Week 5 Lab Report

Nguyen Tien Anh Quach

2024-02-28

Introduction

This report is to analyze residential carbon production per capita data to examine which explanatory variable(s) best predicts the C production. The list of variables includes:

(1) Average annual temperature (degrees F)

(2) Percent of state that voted for President Trump

(3) Standardized climate change google search share

(4) Percent of state population with Bachelor's Degree

(5) Median per capita income (USD)

(6) Renewable portfolio standard

(7) Percent of state population that is urban

(8) The state that is located in the west

Summary statistics of model variables

Summary statistics of the different continuous variables are below (Table 1). From the first look, all continuous variables have low skewness (close to 0). Therefore, they may be normally distributed. However, I will check the assumptions properly later on.

Table 1. Summary statistics of average annual temperature, percent Trump vote, climate change search share, percent Bachelor's degree, income per capita, and percent urban population in U.S. states.

Variable	Mean	Median	SD	IQR	Skewness
Percent Bachelor's Degree	29.51	28.40	6.06	6.10	1.45
Climate Change Search Share	37.18	34.00	14.53	17.00	1.84
Income per Capita (USD)	28786.55	27646.00	4825.45	5913.50	1.17
Avg Annual Temp (deg F)	51.94	51.20	8.71	13.40	-0.01
Percent Trump Vote	48.36	48.67	11.93	16.41	-0.91
Percent Urban Population	74.11	74.20	14.89	22.15	-0.40

Variables' correlations

The correlations (1) among independent variables and (2) between residential carbon production per capita and other independent variables (Fig. 1).

The residential carbon production per capita has a significantly negative correlation with average annual temperature and signficantly positive correlations with renewable portfolio standard, climate change search share, percent Bachelor's degree, and income per capita.

Among the independent variables, there are many significantly correlations, which mean that there may be multicollinearity existing and assumptions may be violated.

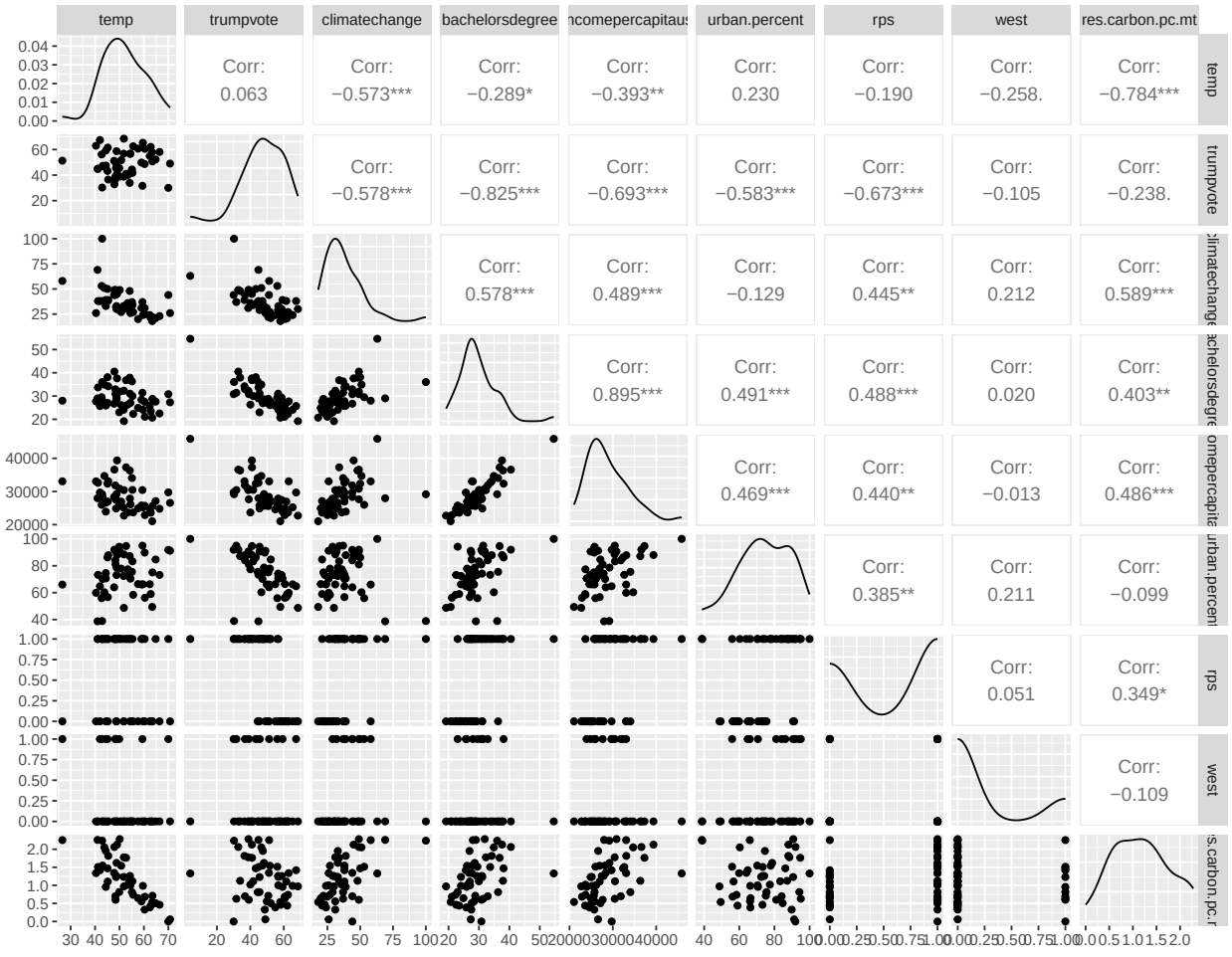


Figure 1. Correlations among continuous variables and between residential carbon production per capita and other variables.

Results

Initial model

The initial model was:

$$\text{res.carbon.pc.mt} \sim \text{temp} + \text{trumpvote} + \text{climatechange} + \text{bachelorsdegree} + \text{incomepercapitaus} + \text{urban.percent} + \text{rps} + \text{west}$$

The initial LM showed that the model was statistically significant (Adjusted $R^2 = 0.78$, $F(8, 41) = 22.54$, $p < 0.001$), but only temperature ($p < 0.001$, df = 41) and states that are in the west ($p < 0.001$, df = 41) had statistically significant coefficients.

