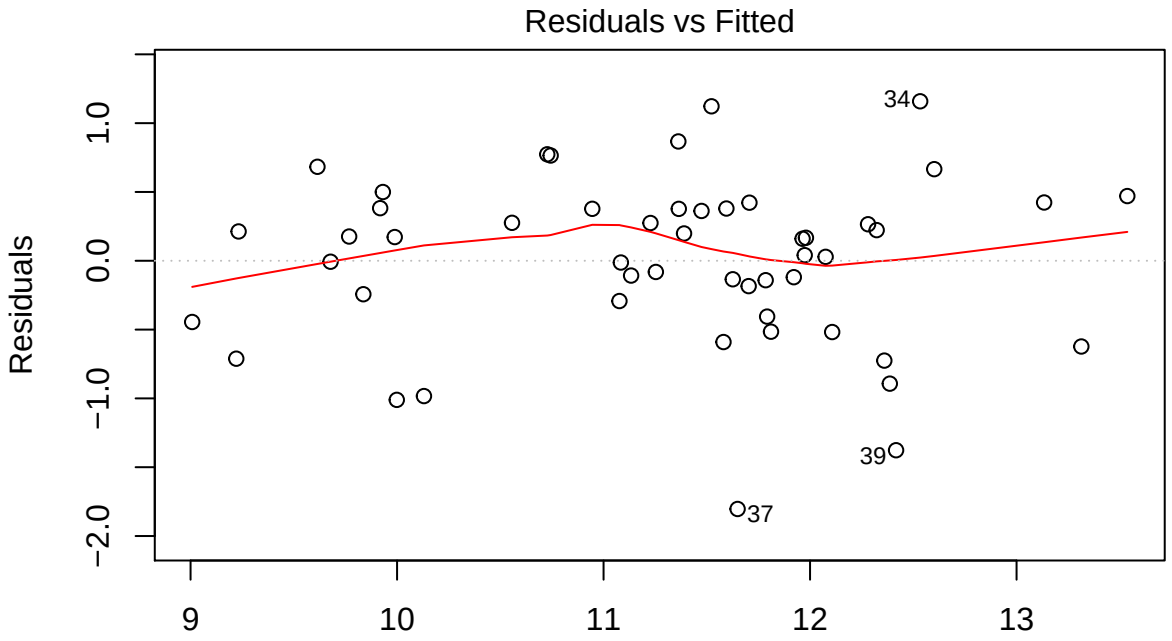
Transformation

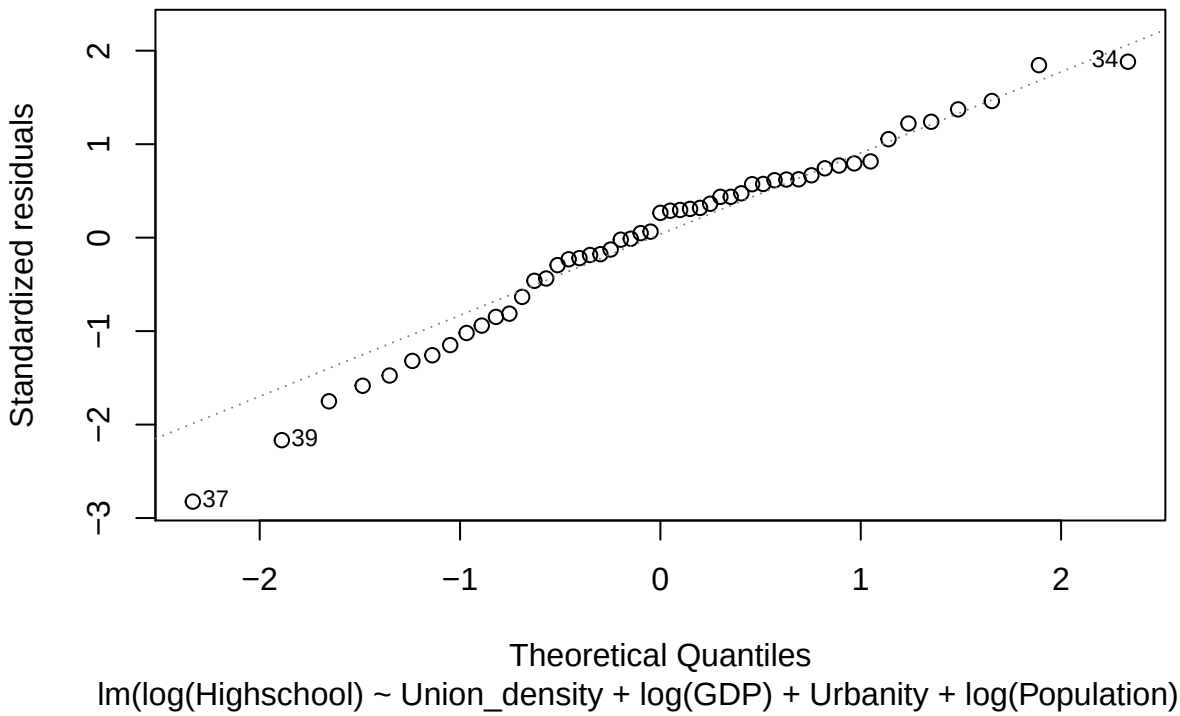
- Use CrPlots to determin which variable to take log transformation
- Model validation

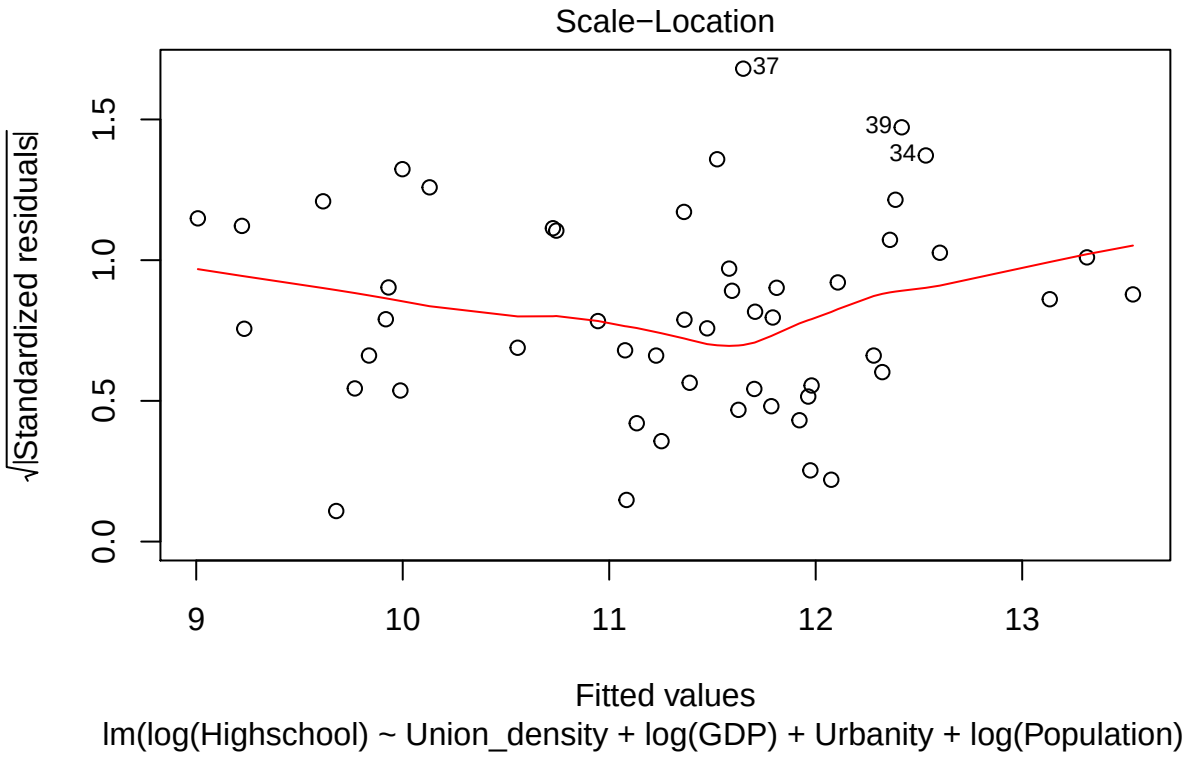
```
##
## Call:
## lm(formula = log(Highschool) ~ Union_density + log(GDP) + Urbanity +
##     log(Population) + log(Export) + Unemployment_rate + log(Cost_of_living) +
##     Gini, data = olddata3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8035 -0.3491  0.1601  0.3786  1.1588
##
## Coefficients:
```

```
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           5.198356   6.701335   0.776    0.442
## Union_density        -0.043265   0.027298  -1.585    0.120
## log(GDP)             -0.438462   0.545691  -0.803    0.426
## Urbanity              0.007762   0.009767   0.795    0.431
## log(Population)       0.950302   0.146685   6.479 8.14e-08 ***
## log(Export)          -0.005278   0.092754  -0.057    0.955
## Unemployment_rate     0.018547   0.112717   0.165    0.870
## log(Cost_of_living) -0.632347   0.983668  -0.643    0.524
## Gini                -1.585547   5.110432  -0.310    0.758
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6602 on 42 degrees of freedom
## Multiple R-squared:  0.7649, Adjusted R-squared:  0.7202
## F-statistic: 17.09 on 8 and 42 DF,  p-value: 5.888e-11
```




Residuals vs Fitted
Residuals
1.0
0.0
−1.0
−2.0
34
39
37
9
10
11
12
13
Fitted values
lm(log(Highschool) ~ Union_density + log(GDP) + Urbanity + log(Population) ...


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
34
39
37
−2
−1
0
1
2
Theoretical Quantiles
lm(log(Highschool) ~ Union_density + log(GDP) + Urbanity + log(Population) ...

