US Census Demographic Data

TERM PROJECT FINAL REPORT

DEPARTMENT OF STATISTICS OF MIDDLE EAST TECHNICAL UNIVERSITY

5/3/2019

Bennur Kaya
Beste Karaçay
Gizem Ceylan
Merve Erşahin

Middle East Technical University
Department of Statistics
Spring 2018-2019

TABLE OF CONTENT

Middle East Technical University
Department of Statistics
Spring 2018-2019

1. INTRODUCTION

1.1. Abstract

A census is an official survey to obtain details such as the number of people living in the countries/states, their ages, races and work types. The US Census Demographic Data were collected by the US Census Bureau that has an aim to provide objective and high quality data regarding its people and economy. The data is taken from the 2015 American Community Survey as 5 year estimates. Income disparities have become so pronounced that America's top 10 percent now average more than nine times as much income as the bottom 90 percent (UC Berkeley, 2019). This project was conducted to compare the income levels of people in North Carolina and New York in terms of gender, race, employment type, transportation preferences, poverty and work type. In addition, some analysis results do not only include differences in income levels of races, but also the effect of employment type. According to analysis results, the average income differs considering states where people live in. Moreover, the change in the number of hispanic people in the population affects the average income although the change in the number of white people do not.

1.2. Literature of Review

According to US Census Bureau Data regardless of race and ethnicity, female workers earn 80 cents for every dollar that is paid to their opposite sex colleagues (American Association of University Women, 2019). Even when the race and ethnicity are involved in the process, income inequality increases much more. In a raport which is conducted in New York City, it is indicated that the black women who are a full-time workers have earned 57 cents for each dollar which is earned by both white and non-hispanic men. On the other hand, is it seen that the gap between average incomes is highly increasing for hispanic women in NYC. They earn 49 cents in the same conditions (The City Comptroller's Office, 2018). In 2017, New York (NY) had 8.62 million people with a median age of 36.6 although North Carolina (NC) had 10.3 million people with a 38.8 age average. Besides those, the economy of NY employs 4.19 million people while there exists 4.62 million employees in NC. New York have $60,879 (± $500) as an average annual income. Also, North Carolina have a median annual income of $52,752 (± $510). Since the median income of the United States equals to $60,336, it can be indicated that the average annual income of NY is more than the entire United States. Unlike NY, NC is under the average. Furthermore, full-time male employees' salaries in North Carolina were almost $62,048 (± $1,147) although full-time female employees' salaries show diversity between $45,399 and $47,017 per year. That is, males made almost 1.35 times more than females in NC. Even though it is less than in North Carolina, there is also income inequality in New York. Males made $78,550 (± $1,168) as average income per year although female salary changed between $59,818 and $61,526. That is, males made almost 1.30 times more than female workers. In 2017, the highest paid ethnicity of NY workers were people who are white with $67,169 (± $870) on the average. It is followed by native hawaiian and other pacific islander with $62,618 (± $15,800) and asian with $62,444 (± $2,225). The highest paid race of NC workers were asian with $63,340 (± $4,747) on the average. These employees were paid almost 1.20 times more than white employees who are the second highest paid. North Carolina has 1.58 million among 9.8 million people who live under the poverty limit. That equals almost 16% of the population in the state. In NY, it was about 19.5%. Hispanic people had the highest rate who lived below the poverty limit in NY and it was followed by people who are white and black respectively. Moreover, North Carolina had 10.3 million people who can be considered that 95% of them were nationals. On the other hand, New York had a more diverse population. In NY, the almost 32% of the inhabitants were white people as residents. In addition, the 29.2% of them were hispanic/latino which were almost 2.5 million people, the 21.8% of New Yorkers were black or african american residents and the 14.4% of the New York citizens were asian resident which were almost 1.25 million people. 36.9% of the NYers were born outside of the U.S with the most common birthplace as Mexico. This percentage was higher than the average of nation which equals 13.7%. Additionally, the most preferred method of transportation for employees in NY was public transportation which is followed by those who prefered to drive lonely and walk respectively. Unlike NY, in NC, almost 81% of the population chose to drive lonely and this preference was followed by those who carpooled and those who worked at home (Data USA, 2019)

1.3. Data Description

US Census Demographic dataset were collected by the US Census Bureau in 2017. The Census Bureau updates the estimates approximately every year. This dataset expands on earlier *New York City Census* dataset. It also includes data from North Carolina as an addition to New York City. Census tracts are defined by the census bureau and have a consistent size. A typical census tract has around 5000 or so residents.

This dataset consists of 162 observations and 15 variables. One of the variables is dependent and the rest are independent which are defined as one discrete, 12 continuous and one categorical variable.

Name and type of the independent variables:

- State - Categorical
- Women (number of women) - Numerical (Discrete)
- Hispanic (% of population that is Hispanic/Latino) - Numerical (Continuous)
- White (% of population that is White) - Numerical (Continuous)
- Black (% of population that is Black) - Numerical (Continuous)
- Poverty (% under poverty level) - Numerical (Continuous)
- Professional (% employed in management, business, science, and arts) - Numerical (Continuous)
- Service (% employed in service jobs) - Numerical (Continuous)
- Office (% employed in sales and office jobs) - Numerical (Continuous)
- Drive (% commuting alone in a car, van, or truck) - Numerical (Continuous)
- Carpool (% carpooling in a car, van, or truck) - Numerical (Continuous)
- Employed (Number of employed (16+) - Numerical (Continuous)
- Private Work (% employed in private industry) - Numerical (Continuous)
- Public Work (% employed in public jobs) - Numerical (Continuous)

Name and type of the dependent variable:

- Median household income ($) - Numerical (Continuous)

1.4. Research Questions

Median household income may depend on different factors. All of these factors should be taken into consideration while trying to understand whether they affect the median household income or not. Thus, some research questions should be generated and answered to be able examine the effect of the independent variables on dependent variable. For instance, questions such as “Does median household income depends on populations’ race? How does income affected by states (NY, NC)?” should be examined. In addition, since income can also be affected by the work and transportation type (drive, carpool), “Is there a relationship between income, work type (private work, public work) and transportation type (drive, carpool)?” question should also be examined. On the other hand, variables may affect each other, too. That is why, it is also important to examine the question of whether the poverty depends on employment type, gender and transportation type or not.

2. AIMS OF RESEARCH

The project aims to analyze the factors affecting the income distribution of New York and North Carolina’s population. The effect of gender difference and race diversity on income are examined as well as the effect of type of work and transportation, poverty and employment type.

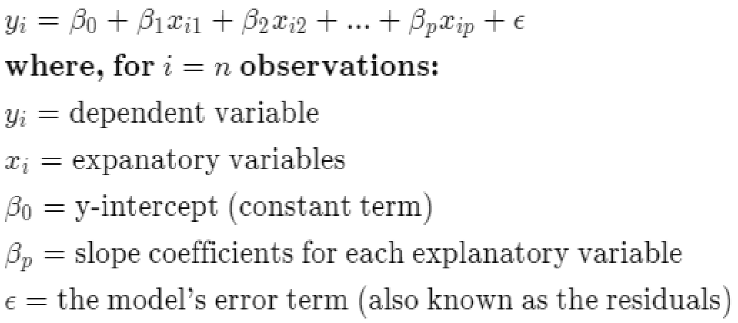
3. SURVEY METHODOLOGY

3.1. Descriptive Statistics

Descriptive statistics is the term which is used for analysis of data that helps describe, show or summarize basic features of the data in a meaningful way. They provide simple summaries, graphics and basic virtual analysis about the sample and the measures of the data. There are two general types of statistic that are used to describe the data which are measures of central tendency and measures of spread. In terms of describing the central tendency we used a number of summary statistics which are the mode, median, and the mean. To be able to describe the spread; the range, quartiles, variance and standard deviation are calculated. In addition, we also made summarization of the data using a combination of tables, graphs and charts and calculated skewness & kurtosis.

3.2. Multiple Linear Regression

Multiple linear regression (MLR), in short, multiple regression, is a statistical technique whose aim is to model the linear relationship between the series of two or more explanatory (independent) variables and a single response (dependent) variable.



$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon$

where, for $i = n$ observations:

y_i = dependent variable

x_i = expanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

There exist some assumptions for the multiple linear regression.

- There should be a linear relationship between the dependent and the independent variables.

- The residuals should be normally distributed
- Independent variables shouldn’t be highly correlated with each other.
- y_i observations are selected randomly and independently.

3.3. Logistic Regression

Logistic regression is to analyze the relationship between a binary outcome such as “the success” and “the failure” & a set of explanatory variables.