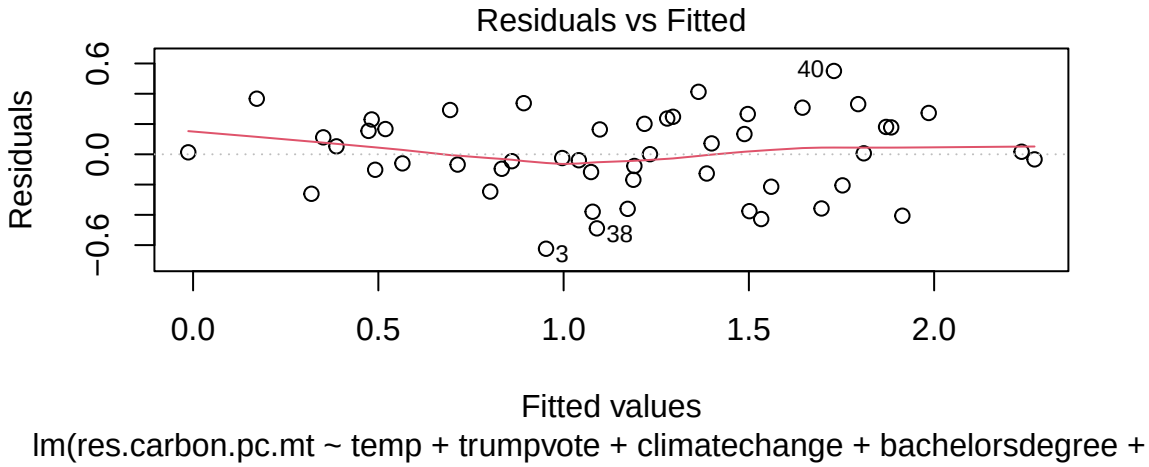
A one degree increase in average annual temperature is associated with a decrease of approximately 0.051 metric tons carbon in residential carbon production per capita, holding other variables constant. In addition, as state are in the west, the residential carbon production per capita is 0.61 metric tons lower, holding other variables constant.

Check model assumptions

1. Multicollinearity

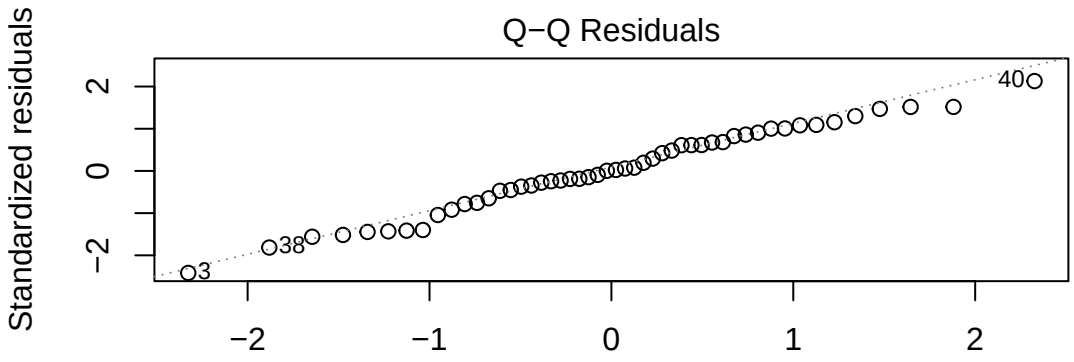
Variance Inflation Factors (VIF) showed that no variable had VIF that was more than 10. However, trumpvote, climatechange, and bachelorsdegree had VIFs more than 5. In addition, incomepercapitaus and urban.percent had VIFs that were marginally 5.

2. Linearity



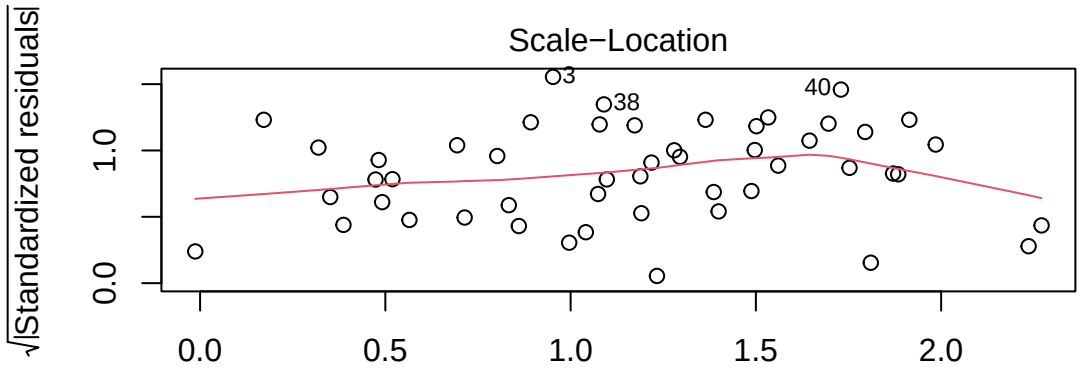
There is a relatively straight line through the fitted values. Therefore, linearity assumption is met!

3. Normal distribution of error



The majority of the data points are along the the standard normal distribution line. Shapiro-Wilk test of the residuals also yielded a nonsignificant p-value of 0.66. Therefore, the normal distribution assumption of error is met!