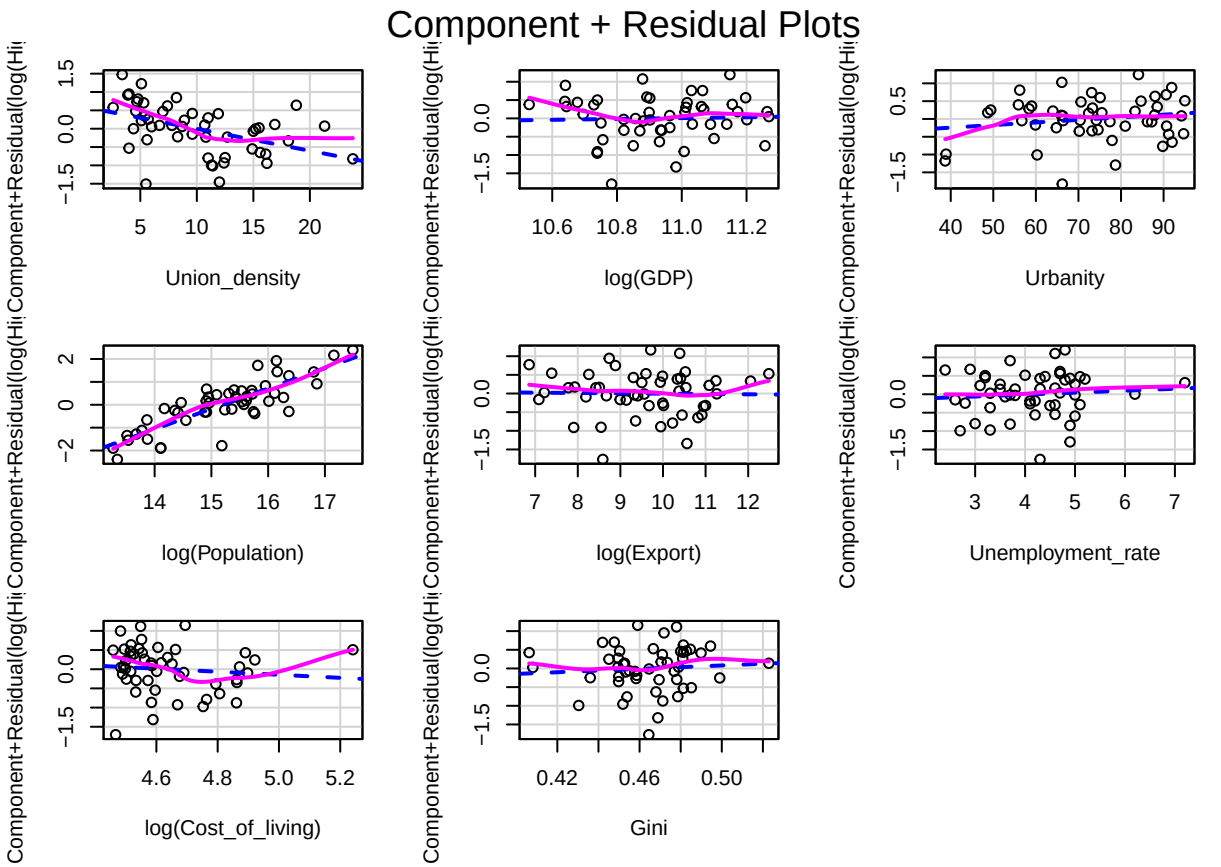


Because the original model does not in line with the assumption, data transformation is used to adjust the model. To use the linear regression model, the most significant assumption is that variables should be in the linear relationship with the outcome. Component residual plot is used to measure if every variable satisfies linearity assumption. The red line shows the pattern of the datapoints included in every variable, and the dashed line indicates the best of the fit. If the red line is close to the dashed line and does not show curve pattern, meaning that the linearity assumption is satisfied for that variable. Based on the initial component residual plot of the original model, leverage points exist in datapoints from different variables including "GDP", "population", "Export", "unemployment rate" and "cost of living". There are issues with leverage points since it would directly change the pattern of the residual plots. For example, looking at the residual plots of "Export", if the leverage points move upwards or downwards, the red line and the dashed line would move accordingly. The movement is not caused by the movement of the most datapoints, rather caused by the leverage points, which is not the ideal case where the red line should indicate the pattern of the majority of datapoints. To avoid the leverage points, log transformation is used to build the first model, "fit1". In model "fit1", log transformation has been implemented for the outcome, "highschool" and variables that have leverage points, "GDP", "population", "Export", "unemployment rate" and "cost of living". After summarizing the model "fit1", the coefficient between the outcome "highschool" and the variable "union density" is -0.04, not significant. Because the dataset is a sample from the population, coefficient for each sample would be different because of the sampling bias arising from picking up different samples within the population. The insignificant coefficient indicates that the negative relationship between vocational training and union density is not conclusive.

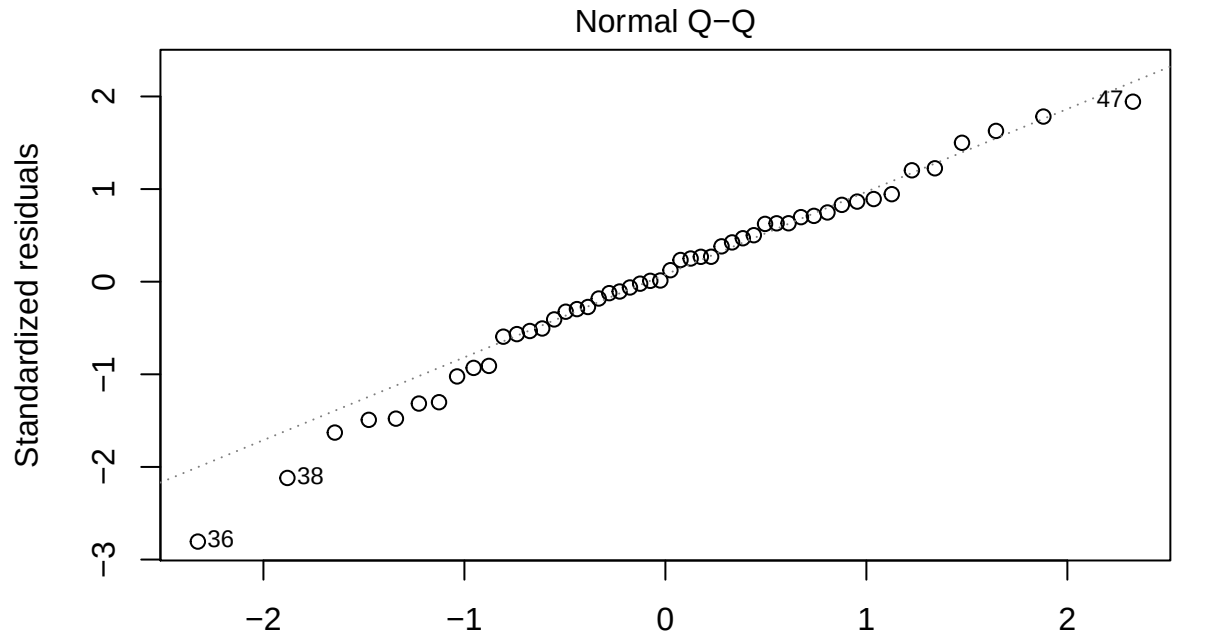
After modeling, residual plots are generated to check the validation of the model. Looking at the residual plot, nearly all the residuals are equally distributed within the horizontal line valued at zero, indicating that the residuals are random, and the linearity assumption is satisfied. Looking at the Q-Q plot, majority of the residual points are aligned with the straight line, indicating the normality assumption is nearly satisfied. Looking at the scale-location plot, most residues are spread equally along the ranges of predictors, indicating the equal variance assumption is satisfied. Based on the residual plot, the model "fit1" is more valid compared to the original model. Check the component and residual plot again, the issue regarding the leverage points have been improved a lot. But two issues still have to be considered. First, there is a leverage point in GDP after taking log transformation. Second, the variable of union density slightly shows a curved pattern,

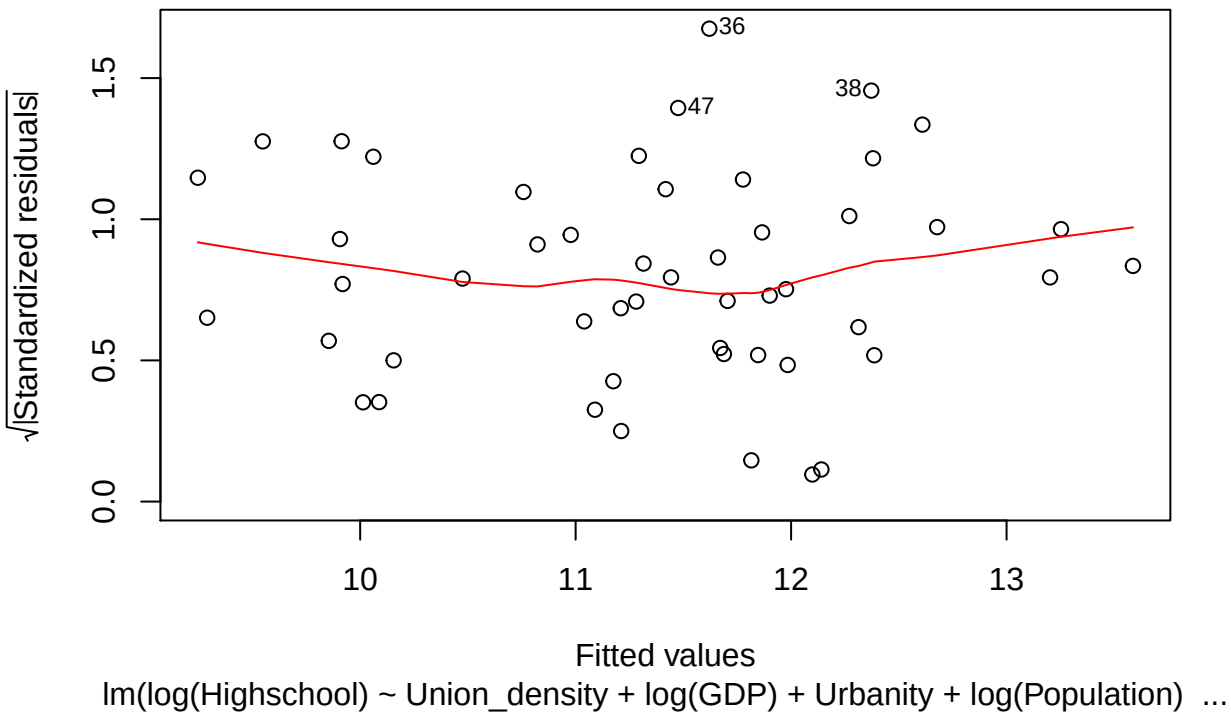
indicating that linear assumption is not well satisfied, and a more robust model might be considered.

```
##
## Call:
## lm(formula = log(Highschool) ~ Union_density + log(GDP) + Urbanity +
##     log(Population) + log(Export) + Unemployment_rate + log(Cost_of_living) +
##     Gini, data = olddata4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77486 -0.31688  0.03988  0.41357  1.16933
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -2.757715   8.934208  -0.309   0.7591
## Union_density       -0.059829   0.029775  -2.009   0.0511 .
## log(GDP)             0.119772   0.684302   0.175   0.8619
## Urbanity             0.007271   0.009686   0.751   0.4571
## log(Population)      0.889431   0.152379   5.837 7.38e-07 ***
## log(Export)         -0.008890   0.091954  -0.097   0.9234
## Unemployment_rate    0.052305   0.114538   0.457   0.6503
## log(Cost_of_living) -0.397213   0.990627  -0.401   0.6905
## Gini                 2.270601   5.834118   0.389   0.6991
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6542 on 41 degrees of freedom
## Multiple R-squared:  0.7502, Adjusted R-squared:  0.7014
## F-statistic: 15.39 on 8 and 41 DF,  p-value: 3.764e-10
```


Component + Residual Plots
Component+Residual(log(Hi
Union_density
log(GDP)
Urbanity
log(Population)
log(Export)
Unemployment_rate
log(Cost_of_living)
Gini