The logistic response function is;

$$E(y) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\mathbf{x}'\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{x}'\boldsymbol{\beta})}$$

And the logistic regression function is;

$$\log\left(\frac{p_i}{1-p_i}\right)= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} +\ldots+ \beta_k x_{ki} \quad \text{for } i=1,2,\ldots,n$$

There exist some assumptions that apply for logistic regression.

- The dependent variable should be binary.
- Observations should be independent of each other.
- The independent variables should not be highly correlated with each other.
- Linearity of independent variables and log odds is assumed.

4. DATA ANALYSIS

We started with examining the descriptive summary statistics of the data.

Summary of Income of New York					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
36593	50863	53674	58309	61260	105744

Table 1. Summary statistics for New York

When we check the mean of income for New York which is 58309, it can be seen that it is really close to the median which is 53674. Furthermore, by looking at the summary statistics, it is possible to say that the half of the income of the participants that are from New York are 53674 or below since the second quartile i.e median is 53674. Thus, the other half are 53674 or above. In addition, 75% of the income are 61259.5 or below.

Summary of Income of North Carolina					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
31287	37726	43562	44552	49258	73577

Table 2. Summary statistics for North Carolina

If we check the mean of income for North Carolina which is 44552, it can be concluded that it is almost equal to median which is 43562. In addition, by looking at the summary statistics, it is possible to say that the half of the income of the participants that are from North Carolina are 43562 or below. Therefore, the other half are 43562 or above.

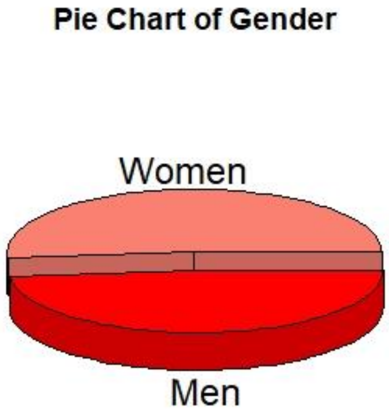
Range of Income of New York
36593 - 105744
Range of Income of North Carolina
31287 - 73577

Table 3. Range of Income of states

Skewness is the measure of symmetry and the direction of skewness is given by the sign. A positive value means that the distribution is positively skewed whereas the negative sign shows the opposite. In addition, the larger the value, the larger the distribution differs from a normal distribution. When we check the skewness of the income variable, we have found it as 1.573535. Thus, we can conclude that it is positively skewed.

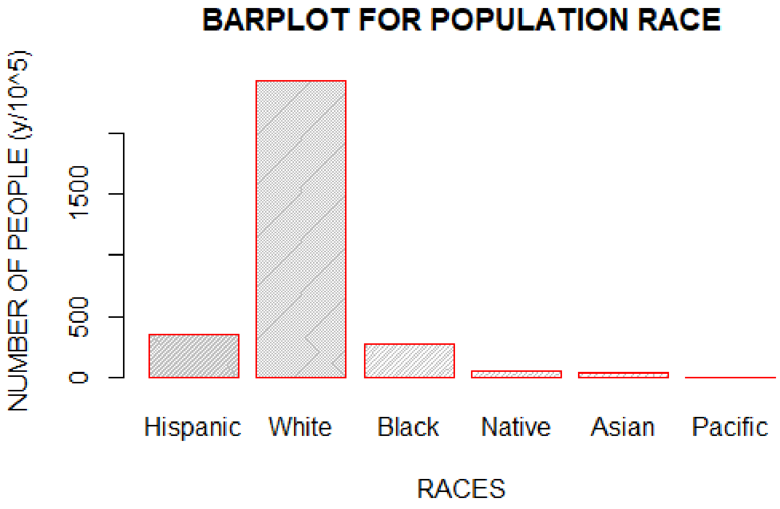
Kurtosis is about existence of outliers. Kurtosis is the measure of whether the data are heavy-tailed (profusion of outliers) or light-tailed (lack of outliers) relative to a normal distribution. If we look at the kurtosis of the income variable, it can be seen that it is equal to 3.575026 which means it is close to normal distribution since normal distribution has a kurtosis of 3.

Now, let’s examine the other variables. Firstly, as it can be seen from the pie chart for gender below, it can be concluded that the number of the respondents who are women is greater than men.



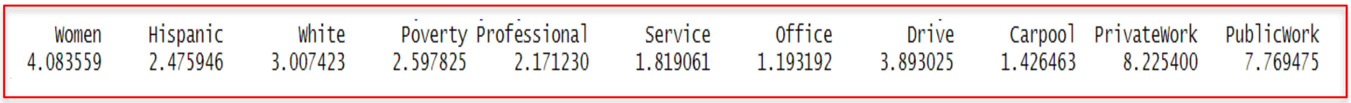
Graph 1. Pie Chart for Gender

Secondly, the bar plot below shows us the respondents’ distribution according to races. When we look at the bar plot below, it is possible to see that the majority of the population belongs to the white’s and the minority of the population belongs to the pacific’s.



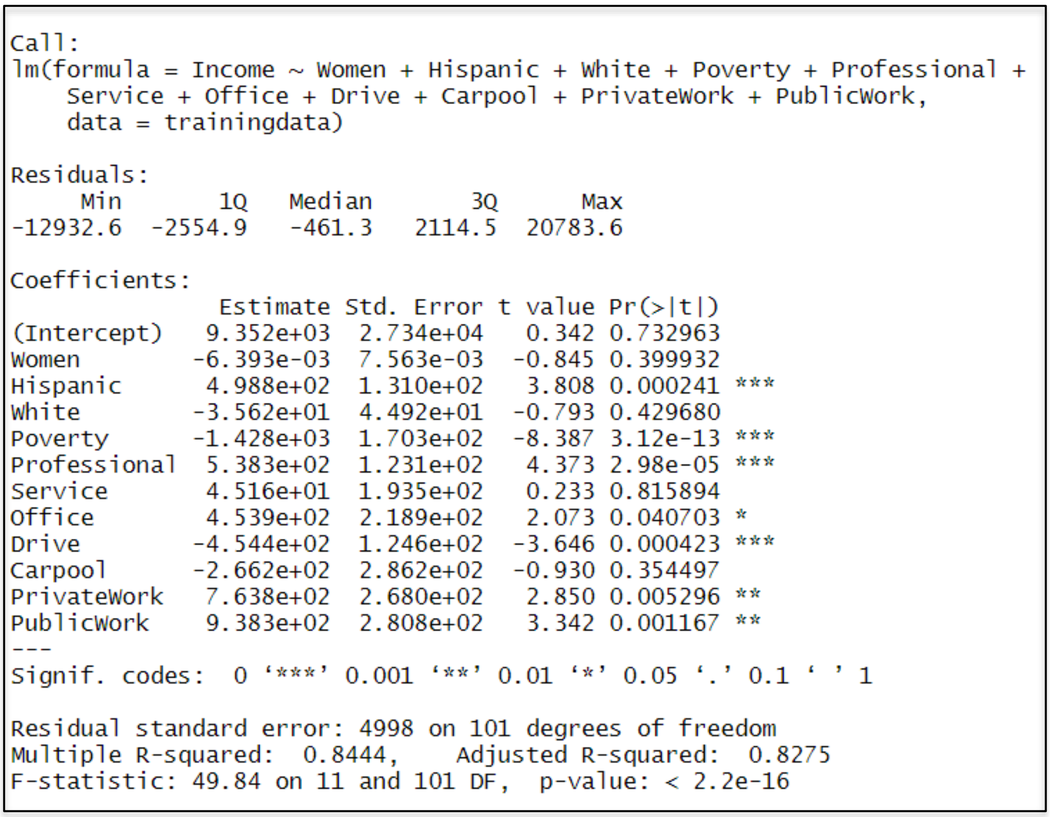
Graph 2. Bar Chart for Races

In order to build full model, firstly, cross-validation is applied. Before cross-validation, VIF values are checked to prevent multicollinearity.



```
   Women     Hispanic       White     Poverty Professional     Service      Office       Drive     Carpool  PrivateWork  PublicWork
4.083559     2.475946    3.007423    2.597825     2.171230    1.819061    1.193192    3.893025    1.426463     8.225400    7.769475
```

None of the variables create multicollinearity problem since they are all less than 10, so we do not need to eliminate them, for now.



```
Call:
lm(formula = Income ~ Women + Hispanic + White + Poverty + Professional +
    Service + Office + Drive + Carpool + PrivateWork + PublicWork,
    data = trainingdata)

Residuals:
     Min       1Q   Median       3Q      Max
-12932.6  -2554.9   -461.3   2114.5  20783.6

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.352e+03  2.734e+04   0.342 0.732963
Women        -6.393e-03  7.563e-03  -0.845 0.399932
Hispanic      4.988e+02  1.310e+02   3.808 0.000241 ***
White        -3.562e+01  4.492e+01  -0.793 0.429680
Poverty      -1.428e+03  1.703e+02  -8.387 3.12e-13 ***
Professional  5.383e+02  1.231e+02   4.373 2.98e-05 ***
Service       4.516e+01  1.935e+02   0.233 0.815894
Office        4.539e+02  2.189e+02   2.073 0.040703 *
Drive        -4.544e+02  1.246e+02  -3.646 0.000423 ***
Carpool      -2.662e+02  2.862e+02  -0.930 0.354497
PrivateWork   7.638e+02  2.680e+02   2.850 0.005296 **
PublicWork    9.383e+02  2.808e+02   3.342 0.001167 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4998 on 101 degrees of freedom
Multiple R-squared:  0.8444,    Adjusted R-squared:  0.8275
F-statistic: 49.84 on 11 and 101 DF,  p-value: < 2.2e-16
```

	Model Performance for Train Set	**Model Performance for Test Set**
R^2	0.8197	0.8275
MSE/RMSE	5479	4998

Table 4. Model performance comparison

In the model performance table, it is seen that the mean square error of the train set and root mean square error of the test set differs positively. Moreover, R^2 value is nearly the same for both. Therefore, validation of the model is satisfied.