4. Constant variance of error



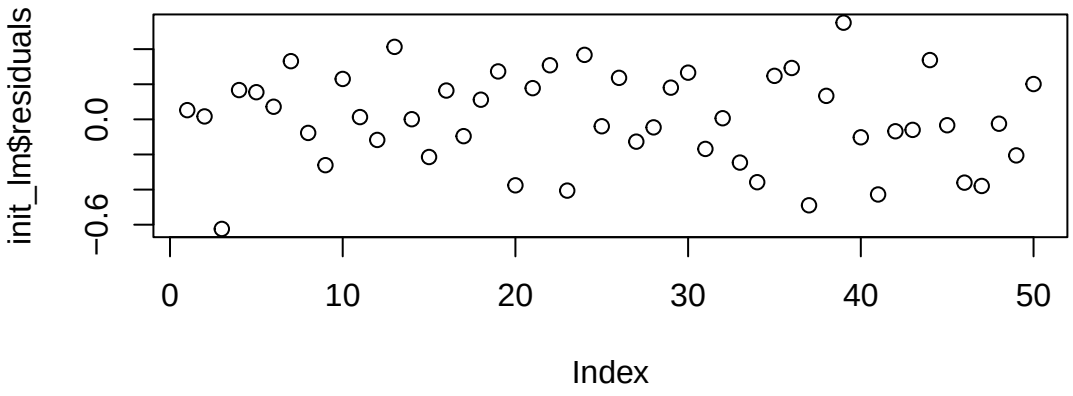
There is a relatively straight line through the fitted values. Studentized Breusch-Pagan test also yielded a nonsignificant p-value of 0.78. Therefore, constant variance assumption is met!

However, it is important to note that there are five leverage points in the model: 3, 25, 35, 40, and 45.

5. Zero error mean

The mean of the residuals is 0. Therefore, this assumption is met!

6. Uncorrelated errors



Errors are uncorrelated as they are randomly scattered around $\epsilon = 0$. The assumption is met!

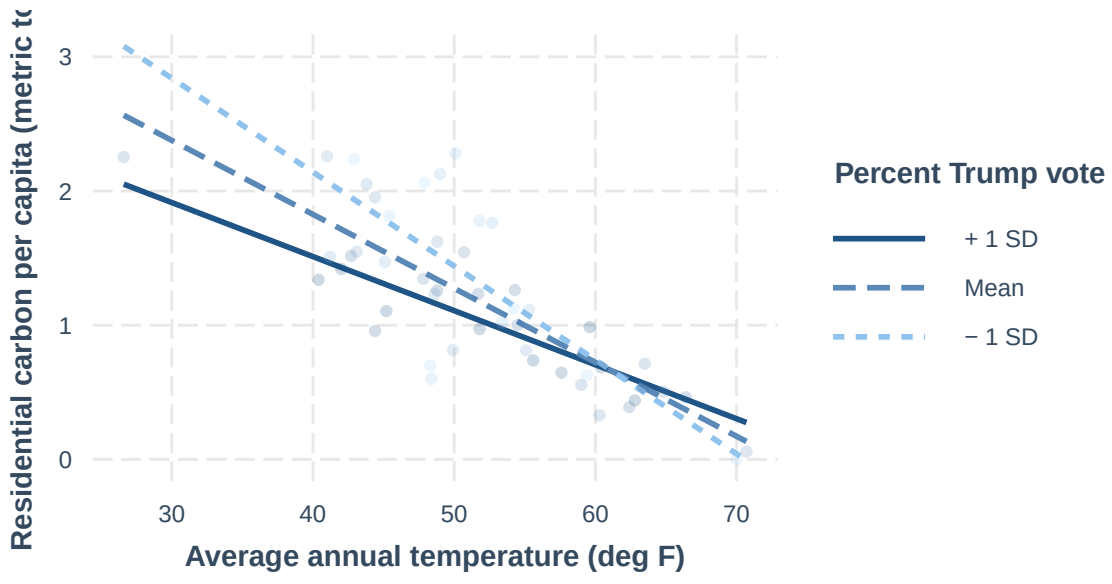
In summary, all the assumptions are met, perhaps except for multicollinearity. The variables, trumpvote, climatechange, and bachelorsdegree, had high VIFs that were more than 5. Eliminating redundant variables is needed for a better model.

Second model

The second model was:

$$\text{res.carbon.pc.mt} \sim \text{temp} + \text{trumpvote} + \text{temp*trumpvote}$$

Below is the plot showing the interactions between temp and trumpvote:



The second LM showed that the model was statistically significant (Adjusted $R^2 = 0.68$, $F(3, 46) = 35.67$, $p < 0.001$). For this model, temperature ($p < 0.001$, df = 46), trumpvote ($p = 0.006$, df = 46), and their

interaction effect (p = 0.017) had statistically significant coefficients.

A one degree increase in average annual temperature is associated with a decrease of approximately 0.13 metric tons carbon in residential carbon production per capita, holding percent Trump vote constant. In addition, a one percent increase in Trump voter, the residential carbon production per capita is 0.09 metric tons lower, assuming average annual temperature is held constant.

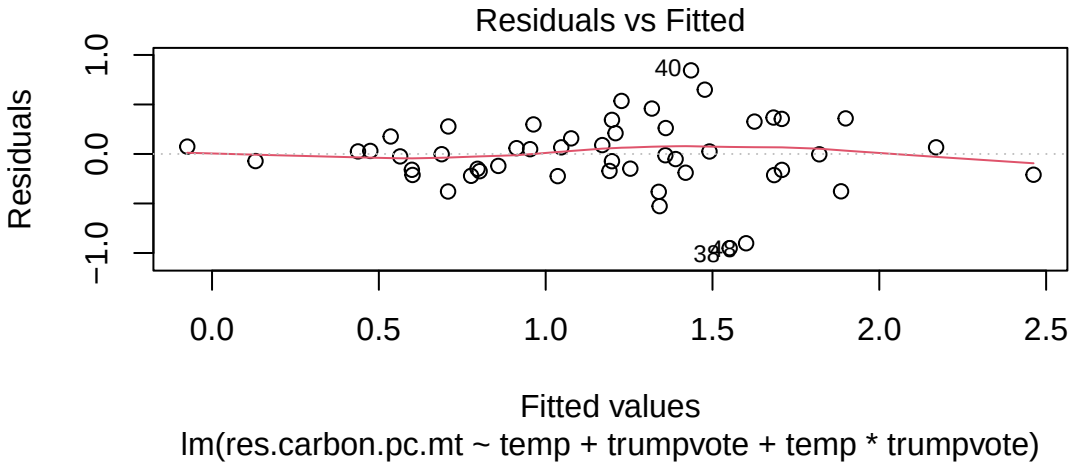
Interestingly, the interaction term between average annual temperature and percent Trump vote resulted in a positive coefficient (0.0015). This means that in average, if percent Trump vote increases one percent, a one degree increase in avg annual temperature will decrease residential carbon production per capita by (0.13-0.0015) 0.1285 metric tons.