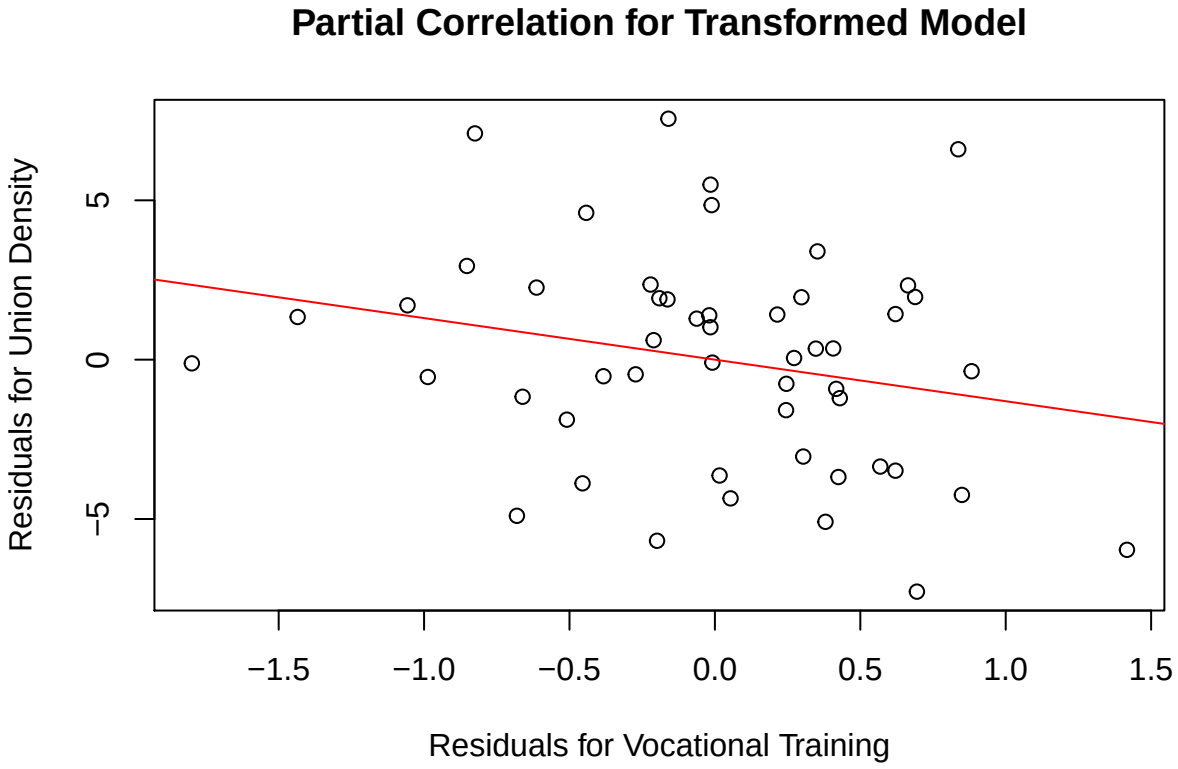

Residuals vs Fitted
Residuals
47
38
36
Fitted values
lm(log(Highschool) ~ Union_density + log(GDP) + Urbanity + log(Population) ...





To deal with the leverage point in log GDP, one of the solutions is to leave it out. The leverage point is Washington DC with highest log GDP. Another model (fit2) with exactly same variables and outcome is created with a new dataset excluding datapoint of Washington DC. After summarizing the model "fit2", the coefficient of union density changes 20%, indicating the difference is not ignorable.Thus, the leverage point of Washington DC cannot be simply leaved out.

Partial Correlation for Transformed Model



Pearson Method

```
## [1] -0.2375583
```

```
## [1] 0.09323562
```

Spearman Method

```
## [1] -0.2179186
```

```
## [1] 0.1244902
```

Partial correlation analysis is implemented again with transformed variables to check the relationship between union density and vocational education. Looking at the partial correlation plot, the red line becomes flatter.
Based on the output from two methods, the partial correlation coefficient between Union density and Vocational training are smaller, and P-values are larger. Thus, in the transformed model, the negative relationship between the two union density and vocational training is not as strong as it in the original model.

Adding Racial Diversity

To analyze if racial diversity can explain the relationship, the next model adds racial diversity as another confounding variable and check if there is any difference on the coefficients.

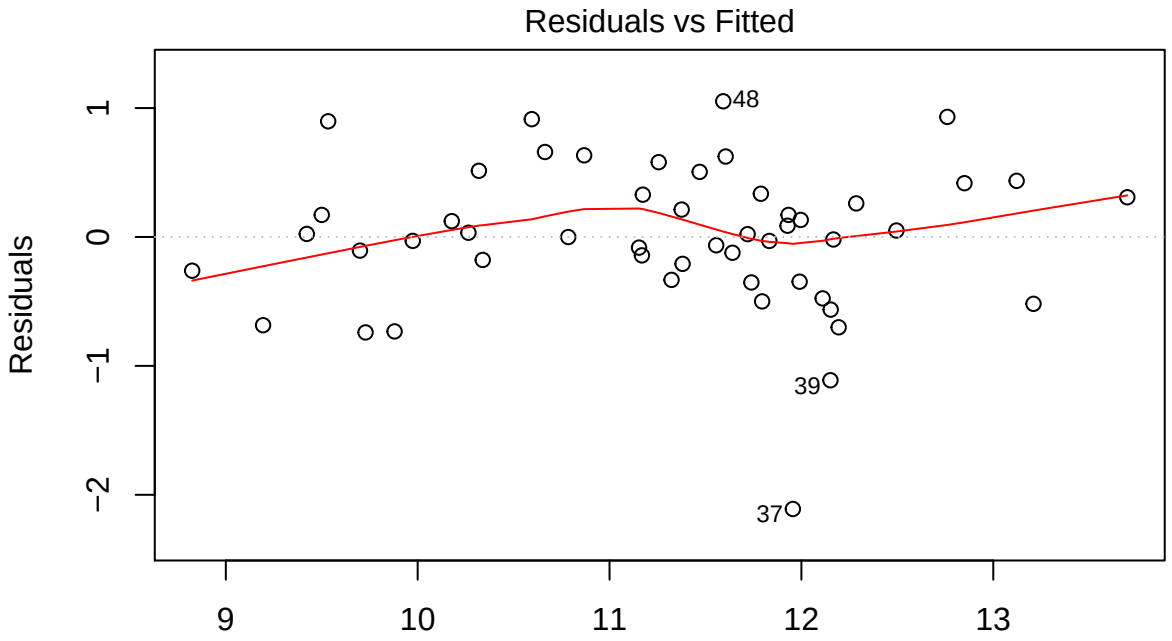
As mentioned before, rather than using the single proportion of Black, White, and Hispanic people, a new variable named normalized racial diversity is used based on the article "How to Measure Density When You Must". The normalized racial diversity is calculated by the proportion of same race which equals to the sum square Black, square White adds, and square Hispanic, then the proportion of different race equals to one minus the proportion of same race. However, in the real world, different states may have different types of races, and different types of the race will affect the result of racial diversity. To alleviate this kind of problem

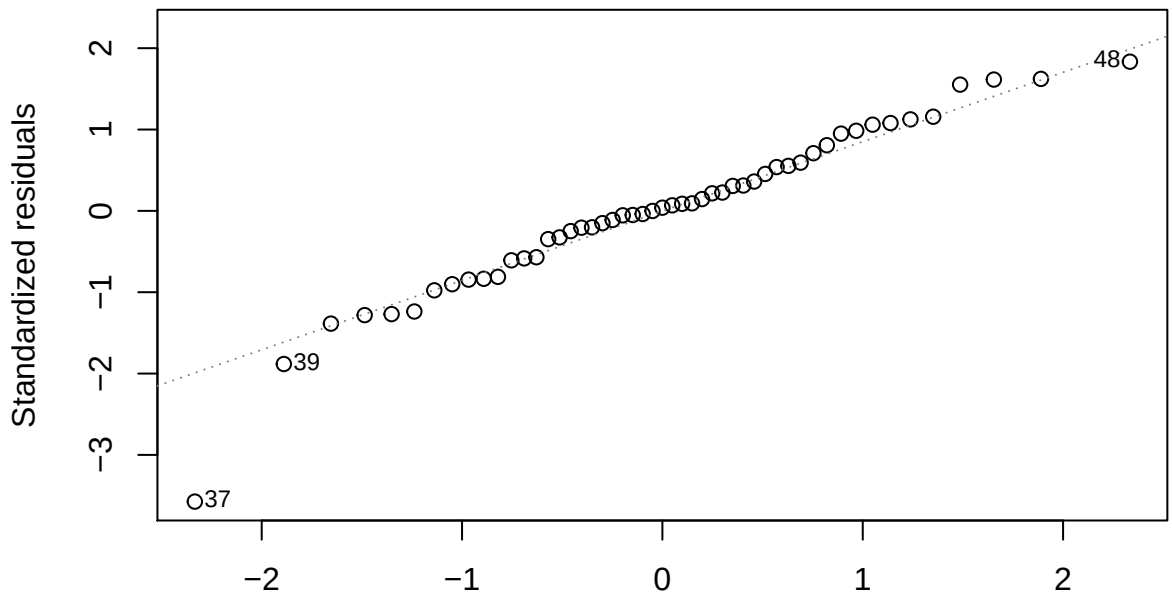
that can arise, racial diversity needs to be normalized before adding to the model. Times the radical diversity by C/(C-1) can get the normalized racial diversity, C means the total type of race. In this case, C is 3.

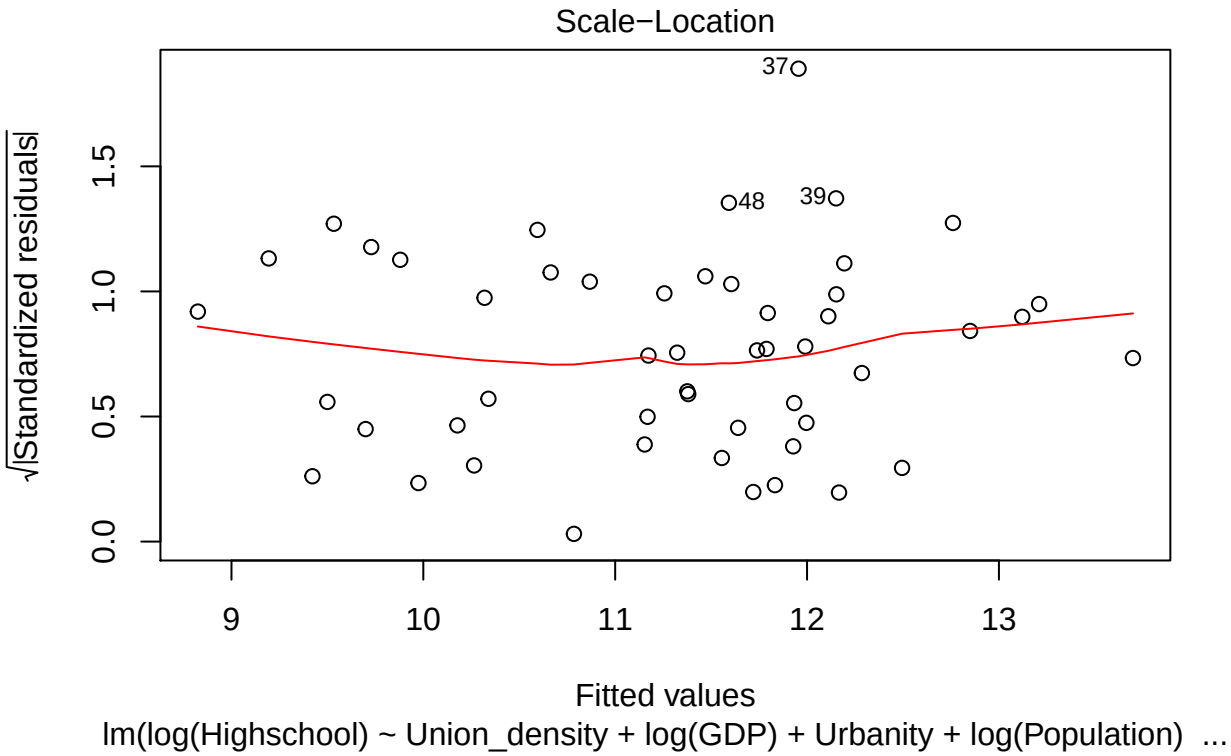
$$RacialDiversity = 1 - (Black^2 + White^2 + Hispanic^2)$$

$$NormRacialDiversity = \frac{C}{C-1} * RacialDiversity$$

```
##
## Call:
## lm(formula = log(Highschool) ~ Union_density + log(GDP) + Urbanity +
##     log(Population) + log(Export) + Unemployment_rate + log(Cost_of_living) +
##     Gini + norm_Racial_diversity, data = olddata3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11000 -0.29731  0.02178  0.33166  1.05237
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.446738   6.396034   1.164   0.2510
## Union_density          -0.024490   0.026891  -0.911   0.3678
## log(GDP)               -0.053305   0.538701  -0.099   0.9217
## Urbanity               -0.006738   0.010946  -0.616   0.5416
## log(Population)         0.798928   0.151591   5.270 4.68e-06 ***
## log(Export)             0.097660   0.097082   1.006   0.3203
## Unemployment_rate      -0.130898   0.122559  -1.068   0.2918
## log(Cost_of_living)    -1.709210   1.026990  -1.664   0.1037
## Gini                   -1.011016   4.833292  -0.209   0.8353
## norm_Racial_diversity   1.590066   0.645692   2.463   0.0181 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6236 on 41 degrees of freedom
## Multiple R-squared:  0.7952, Adjusted R-squared:  0.7503
## F-statistic: 17.69 on 9 and 41 DF,  p-value: 1.711e-11
```