VARIABLE SELECTION

We started by applying forward selection method. After applying forward selection method, the coefficients that can be seen from the table xx are added into the model, respectively. So the best model suggested by the forward selection method consists of Poverty, Professional, Hispanic, Public Work, Private Work, Office, Drive, Service and Carpool.

Then, to double check, we can also use backward selection method. According to the backward selection method, the best model consists of Hispanic, Poverty, Professional, Service, Office, Private Work, Public Work and Carpool.

We can also apply stepwise selection method. After applying stepwise selection method, these coefficients are added into the model, respectively. Hence, the best model consists of Poverty, Professional, Hispanic, Public Work, Private Work, Carpool, Office and Drive.

Forward Selection	**Backward Selection**	**Stepwise Selection**
● Poverty ● Professional ● Hispanic ● Public Work ● Private Work ● Office ● Drive ● Service ● Carpool	● Hispanic ● Poverty ● Professional ● Service ● Office ● Private Work ● Public Work ● Carpool	● Poverty ● Professional ● Hispanic ● Public Work ● Private Work ● Carpool ● Office ● Drive

Table 5. Variable selection results

By looking at these selection methods, forward selection and backward selection gave the same results. Although the selection suggests that “White” and “Women” should be eliminated, we decided to keep them in the model since they are important for comparison and they do not create any multicollinearity problem.

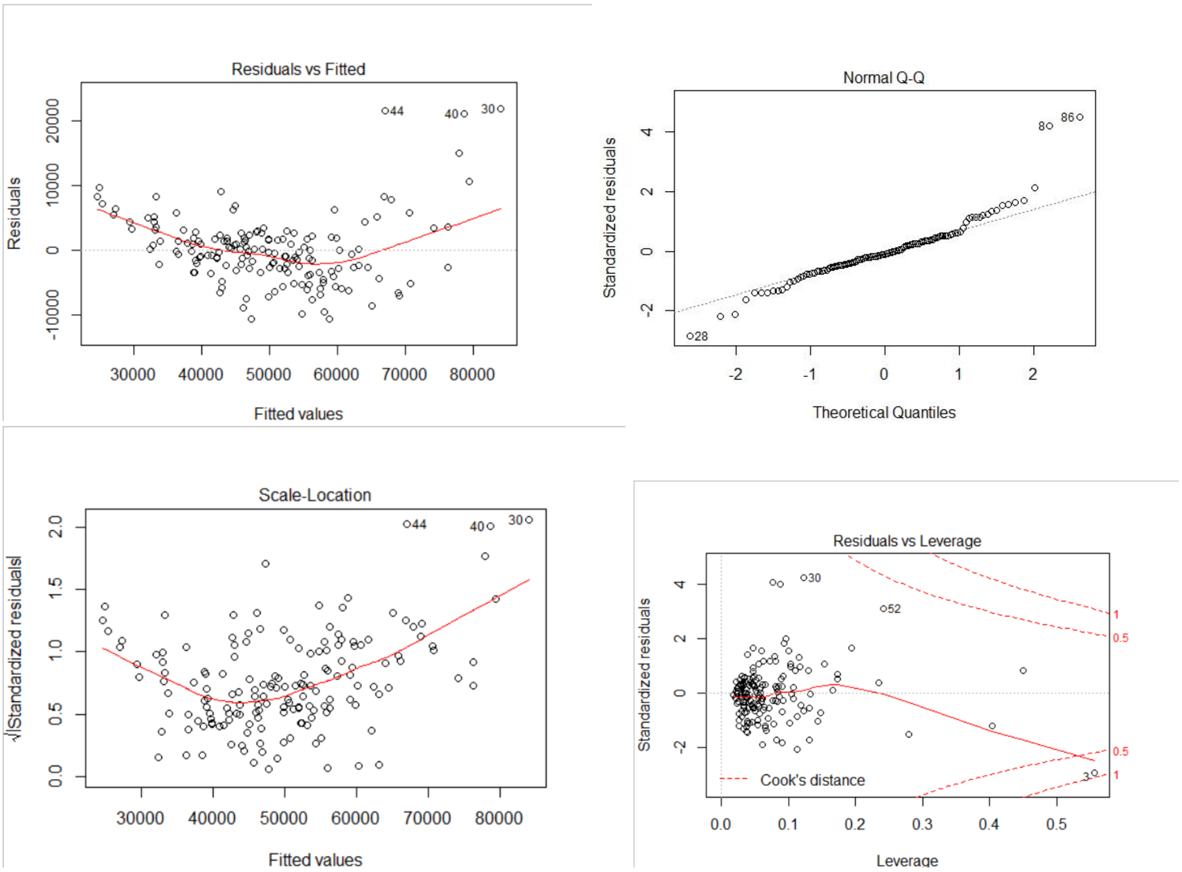
So, the best model is:

$$\textbf{Income} = \beta 0 + \beta 1*\textbf{Women} + \beta 2*\textbf{Hispanic} + \beta 3*\textbf{White} + \beta 4*\textbf{Poverty} + \beta 5*\textbf{Professional} + \beta 6*\textbf{PublicWork} + \beta 7*\textbf{PrivateWork} + \beta 8*\textbf{Office} + \beta 9*\textbf{Carpool} + \beta 10*\textbf{Drive}$$

	Original Model Statistics	**Best Fitted Model**
R^2	0.8197	0.8275

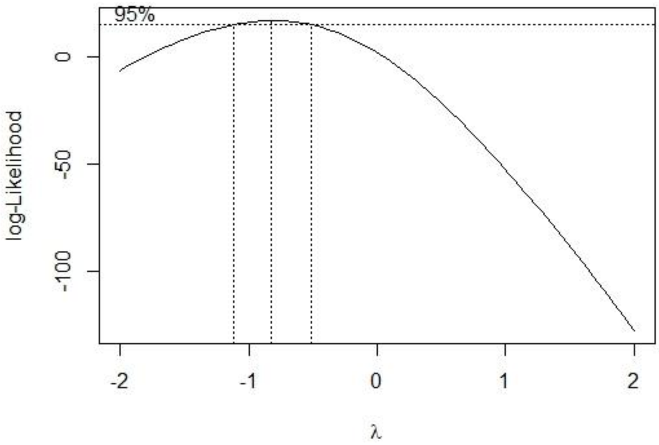
When we check the Adjusted R-squared values of the first model and the best model, it can be seen that it increased by 0.1.

Now, we will check the assumptions and then, if needed, apply residual analysis.