Check model assumptions

1. Multicollinearity

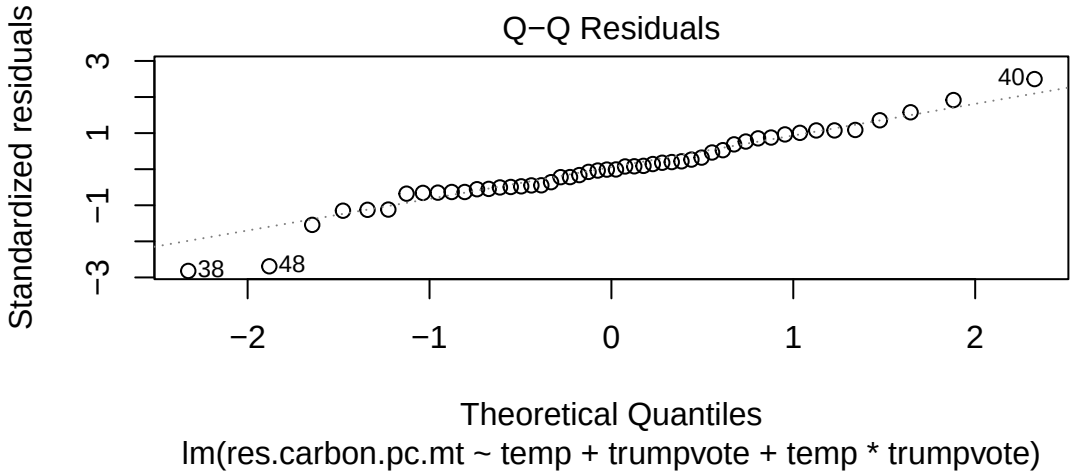
Variance Inflation Factors (VIF) showed that temp, trumpvote, and the interaction between the two variables have VIFs of 27, 42, and 72, respectively. Therefore, all are too high and the multicollinearity assumption is NOT met!

2. Linearity



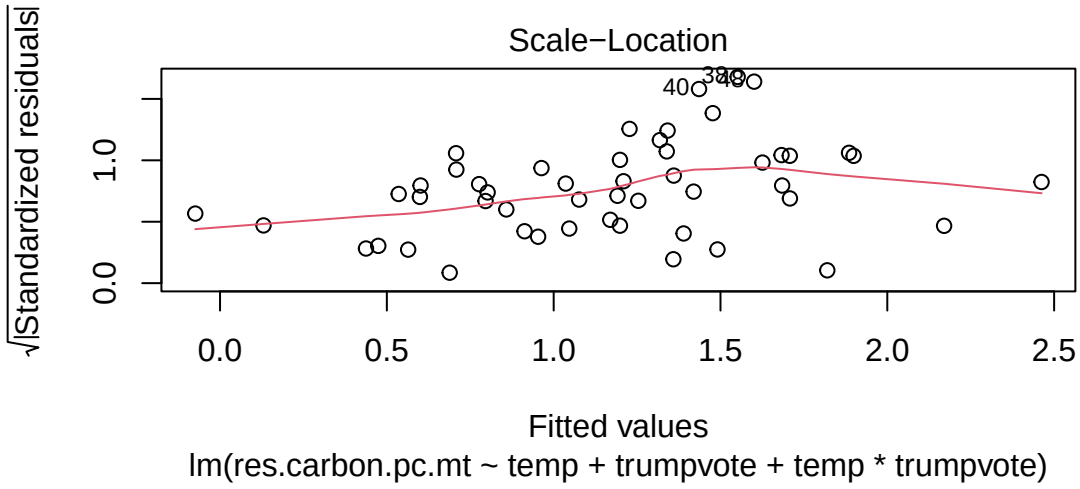
There is a relatively straight line through the fitted values. Therefore, linearity assumption is met!

3. Normal distribution of error



The majority of the data points are along the the standard normal distribution line. Shapiro-Wilk test of the residuals also yielded a nonsignificant p-value of 0.12. Therefore, the normal distribution assumption of error is met!

4. Constant variance of error



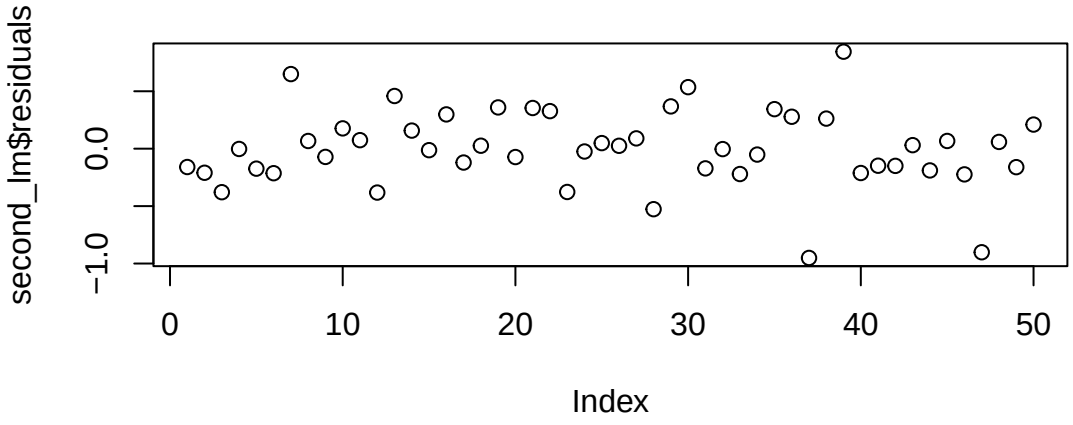
There is a relatively straight line through the fitted values. However, the studentized Breusch-Pagan test yielded a significant p-value of 0.04. Therefore, constant variance assumption is NOT met!