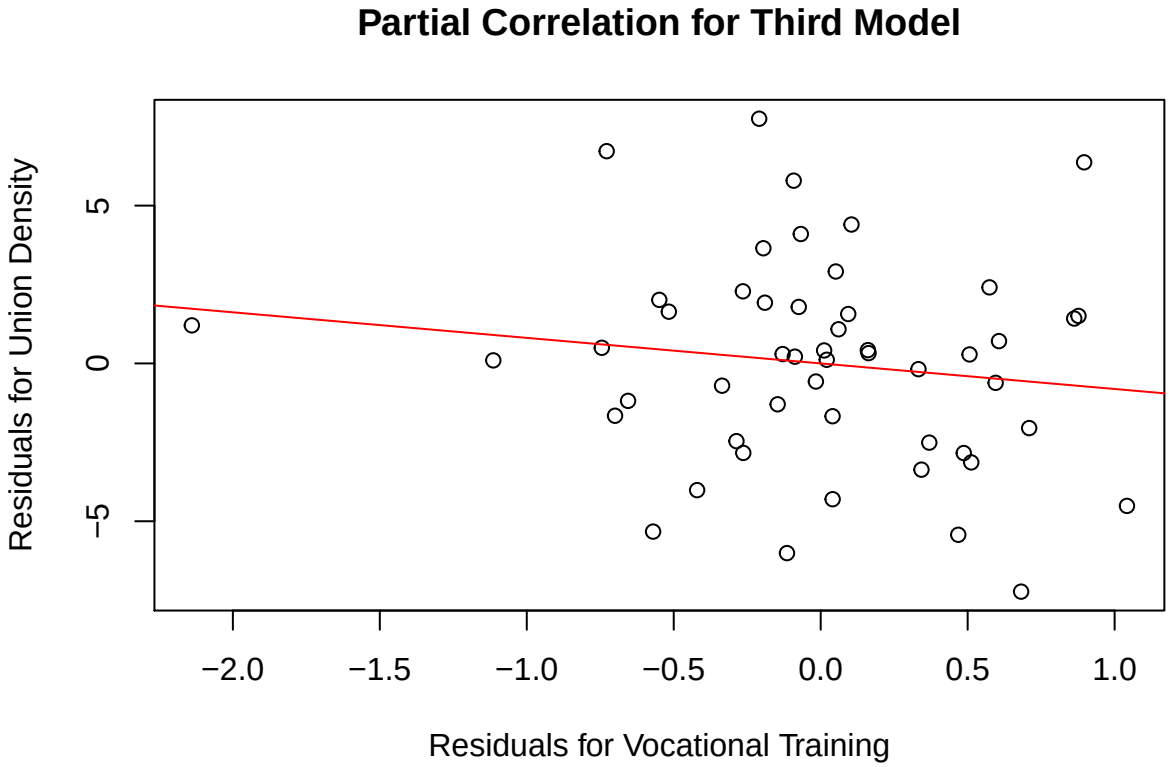

Residuals vs Fitted
Residuals
Fitted values
lm(log(Highschool) ~ Union_density + log(GDP) + Urbanity + log(Population) ...
48
39
37
1
0
−1
−2
9
10
11
12
13


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(Highschool) ~ Union_density + log(GDP) + Urbanity + log(Population) ...
48
39
37
2
1
0
−1
−2
−3
−2
−1
0
1
2



The assumptions of adding the normalized racial diversity model are similar to the transformed model; the only change is that adjusted R-squared increases a little bit, from 0.7202 to 0.7503. After data transformation, the linear regression assumptions have been satisfied. Compared to the previous model, the adjusted R-squared increases a little bit, from 0.7202 to 0.7503, meaning that this model captures more information from the dataset. The adjusted R-squared is a modified version of R-squared and increases only if the new term improves model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. Therefore, after adding normalized racial diversity, the new model fits better than the transformed model.

Partial Correlation for Third Model



By checking the partial correlation of the third model, the red line is pretty close to the horizontal line, indicating that the negative correlation between union density and vocational training is weak.

Pearson Method

```
## [1] -0.1242534
```

```
## [1] 0.3849993
```

Spearman Method

```
## [1] -0.1408109
```

```
## [1] 0.09939286
```

TThen, from the output of different methods, the partial correlations between Union density and Vocational training are close to zero, and P-values are quite large, which indicates this relationship between these two factors is not statistically significant.

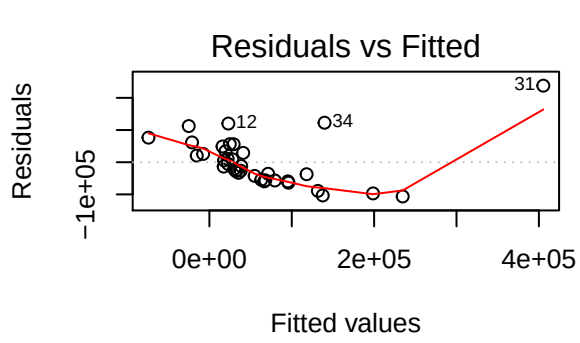
To summarize the results, two findings are found based on the research question. First, there is no significant relationship between union density and vocational education. Second, the racial diversity changes the results of linear model to some extent. But for the question of if the racial diversity explains the relationship, no conclusion can be generalized.

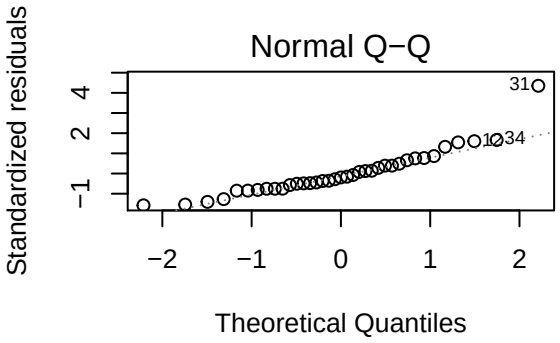
Part TWO

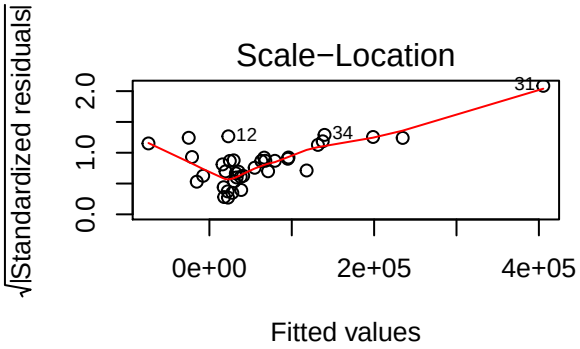
Methodology and Result

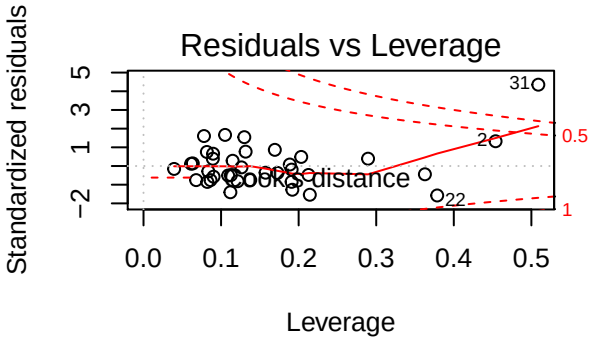
Create a new dataset that contains only the data in the year of 2010 (newdata2)

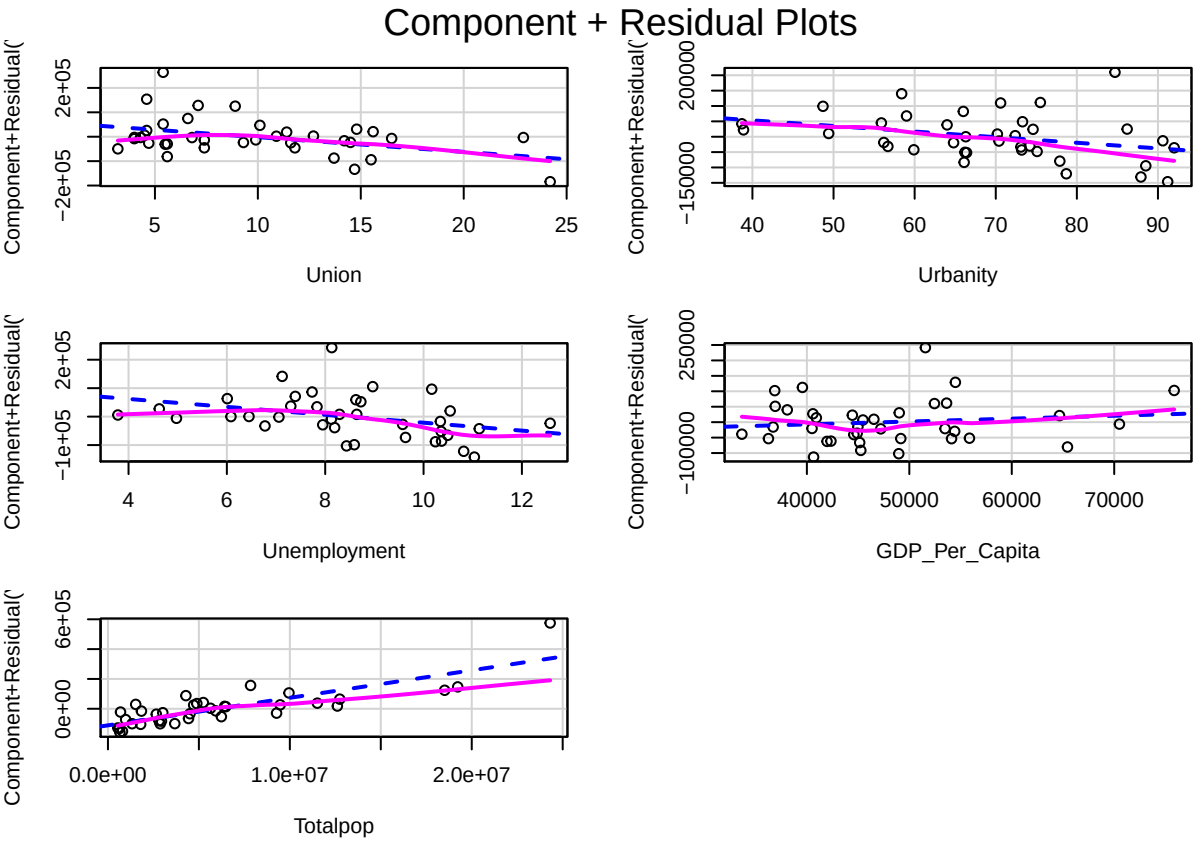
```
##
## Call:
## lm(formula = Voe ~ Union + Urbanity + Unemployment + GDP_Per_Capita +
##     Totalpop, data = newdata2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -106386  -54424  -13568   33851  237874
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.088e+05  1.282e+05   1.629   0.1135
## Union          -6.044e+03  2.940e+03  -2.056   0.0483 *
## Urbanity       -1.824e+03  1.426e+03  -1.279   0.2103
## Unemployment   -1.392e+04  9.042e+03  -1.540   0.1338
## GDP_Per_Capita  9.359e-01  2.052e+00   0.456   0.6515
## Totalpop        1.846e-02  3.265e-03   5.653  3.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78000 on 31 degrees of freedom
## Multiple R-squared:  0.5737, Adjusted R-squared:  0.5049
## F-statistic: 8.343 on 5 and 31 DF,  p-value: 4.365e-05
```


