Assignment 3: Regression

Part 1 Data Preparation and Preliminary Tests

- Explanation for why choosing to operationalize 'need' in such ways
  - Total Official Need ($, untransformed)

    Firstly, we choose to operationalize the 'need' as the original Total Official Need data of each water facility in the units of dollar. After creating a regression model with the Total Official Need as the dependent variable, the model can reveal the relationship between the Total Official Need of each facility and a series of explanatory variables, and also reveal how the Total Official Need of each facility would change in units because of the change of the explanatory variables in units.

  - Log of Total Official Need (log $)

    Secondly, we choose to operationalize the 'need' as the logging value of the original Total Official Need data of each water facility in the unit of logging dollars. In fact, logging the dependent variable would make the whole model more approaching to a normal distribution. After creating a regression model with the Log of Total Official Need as the dependent variable, the model can reveal the relationship between the logging value of Total Official Need of each water facility and the series of explanatory variables, and also reveal how many percentage units the Total Official Need of each facility would change because of the unit change of the explanatory variables.

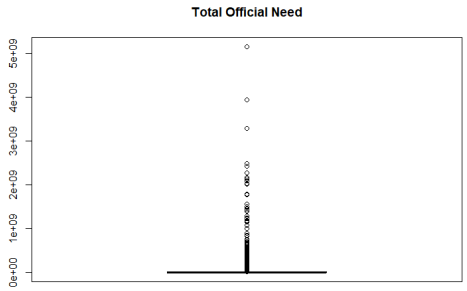
  - Residential Burden ($/person)

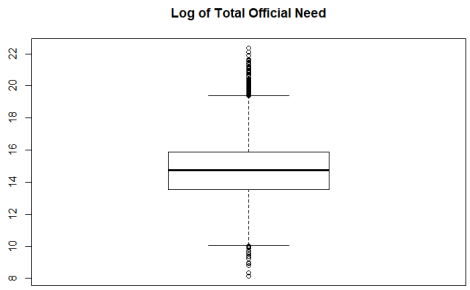
    Thirdly, we choose to operationalize the 'need' as the Residential Burden of each water facility which equals to the result of dividing the Total Official Need of each water facility by the maximum value between Present residential population receiving collection and receiving treatment for each facility. It is in the unit of dollar per person. After creating a regression model with the Residential Burden of each facility as the dependent variable, the model can reveal the relationship between Residential Burden and a series of explanatory variables, and also reveal how the residential burden of each facility would change in units because of the change of the explanatory variables in units.

  - Log of Residential Burden (log $/person)

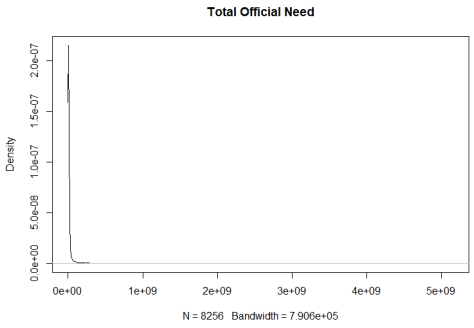
Lastly, we choose to operationalize the 'need' as the logging value of the Residential Burden of each water facility in the unit of the logging value of dollar per person. After creating a regression model with the Logging value of the Residential Burden of each facility as the dependent variable, the model can reveal the relationship between the logging value of Residential Burden and the series of explanatory variables, and also reveal how many percentage units the Residential Burden of each facility would change because of the unit change of the explanatory variables.

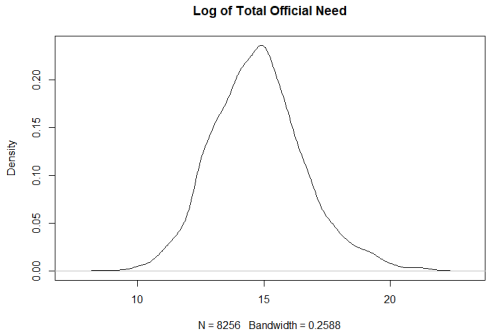
- 1.2 Boxplots, Density plots and Histograms
  - Total Official Need ($, untransformed) vs Log of Total Official Need (log $)
    - ❖ Boxplot



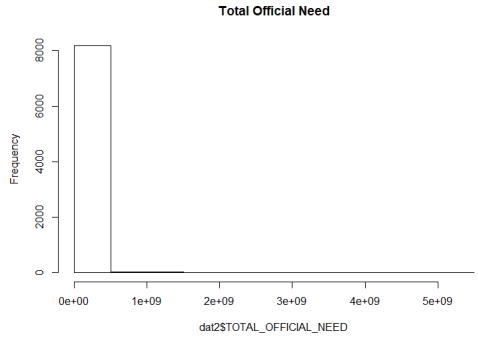


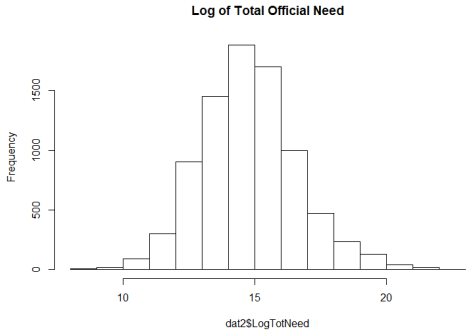
    - ❖ Density plots



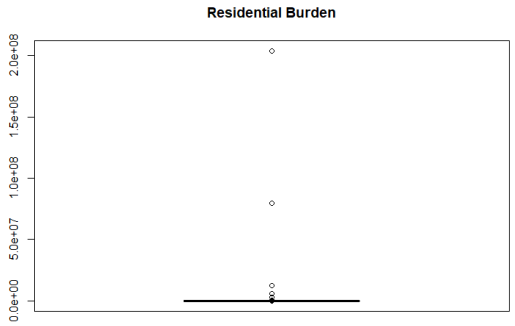


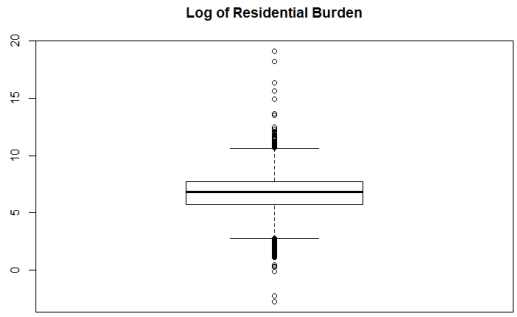
    - ❖ Histograms



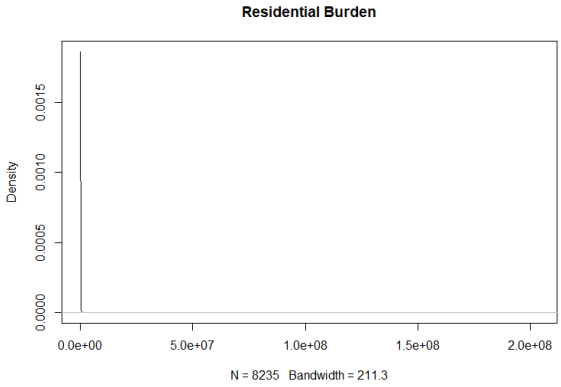


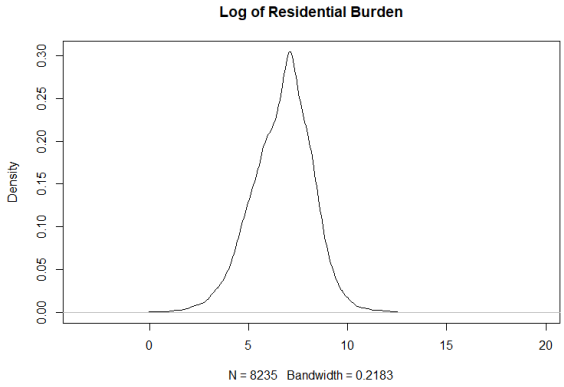
- Residential Burden ($/person) vs Log of Residential Burden (log $/person)
  - ❖ Boxplot



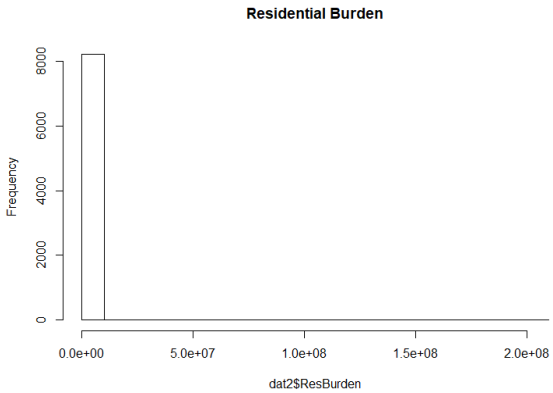


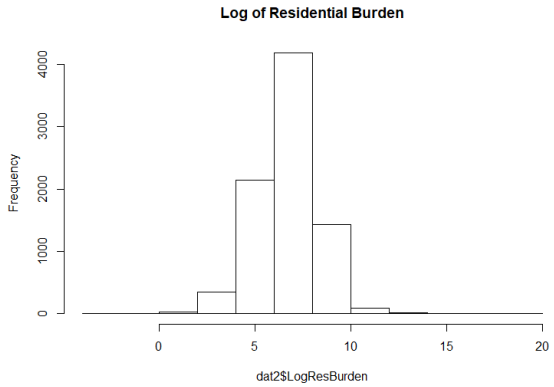
  - ❖ Density plots





  - ❖ Histograms





Analysis of Section 1.2: Based on the graphs, the distribution of the data of the raw values (Total Official Need, & Residential Burden) are skewed while the distribution of their logging values is more like a normal distribution. The reason is that the use of logging values would make the model more approaching to be a normal distribution.

- 1.3 Variables Selection
    - Dummy Variables: if the facility associated with a CSS, if the facility contributing to a TMDL receiving water body, If the city is a (population) Growing City, if the city has non-residential population, if the city from EPA1-10.
    - Other Explanatory Variables: Population Density 1980, 1990, 2000, 2010; Percentage of White Population in 2000, 2010; Present (2012) Residential & non-residential Population receiving service; Median Income in 1968, 1979, 1989, 1999, 2009; Population 1980, 1990, 2000, 2010; logging values of median incomes; logging values of population; percent change of population density from 2000 to 2010.

For selecting the explanatory variables, firstly, we want to include the population variables in the regression. Under our consideration, since water is the rigid demand for people, the higher the population is, the higher the total needs for the water infrastructures would be. So, we would select Population 1980, 1990, 2000, 2010, logging values of population, and Present (2012) Residential & non-residential Population receiving service in regression. What's more, although population density is also about population, because of the difference of county area, we think those variables (Population Density 1980, 1990, 2000, 2010) may play some different effects from the simple population variables on the dependent variable. Additionally, we want to add one percentage change variables which is about the population density change from 2000 to 2010. In theory, to some extents, it would be related to the dependent variable. For example, if the population density decreased a lot in a county, the needs for water infrastructure may be lower.

In addition, we also want to include some variables of median income in different decades. In theory, people would not only have rigid demands of water for supporting our lives, but also have water needs for other uses like recreational use when people are rich enough. So, we think if the median income is high, the needs for water facilities may also be high. Thus, we select Median Income in 1968, 1979, 1989, 1999 and 2009 in the model. What's more, we also think the difference in race may also be an influence factor to the need of water infrastructure. So, we would include the Percentage of White Population in 2000 and 2010 in the model to see if the number of White Population would affect the needs for water infrastructure.

Then, for selecting dummy variables which would only have two numeric values that 1 means the presence and 0 means the absence of some categorical effect that may be expected to shift the outcome. In fact, we want to include if the facility associated with a CSS or if the facility contributing to a TMDL receiving water body in the model, because we speculate that if the facility associated with a CSS or if the facility contributing to a TMDL receiving water body, the reported needs would be higher than other facilities. Furthermore, we also want to include if the city is a (population) Growing City and if the city has non-residential population as two dummy variables in the model. In fact, a population growing city may have an increasing water demand. Additionally, the city with a lot of non-residential population may also have a different situation about water demand from the pattern of the cities only with residential population. Under our consideration, the non-residential population may need less water in their daily life in the city, so that the water needs would be fewer in these cities. Finally, we want to include 10 dummy variables for the 10 EPA Regions in the model. In theory, we think the difference in geographic locations may be necessary to explain the difference in reported needs of water infrastructure between different EPA regions. In fact, in the multivariate model, we may select one EPA region as the control group and the others as the treatment group.

- 1.4 Pairwise Correlation

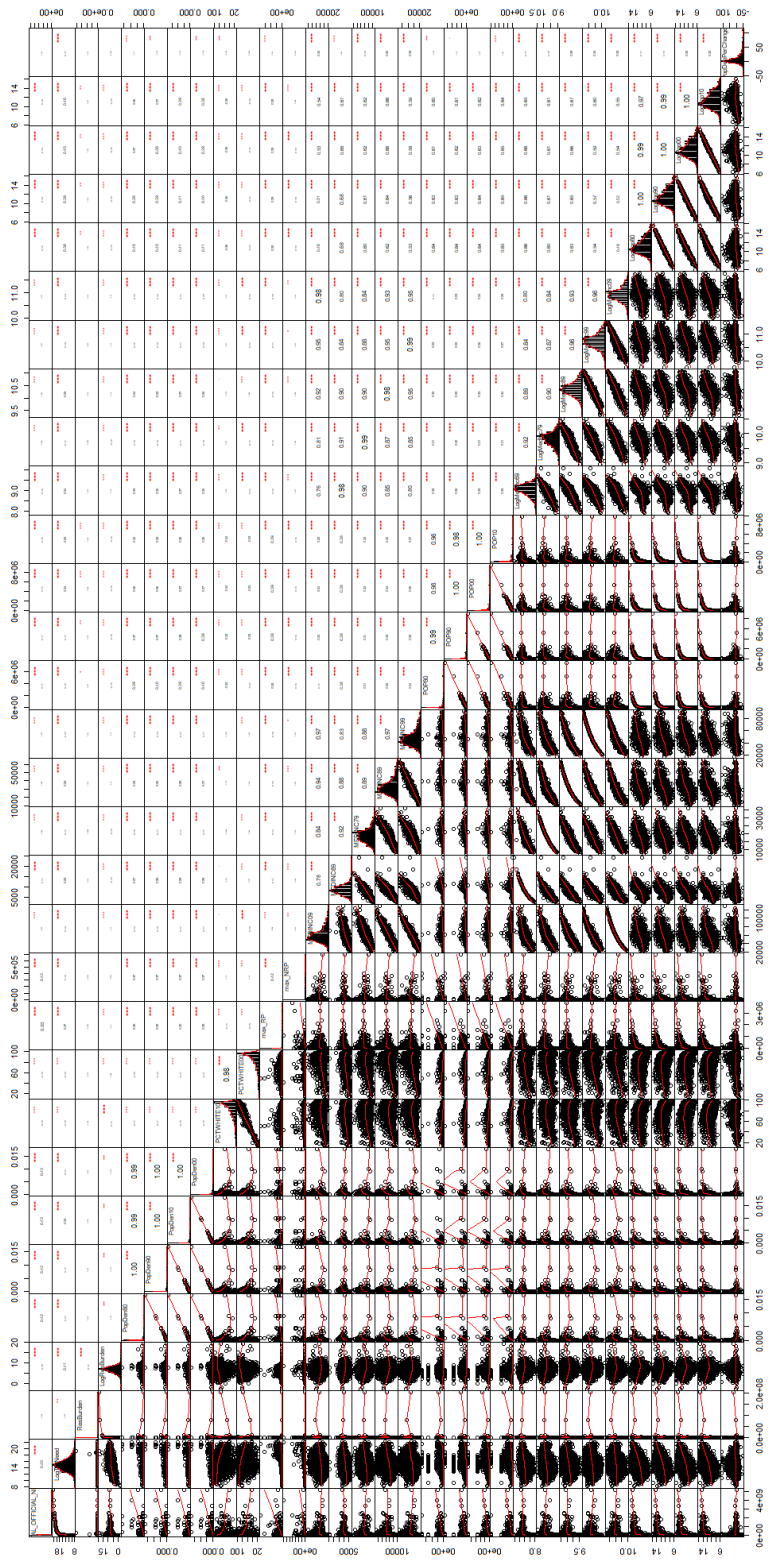


Figure 1.4.1

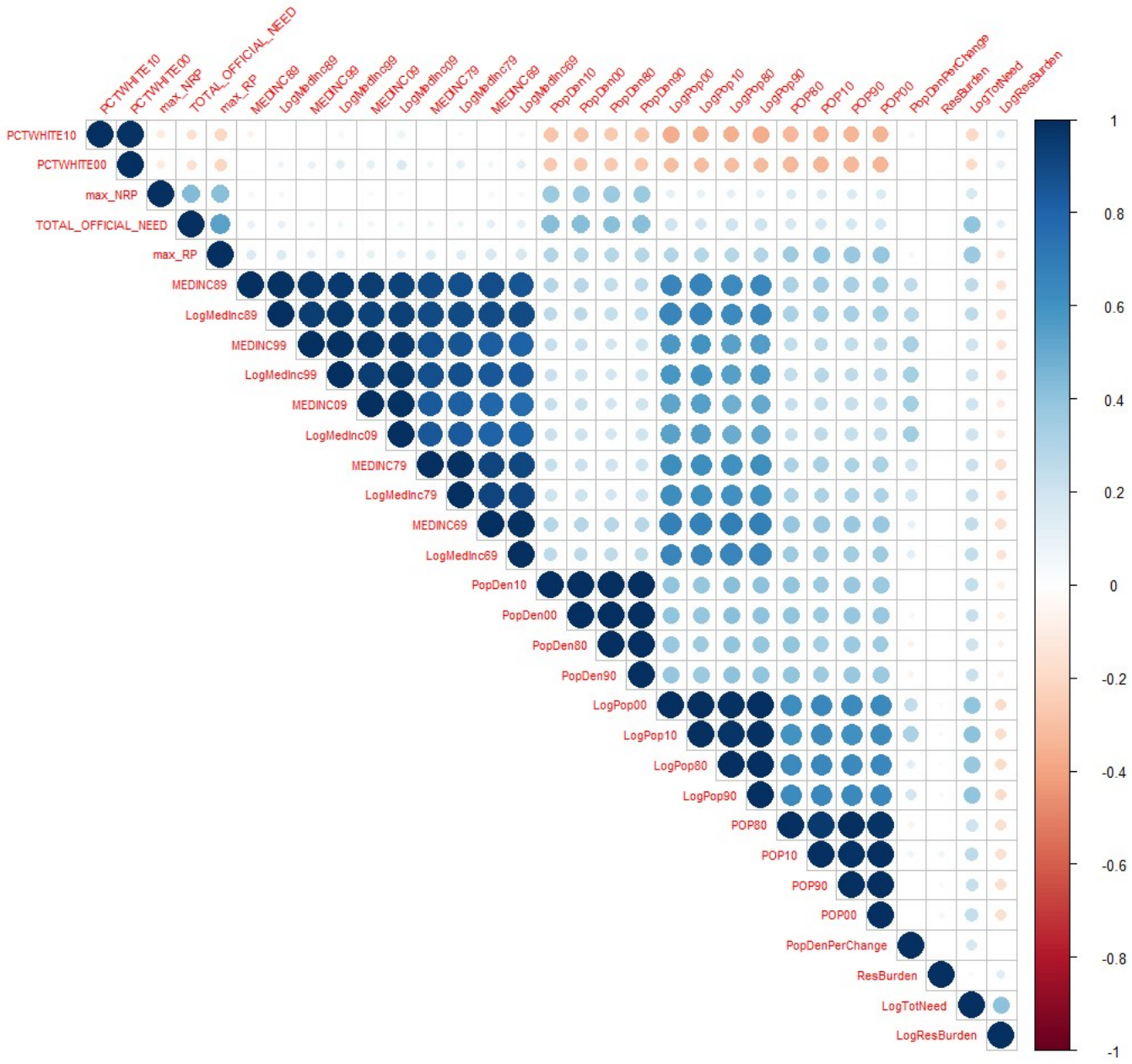

PCTWHITE10
PCTWHITE00
max_NRP
TOTAL_OFFICIAL_NEED
max_RP
MEDINC89
LogMedInc89
MEDINC99
LogMedInc99
MEDINC09
LogMedInc09
MEDINC79
LogMedInc79
MEDINC69
LogMedInc69
PopDen10
PopDen00
PopDen80
PopDen90
LogPop00
LogPop10
LogPop80
LogPop90
POP80
POP10
POP90
POP00
PopDenPerChange
ResBurden
LogTotNeed
LogResBurden
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Figure 1.4.2