Also, it is important to note that there are three leverage points in the model: 38, 40, and 48.

5. Zero error mean

The mean of the residuals is 0. Therefore, this assumption is met!

6. Uncorrelated errors



Errors are uncorrelated as they are randomly scattered around $\epsilon = 0$. The assumption is met!

In summary, all the assumptions are met, except for multicollinearity and homogeneity of variance. From the variable correlation plot, temp and trumpvote have no significant correlation with each other. Therefore, the interaction term between temp and trumpvote is likely the variable driving the high VIF.

Rerun second model without interaction

The revised second model was:

$$\text{res.carbon.pc.mt} \sim \text{temp} + \text{trumpvote}$$

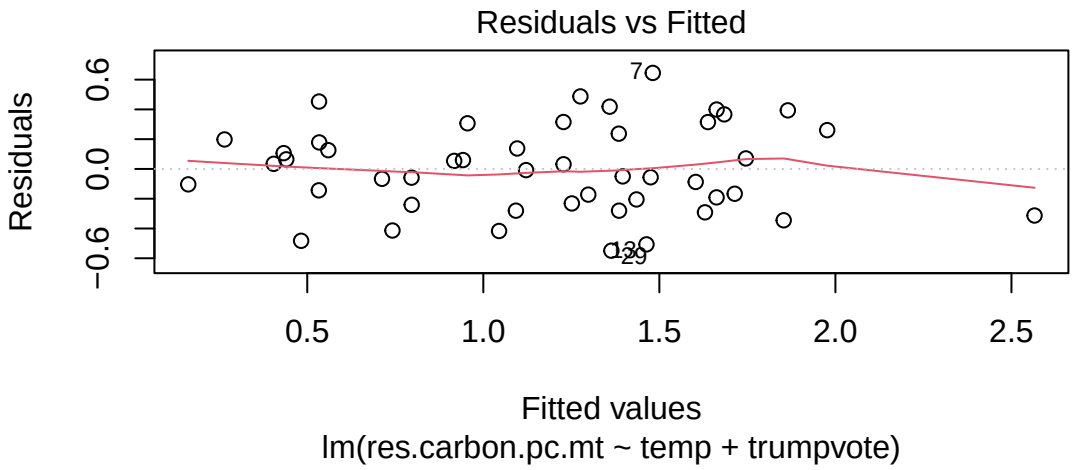
The revised version of the second LM showed that the model was statistically significant (Adjusted $R^2 = 0.75$, $F(2, 44) = 70.38$, $p < 0.001$). For this model, temperature ($p < 0.001$, $df = 47$) and trumpvote ($p = 0.001$, $df = 47$) both have statistically significant coefficients.
A one degree increase in average annual temperature is associated with a decrease of approximately 0.055 metric tons carbon in residential carbon production per capita, holding percent Trump vote constant. In addition, a one percent increase in Trump voter, the residential carbon production per capita is 0.014 metric tons lower, assuming average annual temperature is held constant.

Check model assumptions

1. Multicollinearity

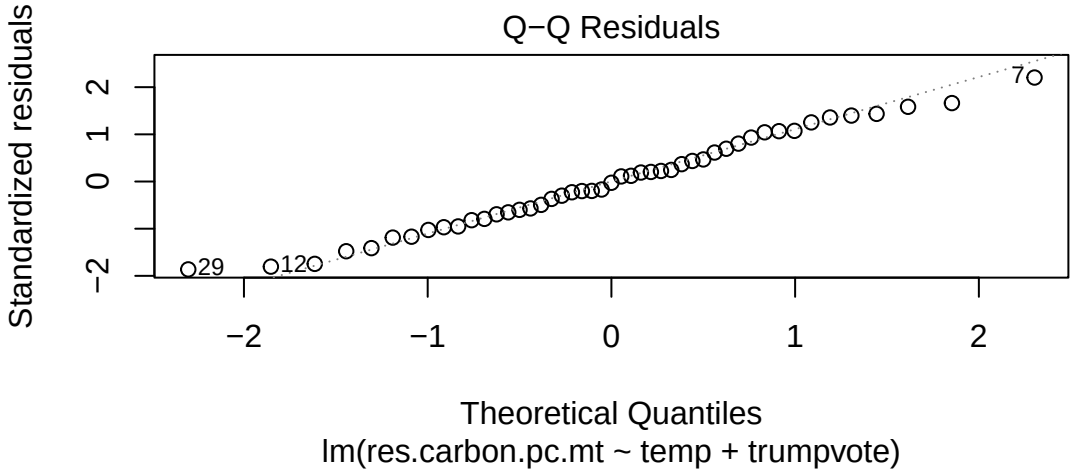
Variance Inflation Factors (VIF) showed that temp and trumpvote both have VIFs of 1. Therefore, the multicollinearity assumption is met!

2. Linearity



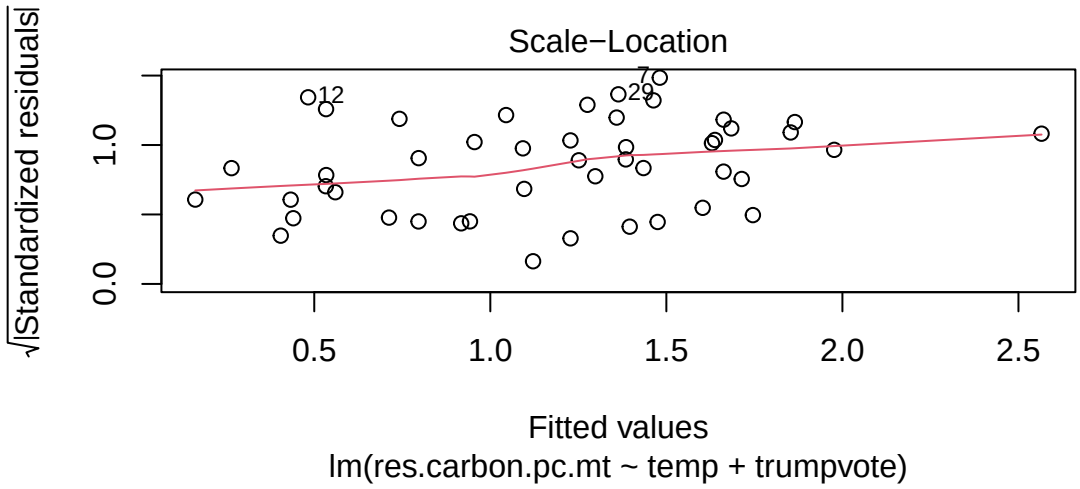
There is a relatively straight line through the fitted values. Therefore, linearity assumption is met!