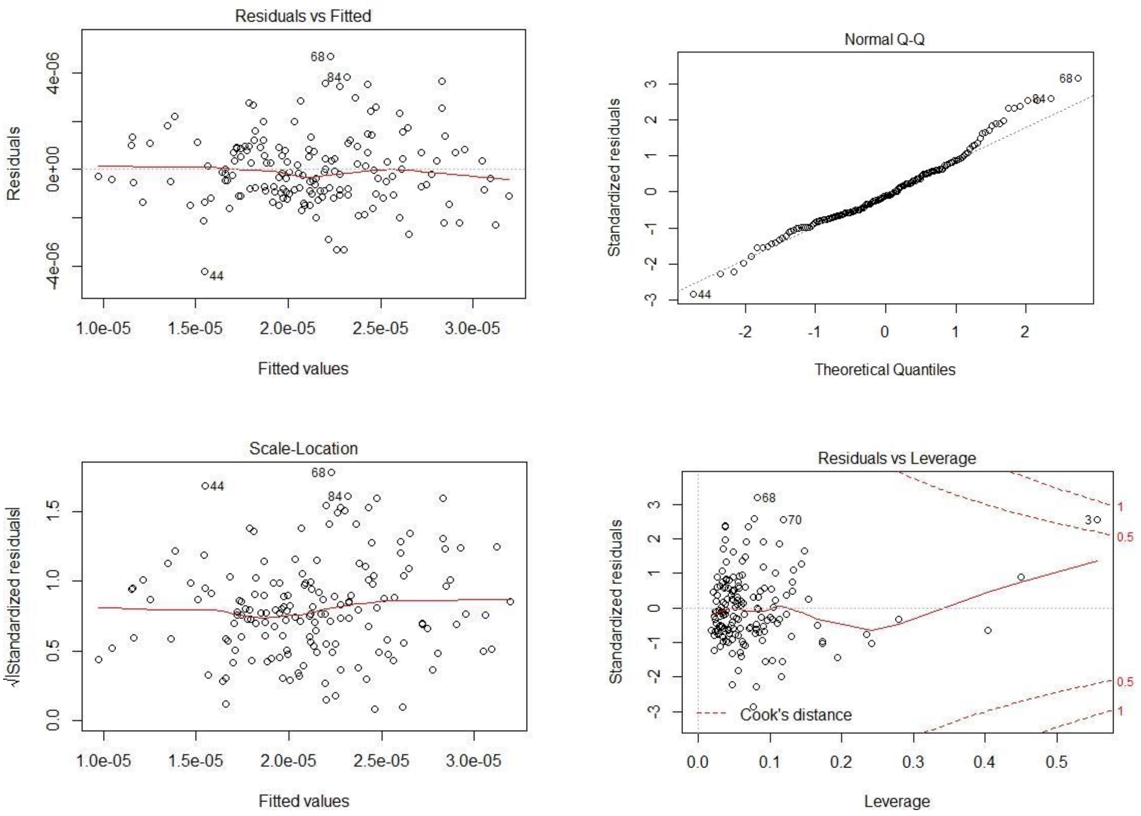
Graphs 3,4,5,6. Residual analysis before transformation

Here, it can be seen that the data does not follow the Normal distribution. It violates the Constant Variance assumption. It seems that there are departures from the Normality line. Residuals do not lie on normal line. Also, in "Residuals vs Leverage" graph there exist some leverage points. So, transformation is required.



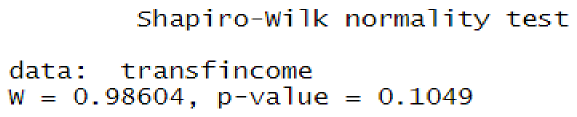
Suggested transformation is reciprocal transformation on Y (income).

After the transformation is applied, we checked the assumptions one more time to see if there are any changes with respect to normality.



Graphs 7,8,9,10. Residual analysis aftertransformation

Now, the Constant Variance assumption is not violated and the data points mostly lie on the normal line. To verify the normality, Shapiro-Wilk test is applied.



```
        Shapiro-Wilk normality test

data:  transfincome
W = 0.98604, p-value = 0.1049
```

Hypothesis of Shapiro-Wilk
H0: The population is normally distributed
H1: The population is not normally distributed

So, by looking to the p-value, it can be said that we fail to reject the null hypothesis. Hence, the population is normally distributed.

The transformed model is then:

1/(Income) = β0 + β1*Women + β2*Hispanic + β3*White + β4*Poverty + β5*Professional + β6*PublicWork + β7*PrivateWork + β8*Office + β9*Carpool + β10*Drive

```
Call:
lm(formula = transfincome ~ Women + Hispanic + White + Poverty +
    Professional + Office + Drive + Carpool + PrivateWork +
    PublicWork, data = censusdata)

Residuals:
       Min         1Q     Median         3Q        Max
-4.187e-06 -9.831e-07 -1.770e-07  8.516e-07  4.652e-06

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.473e-05  6.853e-06   6.527 9.74e-10 ***
Women        -5.407e-13  1.236e-12  -0.438  0.66233
Hispanic     -1.302e-07  2.929e-08  -4.443 1.71e-05 ***
White        -3.054e-08  1.117e-08  -2.734  0.00701 **
Poverty       5.373e-07  3.905e-08  13.758  < 2e-16 ***
Professional -2.541e-07  2.653e-08  -9.576  < 2e-16 ***
Office       -1.751e-07  5.627e-08  -3.112  0.00223 **
Drive         3.599e-08  2.187e-08   1.645  0.10201
Carpool      -4.764e-08  6.283e-08  -0.758  0.44947
PrivateWork  -2.019e-07  6.903e-08  -2.924  0.00399 **
PublicWork   -3.424e-07  7.369e-08  -4.647 7.31e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.53e-06 on 150 degrees of freedom
Multiple R-squared:  0.9005,     Adjusted R-squared:  0.8932
F-statistic: 123.4 on 11 and 150 DF,  p-value: < 2.2e-16
```

When we look at the above statistics, we can see that the p-value is too small, thus we reject the null hypothesis which claims that the model is insignificant. Thus, the model looks significant. Moreover, the adjusted R-squared value is now higher as it is expected.

RESEARCH QUESTIONS

Question 1: Does income depend on populations’ race?

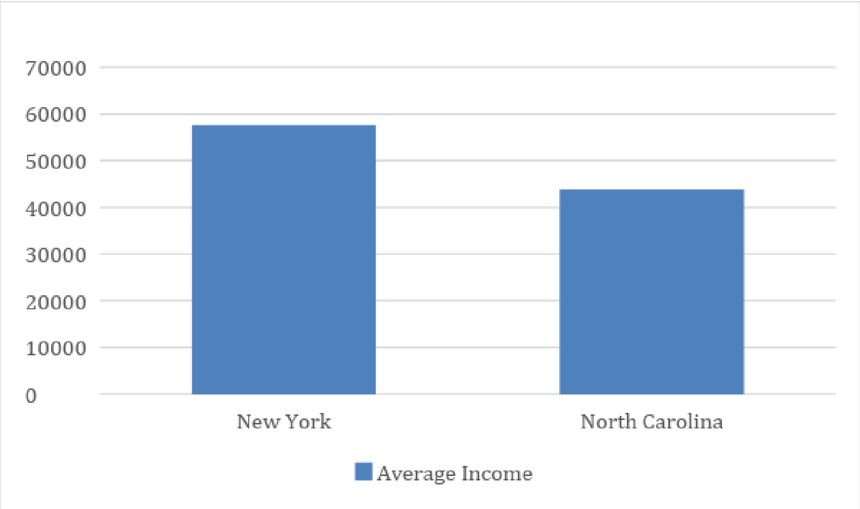
Hypothesis:
H0= Income is not associated with race
H1= Income is associated with race