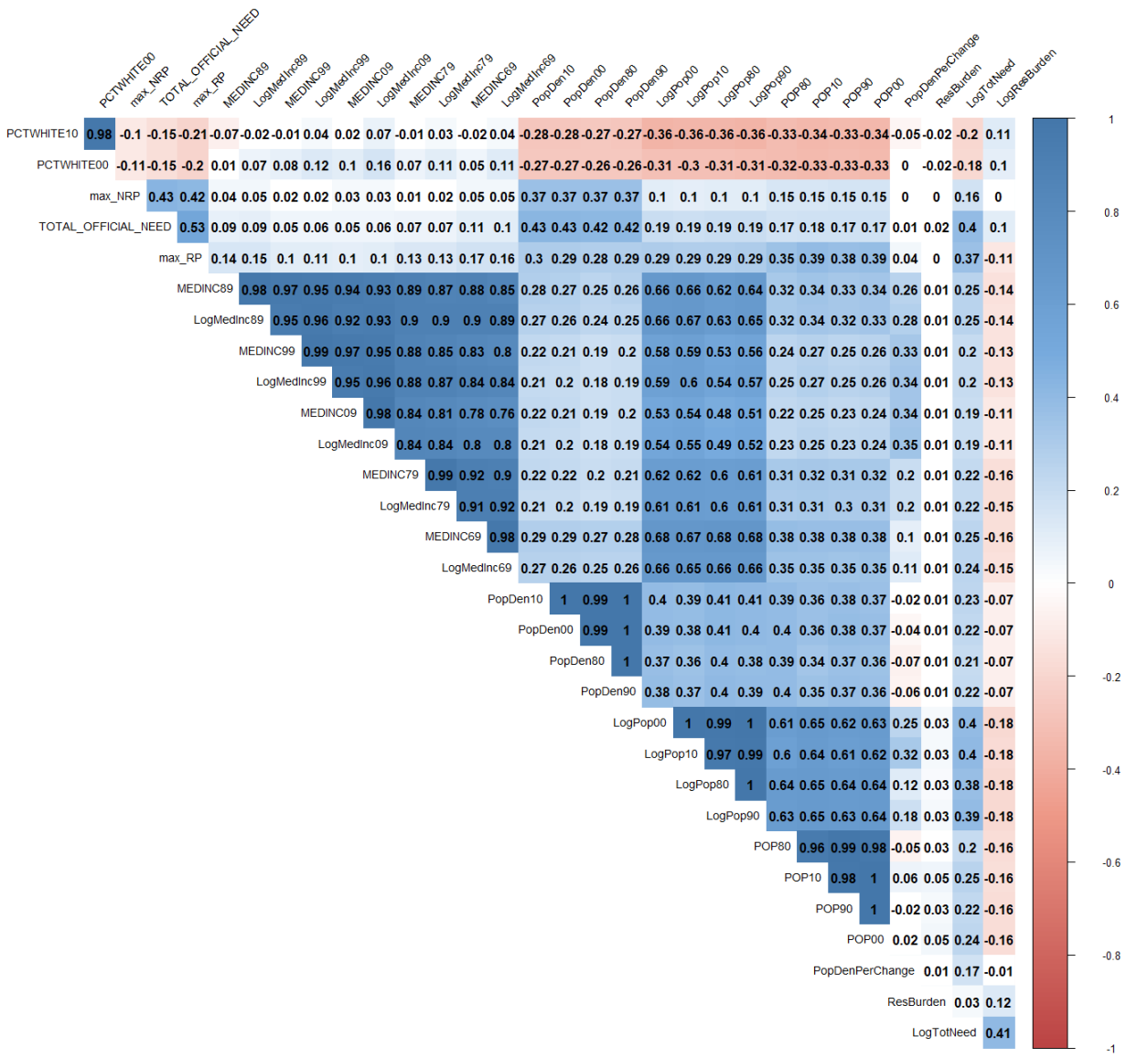


Figure 1.4.3

- Analysis Criteria

  In this section, we ran pairwise correlations between each pair of all variables (both dependent and explanatory variables) we want to include in the regression models to avoid multicollinearity. So, we would not include highly correlated variables in the regressions. Our criteria for the correlation coefficients for the correlation is that: if the absolute coefficient is higher than 0.5, these two variables would be considered as highly correlated and should at most keep one of them in the models.

- Analysis

  We ran some R code to produce these Figure 10. above to indicate the correlation coefficient between the 4 dependent variables of "Need" and all explanatory and dummy variables we want to include in the regression models we reported in Section 1.3.

  According to the Figure 1.4.3, the correlation coefficient between every two of the median income variables also with their logging value variables for different years are very high. What's more, the correlation coefficients between every two of population variables and also with their logging values are also high. **So, we may only include one population variable and one median income variable in the model (e.g. Pop10, MedInc09)** Besides, we should also avoid including both median income variables (raw data & logging data) and logging values of the population density in different years in the models at the same time.

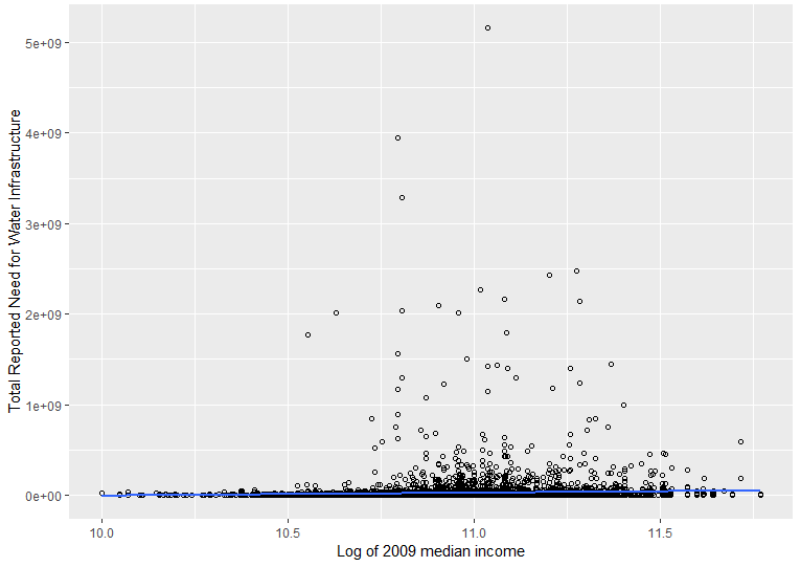
Part 2 Bivariate Regressions

To do a Bivariate Regression, we selected 10 explanatory variables including 4 dummy variables and 6 numeric variables, which are: Logging value of median income in 2009 in county, Logging value of 2010 population in county, If the city's population has grown from 2000 to 2010 (Dummy), If a city has non-residential population (Dummy), If a facility is associated with a combined sewer system (Dummy), If a facility contributes to a TMDL receiving water body (Dummy), Percent of county population identifying as white in 2010, Percent Change of Population Density from 2000 to 2010, Present (2012) residential population receiving services, and Population Density in 2010.

- **Analysis Methods and Criteria**

  Firstly, we would build 10 bivariate regression models for each of the 4 dependent variables. Then, we will check the P value for the explanatory variables. If the P value is less than the selected significant level (0.01), this variable would be significantly correlated to the dependent variable and could be include in the multivariate regression model, and vice versa.

  - **Analysis of the bivariate regression models**
    - Total Official Need ($, untransformed)
      i. Logging value of Median income in 2009 in county



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable LogMedInc09 is 31522537, which means that for every additional one percentage unit increase from the median income in 2009, the total need for water infrastructure would increase by $31,522,537.

- How well predict the model

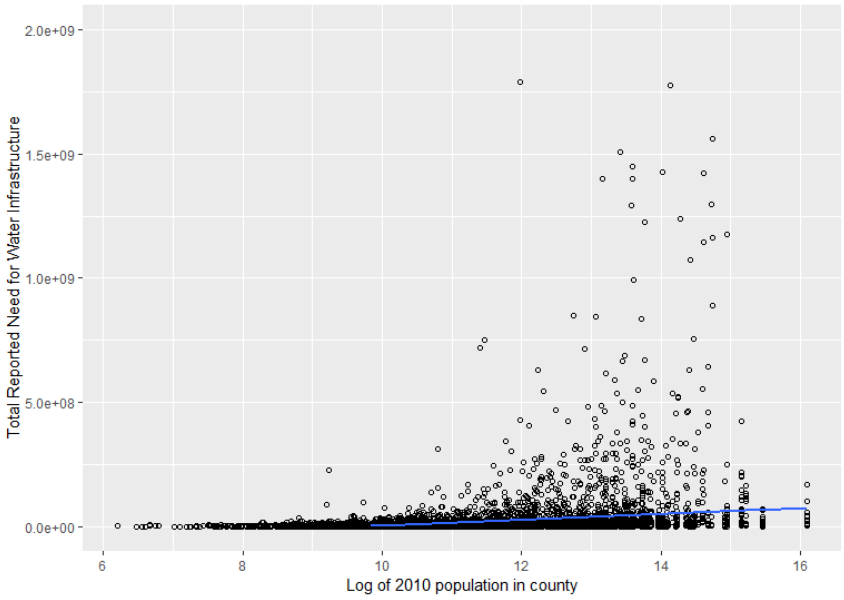
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the logging value of median income in 2009 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Total Reported Needs).

  It's worth mentioning that the adjusted R-squared value for this regression is 0.003176, which means that only about 0.3% of the variance of the Total Report Need could be explained by the logging value of the median income in 2009. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph reveals that there are lots of outlier data and the linear relationship is not significant between the dependent and explanatory variables in this model.

ii. Logging value of 2010 population in county



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable LogPop10 is 16045067, which means that for every additional one percentage unit increase from the county populaion in 2010, the total need for water infrastructure would increase by $16,045,067.

- How well is the model prediction

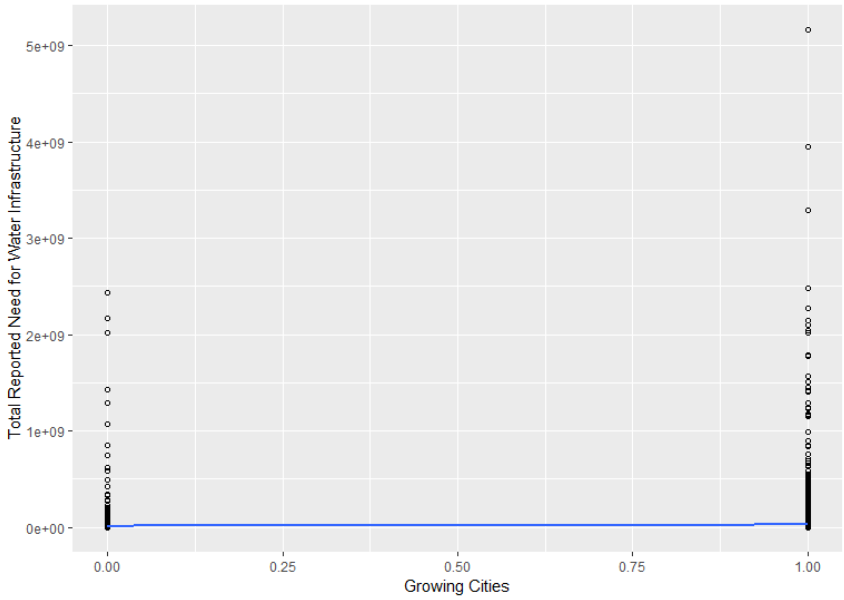
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the logging value of county population in 2010 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Total Reported Needs).

  It's worth mentioning that the adjusted R-squared value for this regression is 0.03519, which means that only about 3.5% of the variance of the Total Report Need could be explained by the logging value of the population in 2010. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

The scatterplot graph also reveals that there are lots of outlier data and the linear relationship is not significant between the dependent and explanatory variables in this model.

iii. Growing Cities



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable GrowingCities is 13127286, which means that if the city's population has grown from 2000 to 2010, the total need for water infrastructure would be $13,127,286 higher than the cities whose population didn't grow.

- How well is the model prediction

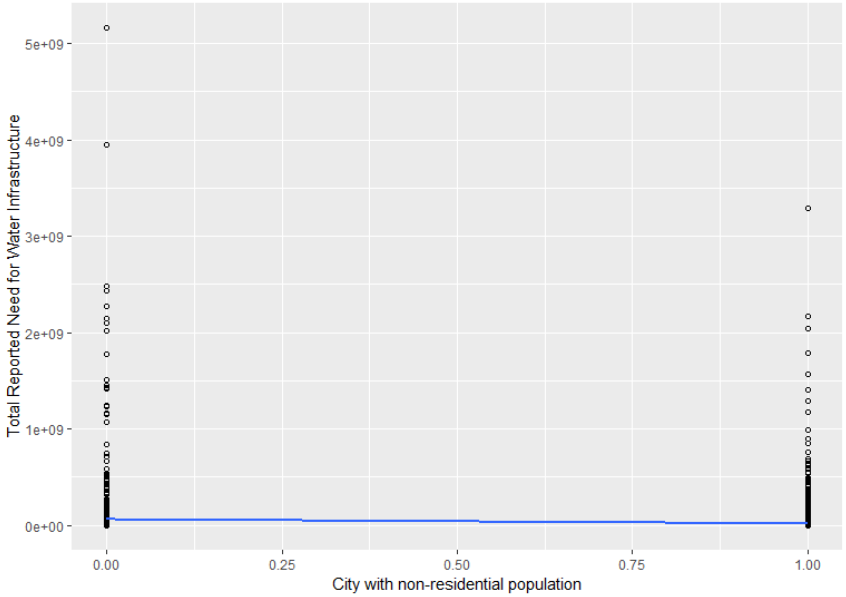
  The result summary of this bivariate regression model reveals that the P value for this dummy variable is far less than 0.01 which is our choosing significant level. It means that if the city is a population growing city from 2000 to 2010 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Total Reported Needs).

  It's worth mentioning that the adjusted R-squared value for this regression is 0.001774, which means that only about 0.2% of the variance of the Total Report Need could be explained by this dummy variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

iv. City with non-residential population



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable City_nR is -44526362, which means that if a city has non-residential population, the total need for water infrastructure would be $44,526,362 lower than the cities which only have residential population.

- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for this dummy variable is far less than 0.01 which is our choosing significant level. It means that if the city has non-residential population is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Total Reported Needs).

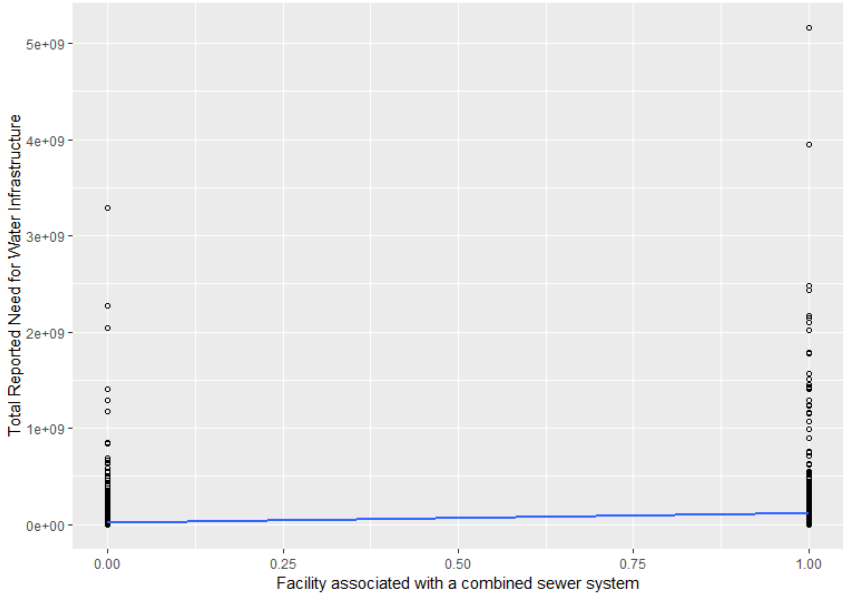
  It's worth mentioning that the adjusted R-squared value for this regression is 0.01264, which means that only about 1.3% of the variance of the Total Report Need could be explained by this dummy variable. However, although the R

squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

v. Facility associated with a combined sewer system



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable CSS is 99852428, which means that if a facility is associated with a combined sewer system, the total need for water infrastructure would be $99,852,428 higher than the facilities aren't associated with a combined sewer system.