3. Normal distribution of error



The majority of the data points are along the the standard normal distribution line. Shapiro-Wilk test of the residuals also yielded a nonsignificant p-value of 0.75. Therefore, the normal distribution assumption of error is met!

4. Constant variance of error



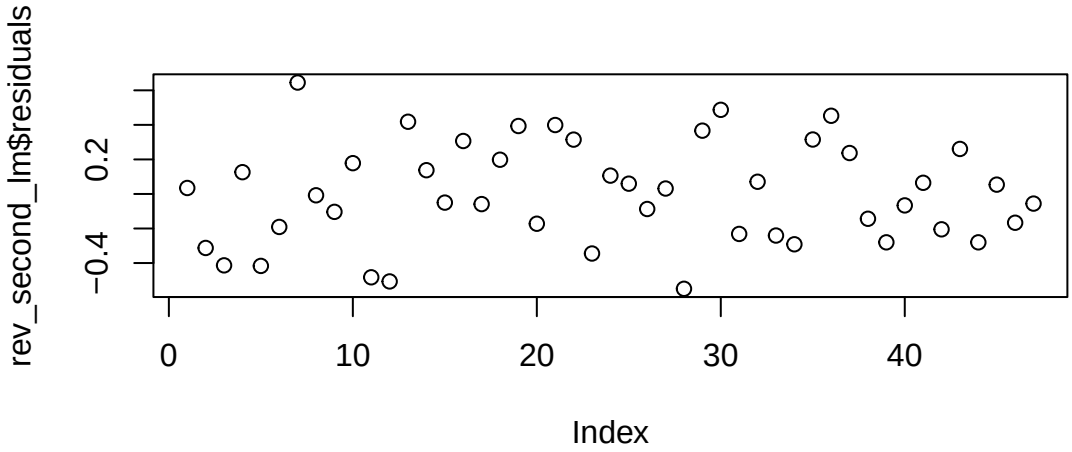
There is a relatively straight line through the fitted values. However, the studentized Breusch-Pagan test yielded a significant p-value of 0.04. Therefore, constant variance assumption is NOT met!

Also, it is important to note that there are three more leverage points in the model: 2, 5, and 12.

5. Zero error mean

The mean of the residuals is 0. Therefore, this assumption is met!

6. Uncorrelated errors



Errors are uncorrelated as they are randomly scattered around $\epsilon = 0$. The assumption is met!

In summary, all the assumptions are met, except for homogeneity of variance. However, removing three leverage points seemed to help as the p-value of the BP test increased. Removing more outliers could help, but data are discarded without knowing whether or not this is the best model to predict residential carbon per capita. Therefore, choosing the best model is needed.

Best model

Using repeated 3-fold cross validation with 100 repeats, the best model was:

$$\text{res.carbon.pc.mt} \sim \text{temp} + \text{climatechange} + \text{urban.percent} + \text{west}$$

Note: The code for model training is included in the Model Training R Script to reduce processing time of the Rmd PDF document.

This model was reported to have the highest R-square (0.78) and lowest RMSE (0.295).

The third (and the best) LM showed that the model was statistically significant (Adjusted $R^2 = 0.80$, $F(4, 45) = 48.57$, $p < 0.001$). For this model, temperature ($p < 0.001$, df = 45), climatechange ($p < 0.001$, df = 45), urban.percent ($p = 0.002$, df =45), and west ($p < 0.001$, df = 45) all have statistically significant coefficients.

On average:

A one degree increase in average annual temperature is associated with a decrease of approximately 0.054 metric tons carbon in residential carbon production per capita, holding other variables constant.

A one percent increase in climate change search, the residential carbon production per capita is 0.013 metric tons higher, assuming other variables are held constant.

A one percent increase in urban population, the residential carbon production per capita is 0.009 metric tons higher, assuming other variables are held constant.

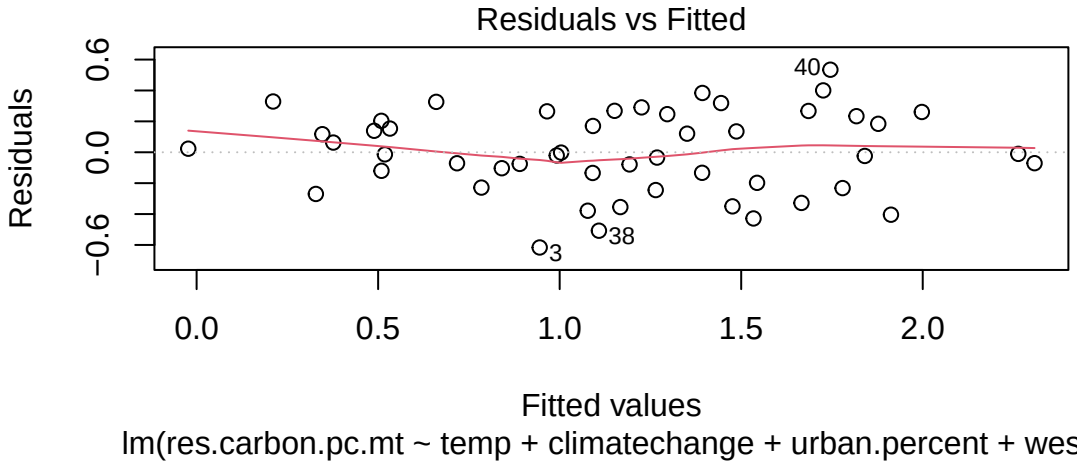
States in the west are associated with a decrease of approximately 0.635 metric tons carbon in residential carbon production per capita, holding other variables constant.