Residuals vs Fitted
Residuals
Fitted values
31
12
34
−1e+05
0e+00
2e+05
4e+05


Normal Q−Q
Standardized residuals
Theoretical Quantiles
31
12
34
−2
−1
0
1
2

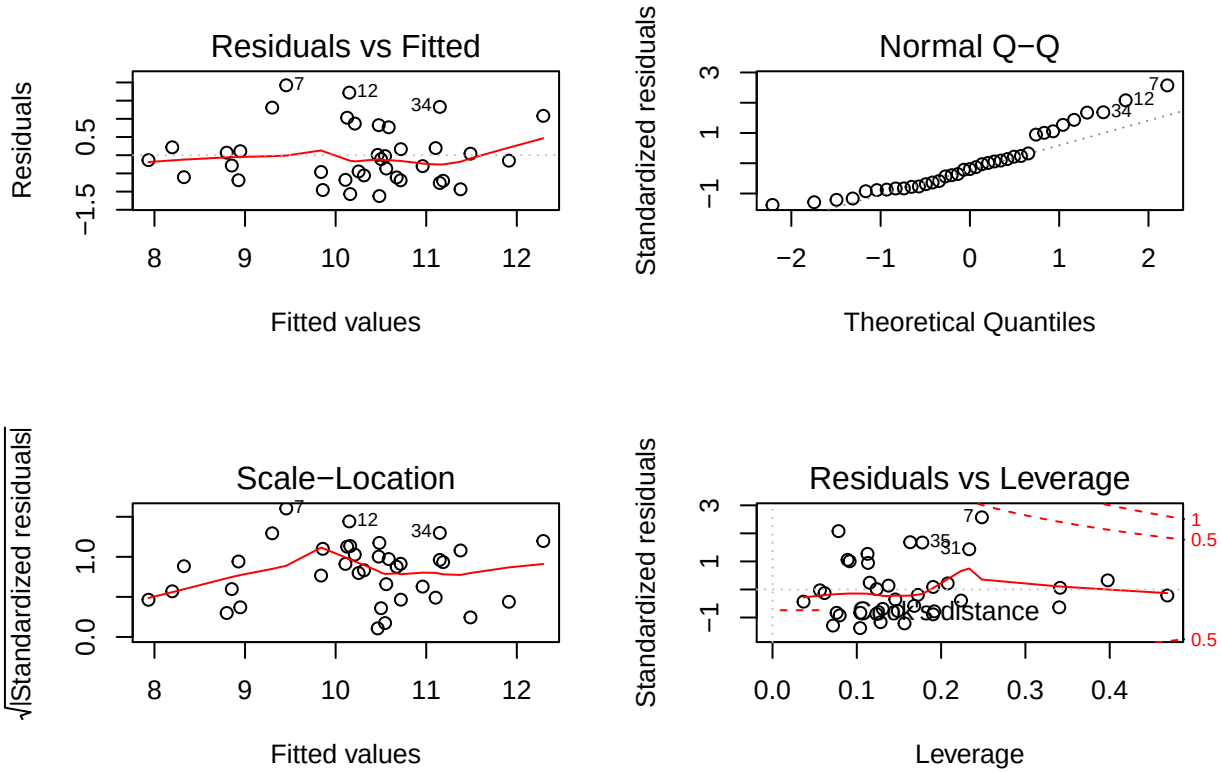

Scale−Location
√|Standardized residuals|
Fitted values
31
12
34
0.0
1.0
2.0
0e+00
2e+05
4e+05

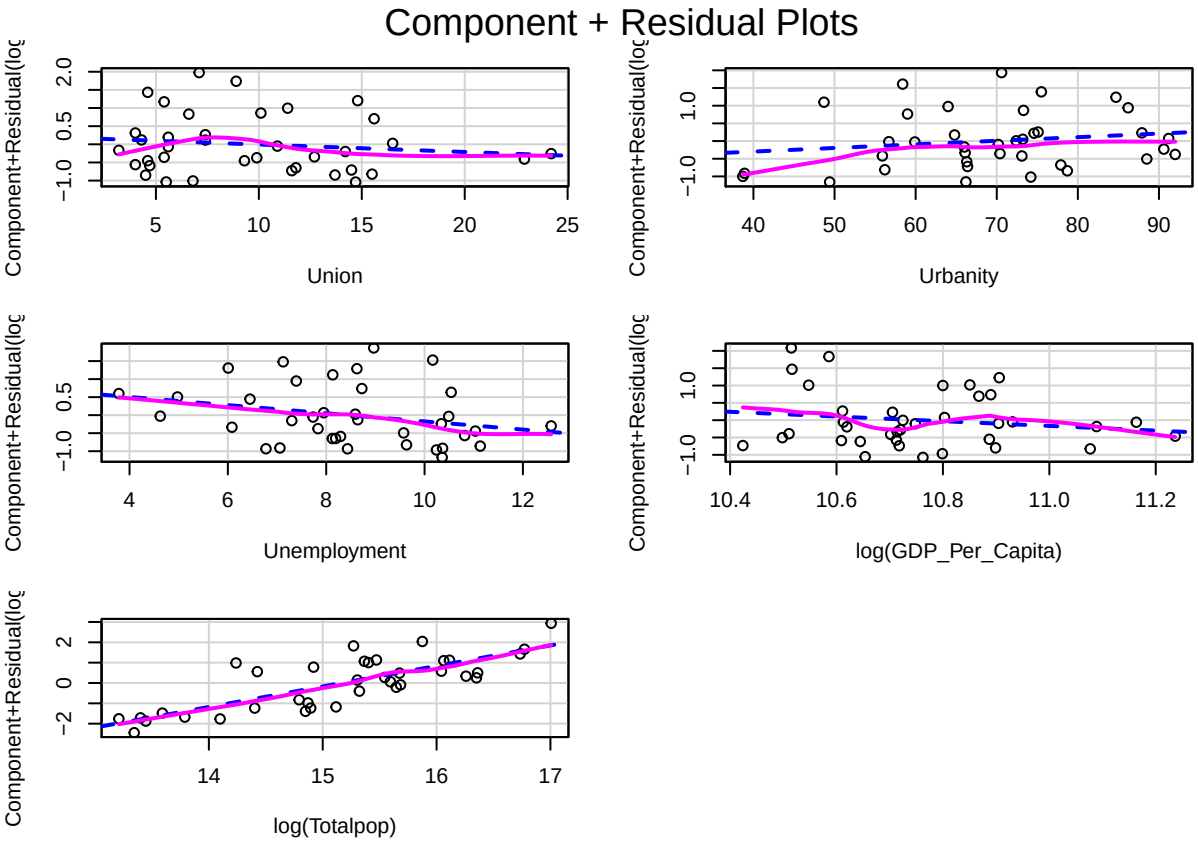

Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
31
12
22
0.5
1
0.0
0.1
0.2
0.3
0.4
0.5


Component + Residual Plots
Component+Residual(
Union
Urbanity
Unemployment
GDP_Per_Capita
Totalpop

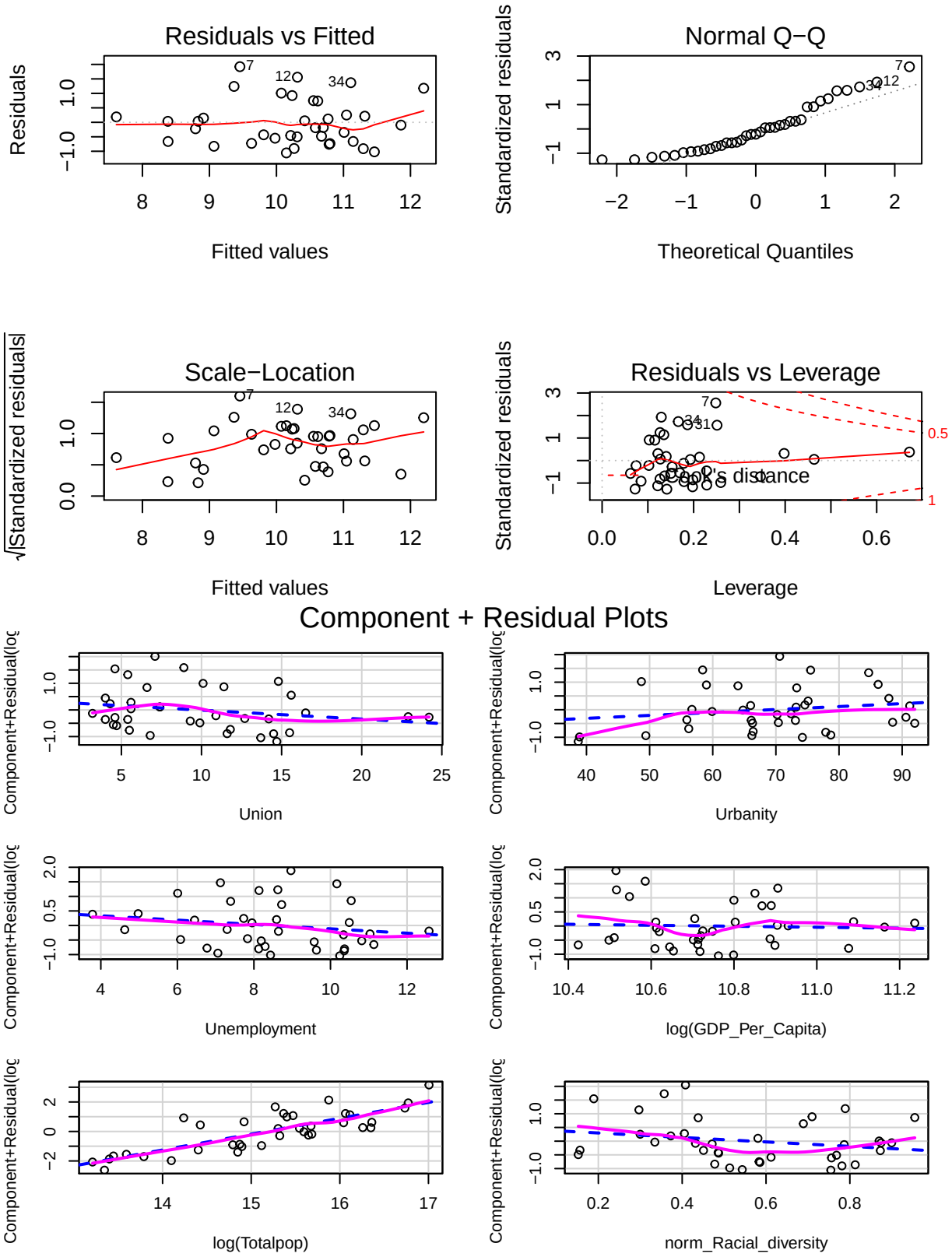
Transformation

```
##
## Call:
## lm(formula = log(Voe) ~ Union + Urbanity + Unemployment + log(GDP_Per_Capita) +
##     log(Totalpop), data = newdata2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.1238 -0.6101 -0.1354  0.2171  1.9219
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.65379   13.76133   0.193 0.848339
## Union                -0.02057    0.03218  -0.639 0.527373
## Urbanity              0.01019    0.01839   0.554 0.583630
## Unemployment         -0.11309    0.11422  -0.990 0.329785
## log(GDP_Per_Capita) -0.67745    1.23285  -0.549 0.586604
## log(Totalpop)         1.00888    0.25223   4.000 0.000365 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8618 on 31 degrees of freedom
## Multiple R-squared:  0.6182, Adjusted R-squared:  0.5566
## F-statistic: 10.04 on 5 and 31 DF,  p-value: 8.748e-06
```






```
##
## Call:
## lm(formula = log(Voe) ~ Union + Urbanity + Unemployment + log(GDP_Per_Capita) +
##     log(Totalpop) + norm_Racial_diversity, data = newdata2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0568 -0.6616 -0.1787  0.2531  1.9217
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -3.52911   15.72551  -0.224 0.823952
## Union                   -0.03351    0.03594  -0.932 0.358573
## Urbanity                 0.01069    0.01849   0.578 0.567653
## Unemployment            -0.07483    0.12378  -0.605 0.550046
## log(GDP_Per_Capita)     -0.17485    1.38037  -0.127 0.900049
## log(Totalpop)            1.07513    0.26590   4.043 0.000339 ***
## norm_Racial_diversity -0.81309    0.98377  -0.827 0.415040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8663 on 30 degrees of freedom
## Multiple R-squared:  0.6267, Adjusted R-squared:  0.552
## F-statistic: 8.393 on 6 and 30 DF,  p-value: 2.193e-05
```