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.062e+04  2.666e+04   0.398 0.691215
Hispanic     4.938e+02  1.286e+02   3.839 0.000214
White       -3.385e+01  4.407e+01  -0.768 0.444217
```

As we check the p-value of the coefficients, it is seen that, for the states New York and North Carolina, Hispanic variable is significant but White is not significant at 0.05 alpha level. So, we can conclude that the White has not a significant effect on income whereas Hispanic has.

When we increase Hispanic population by 1%, the income increases by 493.8$.

Question 2: How does income affected by the states (NY, NC)?



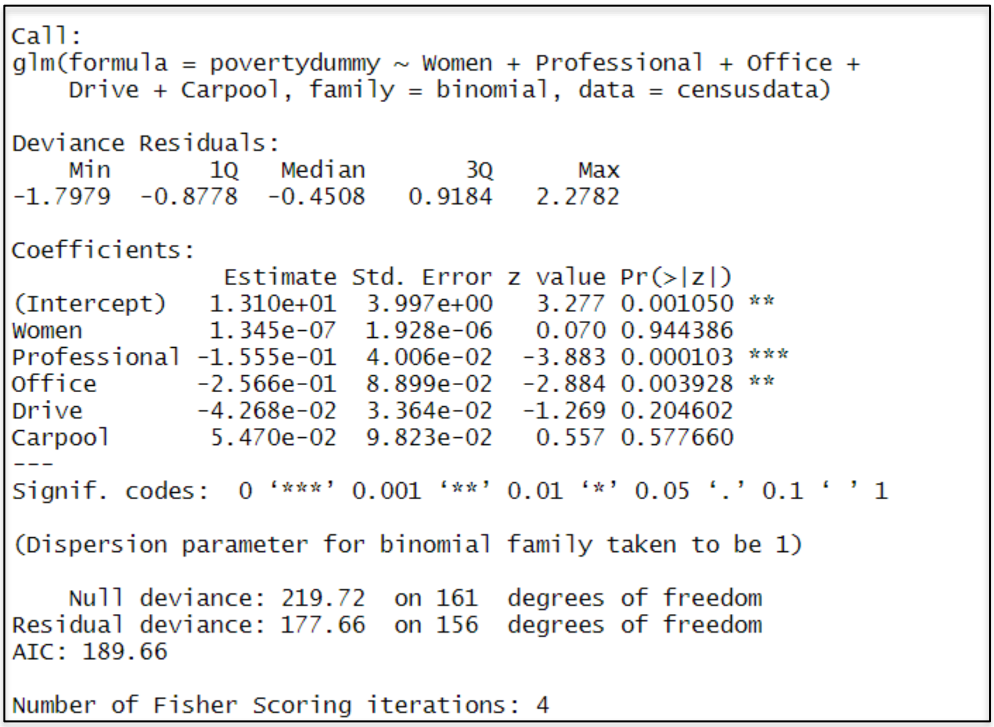
Graph 11. Bar plot of Income with respect to states

In order to examine the relationship between the income and the states, we created a bar plot. This plot shows us that the average income in New York is higher than in North Carolina. The result is as expected since New York is the most populous city in the US and one of the most populous metropolitan cities in the

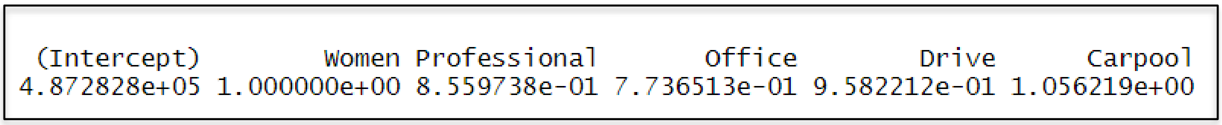
Question 3: Does gender, transformation and employment type affects Poverty?

In order to answer this question, the Poverty variable has been turned into a dummy variable. We fitted a logistic regression to check the odds.

Poverty = β0 + β1*Women + β2*Professional + β3*Office + β4*Drive+ β5*Carpool



```
Call:
glm(formula = povertydummy ~ Women + Professional + Office +
    Drive + Carpool, family = binomial, data = censusdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7979  -0.8778  -0.4508   0.9184   2.2782

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.310e+01  3.997e+00   3.277 0.001050 **
Women         1.345e-07  1.928e-06   0.070 0.944386
Professional -1.555e-01  4.006e-02  -3.883 0.000103 ***
Office       -2.566e-01  8.899e-02  -2.884 0.003928 **
Drive        -4.268e-02  3.364e-02  -1.269 0.204602
Carpool       5.470e-02  9.823e-02   0.557 0.577660
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 219.72  on 161  degrees of freedom
Residual deviance: 177.66  on 156  degrees of freedom
AIC: 189.66

Number of Fisher Scoring iterations: 4
```



```
 (Intercept)        Women Professional       Office        Drive      Carpool
4.872828e+05 1.000000e+00 8.559738e-01 7.736513e-01 9.582212e-01 1.056219e+00
```

Here, from the odds table it can be seen that the gender (Women) has no effect on Poverty since the odds (e^β1) is equal to 1.

However, when we look at the employment types which are Professional and Office, since they are less than 1, we will take 1/e^β2=1.169591 and 1/e^β3=1.293661. So, it can be concluded that in every 1% unit decrease in Professional variable, odds of being under the poverty level increases by 1.169 when all the other variables are held constant. Similarly, with each additional 1% unit decrease in Office variable, odds of being under the poverty level increases by 1.29.

Also, Carpool seems to have no effect on Poverty, too. It is because the odds of it is almost equal to 1, which means there is no association between Poverty and Carpool.

Question 4: Is there a relationship between income, work type (public work and private work) and transportation type (drive and carpool)?

Hypothesis:
H0= There is a relationship between income, work and transportation type
H1= There is no relationship between income, work and transportation type

```
Call:
lm(formula = transfincome ~ Women + Hispanic + White + Poverty +
    Professional + Office + Drive + Carpool + PrivateWork +
    PublicWork, data = censusdata)

Residuals:
       Min         1Q     Median         3Q        Max
-4.187e-06 -9.831e-07 -1.770e-07  8.516e-07  4.652e-06