- How well is the model prediction

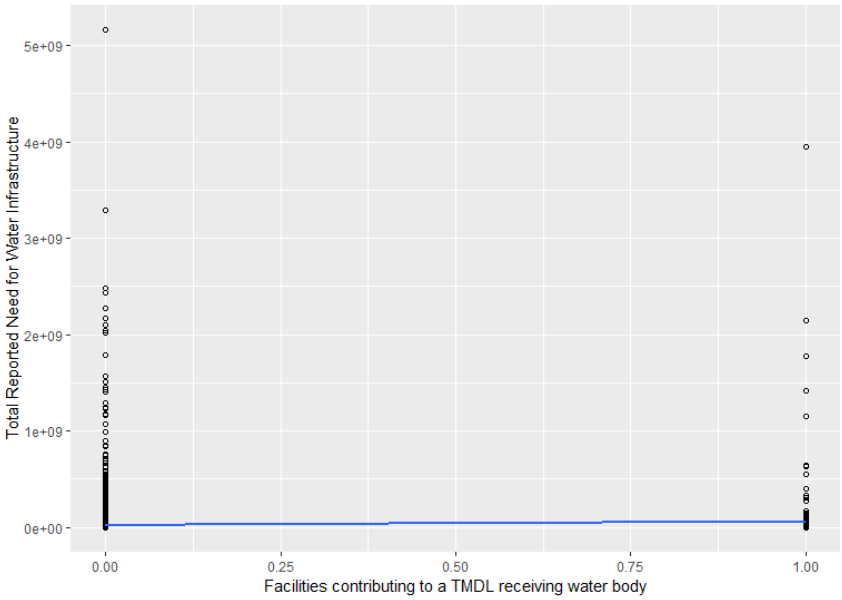
  The result summary of this bivariate regression model reveals that the P value for this dummy variable is far less than 0.01 which is our choosing significant level. It means that if a facility is associated with a combined sewer system is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Total Reported Needs).

It's worth mentioning that the adjusted R-squared value for this regression is 0.04149, which means that only about 4.1% of the variance of the Total Report Need could be explained by this dummy variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

vi. Facilities contributing to a TMDL receiving water body



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable TMDL_Dummy is 38660991, which means that if a facility contributes to a TMDL receiving water body, the total need for water infrastructure would be $38,660,991 higher than the facilities don't contributes to a TMDL receiving water body.

- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for this dummy variable is far less than 0.01 which is our choosing significant level. It means that if a facility contributes to a TMDL receiving water body is highly significantly correlated to the dependent value and it could be included

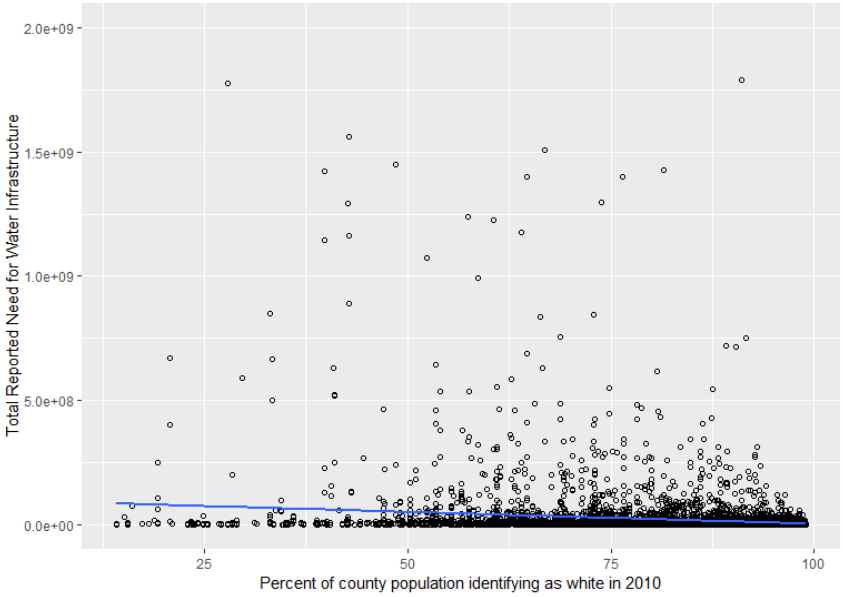
in the multivariate regression model to predict the variation of the dependent variable (Total Reported Needs).

It's worth mentioning that the adjusted R-squared value for this regression is 0.002729, which means that only about 0.3% of the variance of the Total Report Need could be explained by this dummy variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

vii. Percent of county population identifying as white in 2010



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable PCTWHITE10 is -1378106, which means that for every additional one unit increase in the percentage of county population identified as White in 2010, the total need for water infrastructure would decrease by $1,378,106.

- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing

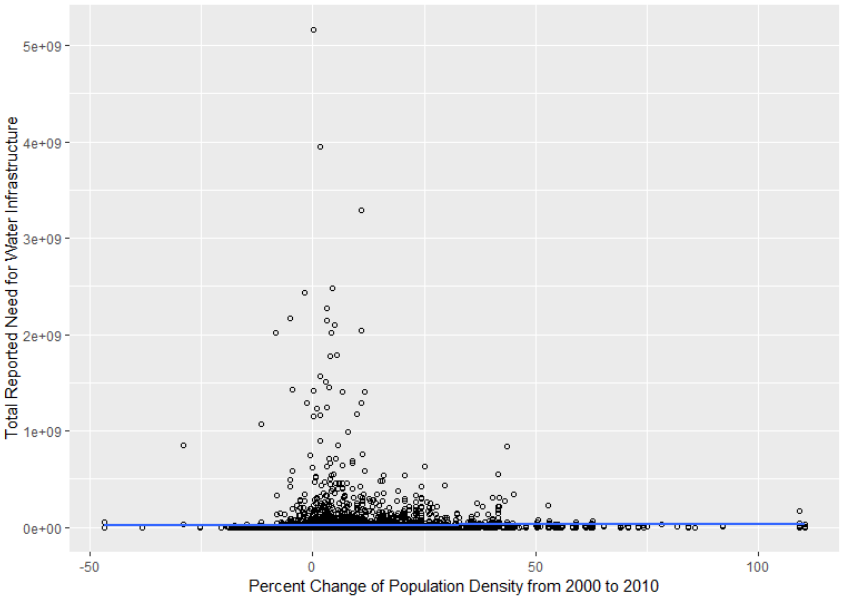
significant level. It means that the percentage of county population identified as White in 2010 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Total Reported Needs).

It's worth mentioning that the adjusted R-squared value for this regression is 0.02353, which means that only about 2.4% of the variance of the Total Report Need could be explained by percentage of county population identified as White in 2010. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph also reveals that there are lots of outlier data and the linear relationship is not significant between the dependent and explanatory variables in this model.

viii. Percent Change of Population Density from 2000 to 2010



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable PopDenPerChange is 109695, which means that for every additional one percentage unit increase in the change of Population Density from 2000 to 2010, the total need for water infrastructure would increase by $109,695.

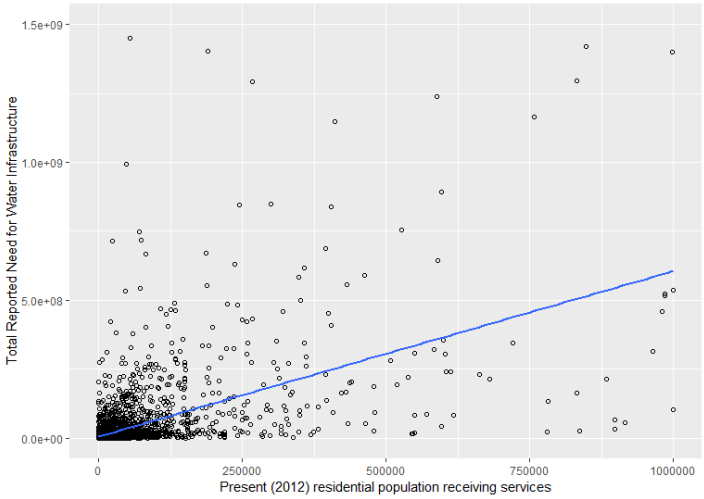
- How well is the model prediction

The result summary of this bivariate regression model reveals that the P value for the explanatory variable is 0.315, which is larger than 0.01 which is our choosing significant level. It means that the percentage change of Population Density from 2000 to 2010 is not correlated to the dependent value and it may not be included in the multivariate regression model.

- Scatterplot Characteristics

  The scatterplot graph also reveals that there are lots of outlier data and it seems that there's no linear relationship between the dependent and explanatory variables in this model.

ix. Present (2012) residential population receiving services



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable max_RP is 671.6, which means that for every additional one unit increase in the residential population who receiving services in 2012, the total need for water infrastructure would increase by $671.6.

- How well is the model prediction

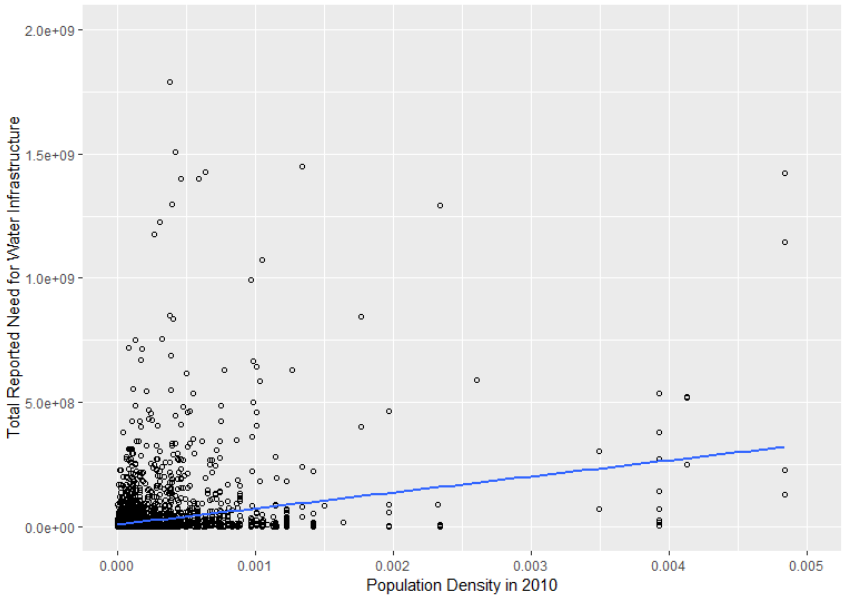
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the present (2012) residential population receiving services is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Total Reported Needs).

The adjusted R-squared value for this regression is 0.2825, which means that about 28.25% of the variance of the Total Report Need could be explained by the present (2012) residential population receiving services. So, this is a good model.

- Scatterplot Characteristics

  The scatterplot graph reveals that the relationship between the dependent and explanatory variables in this model is similar to a linear relationship.

x. Population Density in 2010



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable PopDen10 is 1.076*10^11, which means that for every additional one unit increase from the population density in 2010, the total need for water infrastructure would increase by $1.076*10^11.

- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the population density in 2010 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Total Reported Needs).

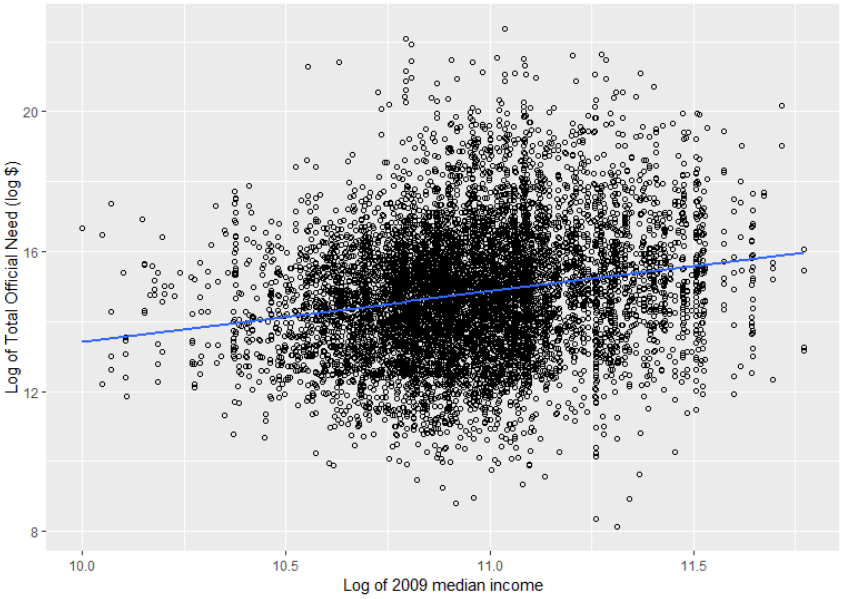
The adjusted R-squared value for this regression is 0.1807, which means that about 18.07% of the variance of the Total Report Need could be explained by this variable. So, this is a good model.

- Scatterplot Characteristics

  The scatterplot graph reveals that the relationship between the dependent and explanatory variables in this model is similar to a linear relationship.

b. Log of Total Official Need (log $)

i. Logging value of Median income in 2009 in county



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable LogMedInc09 is 1.44534, which means that for every additional one percentage unit increase from the median income in 2009, the total need for water infrastructure would increase by 1.44534%.

- How well is the model prediction

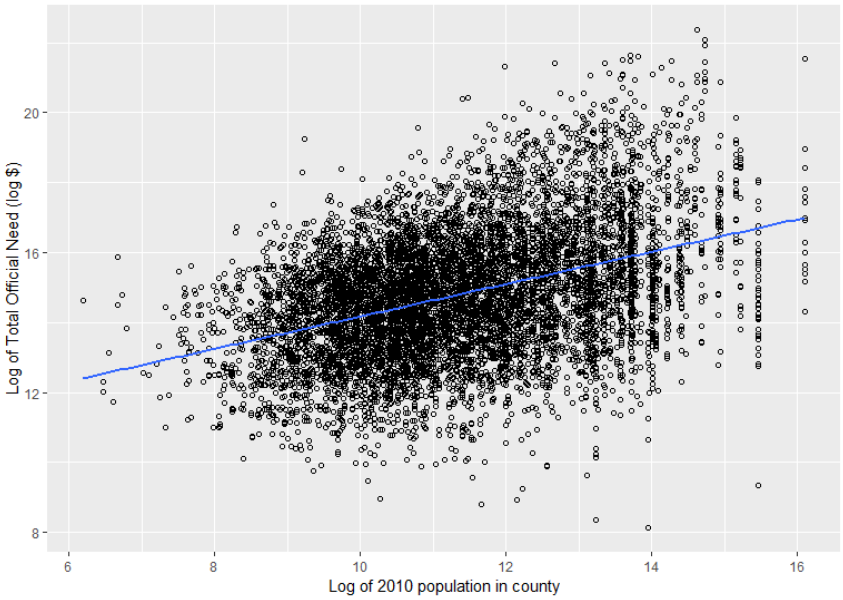
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the logging value of median income in 2009 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

It's worth mentioning that the adjusted R-squared value for this regression is 0.0381, which means that only about 3.8% of the variance of the Logging value of Total Report Need could be explained by the logging value of the median income in 2009. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph reveals that the linear relationship is not significant between the dependent and explanatory variables in this model.

ii. Logging value of 2010 population in county



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable LogPop10 is 0.46297, which means that for every additional one percentage unit increase from the county population in 2010, the total need for water infrastructure would increase by about 0.46%.

- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the logging value of county population in 2010 is highly significantly correlated to the dependent value and it could be

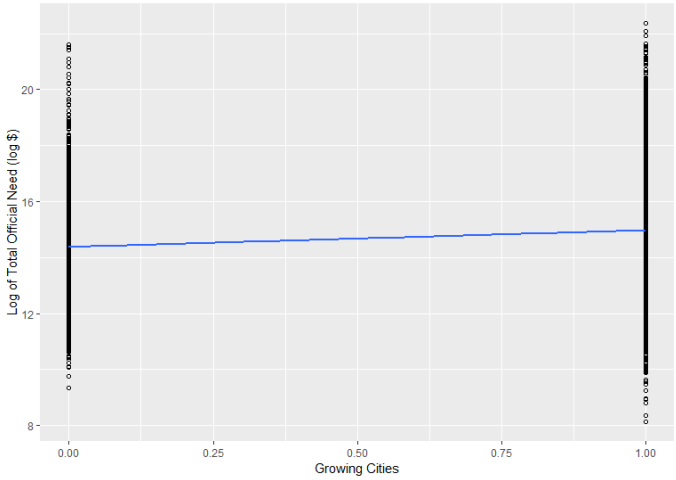
included in the multivariate regression model to predict the variation of the dependent variable (Logging value of Total Reported Needs).

It's worth mentioning that the adjusted R-squared value for this regression is 0.162, which means that about 16.2% of the variance of the logging value of Total Report Need could be explained by the logging value of the population in 2010. So, it is a good model.

- Scatterplot Characteristics

  The scatterplot graph reveals that the relationship between the dependent and explanatory variables in this model is similar to a linear relationship.

iii. Growing Cities



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable GrowingCities is 0.59507, which means that if the city's population has grown from 2000 to 2010, the total need for water infrastructure would be about 0.6% higher than the cities whose population didn't grow.

- How well is the model prediction

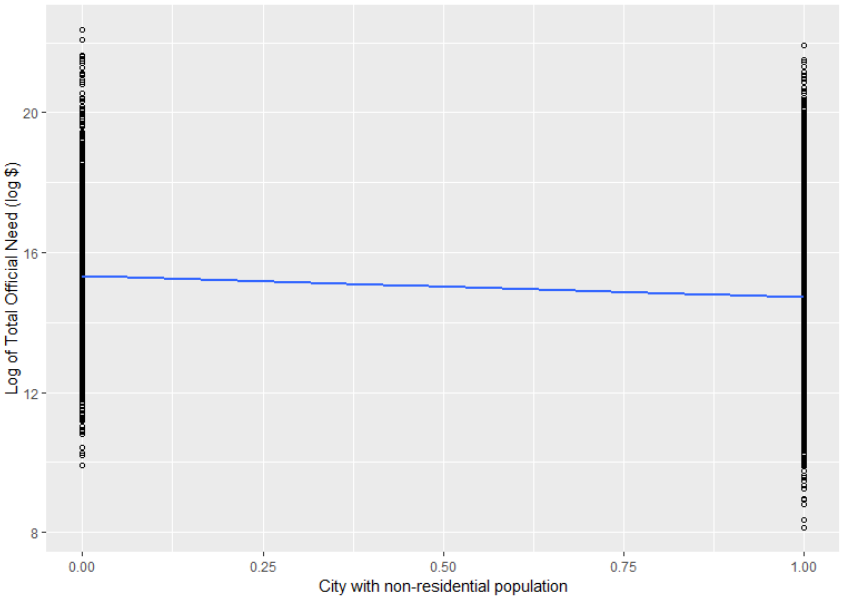
  The result summary of this bivariate regression model reveals that the P value for this dummy variable is far less than 0.01 which is our choosing significant level. It means that if the city is a population growing city from 2000 to 2010 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Logging value of Total Reported Needs).

It's worth mentioning that the adjusted R-squared value for this regression is 0.02142, which means that only about 2.1% of the variance of the logging value of Total Report Need could be explained by this dummy variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

iv. City with non-residential population



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable City_nR is -0.60994, which means that if a city has non-residential population, the total need for water infrastructure would be about 0.6% lower than the cities which only have residential population.

- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for this dummy variable is far less than 0.01 which is our choosing significant level. It means that if the city has non-residential population is highly significantly correlated to the dependent value and it could be included in the

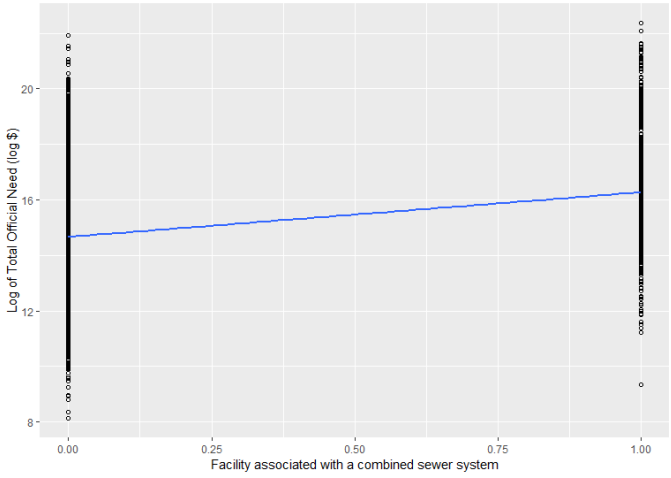
multivariate regression model to predict the variation of the dependent variable.

It's worth mentioning that the adjusted R-squared value for this regression is 0.01313, which means that only about 1.3% of the variance of the logging value of Total Report Need could be explained by this dummy variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

v. Facility associated with a combined sewer system



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable CSS is 1.60763, which means that if a facility is associated with a combined sewer system, the total need for water infrastructure would be 1.60763% higher than the facilities aren't associated with a combined sewer system.

- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for this dummy variable is far less than 0.01 which is our choosing significant level. It means that if a facility is associated with a combined sewer system is

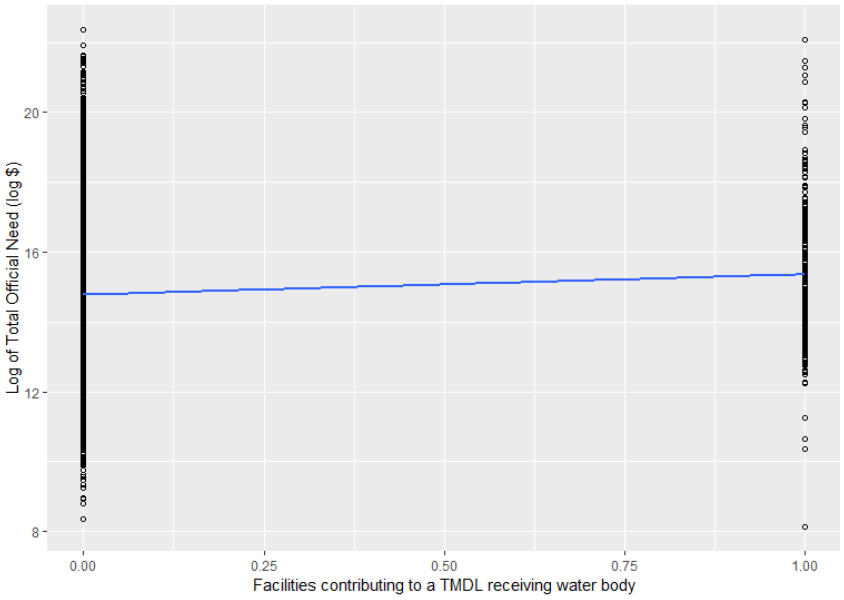
highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

It's worth mentioning that the adjusted R-squared value for this regression is 0.05951, which means that only about 6% of the variance of the logging value of Total Report Need could be explained by this dummy variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

vi. Facilities contributing to a TMDL receiving water body



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable TMDL_Dummy is 0.58474, which means that if a facility contributes to a TMDL receiving water body, the total need for water infrastructure would be about 0.6% higher than the facilities don't contributes to a TMDL receiving water body.

- How well is the model prediction

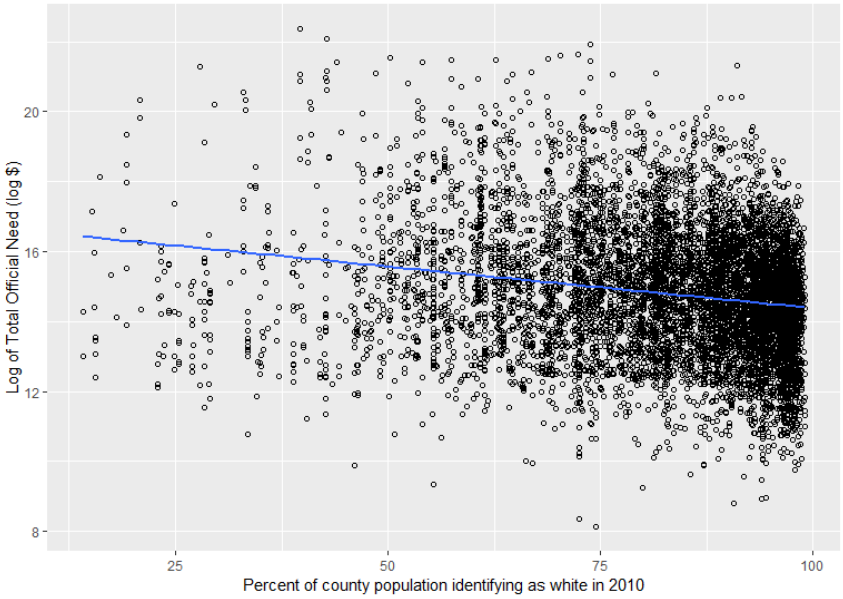
The result summary of this bivariate regression model reveals that the P value for this dummy variable is far less than 0.01 which is our choosing significant level. It means that if a facility contributes to a TMDL receiving water body is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

It's worth mentioning that the adjusted R-squared value for this regression is 0.003486, which means that only about 0.3% of the variance of the logging value of Total Report Need could be explained by this dummy variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

vii. Percent of county population identifying as white in 2010



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable PCTWHITE10 is -1378106, which means that for every additional one

unit increase in the percentage of county population identified as White in 2010, the total need for water infrastructure would decrease by $1,378,106.

- How well is the model prediction

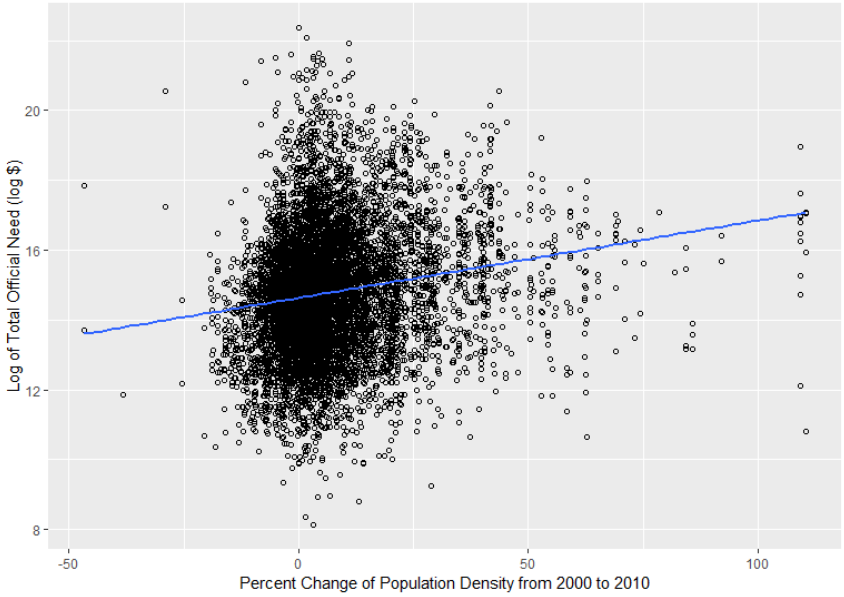
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the percentage of county population identified as White in 2010 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Total Reported Needs).

  It's worth mentioning that the adjusted R-squared value for this regression is 0.02353, which means that only about 2.4% of the variance of the logging value of Total Report Need could be explained by percentage of county population identified as White in 2010. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph also reveals that there are lots of outlier data and the linear relationship is not significant between the dependent and explanatory variables in this model.

viii. Percent Change of Population Density from 2000 to 2010



- Interpretation of coefficient of the variables

Based on the results of the estimates of coefficients, the coefficient of the variable PopDenPerChange is 0.022145, which means that for every additional one percentage unit increase in the change of Population Density from 2000 to 2010, the total need for water infrastructure would increase by 0.022145%.

- How well is the model prediction

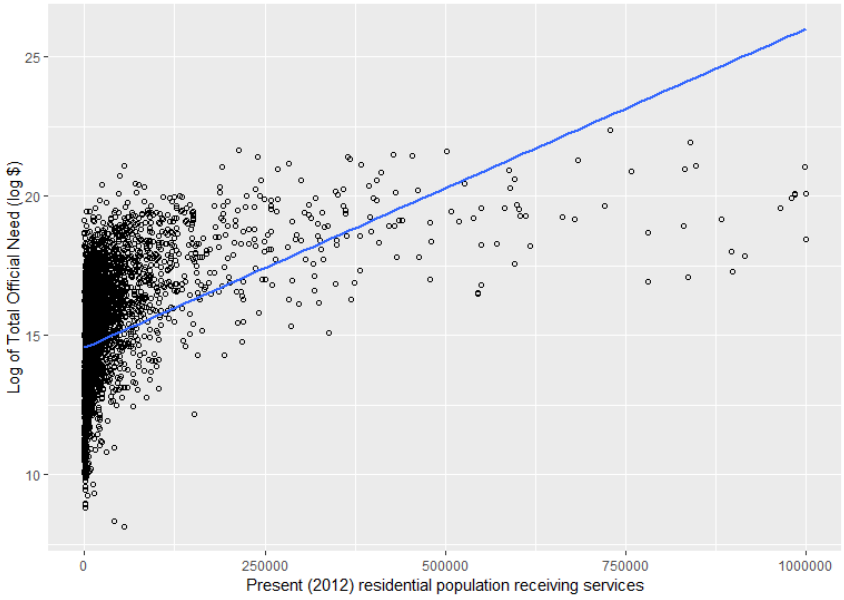
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the percentage change of Population Density from 2000 to 2010 is highly significantly correlated to the dependent value and it may not be included in the multivariate regression model.

  It's worth mentioning that the adjusted R-squared value for this regression is 0.02761, which means that only about 2.8% of the variance of the logging value of Total Report Need could be explained by this dummy variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well

- Scatterplot Characteristics

  The scatterplot graph reveals that there's no obviously linear relationship between the dependent and explanatory variables in this model.

ix. Present (2012) residential population receiving services



- Interpretation of coefficient of the variables

Based on the results of the estimates of coefficients, the coefficient of the variable max_RP is 6.268*10^-6, which means that for every additional one unit increase in the residential population who receiving services in 2012, the total need for water infrastructure would increase by 6.286*10^-6%.

- How well is the model prediction

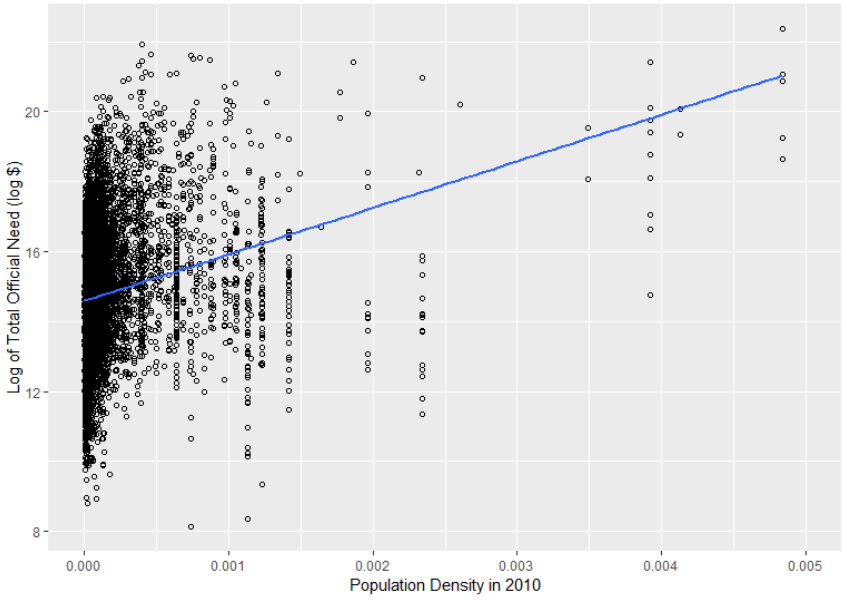
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the present (2012) residential population receiving services is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

  The adjusted R-squared value for this regression is 0.1361, which means that about 13.6% of the variance of the logging value of Total Report Need could be explained by the present (2012) residential population receiving services. So, to some extent, this is a good model.

- Scatterplot Characteristics

  The scatterplot graph reveals that the relationship between the dependent and explanatory variables in this model is similar to a linear relationship.

x. Population Density in 2010



- Interpretation of coefficient of the variables

Based on the results of the estimates of coefficients, the coefficient of the variable PopDen10 is 787.09235, which means that for every additional one unit increase from the population density in 2010, the total need for water infrastructure would increase by about 787%.

- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the population density in 2010 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

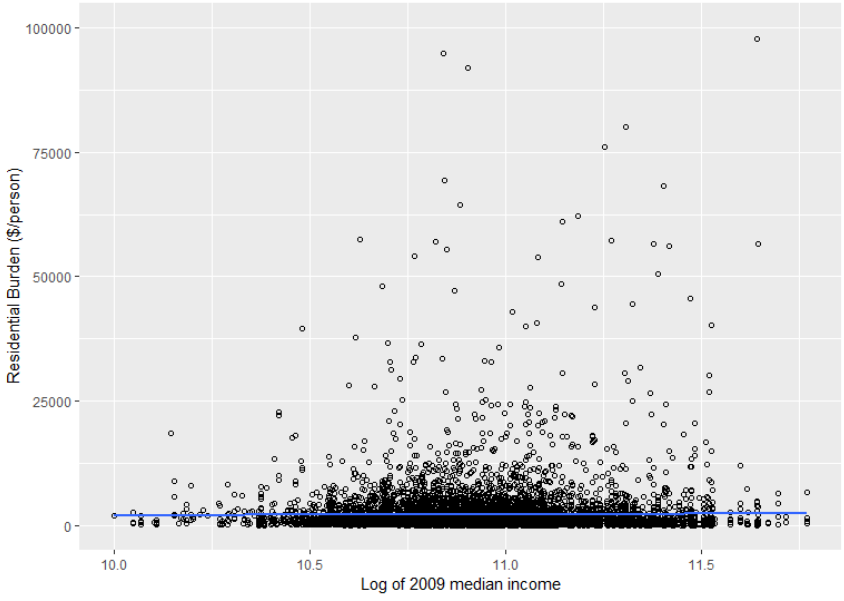
  The adjusted R-squared value for this regression is 0.05329, which means that about 5.3% of the variance of the logging value of Total Report Need could be explained by this variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph also reveals that there are lots of outlier data and the linear relationship is not significant between the dependent and explanatory variables in this model.

c. Residential Burden ($/person)

i. Logging value of Median income in 2009 in county



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable LogMedInc09 is 130117, which means that for every additional one percentage unit increase from the median income in 2009, the Residential Burden would increase by $130,117 per person.

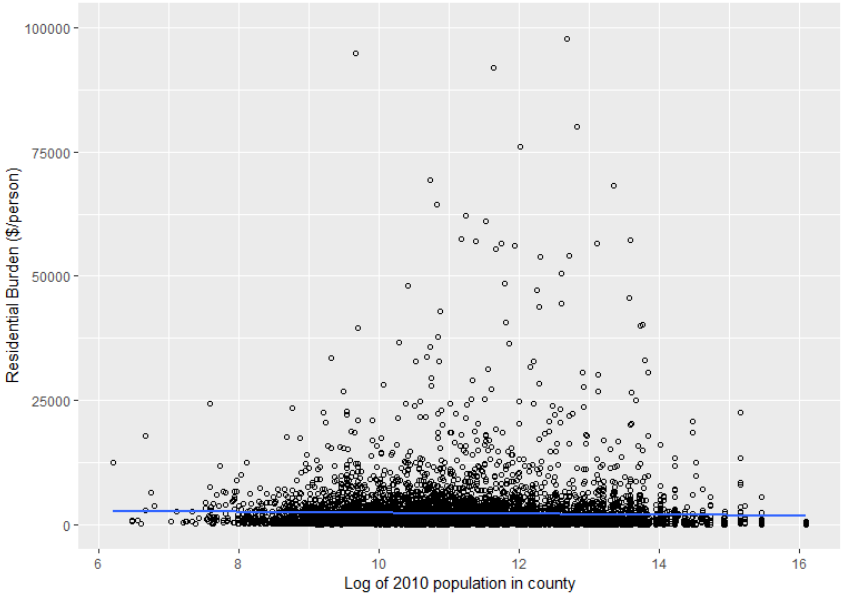
- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is 0.227, which is larger than 0.01 which is our choosing significant level. It means that the logging value of median income in 2009 is not correlated to the dependent value and it could not be included in the multivariate regression model to predict the variation of the dependent variable.

- Scatterplot Characteristics

  The scatterplot graph reveals that there's no linear relationship between the dependent and explanatory variables in this model.

ii. Logging value of 2010 population in county



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable LogPop10 is 48472, which means that for every additional one percentage unit increase from the county population in 2010, the Residential Burden would increase by about $48472 per person.

- How well is the model prediction

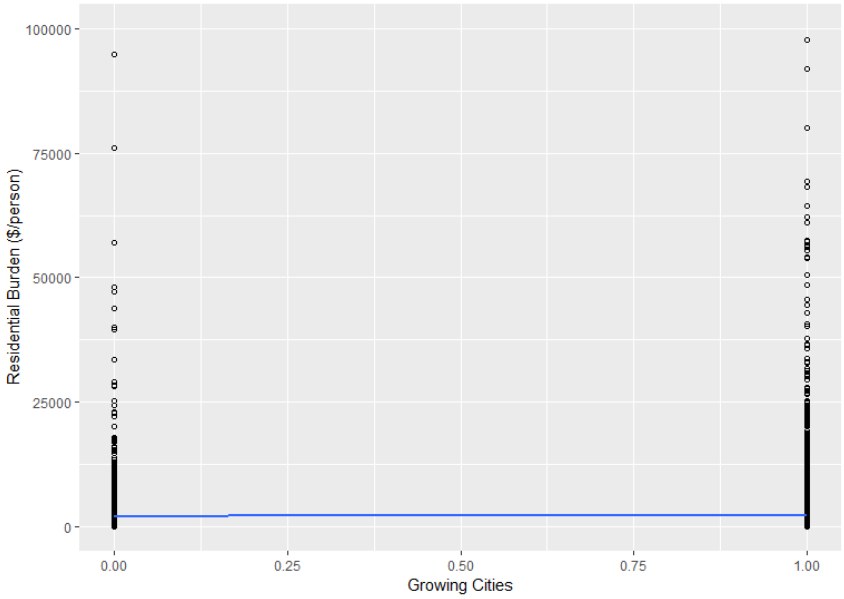
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is less than 0.01 which is our choosing significant level. It means that the logging value of county population in 2010 is correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

  It's worth mentioning that the adjusted R-squared value for this regression is 0.0008973, which is extremely small. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph reveals that there are lots of outlier data and the linear relationship is not significant between the dependent and explanatory variables in this model.

iii. Growing Cities



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable GrowingCities is 51935, which means that if the city's population has grown from 2000 to 2010, the Residential Burden would be $51935 per person higher than the cities whose population didn't grow.