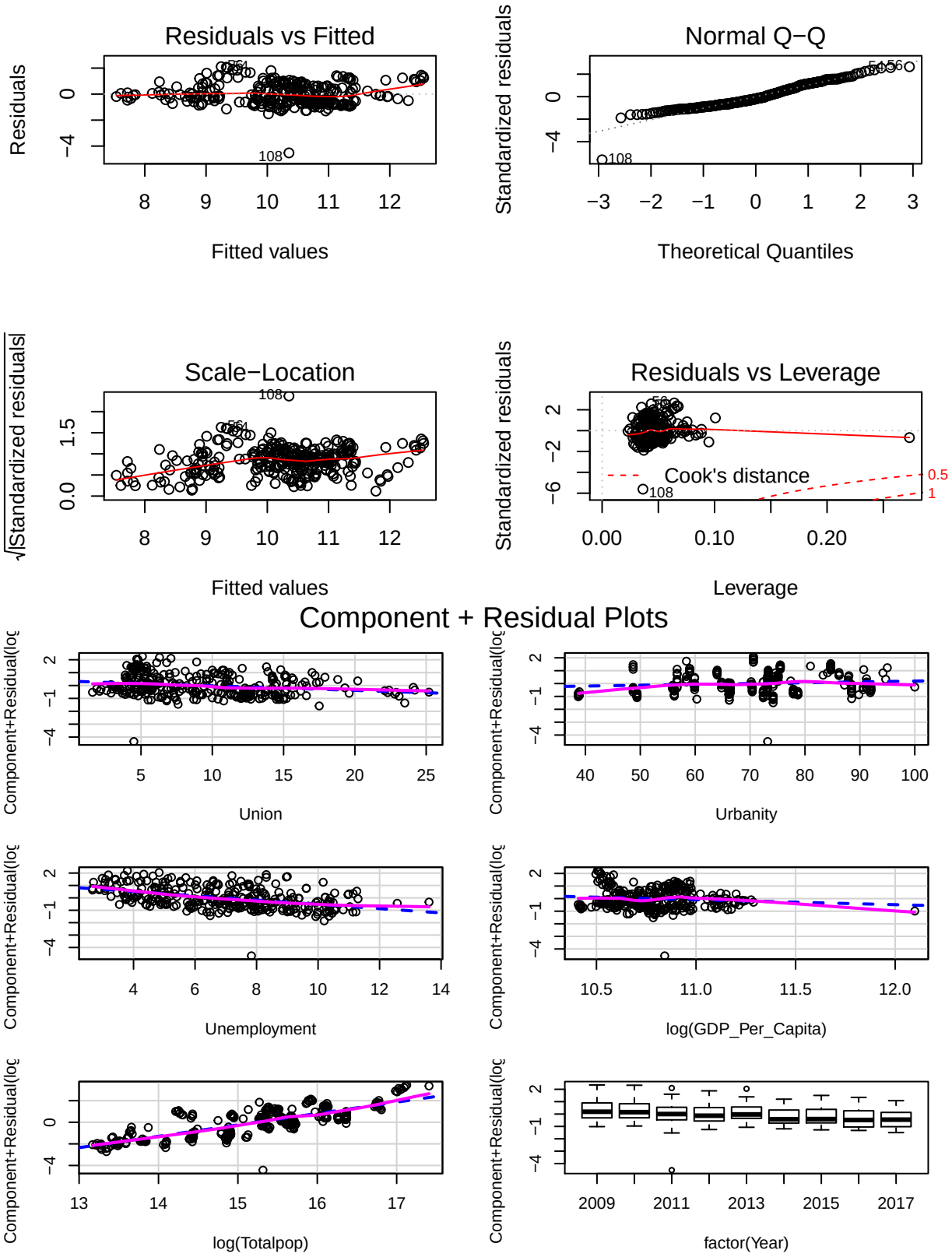
Model3 is built based on the transformed model and it considers the normalized racial diversity. Although the residual plot and QQ-plot shows the residuals are normally distributed in this model, the p-value of union density still large and the coefficient of union density is closed to 0. Racial diversity seems not able to explain

the negative relationship between union diversity and vocational training.

The data set, which is used to build all the years model, only includes the data from 2009 to 2017. Since in the raw data set, there is some missing value from 2003 to 2008. And due to the new data only have one measurement for urbanity, so this model is assuming that for all the years, urbanity is the same.

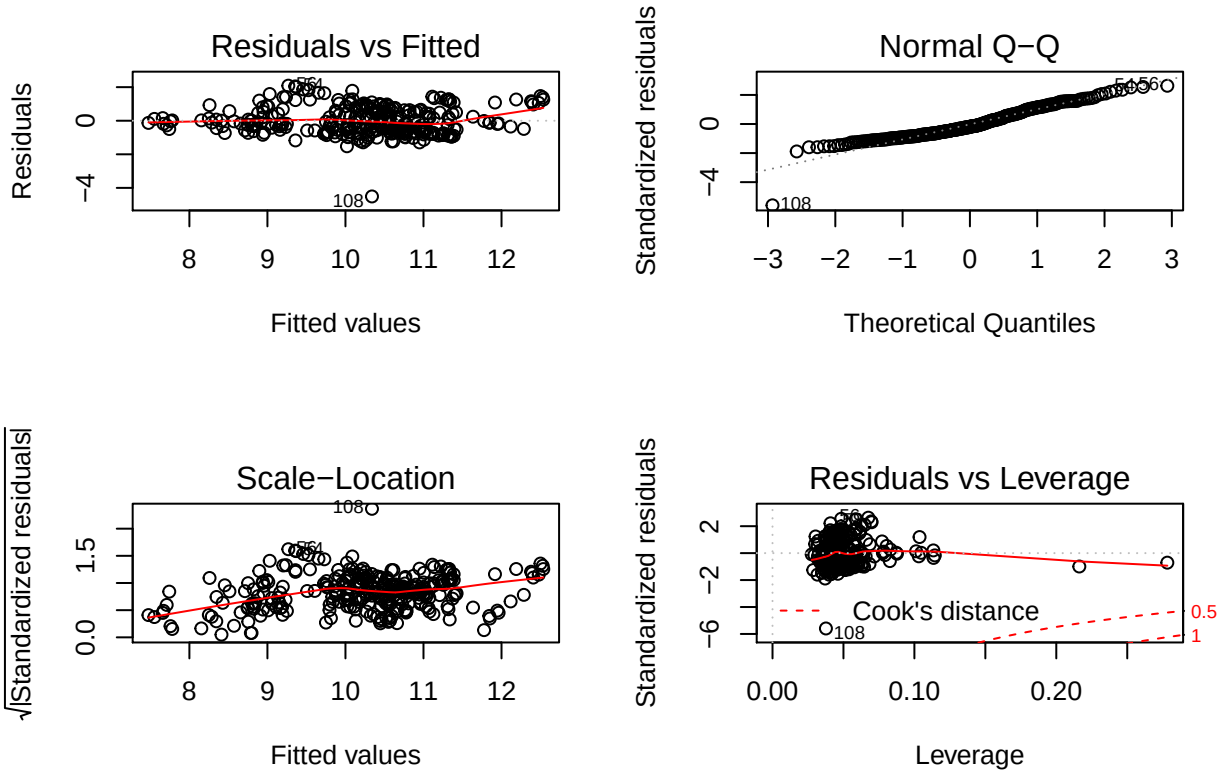
```
##
## Call:
## lm(formula = log(Voe) ~ Union + Urbanity + Unemployment + log(GDP_Per_Capita) +
##     log(Totalpop) + factor(Year), data = newdata3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5245 -0.5615 -0.1621  0.5521  2.1095
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.156130   4.211933  -0.037  0.97046
## Union               -0.035833   0.010926  -3.280  0.00117 **
## Urbanity             0.006495   0.006699   0.970  0.33305
## Unemployment        -0.172147   0.039589  -4.348 1.91e-05 ***
## log(GDP_Per_Capita) -0.397262   0.360667  -1.101  0.27163
## log(Totalpop)        1.053662   0.080729  13.052  < 2e-16 ***
## factor(Year)2010     0.020848   0.180272   0.116  0.90801
## factor(Year)2011    -0.286825   0.187695  -1.528  0.12759
## factor(Year)2012    -0.242713   0.199784  -1.215  0.22542
## factor(Year)2013    -0.124493   0.196571  -0.633  0.52703
## factor(Year)2014    -0.492981   0.213243  -2.312  0.02150 *
## factor(Year)2015    -0.452687   0.221018  -2.048  0.04146 *
## factor(Year)2016    -0.585156   0.246768  -2.371  0.01839 *
## factor(Year)2017    -0.653285   0.262517  -2.489  0.01340 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.821 on 284 degrees of freedom
## Multiple R-squared:  0.608,  Adjusted R-squared:   0.59
## F-statistic: 33.88 on 13 and 284 DF,  p-value: < 2.2e-16
```

Compared the all the year model and one year model, the coefficient of union density does not change a lot, -0.02 means every 1 unit increases on Union density, the Voe expect to decrease 2%. The coefficient of 2009 is the intercept, and the c of 2010 indicates the Voe in 2012 is 0.058 unit higher than the Voe coefficients in 2009; the coefficient of 2011 means the Voe in 2012 is 0.18 unit lower than the Voe on 2009; and so on.



By checking the QQ plot, there is one outlier in the lower left, and it was the data for Louisiana in 2010. However, it is unreasonable to remove one year data of one state, so this outlier still exists in the following model.