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.473e-05  6.853e-06   6.527 9.74e-10 ***
Women        -5.407e-13  1.236e-12  -0.438  0.66233
Hispanic     -1.302e-07  2.929e-08  -4.443 1.71e-05 ***
White        -3.054e-08  1.117e-08  -2.734  0.00701 **
Poverty       5.373e-07  3.905e-08  13.758  < 2e-16 ***
Professional -2.541e-07  2.653e-08  -9.576  < 2e-16 ***
Office       -1.751e-07  5.627e-08  -3.112  0.00223 **
Drive         3.599e-08  2.187e-08   1.645  0.10201
Carpool      -4.764e-08  6.283e-08  -0.758  0.44947
PrivateWork  -2.019e-07  6.903e-08  -2.924  0.00399 **
PublicWork   -3.424e-07  7.369e-08  -4.647 7.31e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.53e-06 on 150 degrees of freedom
Multiple R-squared:  0.9005,     Adjusted R-squared:  0.8932
F-statistic: 123.4 on 11 and 150 DF,  p-value: < 2.2e-16
```

When we look at the p-values of the coefficients from the outputs of transformed best model, Private Work and Public Work variables are significant at 0.05 alpha level. So, we can say that those two variables have a significant effect on the income.
When we increase Private Work population by 1%, the income decreases by -2.01e^-07$.
When we increase Public Work population by 1%, the income decreases by -3.42e^-07$.

In the same way, when we look at the p-value of the coefficients, drive and carpool variables are not significant at 0.05 alpha level. So, transportation types do not affect the income.

5. CONCLUSION

According to analysis results, change in number of Hispanic people in the population affects the income where the change in White people do not. Furthermore, this project shows that the mean of income changes according to states that people live. For example, people who live in New York earn much more money than people who live in North Carolina. Also, we can say that the income variable depends on work type such as Private Work and Public Work but it does not affected by the transportation type. For poverty, as it is expected, gender is not a factor. However, employment types create a difference for odds of being under poverty level or not.

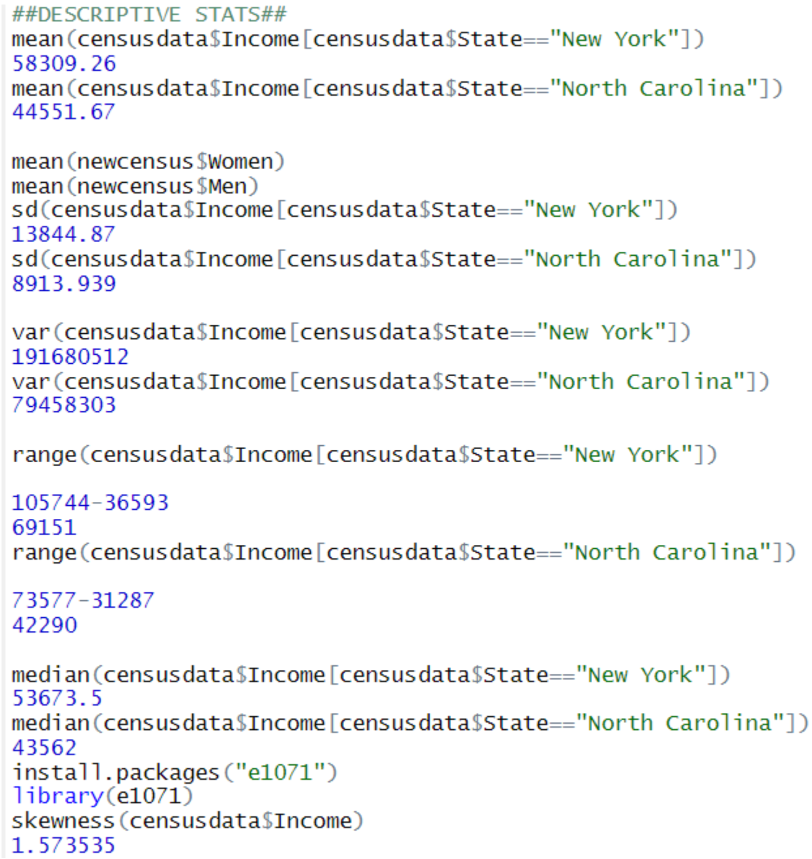
6. REFERENCES

1. Berkeleyedu. (2019). Berkeleyedu. Retrieved 3 May, 2019, from https://eml.berkeley.edu/~saez/saez-UStopincomes-2017.pdf

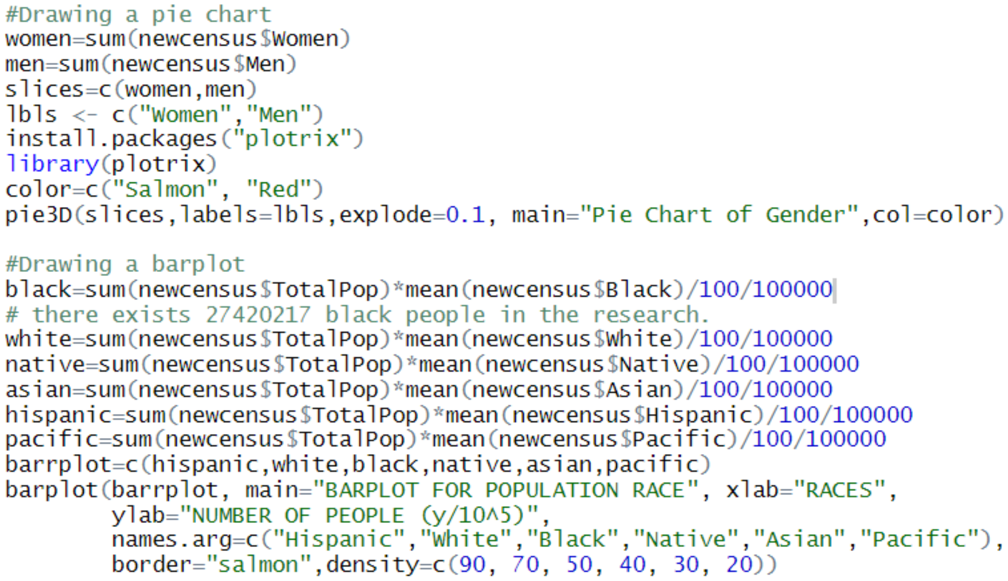
2. Datausaio. (2019). Datausaio. Retrieved 3 May, 2019, from https://datausa.io/profile/geo/north-carolina/

3. Nytimescom. (2019). Nytimescom. Retrieved 3 May, 2019, from https://www.nytimes.com/2019/04/02/nyregion/newyorktoday/nyc-news-women-equal-pay-day.html

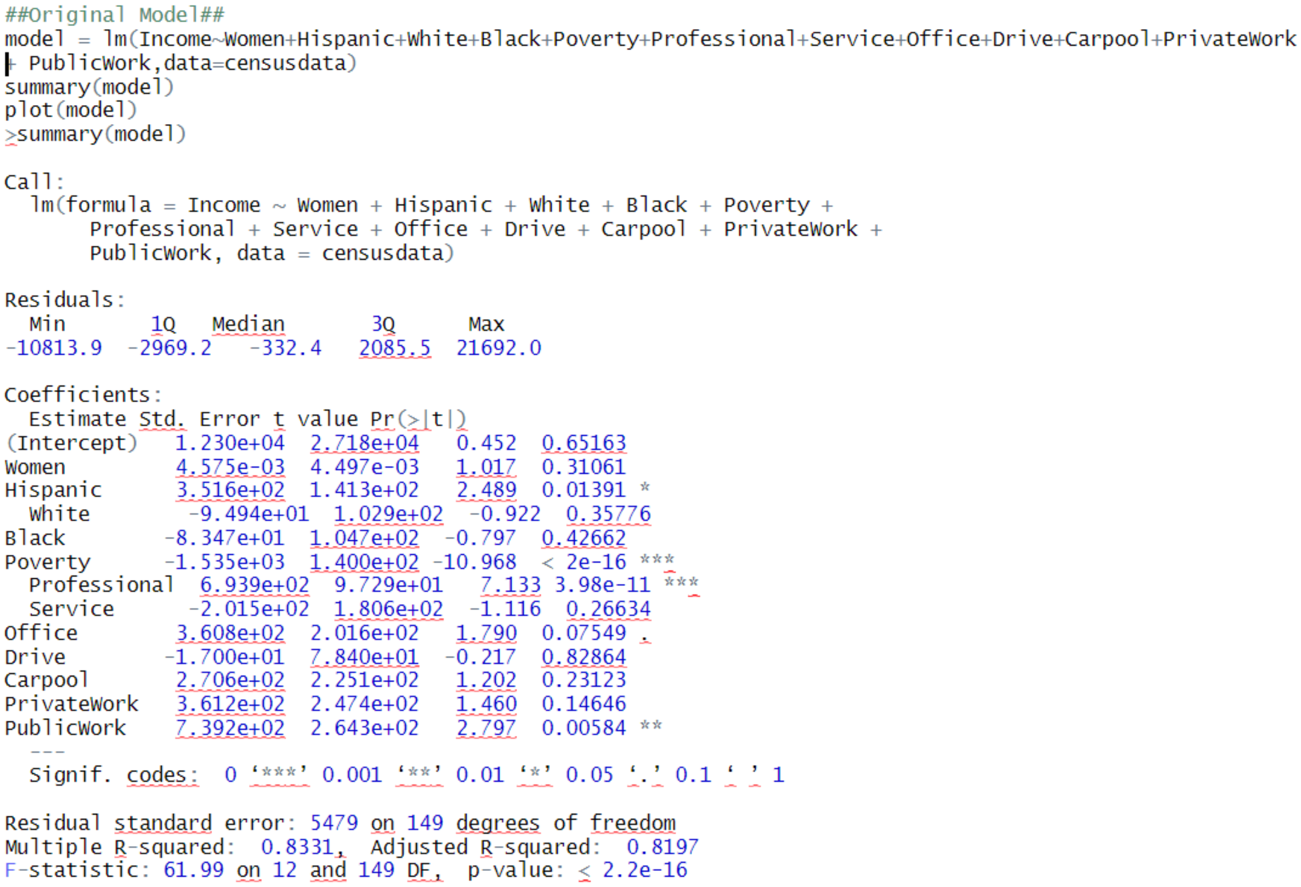
7. APPENDICES



```
##DESCRIPTIVE STATS##
mean(censusdata$Income[censusdata$State=="New York"])
58309.26
mean(censusdata$Income[censusdata$State=="North Carolina"])
44551.67

mean(newcensus$Women)
mean(newcensus$Men)
sd(censusdata$Income[censusdata$State=="New York"])
13844.87
sd(censusdata$Income[censusdata$State=="North Carolina"])
8913.939

var(censusdata$Income[censusdata$State=="New York"])
191680512
var(censusdata$Income[censusdata$State=="North Carolina"])
79458303

range(censusdata$Income[censusdata$State=="New York"])

105744-36593
69151
range(censusdata$Income[censusdata$State=="North Carolina"])