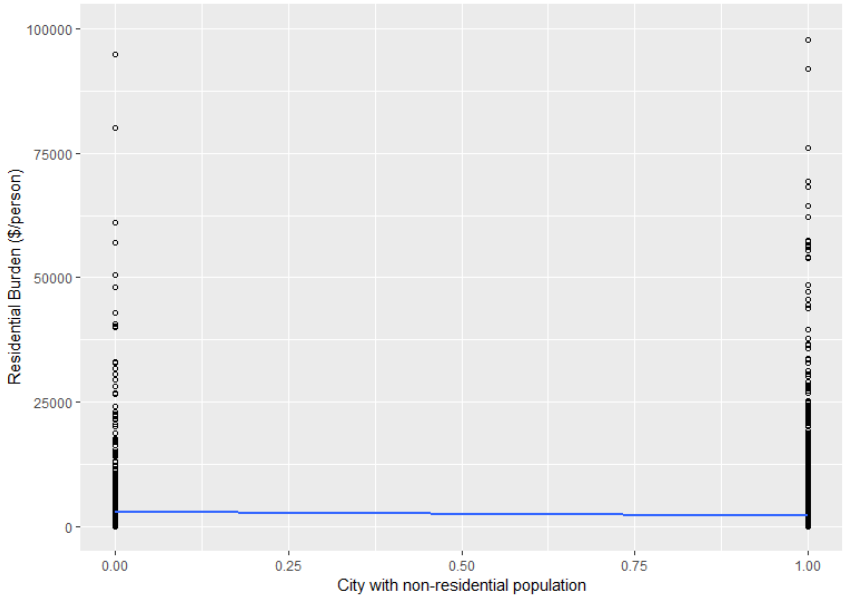
- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for this dummy variable is 0.379, which is higher than 0.01 which is our choosing significant level. It means that if the city is a population growing city from 2000 to 2010 is not correlated to the dependent value and it could not be included in the multivariate regression model to predict the variation of the dependent variable.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

iv. City with non-residential population



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable City_nR is -38958, which means that if a city has non-residential population, the Residential Burden would be about $38958 per person lower than the cities which only have residential population.

- How well is the model prediction

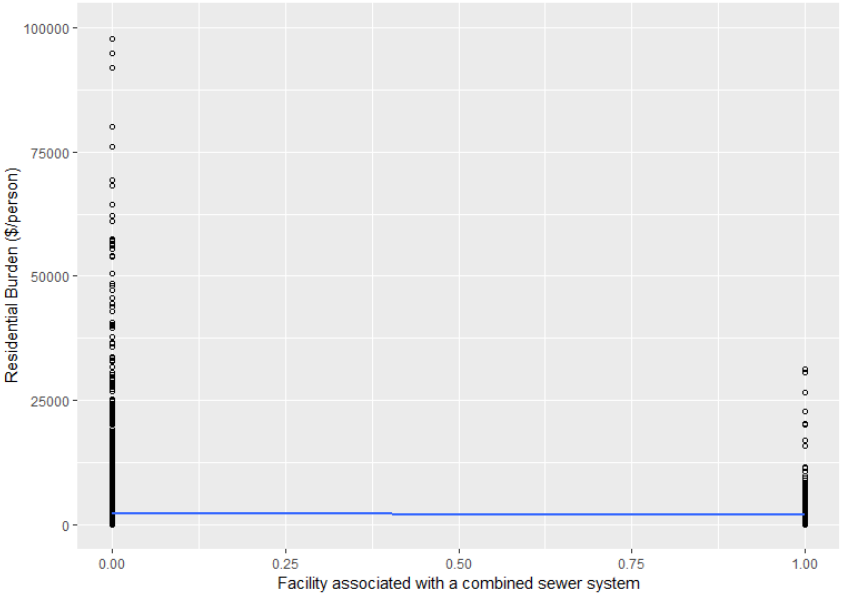
  The result summary of this bivariate regression model reveals that the P value for this dummy variable is 0.614, which is higher than 0.01 which is our

choosing significant level. It means that if the city has non-residential population is not correlated to the dependent value and it could not be included in the multivariate regression model to predict the variation of the dependent variable.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

v. Facility associated with a combined sewer system



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable CSS is 84643, which means that if a facility is associated with a combined sewer system, Residential Burden would be $84643 per person higher than the facilities aren't associated with a combined sewer system.

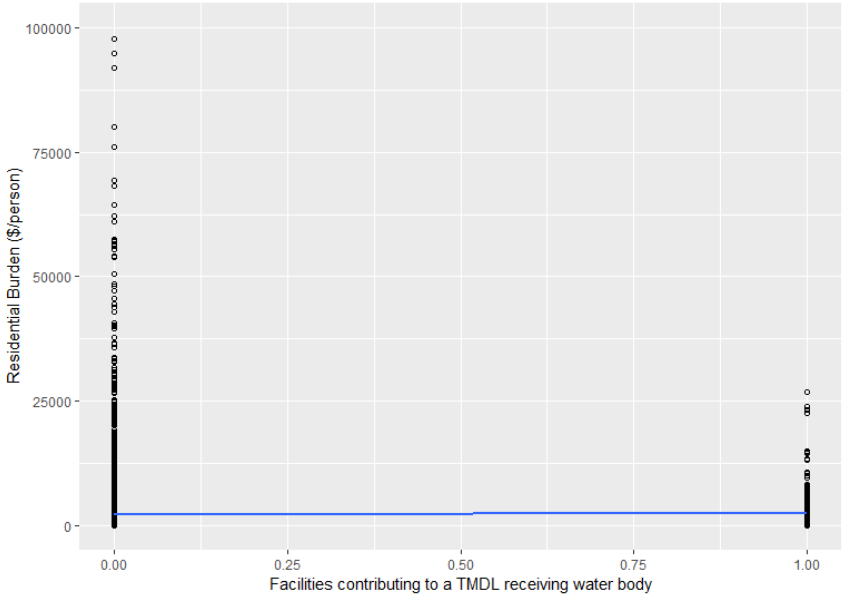
- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for this dummy variable is 0.377, which is higher than 0.01 which is our choosing significant level. It means that if a facility is associated with a combined sewer system is not correlated to the dependent value and it could not be included in the multivariate regression model to predict the variation of the dependent variable.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

vi. Facilities contributing to a TMDL receiving water body



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable TMDL_Dummy is -37831, which means that if a facility contributes to a TMDL receiving water body, the Residential Burden would be about $37831 per person higher than the facilities don't contributes to a TMDL receiving water body.

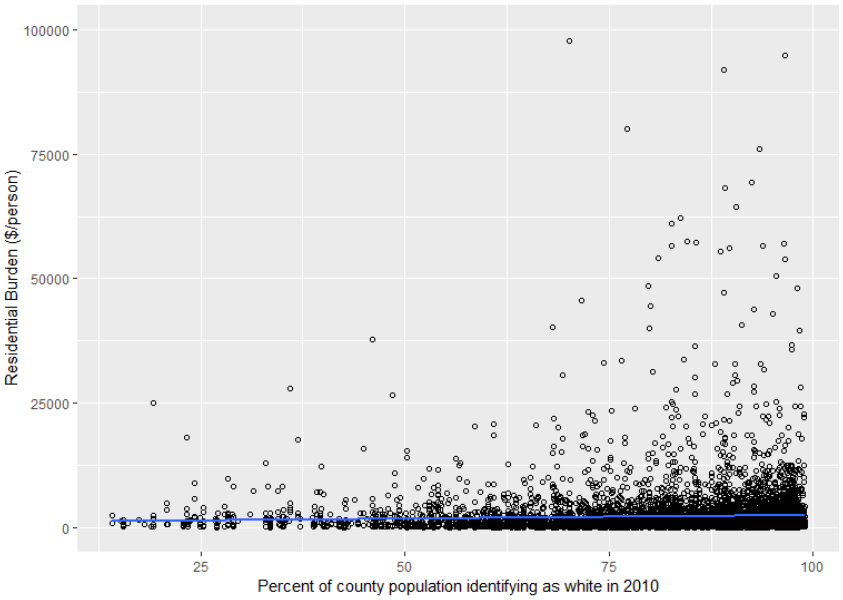
- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for this dummy variable is 0.79, which is far higher than 0.01 which is our choosing significant level. It means that if a facility contributes to a TMDL receiving water body is not correlated to the dependent value and it could not be included in the multivariate regression model to predict the variation of the dependent variable.

- Scatterplot Characteristics

Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

vii. Percent of county population identifying as white in 2010



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable PCTWHITE10 is -2836, which means that for every additional one unit increase in the percentage of county population identified as White in 2010, the Residential Burden would decrease by $2836 per person.

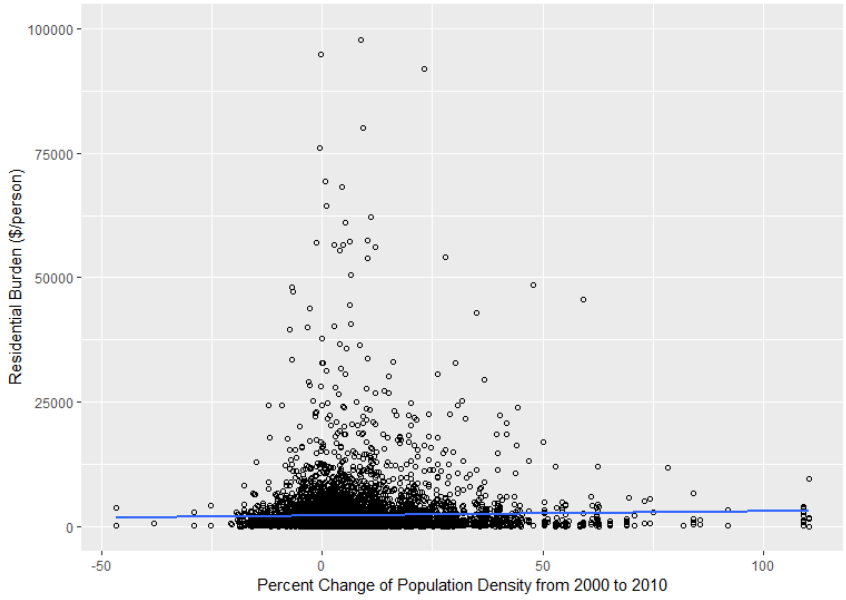
- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is 0.1069 which is higher than 0.01 which is our choosing significant level. It means that the percentage of county population identified as White in 2010 is not correlated to the dependent value and it could not be included in the multivariate regression model to predict the variation of the dependent variable.

- Scatterplot Characteristics

  The scatterplot graph also reveals that there are lots of outlier data and the linear relationship is not significant between the dependent and explanatory variables in this model.

viii. Percent Change of Population Density from 2000 to 2010



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable PopDenPerChange is 1949, which means that for every additional one percentage unit increase in the change of Population Density from 2000 to 2010, the Residential Burden would increase by $1949 per person.

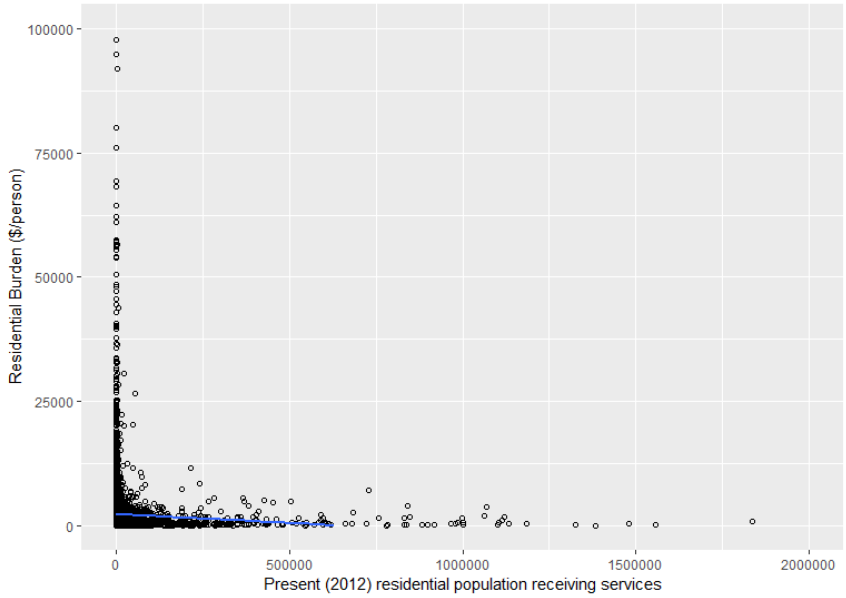
- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is 0.314, which is larger than 0.01 which is our choosing significant level. It means that the percentage change of Population Density from 2000 to 2010 is not correlated to the dependent value and it may not be included in the multivariate regression model.

- Scatterplot Characteristics

  The scatterplot graph also reveals that there are lots of outlier data and it seems that there's no linear relationship between the dependent and explanatory variables in this model.

ix. Present (2012) residential population receiving services



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable max_RP is -7.827*10^-2, which means that for every additional one unit increase in the residential population who receiving services in 2012, the Residential Burden would increase by -$7.827*10^-2 per person, which doesn't make any sense.

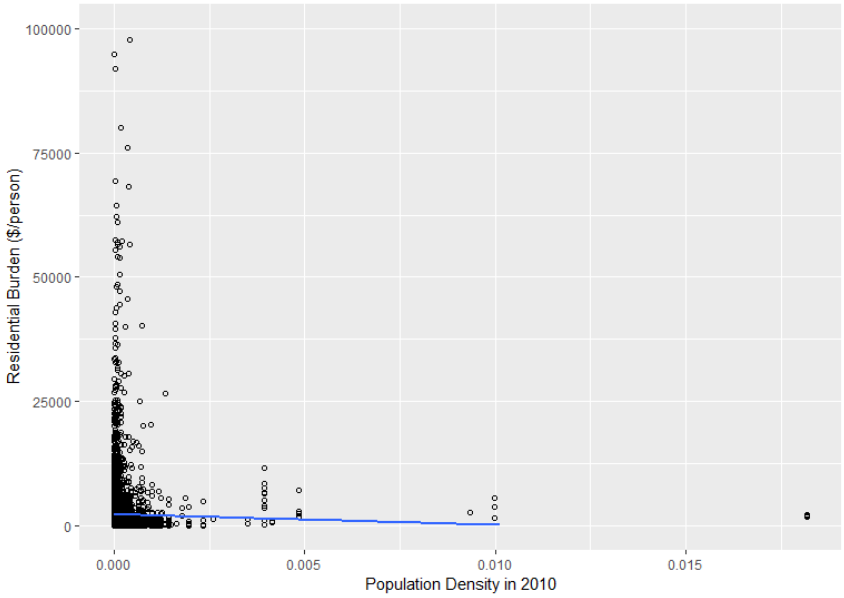
- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is 0.752, which is higher than 0.01 which is our choosing significant level. It means that the present (2012) residential population receiving services is highly significantly correlated to the dependent value and it could not be included in the multivariate regression model to predict the variation of the dependent variable.

- Scatterplot Characteristics

  The scatterplot graph also reveals that there are lots of outlier data and it seems that there's no linear relationship between the dependent and explanatory variables in this model.

x. Population Density in 2010



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable PopDen10 is 64569835, which means that for every additional one unit increase from the population density in 2010, the Residential Burden would increase by $64569835 per person.

- How well is the model prediction

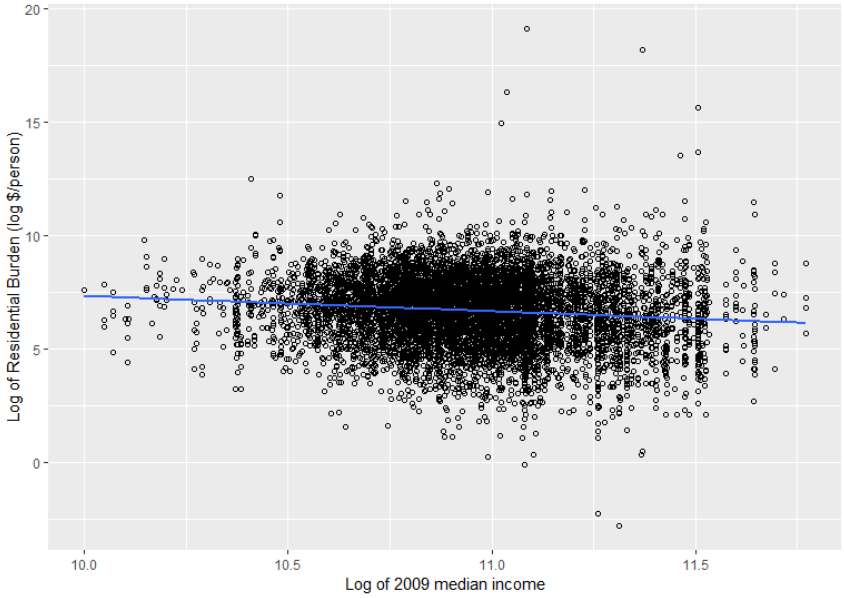
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is 0.192, which is higher than 0.01 which is our choosing significant level. It means that the population density in 2010 is highly significantly correlated to the dependent value and it could not be included in the multivariate regression model to predict the variation of the dependent variable.

- Scatterplot Characteristics

  The scatterplot graph also reveals that there are lots of outlier data and it seems that there's no linear relationship between the dependent and explanatory variables in this model.

d. Log of Residential Burden (log $/person)

i. Logging value of Median income in 2009 in county



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable LogMedInc09 is -0.6718, which means that for every additional one percentage unit increase from the median income in 2009, the Residential Burden would decrease by 0.6718%.