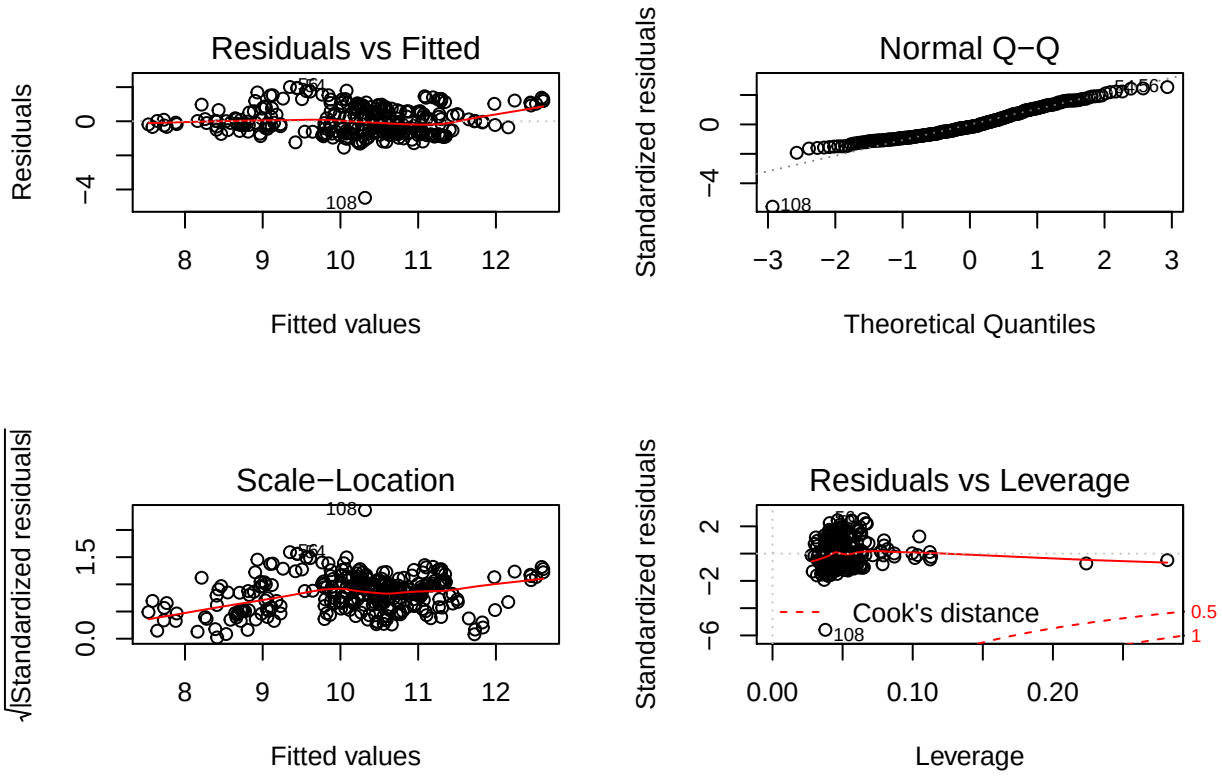
```
##
## Call:
## lm(formula = log(Voe) ~ Union + Urbanity + Unemployment + log(GDP_Per_Capita) +
##     log(Totalpop) + norm_Racial_diversity + factor(Year), data = newdata3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5110 -0.5791 -0.1541  0.5455  2.0881
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -0.621067   4.343951  -0.143 0.886413
## Union                 -0.037625   0.011651  -3.229 0.001387 **
## Urbanity               0.007283   0.006935   1.050 0.294545
## Unemployment          -0.165275   0.042514  -3.888 0.000126 ***
## log(GDP_Per_Capita)   -0.362325   0.369517  -0.981 0.327660
## log(Totalpop)          1.058677   0.081616  12.971  < 2e-16 ***
## norm_Racial_diversity -0.137486   0.307229  -0.448 0.654854
## factor(Year)2010       0.016554   0.180781   0.092 0.927107
## factor(Year)2011      -0.285283   0.187991  -1.518 0.130248
## factor(Year)2012      -0.237300   0.200431  -1.184 0.237426
## factor(Year)2013      -0.123375   0.196864  -0.627 0.531360
## factor(Year)2014      -0.481869   0.214982  -2.241 0.025773 *
## factor(Year)2015      -0.434703   0.224949  -1.932 0.054301 .
## factor(Year)2016      -0.561973   0.252488  -2.226 0.026820 *
## factor(Year)2017      -0.626144   0.269793  -2.321 0.021006 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8222 on 283 degrees of freedom
## Multiple R-squared:  0.6082, Adjusted R-squared:  0.5889
## F-statistic: 31.38 on 14 and 283 DF,  p-value: < 2.2e-16
```



After adding the normalized racial diversity in the all year model, the coefficient of union density only decreases 0.02, and the four plots of the model are similar to the previous model. Therefore, it is hard to say that racial diversity can explain the negative relationship between union density and Voe.

```
##
## Call:
## lm(formula = log(Voe) ~ Union + Urbanity + Unemployment + log(GDP_Per_Capita) +
##     log(StudentpopE) + norm_Racial_diversity + factor(Year),
##     data = newdata3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4872 -0.5756 -0.1443  0.5649  2.0057
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.788405   4.217920   0.424 0.671887
## Union                   -0.040978   0.011600  -3.533 0.000480 ***
## Urbanity                 0.006842   0.006902   0.991 0.322404
## Unemployment            -0.158301   0.042165  -3.754 0.000211 ***
## log(GDP_Per_Capita)     -0.296511   0.369330  -0.803 0.422743
## log(StudentpopE)         1.059516   0.080667  13.134  < 2e-16 ***
## norm_Racial_diversity -0.256329   0.307110  -0.835 0.404619
## factor(Year)2010         0.002072   0.180060   0.012 0.990828
## factor(Year)2011        -0.274554   0.186894  -1.469 0.142933
## factor(Year)2012        -0.205428   0.198837  -1.033 0.302417
## factor(Year)2013        -0.080551   0.195300  -0.412 0.680325
## factor(Year)2014        -0.410924   0.212542  -1.933 0.054187 .
## factor(Year)2015        -0.352169   0.222071  -1.586 0.113892
## factor(Year)2016        -0.465116   0.248963  -1.868 0.062766 .
```

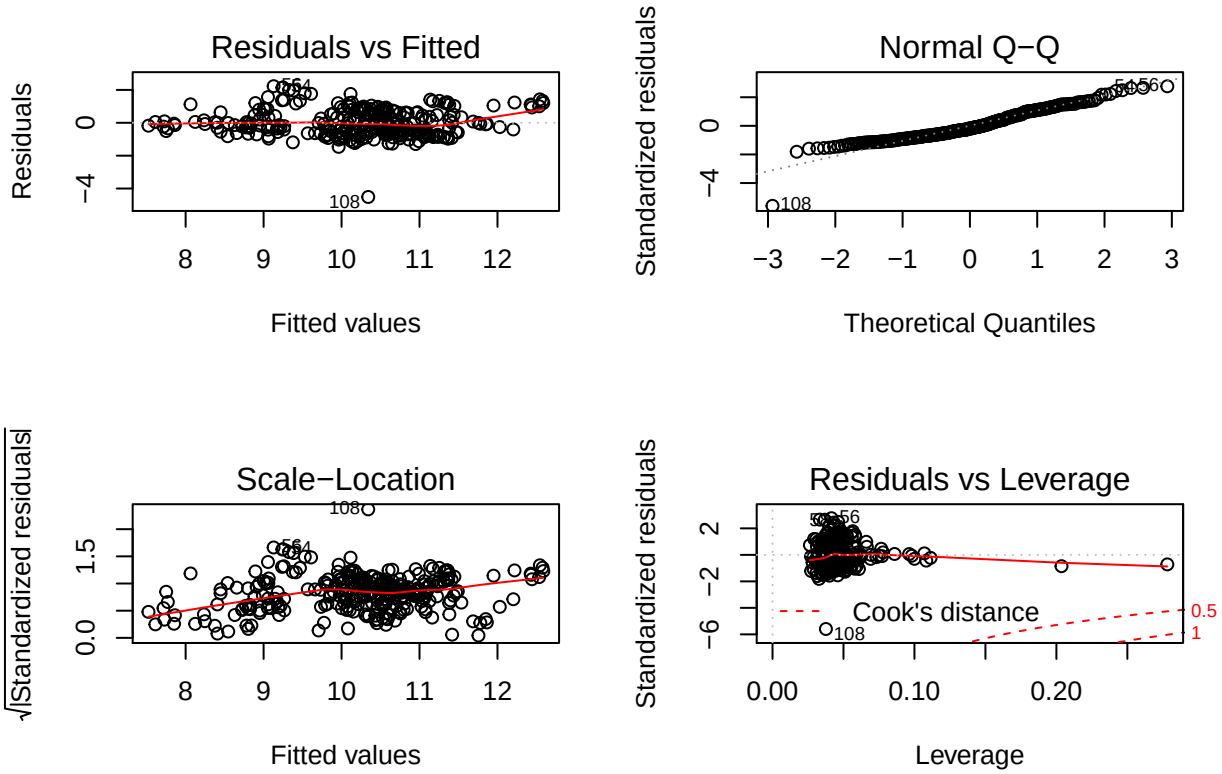
```
## factor(Year)2017      -0.520193   0.265760  -1.957 0.051286 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8183 on 283 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.5927
## F-statistic: 31.87 on 14 and 283 DF,  p-value: < 2.2e-16
```



The dataset contains total population and student population. In order to see how population change could influence the result, replace the total population with student population and add normalized racial diversity into the model6. While, the result does not change much compared with model4. Cr plots do not show much change either so the conclusion is same as the former one.

```
##
## Call:
## lm(formula = log(Voe) ~ Union + Unemployment + log(GDP_Per_Capita) +
##     log(Totalpop) + norm_Racial_diversity + factor(Year), data = newdata3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5132 -0.5629 -0.1436  0.5840  2.2273
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -3.231353   3.563164  -0.907 0.365240
## Union                  -0.036636   0.011615  -3.154 0.001782 **
## Unemployment           -0.166994   0.042491  -3.930 0.000107 ***
## log(GDP_Per_Capita)    -0.151089   0.310028  -0.487 0.626394
## log(Totalpop)           1.113176   0.063003  17.669  < 2e-16 ***
## norm_Racial_diversity -0.055608   0.297225  -0.187 0.851724
```

```
## factor(Year)2010      -0.009004   0.179168  -0.050 0.959954
## factor(Year)2011      -0.320735   0.184968  -1.734 0.084004 .
## factor(Year)2012      -0.275745   0.197095  -1.399 0.162891
## factor(Year)2013      -0.157177   0.194250  -0.809 0.419108
## factor(Year)2014      -0.526845   0.210711  -2.500 0.012972 *
## factor(Year)2015      -0.478987   0.221001  -2.167 0.031039 *
## factor(Year)2016      -0.606833   0.248893  -2.438 0.015377 *
## factor(Year)2017      -0.682200   0.264507  -2.579 0.010408 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8223 on 284 degrees of freedom
## Multiple R-squared:  0.6067, Adjusted R-squared:  0.5887
## F-statistic:  33.7 on 13 and 284 DF,  p-value: < 2.2e-16
```



The issue exists in dataset that it uses 2010 urbanity value stands for all years' value and it is acknowledged that urbanity doesn't have to be included in the model, so model8 removes urbanity, while it turns out to be same result and conclusion as before.