73577-31287
42290

median(censusdata$Income[censusdata$State=="New York"])
53673.5
median(censusdata$Income[censusdata$State=="North Carolina"])
43562
install.packages("e1071")
library(e1071)
skewness(censusdata$Income)
1.573535
```

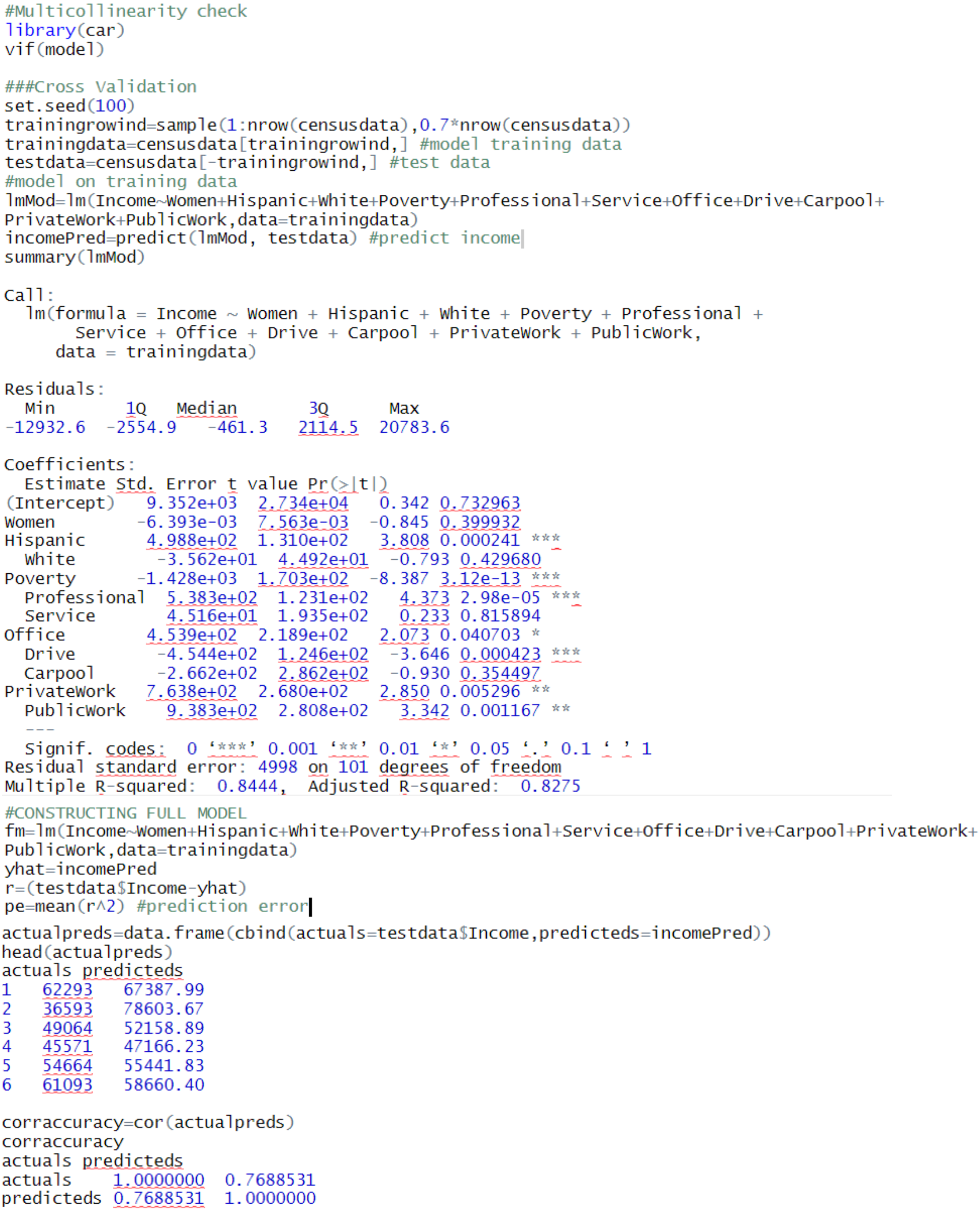


```
#Drawing a pie chart
women=sum(newcensus$Women)
men=sum(newcensus$Men)
slices=c(women,men)
lbls <- c("Women","Men")
install.packages("plotrix")
library(plotrix)
color=c("Salmon", "Red")
pie3D(slices,labels=lbls,explode=0.1, main="Pie Chart of Gender",col=color)

#Drawing a barplot
black=sum(newcensus$TotalPop)*mean(newcensus$Black)/100/100000
# there exists 27420217 black people in the research.
white=sum(newcensus$TotalPop)*mean(newcensus$White)/100/100000
native=sum(newcensus$TotalPop)*mean(newcensus$Native)/100/100000
asian=sum(newcensus$TotalPop)*mean(newcensus$Asian)/100/100000
hispanic=sum(newcensus$TotalPop)*mean(newcensus$Hispanic)/100/100000
pacific=sum(newcensus$TotalPop)*mean(newcensus$Pacific)/100/100000
barrplot=c(hispanic,white,black,native,asian,pacific)
barplot(barrplot, main="BARPLOT FOR POPULATION RACE", xlab="RACES",
        ylab="NUMBER OF PEOPLE (y/10^5)",
        names.arg=c("Hispanic","White","Black","Native","Asian","Pacific"),
        border="salmon",density=c(90, 70, 50, 40, 30, 20))
```

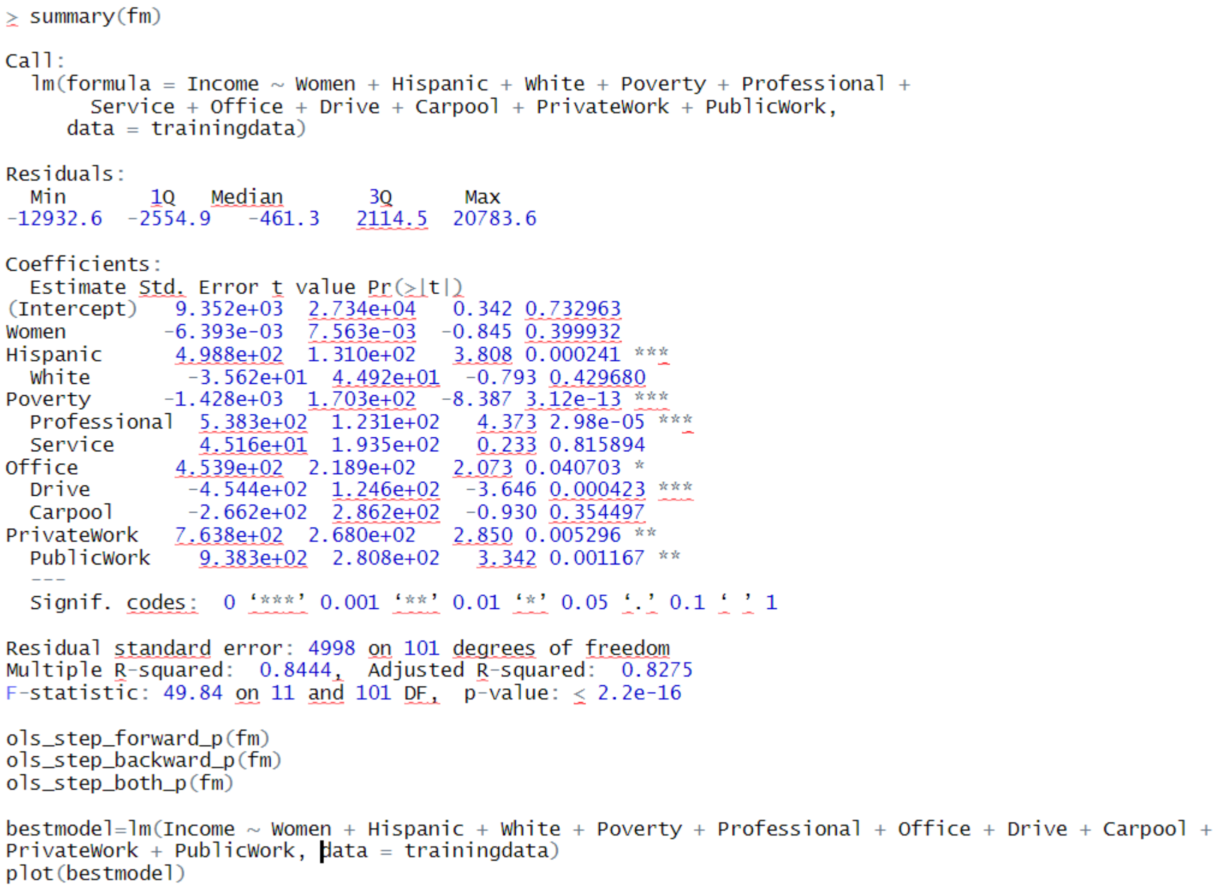


```
##Original Model##
model = lm(Income~Women+Hispanic+White+Black+Poverty+Professional+Service+Office+Drive+Carpool+PrivateWork
+ PublicWork,data=censusdata)
summary(model)
plot(model)
>summary(model)

Call:
  lm(formula = Income ~ Women + Hispanic + White + Black + Poverty +
       Professional + Service + Office + Drive + Carpool + PrivateWork +
       PublicWork, data = censusdata)

Residuals:
  Min        1Q    Median       3Q      Max
-10813.9  -2969.2   -332.4   2085.5  21692.0