- How well is the model prediction

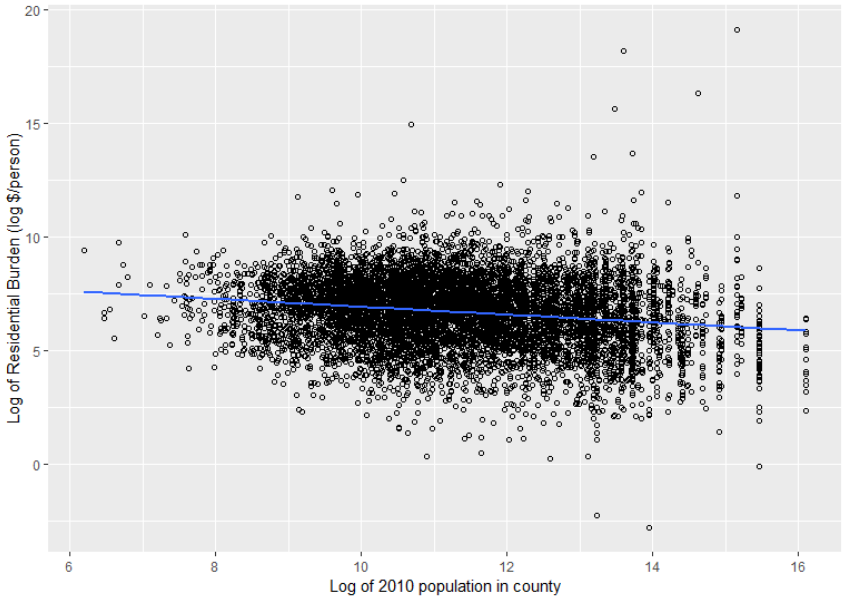
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the logging value of median income in 2009 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable (Logging values of Total Reported Needs).

  It's worth mentioning that the adjusted R-squared value for this regression is 0.01147, which means that only about 1.15% of the variance of the Logging value of Residential Burden could be explained by the logging value of the median income in 2009. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph also reveals that the relationship between dependent and explanatory variables is close to a linear relationship.

ii. Logging value of 2010 population in county



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable LogPop10 is -0.1728, which means that for every additional one percentage unit increase from the county population in 2010, the Residential Burden would decrease by about 0.1728%.

- How well is the model prediction

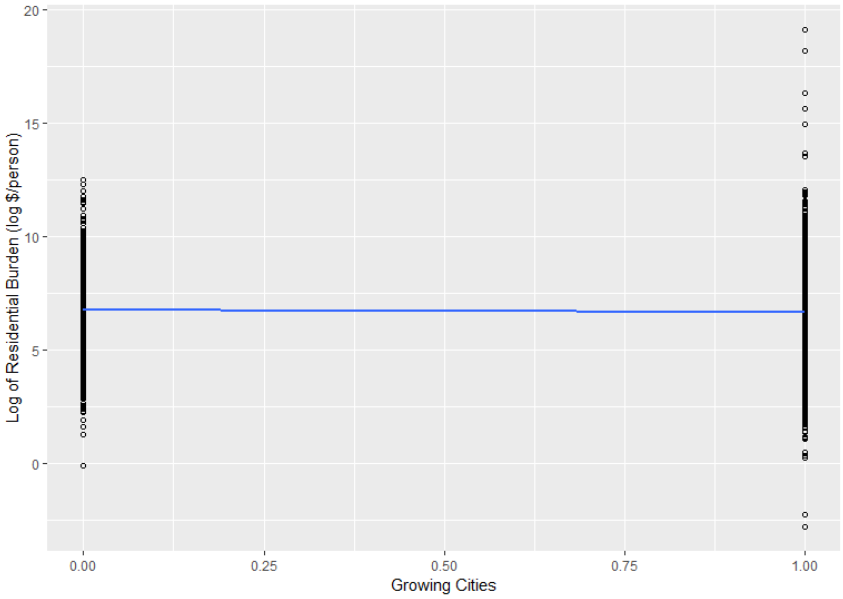
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the logging value of county population in 2010 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

  It's worth mentioning that the adjusted R-squared value for this regression is 0.03163, which means that only about 3.2% of the variance of the logging value of Residential Burden could be explained by the logging value of the population in 2010. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph reveals that the relationship between the dependent and explanatory variables in this model is similar to a linear relationship.

iii. Growing Cities



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable GrowingCities is -0.10095, which means that if the city's population has grown from 2000 to 2010, the Residential Burden would be about 0.1% higher than the cities whose population didn't grow.

- How well is the model prediction

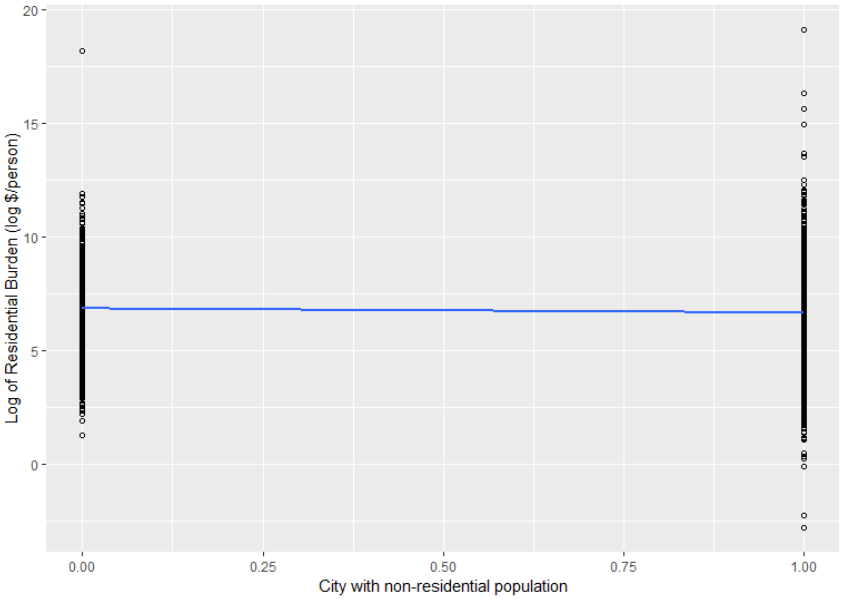
  The result summary of this bivariate regression model reveals that the P value for this dummy variable is less than 0.01 which is our choosing significant level. It means that if the city is a population growing city from 2000 to 2010 is correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

  It's worth mentioning that the adjusted R-squared value for this regression is 0.0007488, which means that only about 0.07% of the variance of the logging value of Residential Burden could be explained by this dummy variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

iv. City with non-residential population



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable City_nR is -0.18321, which means that if a city has non-residential population, the Residential Burden would be about 0.2% lower than the cities which only have residential population.

- How well is the model prediction

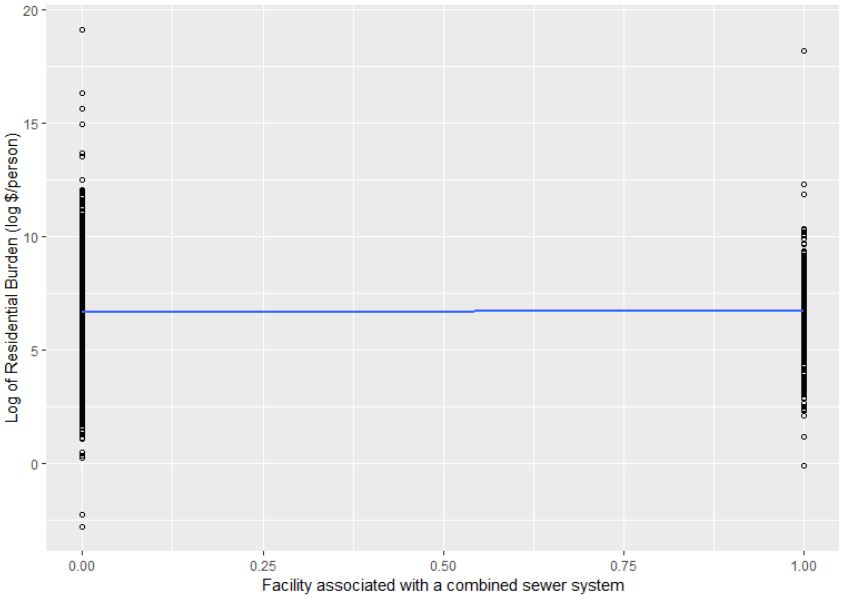
  The result summary of this bivariate regression model reveals that the P value for this dummy variable is far less than 0.01 which is our choosing significant level. It means that if the city has non-residential population is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

  It's worth mentioning that the adjusted R-squared value for this regression is 0.001557, which means that only about 0.16% of the variance of the logging of Residential Burden could be explained by this dummy variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

v. Facility associated with a combined sewer system



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable CSS is 0.006638, which means that if a facility is associated with a combined sewer system, the Residential Burden would be 0.006638% higher than the facilities aren't associated with a combined sewer system.

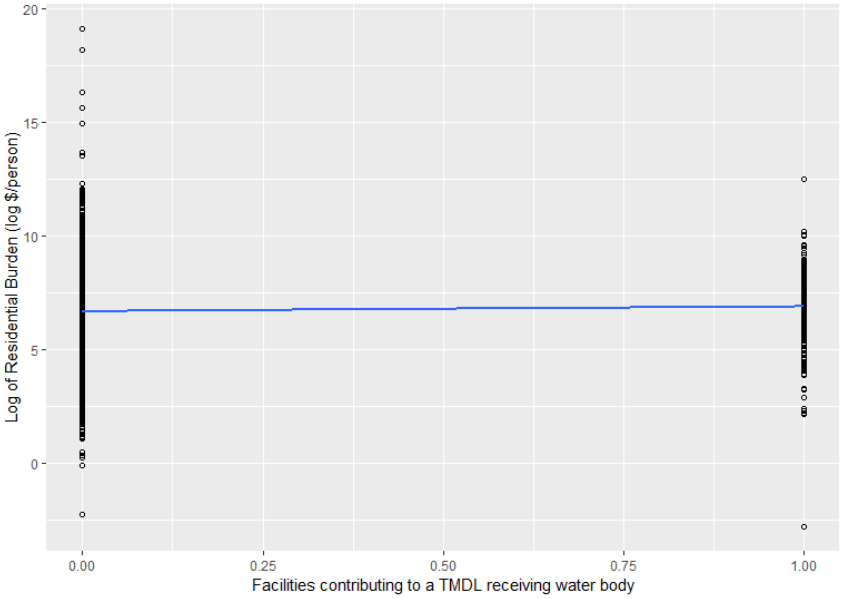
- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for this dummy variable is 0.914, which is far higher than 0.01 which is our choosing significant level. It means that if a facility is associated with a combined sewer system is not correlated to the dependent value and it could not be included in the multivariate regression model to predict the variation of the dependent variable.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

vi. Facilities contributing to a TMDL receiving water body



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable TMDL_Dummy is 0.2113, which means that if a facility contributes to a TMDL receiving water body, Residential Burden would be about 0.2113% higher than the facilities don't contributes to a TMDL receiving water body.

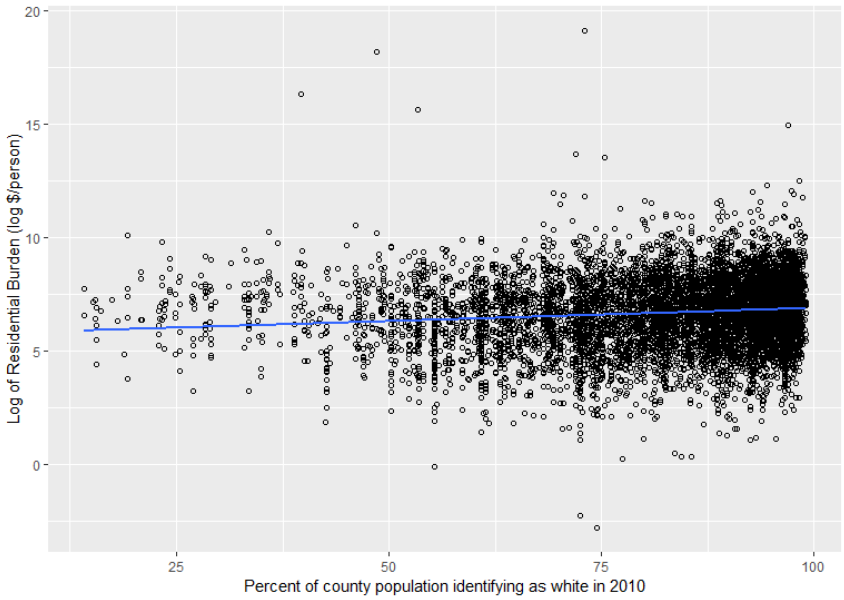
- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for this dummy variable is just about 0.01 which is our choosing significant level. It means that if a facility contributes to a TMDL receiving water body is just slightly correlated to the dependent value and it may not be included in the multivariate regression model to predict the variation of the dependent variable.

- Scatterplot Characteristics

  Since a dummy variable only have numeric values in 0 and 1, the scatterplot graph only reveals 2 vertical lines along x=0 and x=1. So, it's not useful to show the relationship between the dependent and the dummy variable.

vii. Percent of county population identifying as white in 2010



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable PCTWHITE10 0.011479, which means that for every additional one unit increase in the percentage of county population identified as White in 2010, the Residential Burden would increase by 0.011479%.

- How well is the model prediction

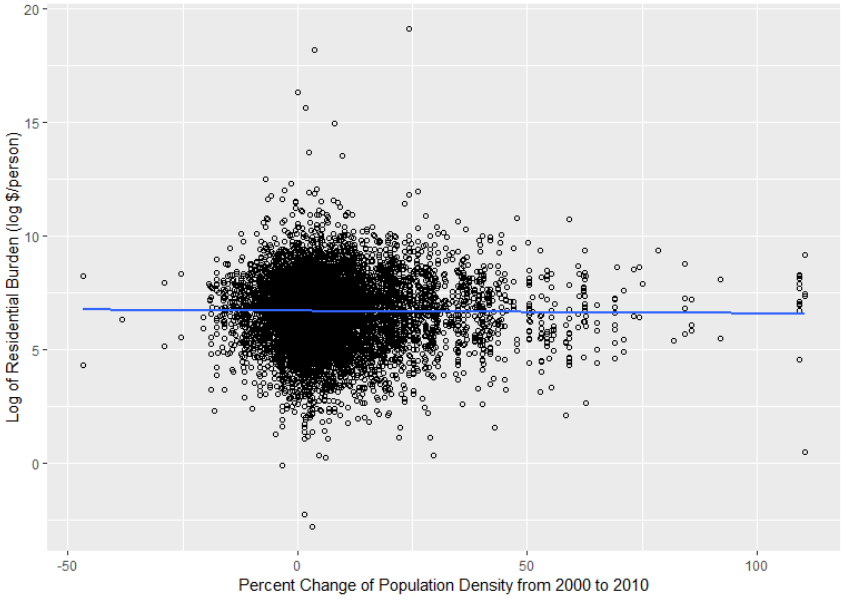
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the percentage of county population identified as White in 2010 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

  It's worth mentioning that the adjusted R-squared value for this regression is 0.01256, which means that only about 1.26% of the variance of the logging value of Residential Burden could be explained by percentage of county population identified as White in 2010. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph also reveals that the relationship between the dependent and explanatory variables in this model is similar to a linear relationship.

viii. Percent Change of Population Density from 2000 to 2010



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable PopDenPerChange is -0.001089, which means that for every additional one percentage unit increase in the change of Population Density from 2000 to 2010, the Residential Burden would decrease by 0.001089%.

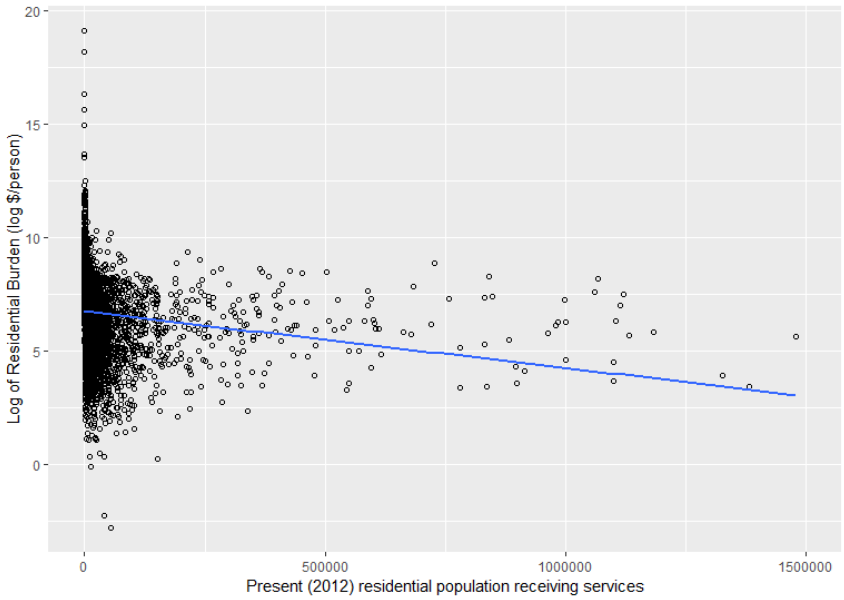
- How well is the model prediction

  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is 0.379, which is higher than 0.01 which is our choosing significant level. It means that the percentage change of Population Density from 2000 to 2010 is not correlated to the dependent value and it may not be included in the multivariate regression model.

- Scatterplot Characteristics

  The scatterplot graph reveals that there's no obviously linear relationship between the dependent and explanatory variables in this model.

ix. Present (2012) residential population receiving services



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable max_RP is -1.62*10^-6, which means that for every additional one unit increase in the residential population who receiving services in 2012, the Residential Burden would decrease by 1.62*10^-6%.

- How well is the model prediction

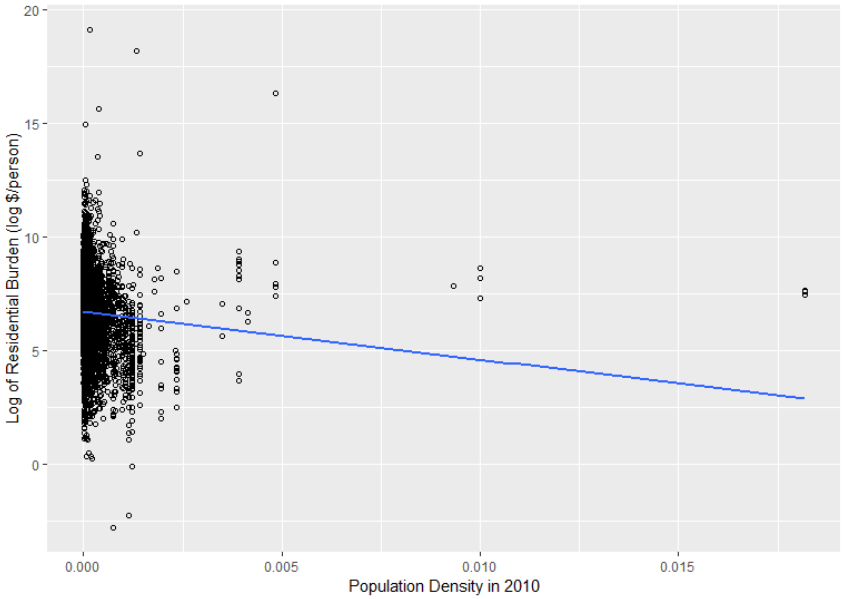
  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the present (2012) residential population receiving services is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

  The adjusted R-squared value for this regression is 0.01265, which means that about only 1.27% of the variance of the logging value of Residential Burden could be explained by the present (2012) residential population receiving services. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph reveals that there's no obviously linear relationship between the dependent and explanatory variables in this model.

x. Population Density in 2010



- Interpretation of coefficient of the variables

  Based on the results of the estimates of coefficients, the coefficient of the variable PopDen10 is -211.66061, which means that for every additional one unit increase from the population density in 2010, the Residential Burden would decrease by about 211.7%.