Mediation Analysis

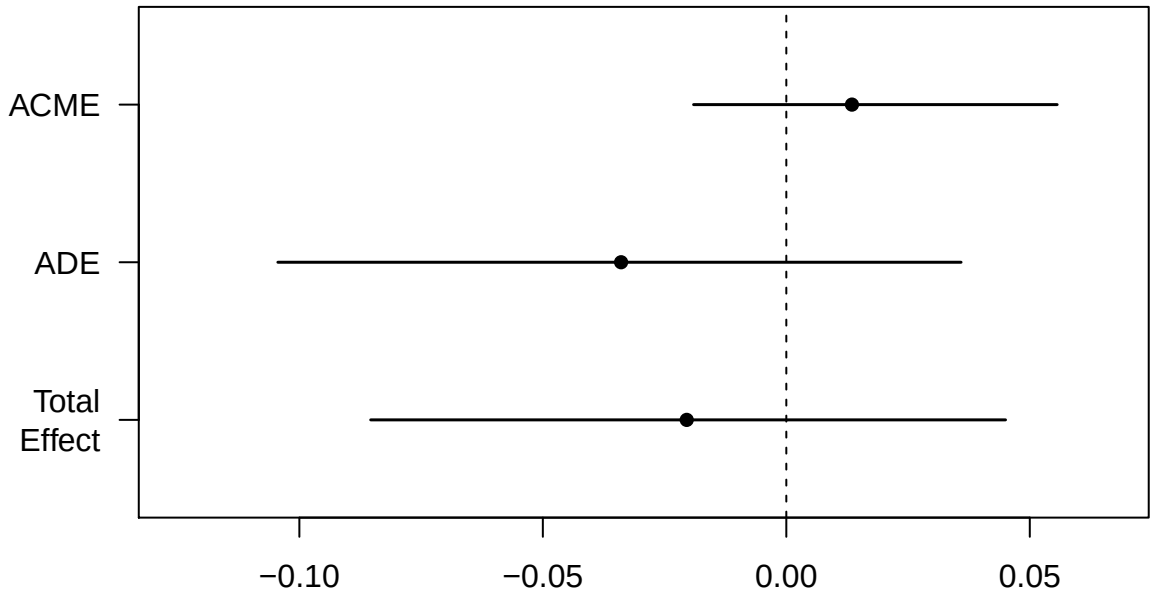
To investigate if racial diversity actually mediates the relationship between union density and vocational education, mediation analysis is conducted. A mediation model proposes that the independent variable influences the mediator variable, which in turn influences the dependent variable. In this project, the union density is the independent variable served as the treatment effect, vocational education is the dependent variable acts as the outcome. Racial diversity is the mediator. This mediation analysis is exploring if union density affects racial diversity first, and racial diversity then affects vocational education. Because causal inference is used in the mediation analysis, assumptions of no other unmeasured confounding variables in the model is made. The direct effect is the effect between treatment (independent variable) and outcome (dependent variable). The mediation effect is the effect between treatment and mediators. In the project,

the direct effect is the effect between union density and vocational education; the mediation effect is the effect between union density and racial diversity. In the mediation analysis, the sum of the direct effect and the mediation effect is the total effect. To interpret the results of mediation analysis, the ACME stands for mediation effect, and ADE stands for direct effect.

```
#Mediation Analysis: Using data on 2010.
model.tm1 <- lm(norm_Racial_diversity~Union+Urbanity+Unemployment+log(GDP_Per_Capita)+log(Totalpop),data
model.ty1 <- lm(log(Voe)~Union+norm_Racial_diversity+Urbanity+Unemployment+log(GDP_Per_Capita)+log(Total
out.1 <- mediate(model.tm1,model.ty1,sims=1000,treat = "Union",mediator = "norm_Racial_diversity")
summary(out.1)
```

```
##
## Causal Mediation Analysis
##
## Quasi-Bayesian Confidence Intervals
##
##                Estimate 95% CI Lower 95% CI Upper p-value
## ACME             0.0135      -0.0190         0.06    0.42
## ADE             -0.0339      -0.1044         0.04    0.36
## Total Effect    -0.0205      -0.0854         0.05    0.55
## Prop. Mediated  -0.1542      -6.3887         7.44    0.76
##
## Sample Size Used: 37
##
##
## Simulations: 1000
```

```
plot(out.1)
```



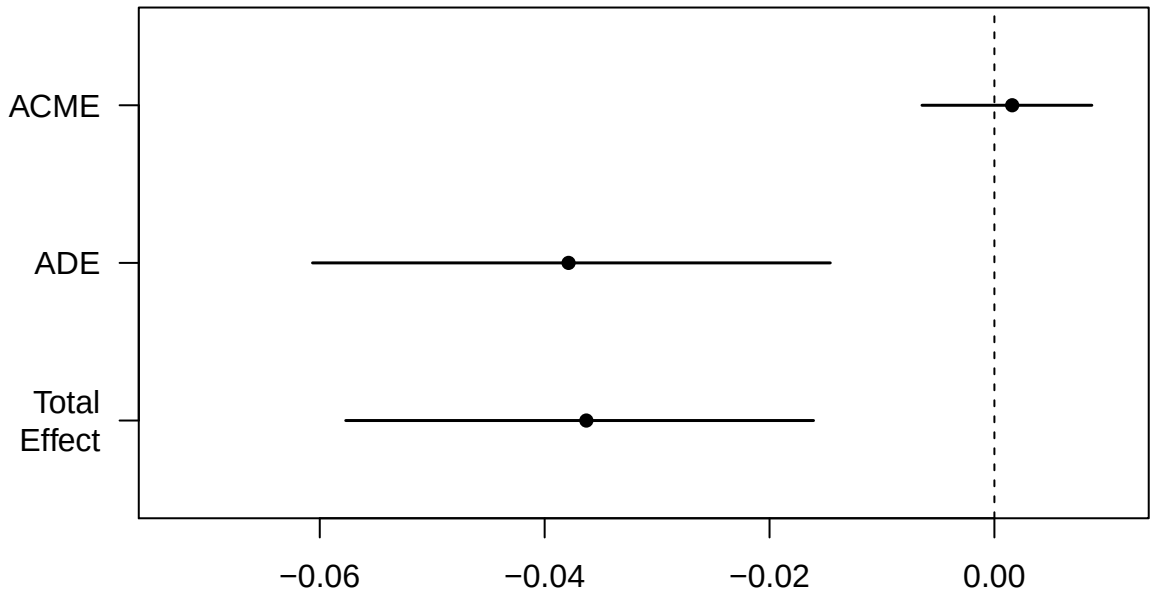

```
#In this case, we find the racial diversity does not have a mediation effect.
```

Based on the summarized results of mediation analysis using the data on 2010, the direct effect is -0.02, which is not significant, indicating that union density and vocational training does not have strong relationship. Furthermore, the estimate for mediation effect is 0.002, which is very close to 0, indicating that there is nearly no mediation effect. The proportion of mediation effects is -0.02. Looking at the plot, a large portion of the total effect is caused by direct effect, and the mediation effect is around zero, which corresponds to the prop.mediated result at -0.02 (2%).

```r
#Mediation Analysis: Using data from 2009 to 2017 (data3).
#Mediation Analysis
model.tm <- lm(norm_Racial_diversity~Union+Urbanity+Unemployment+log(GDP_Per_Capita)+log(Totalpop)+facto
model.ty <- lm(log(Voe)~Union+norm_Racial_diversity+Urbanity+Unemployment+log(GDP_Per_Capita)+log(Totalp
out.2 <- mediate(model.tm,model.ty,sims=1000,treat = "Union",mediator = "norm_Racial_diversity")
summary(out.2)
```

```
##
## Causal Mediation Analysis
##
## Quasi-Bayesian Confidence Intervals
##
##                Estimate 95% CI Lower 95% CI Upper p-value
## ACME            0.00158     -0.00644         0.01   0.664
## ADE            -0.03787     -0.06064        -0.01   0.004 **
## Total Effect   -0.03629     -0.05769        -0.02   0.002 **
## Prop. Mediated -0.04316     -0.32125         0.23   0.662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sample Size Used: 298
##
##
## Simulations: 1000
```

```r
plot(out.2)
```



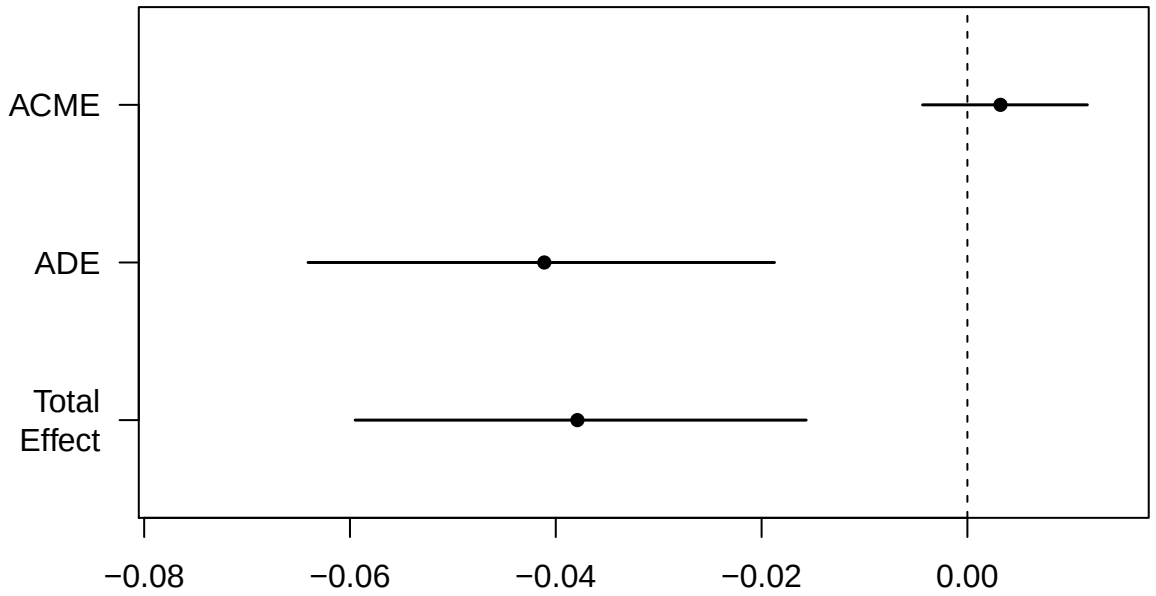

```r
#In this case, we find the racial diversity does not have a mediation effect.
```

Using the data from 2009 to 2017, the direct effect is -0.02 but this time it becomes significant, this might indicate that union density and vocational education have negative relationship. But because this finding is based on two assumptions, the one is urbanity is same value as 2010 for all the years, the other is all the confounding variables are included in the model. However, the mediation effect is 0.001, which is around zero. And based on the plot, most part of the effect is caused by direct effect and mediation effect ranged around zero, indicating that racial diversity does not mediate the relationship between union density and vocational education.

```r
#Mediation Analysis
model.sm <- lm(norm_Racial_diversity~Union+Urbanity+Unemployment+log(GDP_Per_Capita)+log(StudentpopE)+fa
model.sy <- lm(log(Voe)~Union+norm_Racial_diversity+Urbanity+Unemployment+log(GDP_Per_Capita)+log(Studer
out.3 <- mediate(model.sm,model.sy,sims=1000,treat = "Union",mediator = "norm_Racial_diversity")
summary(out.3)
```

```
##
## Causal Mediation Analysis
##
## Quasi-Bayesian Confidence Intervals
##
##                Estimate 95% CI Lower 95% CI Upper p-value
## ACME            0.00322     -0.00435         0.01    0.43
## ADE            -0.04112     -0.06407        -0.02  <2e-16 ***
## Total Effect   -0.03790     -0.05950        -0.02  <2e-16 ***
## Prop. Mediated -0.08309     -0.39995         0.12    0.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Sample Size Used: 298
##
##
## Simulations: 1000
```

```r
plot(out.3)
```




```r
#In this case, we find the racial diversity does not have a mediation effect.
```

For this mediation analysis, student population has replaced the total population. Again, most of the effect is caused by direct effect rather than mediation effect, indicating that racial diversity does not mediate the relationship between union density and vocational education.

Conclusion

- According to series of models built above, there is no evidence to say racial diversity explains relationship between vocational training and union density.

- These models are based on the dataset which mainly includes union density, GDP per capita, and other seven independent variables. There could be other factors which influence the vocation training level, so it might turn out to be different result when other factors are considered in model.