Check model assumptions

1. Multicollinearity

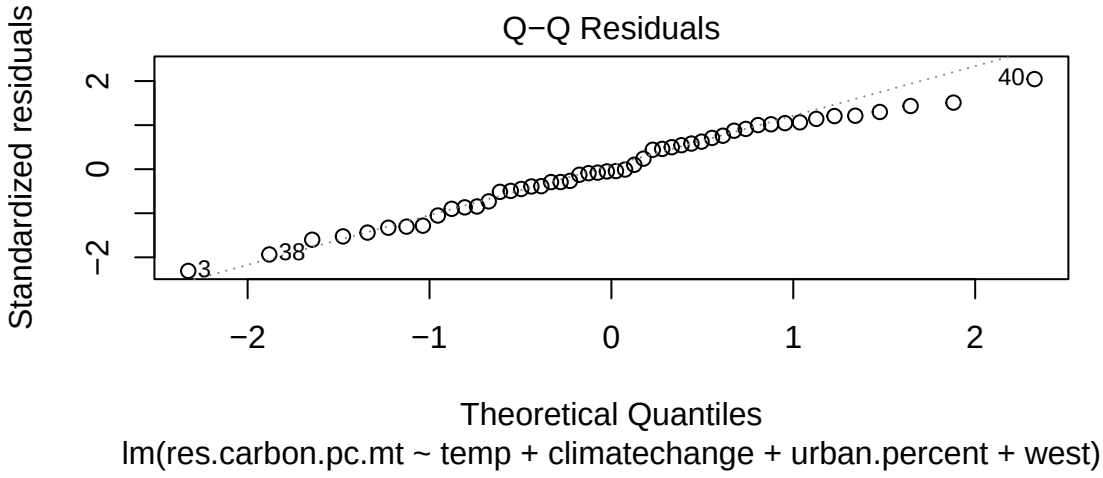
Variance Inflation Factors (VIF) showed that all variables have low VIFs (between 1 and 2). Therefore, the multicollinearity assumption is met!

2. Linearity



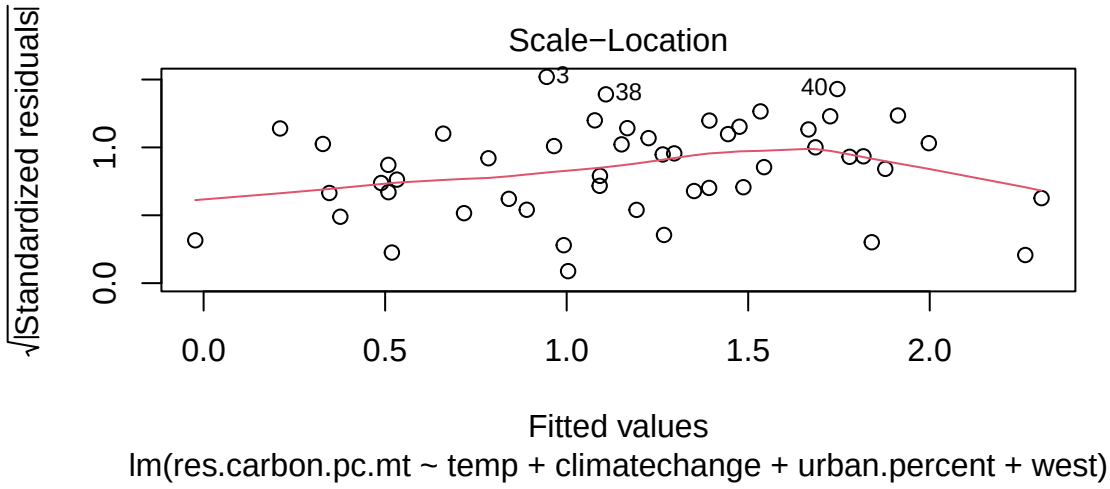
There is a relatively straight line through the fitted values. Therefore, linearity assumption is met!

3. Normal distribution of error



The majority of the data points are along the the standard normal distribution line. Shapiro-Wilk test of the residuals also yielded a nonsignificant p-value of 0.55. Therefore, the normal distribution assumption of error is met!

4. Constant variance of error



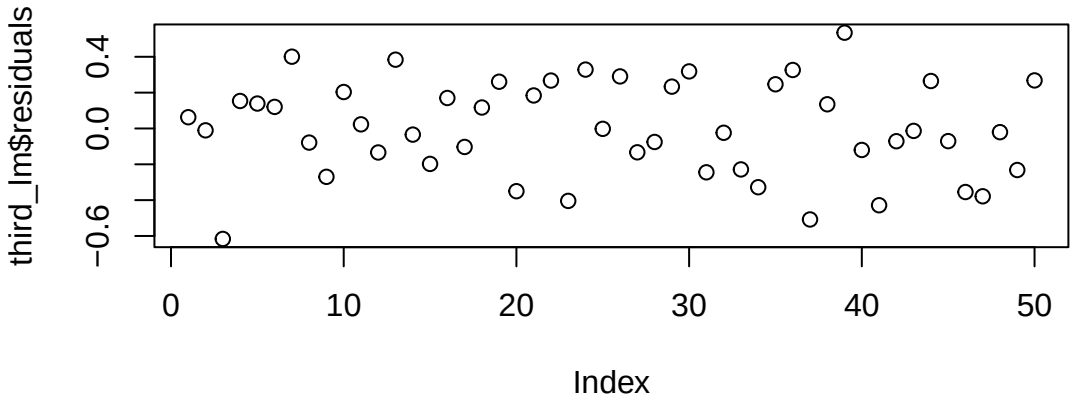
There is a relatively straight line through the fitted values. Also, the studentized Breusch-Pagan test yielded a non-significant p-value of 0.33. Therefore, constant variance assumption is met!

Also, it is important to note that there are three leverage points in the model: 3, 25, 38, and 40.