- How well predict the model

  The result summary of this bivariate regression model reveals that the P value for the explanatory variable is far less than 0.01 which is our choosing significant level. It means that the population density in 2010 is highly significantly correlated to the dependent value and it could be included in the multivariate regression model to predict the variation of the dependent variable.

  The adjusted R-squared value for this regression is 0.005317, which means that about 0.5% of the variance of the logging value of Residential Burden could be explained by this variable. However, although the R squared value is so small, it is reasonable because including only one explanatory variable definitely could not predict the model well.

- Scatterplot Characteristics

  The scatterplot graph also reveals that there are lots of outlier data and the linear relationship is not significant between the dependent and explanatory variables in this model.

Part 3 Multivariate Regressions

- 3.1 Analytical Method and Logic
  - Building up null hypothesis and alternative hypothesis
    In this section, a pair of null hypothesis and alternative hypothesis will be made for each dependent variable. The null hypothesis is usually an assumption to put no regression function between dependent variables and other factors. In another word, in null hypothesis, dependent variable would remain a constant number under the influence of a group of possible independent variables. The alternative hypothesis is an assumption to build up regression between dependent variable and possible independent variables. By doing F test and comparing F value, we could decide if we reject the null hypothesis and accept the alternative hypothesis.
  - Forward and Backward Selection
    - ❖ Forward Selection
      With forward selection, the model starts when there is no predictors (null hypothesis) and gradually adds the most contributive ones. When the improvement to the model is no longer significant, the adding process will stop.
    - ❖ Backward Selection
      With backward selection, the model starts with all predictors in the model and gradually removes the least contributive ones. When all the predictors in the model are statistically significant, the removal would stop.
    - ❖ Stepwise Selection
      Stepwise selection is a combination of forward and backward selections. Starting without predictors (null hypothesis), the most contributive predictors would be added into the model as in the forward selection. After that, any variables that no longer improve the model would be removed.
    - ❖ In this Analysis
      According to the analysis in previous parts of this report, it is clear that the analysis is based on a high-dimensional data which contains multiple predictor variables. In order to choose a model which can predict the water infrastructure need accurately when keep its simplicity, **the stepwise regression** is the best solution.
  - Linear Regression Model Build Up
    Regression model usually requires numeric inputs rather than categorical data. However, in this dataset of water infrastructure, a lot of data is categorical data such as CSS (whether the infrastructure is in a combined sewer system), Growing Cities (whether the population in a city is growing),

City nR (whether city is with residential population), EPA regions(10 regions in total), etc. To use these variables in regression model, they need to be transformed into binary variables through dummy coding. In the analysis, each dual-categorical variable has been processed into a binary variable with the value of 0 or 1. The variable about EPA region has been made into ten different dummy variables. One thing to note here is, before stepwise selection, 10 of the EPA regions are all added into the start-up models. After the selection in R, only several of them would be left in the refined models. In order to clarify the interpretation of these region variables, EPA Region10 is set as a standard comparison between those regions. And all other 9 regions have been added to the final models for clarifying the comparison with EPA10.
The regression model we are going to build here is multiple linear regression following the equation below:

```
y = b0 + b1*x1 + b2*x2 + b3*x3 + …
```

While x represents different independent variables, b represents the coefficient for each independent variable.
The model building process is separated into two parts, the first part of model building uses every original forms of numeric independent variables while the second part tests their log forms. The reason for separating them is to better explain the model theoretically through either unit change in predictors or the percent change in predictors. After stepwise selection, our refined model could be judged by adjusted r square. Choosing the model with a larger adjusted r square is our principle here.

- Correlation Tests to Avoid Multicollinearity
  Before getting the final regression model for each dependent variable, some correlation tests are necessary to avoid collinearity between variables and to improve the model's accuracy and simplicity. Since an overall correlation test has been conducted in previous part, the value of the cor-coefficients could help with this process.
  After removing some independent variables which are highly correlated with other variable while contribute less to the model, the final model for each dependent variable is produced.
- Model Interpretation
  - ❖ Check on F-test
    To interpret the model, the first step is to analyze the results of F-test and the associated p-value. When the **p-value** of an independent variable is smaller than **0.01**, this variable is significantly related to the dependent variable. Besides, if the f-value of the model is large associated with a small p-value, we could reject null hypothesis and accept the alternative hypothesis, which is our final model.
- Model Accuracy Assessment

- ❖ Check on R Square
  In the multiple linear regression, R square is an important measure of how accuracy the model can be in predicting dependent variable. R square represents the correlation coefficient between the observed values of the dependent variable and the predicted values of dependent variable. Since R square reveals the proportion of variance in the dependent variable that could be predicted by the value of independent variables, a value of R-sq close to 1 indicates the higher accuracy of a model in predicting dependent variable.
  Since any more variables adding to the model could result in an increased R-sq, adjusted R-sq is used in this analysis instead by considering the number of independent variables.
- ❖ Check on Residual Standard Error (RSE)
  In multiple linear regression, residual represents the difference between predicted(fitted) values of dependent variables and the observed values of dependent variables. Thus, the RSE could offer a measure of error of prediction. The model is more accurate when it has a lower RSE.
- ❖ Check on VIF
  The variance inflation factor (VIF) could be used to measure the multicollinearity. So it is used here to test the multicollinearity between the variables. The square root of the VIF represents how much times the standard error of the predictor's coefficient is as large as it should be when there is no correlation between this predictor and the other ones. In the analysis, VIF value from 1 to 10 is considered the indication for low multicollinearity.

- 3.2.1 Model 1: Regression Model for Total Official Need
  - Variables in the Final Model:

	Variables	Coefficient	Significance (p-value)
Numeric Variables	The residential population receiving water service	5.202e+02	<2e-16 ***
	The non-residential population receiving water service	1.448e+03	<2e-16 ***
	Population density in 2010	7.045e+10	<2e-16 ***
	Median income in 2009	-1.618e+02	0.0978 .
	% White population in 2000	-1.883e+05	0.0663 .
Dummy Variables	CSS (Whether the infrastructure is in a combined sewer system)	4.624e+07	<2e-16 ***

	EPA Region 1 (whether the infrastructure is in EPA1)	6.435e+06	0.4491
	EPA Region 2	-1.179e+07	0.1370
	EPA Region 3	-1.014e+07	0.1565
	EPA Region 4	3.895e+06	0.5803
	EPA Region 5	-6.711e+06	0.3253
	EPA Region 6	-4.624e+06	0.5213
	EPA Region 7	7.886e+06	0.2695
	EPA Region 8	-1.977e+06	0.8005
	EPA Region 9	-6.185e+06	0.4994
	Whether the city is growing in population	6.435e+06	0.0253 *
	Whether the city has larger non-residential population have water needs than residential population	-6.483e+06	0.0635 .

Judged by the p-value in the table, for the final model to total official need, **significant variables** are: the residential population receiving water service, the non-residential population receiving water service, population density in 2010, CSS (Whether the infrastructure is in a combined sewer system) and whether the city is growing in population.

- Model Interpretation

  When **the number of residential population receiving water service** increase by 1, the total official need for water infrastructure would increase 5.202e+02. This result indicates the water need is a necessity for residential population, especially those who has been receiving service (either collection or treatment) from the water infrastructure.

  When **the number of non-residential population receiving water service** increase by 1, the total official need for water infrastructure would increase 1.448e+03. Theoretically, non-residential population also has a large need on water. So the increasing of non-residential population receiving water service would lead to the increase of total official need.

  When **the population density (persons/ square m) in 2010** increases by 1 person/ square m, the total official need for water infrastructure would increase 7.045e+10. It is easy to understand that when there is a higher population density in a service area of a water infrastructure, there would be larger need for the infrastructure.

  When **the infrastructure is in a combined sewer system**, the total official need for water infrastructure would increase 4.624e+07. If it is not in a combined sewer system, the total official need will not change. For the service level of

water infrastructure, it is common that those in a combined sewer system have a higher ability to offer more service. Thus, those infrastructures have higher need.

When **the city is growing in its population**, the total official need for water infrastructure would increase 6.435e+06. If the population is stable or declining, the total official need would not increase. For a city with larger population between years, the need for water infrastructures would increase because of the additional populations.

The rest variables are insignificant. Their interpretations related to the model could be summarized as follows while the specific coefficients listed in the table are not mentioned again: when the median income in 2009 increases, or when the percentage of white population in 2000 increased, the total official need for water infrastructure would drop. For median income predictor, a possible explanation is when people get richer, they have other ways to access water service such as using other sources of drinkable water. For percent white predictor, it is possible that they also have other habit of using water. For infrastructures in EPA Regions 1, 4, 7, their total official need would be larger compared with EPA region 10. For the rest of the EPA Regions, their needs would be smaller than infrastructures in EPA Region 10. This difference is caused by different water needs and service levels across the States. If a city has a larger need from non-residential population, the total official need would drop. This is probably because the need from per unit non-residential population is smaller than that from per unit residential population.

- Model Accuracy Assessment

Model 1	Adjusted R-sq	F-statistics	p-value	RSE
	0.4062	313.9	< 2.2e-16	104700000

According to the table above, the adjusted R-sq indicates that the model could predict the total official need well with the current predictors. The value of 0.4062 suggests that 40.62% of the total official need could be explained through the current predictors.

The F-statistics as well as p-value both indicates the null hypothesis could be rejected. Because the f-value is large while the p-value is very small. The small p-value suggests that the predictors used in the model is highly significant to the model while large f-value indicates those predictors contribute a lot of information to the model. So we could reject the null hypothesis in which there is no relationship between predictors and the total official need.

The RSE value is relatively not small. But considering the water infrastructure need is originally in large numbers, the large RSE is also acceptable.

Variables	**F-value**	**VIF**

The residential population receiving water service	1622.9539	1.448983
The non-residential population receiving water service	261.2139	1.382964
Population density in 2010	661.1780	1.550284
Median income in 2009	2.7409	1.624598
% White population in 2000	3.3737	1.680106
CSS (Whether the infrastructure is in a combined sewer system)	108.0443	1.145245
EPA Region 1 (whether the infrastructure is in EPA1)	0.5730	2.267919
EPA Region 2	2.2119	3.051936
EPA Region 3	2.0084	4.175393
EPA Region 4	0.3058	5.053839
EPA Region 5	0.9676	5.786003
EPA Region 6	0.4113	4.348190
EPA Region 7	1.2193	4.345722
EPA Region 8	0.0639	2.664286
EPA Region 9	0.4562	2.252037
Whether the city is growing in population	5.0042	1.261551
Whether the city has larger non-residential population have water needs than residential population	3.4447	1.089412

According to the table above, larger f-value is related with higher significance level in the previous table. Variables with higher significance to total official need receive larger f-value. The VIF from the model for each variable indicate the low correlations between variables which further provide evidence for low multicollinearity in model 1.

- 3.2.2 Model 2: Regression Model for log of Total Official Need
  - Variables in the Final Model:

	Variables	Coefficient	Significance (p-value)
Numeric Variables	The residential population receiving water service	4.181e-06	< 2e-16 ***
	Population Change from 1980 to 2010	9.550e-07	<2e-16 ***

	Population in 2000	-1.716e-07	2.62e-06 ***
	Population Density in 2000	2.231e+02	1.27e-08 ***
	% White population in 2010	-7.494e-03	2.64e-07 ***
	Median Income in 2009	1.619e-05	< 2e-16 ***
Dummy Variables	CSS (Whether the infrastructure is in a combined sewer system)	1.390e+00	<2e-16 ***
	EPA Region 1 (whether the infrastructure is in EPA1)	5.798e-02	0.650578
	EPA Region 2	-4.784e-01	6.51e-05 ***
	EPA Region 3	7.422e-03	0.945465
	EPA Region 4	1.579e-01	0.139834
	EPA Region 5	-5.099e-01	6.77e-07 ***
	EPA Region 6	-3.866e-01	0.000359 ***
	EPA Region 7	-8.846e-01	3.84e-16 ***
	EPA Region 8	-6.476e-01	4.51e-08 ***
	EPA Region 9	4.108e-01	0.004048 **
	Whether the infrastructure is connected to TMDL	4.051e-01	2.38e-05 ***
	Whether the city is growing in population	1.443e-01	0.000885 ***
	Whether the city has larger non-residential population have water needs than residential population	-3.021e-01	3.14e-09 ***

Judged by the p-value in the table, for the final model to total official need, **significant variables** are: the residential population receiving water service, population Change from 1980 to 2010, population in 2000, population density in 2000, % white population in 2010, median Income in 2009, CSS (Whether the infrastructure is in a combined sewer system), whether the infrastructure is in EPA Region 2, 5,6,7,8,9, whether the infrastructure is connected to TMDL, whether the city is growing in population and whether the city has larger non-residential population have water needs than residential population.

- Model Interpretation
  When **the number of residential population receiving water service** increase by 1, the total official need for water infrastructure would increase by 4.181e-06 in its percentage. This result indicates the water need is a necessity for

residential population, especially those who has been receiving service (either collection or treatment) from the water infrastructure.

When **the population change from 1980 to 2010** increase by 1, the total official need for water infrastructure would increase by 9.550e-07 in its percentage. While a city is gaining population, it is natural that the need for water infrastructure would increase.

When **population in 2000** increase by 1, the total official need for water infrastructure would decrease by 1.716e-07 in its percentage. This may seem a bit strange since normally the need would increase as population is increasing. Since the coefficient is very small, we could think as this is just some match in statistical level instead of theoretical level.

When **the population density (persons/ square m) in 2000** increases by 1 person/ square m, the total official need for water infrastructure would increase 2.231e+02 in its percentage. Although the dependent variable has been processed into logistic form, the effect of population density is still strong. It is easy to understand that when there is a higher population density in a service area of a water infrastructure, there would be larger need for the infrastructure.

When **the white population percentage in 2010** increased by 1, the total official need for water infrastructure would decrease by 7.494e-03 in its percentage. Probably due to different lifestyle or household organization compared to other races, this predictor affect the need in a negative way.

When **the median income in 2009** increases by 1, the total official need increase by 1.619e-05 in its percentage. The income increase might result in more usage of water in daily life since people will not consider about the costs as they did before.

When **the infrastructure is in a combined sewer system**, the total official need for water infrastructure would increase by 1.390e+00 in its percentage. If it is not in a combined sewer system, the total official need will not change. For the service level of water infrastructure, it is common that those in a combined sewer system have a higher ability to offer more service. Thus, those infrastructures have higher need.

When **the infrastructure is in EPA region 9**, the total need would increase by 4.108e-01 in its percentage compared to those in EPA region 10. While the infrastructure is EPA region 2, 5, 6 7 and 8, the total need would decrease by 4.784e-01, 5.099e-01, 3.866e-01, 8.846e-01 and 6.476e-01 in its percentage. This difference is caused by different water needs and service levels across the States. Compared to model 1, it is clear that the significance of the EPA region has improved a lot. This is because, EPA region has a greater effect on percent change of need. Since different regions differ in the need scale, the percent change could reflect the impact better.

When **the infrastructure is connected to TMDL**, the total official need increases by 4.051e-01 in its percentage.
When **the city is growing in its population**, the total official need for water infrastructure would increase 1.443e-01 in its percentage. If the population is stable or declining, the total official need would not increase. For a city with larger population between years, the need for water infrastructures would increase because of the additional populations.
When **the city has larger non-residential population having water needs,** the total need decreases by 3.021e-01 in its percentage. This is probably because the need from per unit non-residential population is smaller than that from per unit residential population.

- Model Accuracy Assessment

Model 2	Adjusted R-sq	F-statistics	p-value	RSE
	0.2809	170.3	< 2.2e-16	1.549

According to the table above, the adjusted R-sq indicates that the model could predict the total official need well with the current predictors. The value of 0.2809 suggests that 28.09% of the log of total official need could be explained by current predictors.
The F-statistics as well as p-value both indicates the null hypothesis could be rejected. Because the f-value is large while the p-value is very small. The small p-value suggests that the predictors used in the model is highly significant to the model while large f-value indicates those predictors contribute a lot of information to the model. So we could reject the null hypothesis in which there is no relationship between predictors and the log of total official need.
The RSE value is much smaller than that in model 1 which indicates the model's accuracy.

Variables	**F-value**	**VIF**
The residential population receiving water service	538.4323	1.288892
Population Change from 1980 to 2010	76.6262	2.318669
Population in 2000	22.1075	2.216993
Population Density in 2000	32.4422	1.447298
% White population in 2010	26.5443	1.660948
Median Income in 2009	125.2362	1.625510
CSS (Whether the infrastructure is in a combined sewer system)	442.4279	1.154746

EPA Region 1 (whether the infrastructure is in EPA1)	0.2052	3.183426
EPA Region 2	15.9646	4.385474
EPA Region 3	0.0047	5.322556
EPA Region 4	2.1802	5.786003
EPA Region 5	24.7183	5.974152
EPA Region 6	12.7470	4.481839
EPA Region 7	66.5876	4.574581
EPA Region 8	29.9749	2.782519
EPA Region 9	8.2666	2.505153
Whether the infrastructure is connected to TMDL	17.8819	1.108465
Whether the city is growing in population	11.0620	1.311298
Whether the city has larger non-residential population have water needs than residential population	35.1694	1.058194

According to the table above, larger f-value is related with higher significance level in the previous table. Variables with higher significance to log of total official need receive larger f-value. The VIF from the model for each variable is smaller than 10, which indicates the low correlations between variables which further provide evidence for low multicollinearity in model 2.