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.230e+04  2.718e+04   0.452  0.65163
Women          4.575e-03  4.497e-03   1.017  0.31061
Hispanic       3.516e+02  1.413e+02   2.489  0.01391 *
  White         -9.494e+01  1.029e+02  -0.922  0.35776
Black         -8.347e+01  1.047e+02  -0.797  0.42662
Poverty       -1.535e+03  1.400e+02 -10.968  < 2e-16 ***
  Professional  6.939e+02  9.729e+01   7.133 3.98e-11 ***
  Service       -2.015e+02  1.806e+02  -1.116  0.26634
Office         3.608e+02  2.016e+02   1.790  0.07549 .
Drive         -1.700e+01  7.840e+01  -0.217  0.82864
Carpool        2.706e+02  2.251e+02   1.202  0.23123
PrivateWork    3.612e+02  2.474e+02   1.460  0.14646
PublicWork     7.392e+02  2.643e+02   2.797  0.00584 **
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5479 on 149 degrees of freedom
Multiple R-squared:  0.8331,  Adjusted R-squared:  0.8197
F-statistic: 61.99 on 12 and 149 DF,  p-value: < 2.2e-16
```

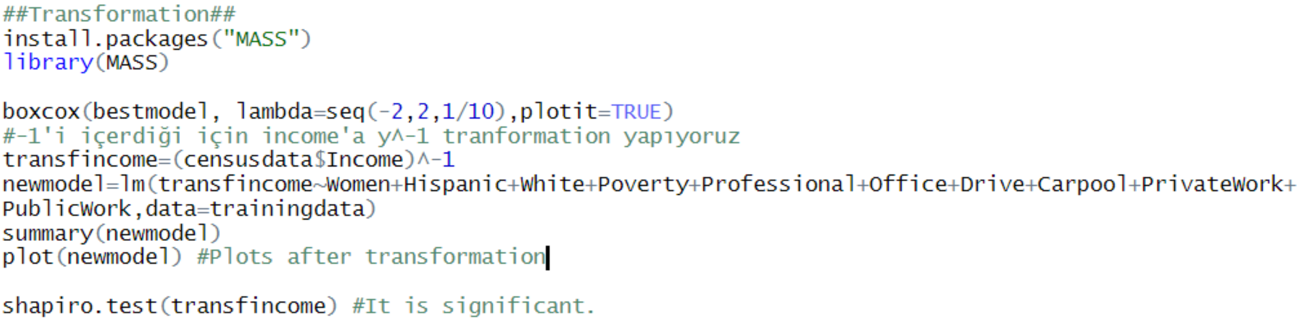
Middle East Technical University
Department of Statistics
Spring 2018-2019



```
#Multicollinearity check
library(car)
vif(model)

###Cross Validation
set.seed(100)
trainingrowind=sample(1:nrow(censusdata),0.7*nrow(censusdata))
trainingdata=censusdata[trainingrowind,] #model training data
testdata=censusdata[-trainingrowind,] #test data
#model on training data
lmMod=lm(Income~Women+Hispanic+White+Poverty+Professional+Service+Office+Drive+Carpool+
PrivateWork+PublicWork,data=trainingdata)
incomePred=predict(lmMod, testdata) #predict income
summary(lmMod)

Call:
  lm(formula = Income ~ Women + Hispanic + White + Poverty + Professional +
       Service + Office + Drive + Carpool + PrivateWork + PublicWork,
     data = trainingdata)

Residuals:
  Min       1Q   Median      3Q      Max
-12932.6  -2554.9    -461.3   2114.5  20783.6

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.352e+03  2.734e+04   0.342 0.732963
Women        -6.393e-03  7.563e-03  -0.845 0.399932
Hispanic      4.988e+02  1.310e+02   3.808 0.000241 ***
  White        -3.562e+01  4.492e+01  -0.793 0.429680
Poverty      -1.428e+03  1.703e+02  -8.387 3.12e-13 ***
  Professional  5.383e+02  1.231e+02   4.373 2.98e-05 ***
  Service       4.516e+01  1.935e+02   0.233 0.815894
Office        4.539e+02  2.189e+02   2.073 0.040703 *
  Drive        -4.544e+02  1.246e+02  -3.646 0.000423 ***
  Carpool      -2.662e+02  2.862e+02  -0.930 0.354497
PrivateWork   7.638e+02  2.680e+02   2.850 0.005296 **
  PublicWork    9.383e+02  2.808e+02   3.342 0.001167 **
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 4998 on 101 degrees of freedom
Multiple R-squared:  0.8444,  Adjusted R-squared:  0.8275
#CONSTRUCTING FULL MODEL
fm=lm(Income~Women+Hispanic+White+Poverty+Professional+Service+Office+Drive+Carpool+PrivateWork+
PublicWork,data=trainingdata)
yhat=incomePred
r=(testdata$Income-yhat)
pe=mean(r^2) #prediction error
actualpreds=data.frame(cbind(actuals=testdata$Income,predicteds=incomePred))
head(actualpreds)
actuals predicteds
1    62293    67387.99
2    36593    78603.67
3    49064    52158.89
4    45571    47166.23
5    54664    55441.83
6    61093    58660.40

corraccuracy=cor(actualpreds)
corraccuracy
actuals predicteds
actuals     1.0000000  0.7688531
predicteds  0.7688531  1.0000000
```



```
> summary(fm)

Call:
  lm(formula = Income ~ Women + Hispanic + White + Poverty + Professional +
       Service + Office + Drive + Carpool + PrivateWork + PublicWork,
     data = trainingdata)

Residuals:
  Min       1Q   Median       3Q      Max
-12932.6  -2554.9   -461.3   2114.5  20783.6

Coefficients:
  Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.352e+03  2.734e+04   0.342 0.732963
Women         -6.393e-03  7.563e-03  -0.845 0.399932
Hispanic       4.988e+02  1.310e+02   3.808 0.000241 ***
  White         -3.562e+01  4.492e+01  -0.793 0.429680
Poverty       -1.428e+03  1.703e+02  -8.387 3.12e-13 ***
  Professional   5.383e+02  1.231e+02   4.373 2.98e-05 ***
  Service        4.516e+01  1.935e+02   0.233 0.815894
Office         4.539e+02  2.189e+02   2.073 0.040703 *
  Drive         -4.544e+02  1.246e+02  -3.646 0.000423 ***
  Carpool       -2.662e+02  2.862e+02  -0.930 0.354497
PrivateWork    7.638e+02  2.680e+02   2.850 0.005296 **
  PublicWork     9.383e+02  2.808e+02   3.342 0.001167 **
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4998 on 101 degrees of freedom
Multiple R-squared:  0.8444,  Adjusted R-squared:  0.8275
F-statistic: 49.84 on 11 and 101 DF,  p-value: < 2.2e-16

ols_step_forward_p(fm)
ols_step_backward_p(fm)
ols_step_both_p(fm)

bestmodel=lm(Income ~ Women + Hispanic + White + Poverty + Professional + Office + Drive + Carpool +
PrivateWork + PublicWork, data = trainingdata)
plot(bestmodel)
```



```
##Transformation##
install.packages("MASS")
library(MASS)

boxcox(bestmodel, lambda=seq(-2,2,1/10),plotit=TRUE)
#-1'i içerdiği için income'a y^-1 tranformation yapıyoruz
transfincome=(censusdata$Income)^-1
newmodel=lm(transfincome~Women+Hispanic+White+Poverty+Professional+Office+Drive+Carpool+PrivateWork+
PublicWork,data=trainingdata)
summary(newmodel)
plot(newmodel) #Plots after transformation

shapiro.test(transfincome) #It is significant.
```

```
############QUESTIONS###########
#1)
> summary(newmodel)

Call:
  lm(formula = transfincome ~ Women + Hispanic + White + Poverty + 
       Professional + Service + Office + Drive + Carpool + PrivateWork + 
       PublicWork, data = censusdata)

Residuals:
    Min         1Q     Median         3Q        Max 
-4.187e-06 -9.831e-07 -1.770e-07  8.516e-07  4.652e-06 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.473e-05  6.853e-06   6.527 9.74e-10 ***
Women        -5.407e-13  1.236e-12  -0.438  0.66233    
Hispanic     -1.302e-07  2.929e-08  -4.443 1.71e-05 ***
White        -3.054e-08  1.117e-08  -2.734  0.00701 ** 
Poverty       5.373e-07  3.905e-08  13.758  < 2e-16 ***
Professional -2.541e-07  2.653e-08  -9.576  < 2e-16 ***
Service       9.164e-08  4.928e-08   1.860  0.06489 .  
Office       -1.751e-07  5.627e-08  -3.112  0.00223 ** 
Drive         3.599e-08  2.187e-08   1.645  0.10201    
Carpool      -4.764e-08  6.283e-08  -0.758  0.44947    
PrivateWork  -2.019e-07  6.903e-08  -2.924  0.00399 ** 
PublicWork   -3.424e-07  7.369e-08  -4.647 7.31e-06 ***
  ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.53e-06 on 150 degrees of freedom
Multiple R-squared:  0.9005,	Adjusted R-squared:  0.8932 
F-statistic: 123.4 on 11 and 150 DF,  p-value: < 2.2e-16
#2)
mean(trainingdata$Income[trainingdata$State=="New York"])
mean(trainingdata$Income[trainingdata$State=="North Carolina"])
#Plot created on Word.

#3)
povertydummy=ifelse(censusdata$Poverty>median(censusdata$Poverty),1,0)
censusdata<-cbind(censusdata,povertydummy)
head(censusdata)
library(ISLR)
glm.fit=glm(povertydummy~Women + Professional + Office + Drive + Carpool, data = censusdata, family = binomial)
summary(glm.fit)

exp(glm.fit$coefficients) #Every decrease of 1 unit in X increases the odds of poverty by about a.
```