5. Zero error mean

The mean of the residuals is 0. Therefore, this assumption is met!

6. Uncorrelated errors



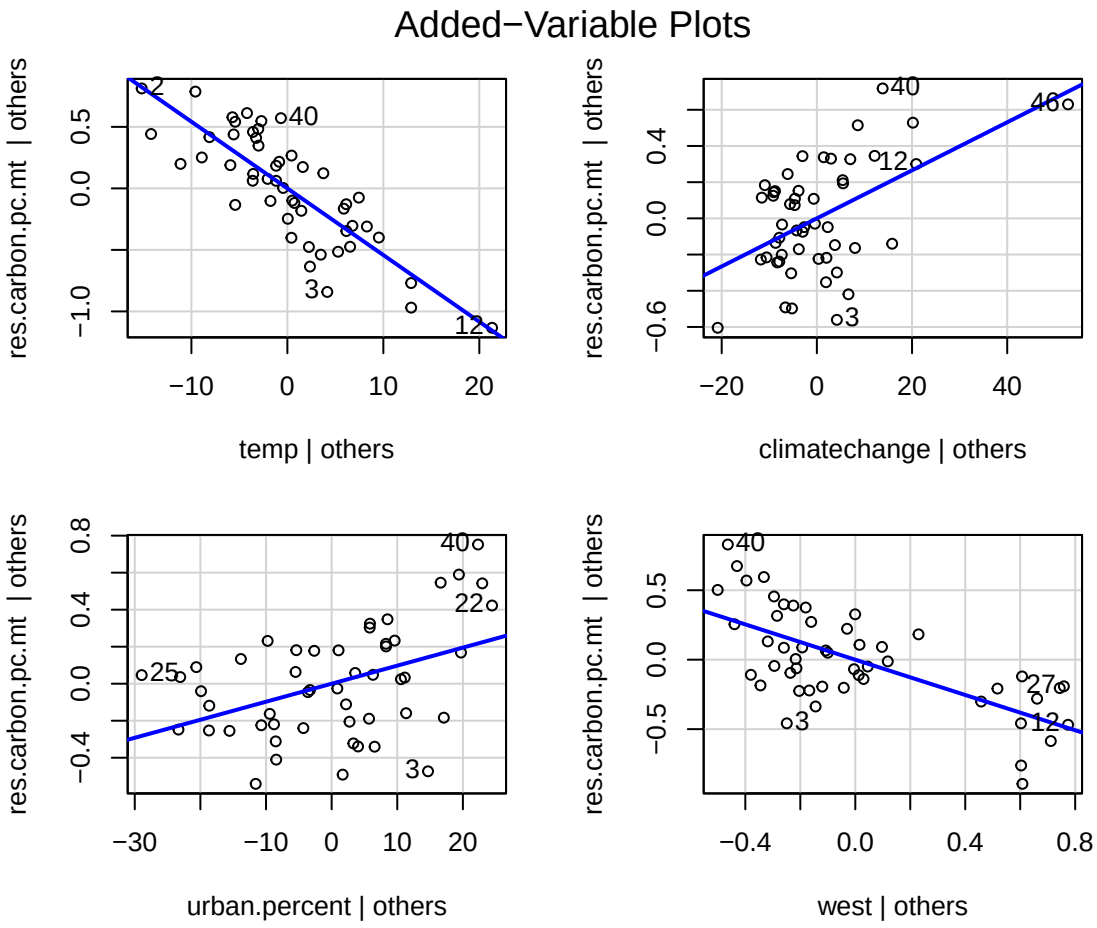
Errors are uncorrelated as they are randomly scattered around $\epsilon = 0$. The assumption is met!

Conclusion:

In summary, all the assumptions are met and this is the best model to predict the residential carbon per capita, based on average annual temperature, percent climate change search, percent urban population, and whether or not a state is in the west.

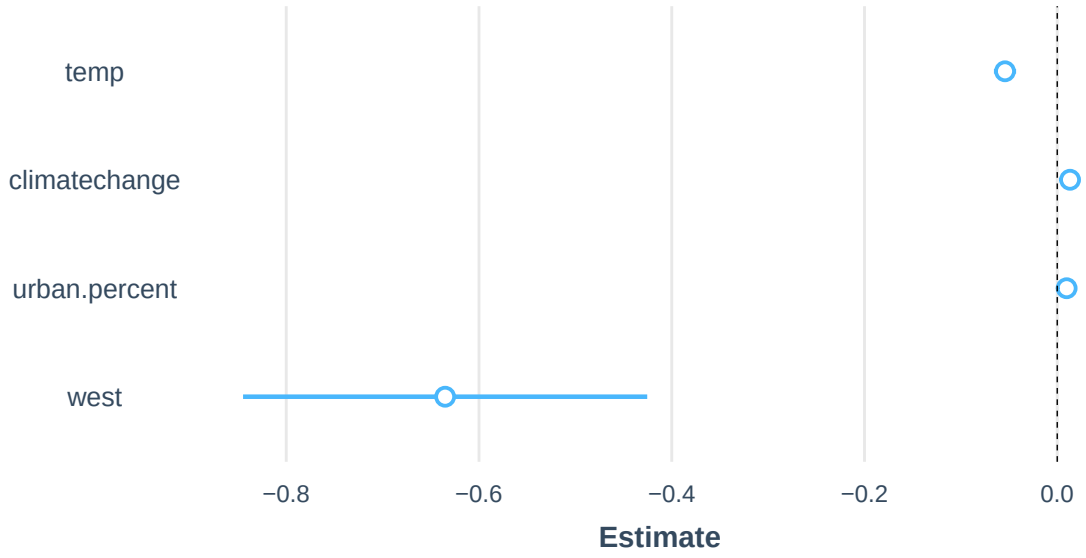
Variable plots of best model

Below is the variable plots:



The relationships match with what was shown from the LM results.

Below is the coefficient plots:



Predicting NC residential carbon per capita

The confidence interval of NC residential carbon per capita is: (0.667, 0.901).

The prediction interval of NC residential carbon per capita is: (0.215, 1.353).