- 3.2.3 Model 3: Regression Model for Residential Burden
  - Variables in the Final Model:

	Variables	Coefficient	Significance (p-value)
Numeric Variables	The residential population receiving water service	-1.047e+00	0.000165 ***
	Population Change from 1980 to 2010	1.128e+00	2.52e-12 ***
	Population in 2000	-8.566e-02	0.109133
	Population Density in 2010	7.366e+07	0.196983
Dummy Variables	CSS (Whether the infrastructure is in a combined sewer system)	2.108e+05	0.038049 *

	EPA Region 1 (whether the infrastructure is in EPA1)	8.584e+03	0.650578
	EPA Region 2	2.343e+04	0.895603
	EPA Region 3	5.356e+04	0.742300
	EPA Region 4	-1.157e+01	0.999941
	EPA Region 5	4.226e+04	0.785954
	EPA Region 6	-3.063e+04	0.849539
	EPA Region 7	5.498e+04	0.733253
	EPA Region 8	2.185e+04	0.902672
	EPA Region 9	4.863e+05	0.023035 *

According to the table above, it is clear that the model for residential burden is not performing so well since most of the variables are not significant to the residential burden. Among these variables, there are four of them that are significant: the residential population receiving water service, the population change from 1980 to 2010, CSS (Whether the infrastructure is in a combined sewer system) and whether the infrastructure is in EPA region 9.

- Model Interpretation

  When **the number of residential population receiving water service** increase by 1, the residential burden would decrease by 1.047e+00. This is because when we are setting up residential burden, the we are dividing water need by the residential population. Since the denominator has been increasing, the value of the dependent variable would naturally be smaller.

  When **the population change from 1980 to 2010** increase by 1, the residential burden would increase by 1.128e+00. While a city is gaining population, it is natural that population need water service would also increase which means the numerator would increase. As a result, the residential burden increases too.

  When **the infrastructure is in a combined sewer system**, the residential burden would increase 2.108e+05. If it is not in a combined sewer system, the total official need will not change. For the service level of water infrastructure, it is common that those in a combined sewer system have a higher ability to offer more service. Thus, those infrastructures have higher need.

  When **the infrastructure is in EPA region 9**, the residential burden would increase by 4.863e+05 compared to those in EPA region 10. This difference is caused by different water needs and service levels across the States.

- Model Accuracy Assessment

Model 3	Adjusted R-sq	F-statistics	p-value	RSE
	0.01139	7.774	< 2.2e-16	2401000

According to the table above, the adjusted R-sq indicates that the model could only predict the residential burden with poorly with the current predictors. The value of 0.01139 suggests that 1.14% of the residential burden could be explained by current predictors.

Although the adjusted r-sq is not optimistic for the model3, the p-value for the model is small enough to reject the null hypothesis. But at the same time, the small f-value makes us remain doubtful about the model's accuracy.

The RSE value is large which indicates the inaccuracy of the model.

The reason why residential burden is harder to predict might be the lack of relevant data in the dataset. Since residential burden take demographic factor into the dependent variable, we might need more data to supplement the explanation.

Variables	**F-value**	**VIF**
The residential population receiving water service	14.2062	1.275234
Population Change from 1980 to 2010	49.1814	2.098492
Population in 2000	2.5673	1.979871
Population Density in 2010	1.6649	1.343571
CSS (Whether the infrastructure is in a combined sewer system)	4.3042	1.136304
EPA Region 1 (whether the infrastructure is in EPA1)	0.0020	2.206041
EPA Region 2	0.0172	2.948211
EPA Region 3	0.1081	4.114681
EPA Region 4	0.0000	4.822798
EPA Region 5	0.0738	5.725834
EPA Region 6	0.0360	4.148910
EPA Region 7	0.1162	4.218474
EPA Region 8	0.0150	2.644308
EPA Region 9	5.1678	2.338233

Compared with previous models, the f-value for this model are very small which indicates the low ability for these data to offer information to this

model. As for VIF, they are all smaller than 10 which suggests low multicollinearity between variables.

- 3.2.4 Model 4: Regression Model for log of Residential Burden
  - Variables in the Final Model:

	Variables	Coefficient	Significance (p-value)
Numeric Variables	The non-residential population receiving water service	4.479e-06	0.000252 ***
	The residential population receiving water service	-1.210e-06	5.16e-11 ***
	Median Income in 1979	-4.849e-05	< 2e-16 ***
	% White in population in 2010	7.546e-03	2.84e-08 ***
	Population in 2010	-1.784e-07	2.98e-10 ***
Dummy Variables	CSS (Whether the infrastructure is in a combined sewer system)	1.271e-01	0.042095 *
	Whether the infrastructure is connected to TMDL	3.772e-01	3.90e-05 ***
	EPA Region 1 (whether the infrastructure is in EPA1)	1.712e-01	0.159556
	EPA Region 2	-3.894e-02	0.727627
	EPA Region 3	2.214e-01	0.032341 *
	EPA Region 4	6.452e-02	0.537649
	EPA Region 5	-2.492e-01	0.010968 *
	EPA Region 6	-3.093e-01	0.003155 **
	EPA Region 7	-6.491e-02	0.531115
	EPA Region 8	2.738e-02	0.809657
	EPA Region 9	2.802e-01	0.034027 *
	Whether the city has larger non-residential population have water needs than residential population	-1.263e-01	0.011191 *

According to the table above, this model has much more significant predictors than the last model. In model 4, significant variables are: the non-residential population receiving water service, the residential population receiving water service, median income in 1979, % white in population in 2010, population in

2010, CSS (Whether the infrastructure is in a combined sewer system), whether the infrastructure is connected to TMDL, whether the infrastructure is in EPA Region 3, 5,6,9 and whether the city has larger non-residential population have water needs than residential population.

- Model Interpretation
  When **the number of non-residential population receiving water service** increase by 1, the residential burden would increase by 4.479e-06 in its percentage. This might because when the need for non-residential population increases, the need from the residential population would also increase.
  When **the number of residential population receiving water service** increase by 1, the residential burden would decrease by 1.210e-06 in its percentage. This is because residential burden is water need divided by the residential population. Since the denominator has been increasing, the value of the dependent variable would naturally be smaller. So the percentage change would decrease too.
  When **median income in 1979** increase by 1, the residential burden would decrease by 4.849e-05 in percentage. While the income is increasing, the water need might be decreasing because of the more popular notion for saving water resource.
  When **the % white in population in 2010** increases by 1, the residential burden would increase by 7.546e-03 in percentage.
  When **population in 2010** increase by 1, the residential burden would decrease by 1.784e-07 in percentage, because residential burden is water need divided by the residential population.
  When **the infrastructure is in a combined sewer system**, the residential burden would increase 2.108e+05. If it is not in a combined sewer system, the total official need will not change. For the service level of water infrastructure, it is common that those in a combined sewer system have a higher ability to offer more service. Thus, those infrastructures have higher need.
  When the **infrastructure is connected to TMDL**, the residential burden would increase by 3.772e-01 in percentage.
  When **the infrastructure is in EPA region 3 or 9**, the residential burden would increase by 2.214e-01 in percentage or 2.802e-01 in percentage compared to those in EPA region 10. When it is in EPA Region 5 or 6, the residential burden would decrease by 2.492e-01 or 3.093e-01 in percentage compared to EPA region 10. This difference is caused by different water needs and service levels across the States.
  When **the city has a larger non-residential population in water need than residential population**, the residential burden would decrease by 1.263e-01 in

percentage. This is because with the same residential population, the need for water has increased.

- Model Accuracy Assessment

Model 4	Adjusted R-sq	F-statistics	p-value	RSE
	0.06384	34.03	< 2.2e-16	1.492

According to the table above, the adjusted R-sq indicates that the model could only predict the residential burden with poorly with the current predictors. The value of 0.06384 suggests that 6.38% of the log of residential burden could be explained by current predictors.
Although the adjusted r-sq is not optimistic for the model3, the p-value for the model is small enough to reject the null hypothesis. The f-value is also larger than the last model about residential burden. So we could have the confidence to reject the null hypothesis.
The RSE value is much smaller than the previous model. And now the prediction of the dependent variable is more accurate although the adjusted r-sq still indicates the lack of useful data.

Variables	**F-value**	**VIF**
The non-residential population receiving water service	13.4079	1.268974
The residential population receiving water service	43.2297	1.448483
Median Income in 1979	89.1871	1.458733
% White in population in 2010	30.8705	1.561551
Population in 2010	39.7894	1.563933
CSS (Whether the infrastructure is in a combined sewer system)	4.1327	1.113671
Whether the infrastructure is connected to TMDL	16.9387	1.093587
EPA Region 1 (whether the infrastructure is in EPA1)	1.9788	2.288457
EPA Region 2	0.1213	2.992551
EPA Region 3	4.5819	4.295002
EPA Region 4	0.3799	5.496833
EPA Region 5	6.4735	5.875216
EPA Region 6	8.7210	4.521127

EPA Region 7	0.3923	4.507726
EPA Region 8	0.0580	2.770909
EPA Region 9	4.4948	2.310166
Whether the city has larger non-residential population have water needs than residential population	6.4376	1.089684

The f-value in this model is larger than the previous one but still small if compared to the models about water needs. The VIF values are all smaller than 10 which suggests low multicollinearity between variables.

- 3.3 Interactions Between Variables

  In the interaction between variables, the numeric variable chosen here is the residential population receiving water service (either collection or treatment). The dummy variables here are whether the infrastructure is connected to a combined sewer system, whether the population of the city is growing. The dependent variable here is total official need for water infrastructure.

  The first pair of models are trying to build the regression between total official need and residential population receiving water service as well as CSS. For simple model when independent variables have the relationship of addition, the adjusted r-sq equals 0.2994. After changing the second item to the multiplication, the r-sq increases to 0.3443.

```
mod101 <- lm(TOTAL_OFFICIAL_NEED ~ max_RP + CSS, data = dat2)

summary(mod101)

#r-sq=0.2994

mod1011 <- lm(TOTAL_OFFICIAL_NEED ~ max_RP + max_RP:CSS, data = dat2)

summary(mod1011)

#r-sq=0.3443
```

  The second pair of models are trying to build the regression between total official need and residential population receiving water service as well as Growing cities. The addition model gets the r-sq of 0.2825 while the model with multiplier get the r-sq of 0.2824.

```
mod102 <- lm(TOTAL_OFFICIAL_NEED ~ max_RP + GrowingCities, data = dat2)

summary(mod102)

#r-sq=0.2825

mod1021 <- lm(TOTAL_OFFICIAL_NEED ~ max_RP + max_RP:GrowingCities, data = dat2)

summary(mod1021)

#r-sq=0.2824
```

- 3.4 Comparison Between Models

Model 1	Adjusted R-sq	F-statistics	p-value	RSE
	0.4062	313.9	< 2.2e-16	104700000

Model 2	Adjusted R-sq	F-statistics	p-value	RSE
	0.2809	170.3	< 2.2e-16	1.549

Model 3	Adjusted R-sq	F-statistics	p-value	RSE
	0.01139	7.774	< 2.2e-16	2401000

Model 4	Adjusted R-sq	F-statistics	p-value	RSE
	0.06384	34.03	< 2.2e-16	1.492

In this section of model comparison, several important measures of the models would be compared to draw a model evaluation.
In terms of adjusted r square, model 1 has the largest value which represents that model 1 has the highest proportion of dependent variables could be explained by its predictors. Model 2 also has a high adjusted r-sq which makes model2 predicts well. But model 3 and 4 about residential burden have a relatively poor ability in setting up the regression between residential burden and the predictors.
For F-values and p-values, they measures whether we could reject the null hypothesis. The p-value for each model is small while the f-value really differs. In general, model 1 and 2 with water needs as dependent variables have larger f-values

which indicates that the predictors have already given a lot of information to dependent variables. But model 3 and 4 has smaller f-value, which indicates their limited ability in predicting dependent variables.
The RSE of the models doesn't follow the same pattern as above. For the two models with original numeric form of data, model 1 and 3, the RSE is pretty large. Considering the observed values of dependent variables are much larger than their logged counterparts, the RSE could be regarded as acceptable. But in another word, the RSE in these models are not really comparable since the range of dependent variables varies a lot.
All in all, among the four multivariate regression models, model 1 and 2 have more meaning statistically and can be put into use to analyze water needs. But model 3 and 4 are less effective than model 1 and 2. To improve model 3 and 4, more demographic data is in need.

Part 4 Conclusion

- 4.1 Bivariate Regression

Most explanatory variables are significantly correlated to the dependent variables. However, we need to exclude the variables which are not correlated to the dependent variables from the multivariate regression models. In fact, according to the analysis of the bivariate regression models above, we found that

1. **Percent Change of Population Density from 2000 to 2010** should not be included in the multivariate regression model of Total Reported Need for water infrastructure.

2. All variables could be included in the multivariate regression model of the Logging value of the total reported need for water infrastructure.

3. **Except for Logging value of population density in 2010 in county**, all of the other 9 variables should not be included in the multivariate regression model of Residential Burden.

4. If a facility is associated with a combined sewer system (Dummy), If a facility contributes to a TMDL receiving water body (Dummy), and Percent Change of Population Density from 2000 to 2010 should not be included in the multivariate regression model of Logging value of Residential Burden.

The table below includes the P value for each bivariate regression model. The highlighted variables would not be included in the multivariate regression models for that dependent variables because the P values are larger than the significant level.

P values		Dependent Variable			
		Total Official Need	Log of Total Official Need	Residential Burden	Log of Residential Burden
Explanatory Variables	Logging value of median income in 2009 in county	1.79e-07***	<2e-16***	0.227	<2e-16***
	Logging value of 2010 population in county	<2e-16***	<2e-16***	0.00377**	<2e-16***
	If the city's population has grown from 2000 to 2010 (Dummy)	7.77e-05***	<2e-16***	0.379	0.00743**
	If a city has non-residential population (Dummy)	<2e-16***	<2e-16***	0.614	2e-04***
	If a facility is associated with a combined sewer system (Dummy)	<2e-16***	<2e-16***	0.377	0.914
	If a facility contributes to a TMDL receiving water body (Dummy)	1.25e-06***	4.93e-08***	0.79	0.0196
	Percent of county	<2e-16***	<2e-16***	0.1069	<2e-16***

	population identifying as white in 2010				
	Percent Change of Population Density from 2000 to 2010	0.314	<2e-16***	0.314	0.379
	Present (2012) residential population receiving services	<2e-16**8	<2e-16***	0.752	<2e-16***
	Population Density in 2010	<2e-16***	<2e-16***	0.192	2.09e-11***

- 4.2 Multivariate Regression

  In multivariate regression models, the predictors of the dependent variables are mainly as following:

Groups	Notes	Type
Population	In 1980, 1990, 2000 and 2010	Numeric
Median Income	In 1969, 1979, 1989, 1999 and 2009	Numeric
Population density	In 1980, 1990, 2000 and 2010	Numeric
Population Change	From 1980 to 2010	Numeric
% white population	In 2000 and 2010	Numeric
Population receiving water service	The max value between receiving collections and receiving treatments; non-residential and residential	Numeric
CSS	Whether connected to combined sewer system	Binary
Growing Cities	Whether the population is increasing	Binary
Cities with non-residential population	Whether the non-residential population is	Binary

	larger than residential population	
EPA Regions	EPA 1~10	Binary (each)

By analyzing some of the key variables' information in Philadelphia, we could get a rough view of why Philadelphia has a high demand for water infrastructure. Since the model measuring total official need has the highest value of adjusted r-sq, we used this model to explore the reasons. In this model, the variables' information are as follows:

	Variables	Coefficient	Significance (p-value)	f-value	VIF
Numeric Variables	The residential population receiving water service	5.202e+02	<2e-16 ***	1622.9539	1.448983
	The non-residential population receiving water service	1.448e+03	<2e-16 ***	261.2139	1.382964
	Population density in 2010	7.045e+10	<2e-16 ***	661.1780	1.550284
	Median income in 2009	-1.618e+02	0.0978 .	2.7409	1.624598
	% White population in 2000	-1.883e+05	0.0663 .	3.3737	1.680106
Dummy Variables	CSS (Whether the infrastructure is in a combined sewer system)	4.624e+07	<2e-16 ***	108.0443	1.145245
	EPA Region 1 (whether the infrastructure is in EPA1)	6.435e+06	0.4491	0.5730	2.267919
	EPA Region 2	-1.179e+07	0.1370	2.2119	3.051936
	EPA Region 3	-1.014e+07	0.1565	2.0084	4.175393
	EPA Region 4	3.895e+06	0.5803	0.3058	5.053839
	EPA Region 5	-6.711e+06	0.3253	0.9676	5.786003
	EPA Region 6	-4.624e+06	0.5213	0.4113	4.348190
	EPA Region 7	7.886e+06	0.2695	1.2193	4.345722
	EPA Region 8	-1.977e+06	0.8005	0.0639	2.664286
	EPA Region 9	-6.185e+06	0.4994	0.4562	2.252037
	Whether the city is growing in population	6.435e+06	0.0253 *	5.0042	1.261551
	Whether the city has larger non-residential	-6.483e+06	0.0635 .	3.4447	1.089412

	population have water needs than residential population				

According to the table above, the 5 variables worth analyzing is the residential population receiving water service, the non-residential population receiving water servicei, population density in 2010, CSS and whether the city is growing in population.

Variables		**Value**	**Rank in the US**
The residential population receiving water service	PHILADELPHIA WATER DEPT (NE)	985611	22
	PHILADELPHIA WATER DEPT (SE)	311218	/
	PHILADELPHIA WATER DEPT (SW)	984773	23
The non-residential population receiving water service	PHILADELPHIA WATER DEPT (NE)	271906	9
	PHILADELPHIA WATER DEPT (SE)	2606	/
	PHILADELPHIA WATER DEPT (SW)	306878	6
Population density in 2010	/	0.004128701	16
CSS (Whether the infrastructure is in a combined sewer system)	/	1 (Yes)	/
Whether the city is growing in population	/	1 (Yes)	/

According to the table above, it is very clear that Philadelphia ranked very high in its non-residential population receiving water service. The population density in 2010 also contributes a lot to the high water demand. Also, the residential population receiving water service is also an important contributor to the demand.
Besides, Philadelphia also meet the requirement of dummy variables to increase the water infrastructure demand. The infrastructures are all connected in a combined sewer system. The city's population is also growing.

In all, Philadelphia has the following conditions that make it have a high water infrastructure demand:
1. large residential population receiving water service;
2. large non-residential population receiving water service;
3. high population density in 2010;
4. connected to a combined sewer system
5. has a growing population trend

This analysis aims to point out the possible reasons for Philadelphia to have a high water infrastructure demand compared with other cities in the United States. There are still several reasons might affect the analysis's accuracy:

1. the dataset used in this analysis is not the recent one;
2. some demographic data had not been added into the analysis;
3. analysis method mainly focused on linear regression, testing other regression model might discover other interesting findings;
4. Due to limited time, some possible dummy variables such as "the state where the infrastructure is in" has not been added to the analysis

By improving above limitations, it is very likely that a more accurate model would help Philadelphia more about its water infrastructure analysis.