

Multiple Regression Computing Project

Your Name: Ajeet Kumar Rai

Stony Brook ID: 116125547

1. Introduction

This project involves the implementation of a linear regression model to predict two target variables: Y1(median household income in the county) and Y2(unemployment rate in the county), using a common set of predictors. The workflow begins with data preprocessing, including missing value imputation and exploratory data analysis (EDA) to understand the data and distribution. Stepwise feature selection was employed to identify the most relevant predictors. Based on the selected features, a baseline linear regression model was constructed using only main effects. Following diagnostic checks, the model was improved by incorporating two-way interaction terms, resulting in better performance. Model evaluation was carried out using R-squared (R^2), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). The complete R code used in this analysis is included in the appendix. Since the distributions of Y1 and Y2 are similar, the same code can be reused for both targets by simply replacing the dependent variable.

2. Methodology

In the Methodology section, we detail the key steps followed during the analysis. These include missing value imputation, exploratory data analysis (EDA), building a baseline model with main effects before diagnostics, refining the model after diagnostic evaluation, and finally incorporating two-way interaction terms to enhance model performance.

2.1 Data

The dataset consists of 3,220 observations, with 34 predictor variables and two response variables: Y1(median household income) and Y2 (unemployment rate). Among the predictors, only "state" and "county" were categorical variables; all others were numeric. During the analysis, it was observed that "state" and "county" did not contribute meaningfully to the model's predictive performance, so they were excluded from the final analysis. Additionally, a large number of predictors contained zero values, which was an important characteristic noted during data summarization.

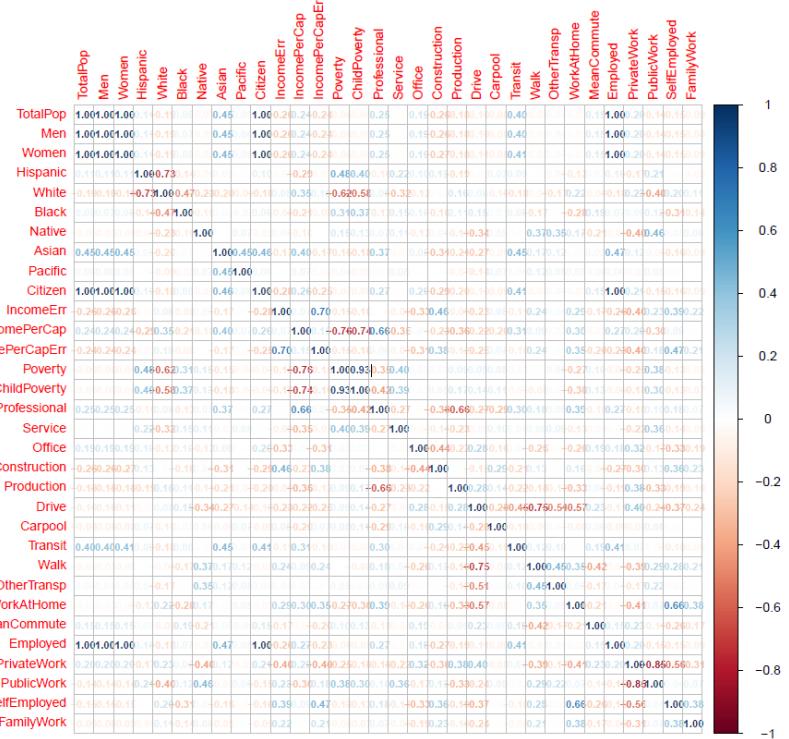
2.2 Missing value

The dataset contained only two missing values, which were not significant and were therefore removed from the analysis.

2.3 Exploratory data analysis

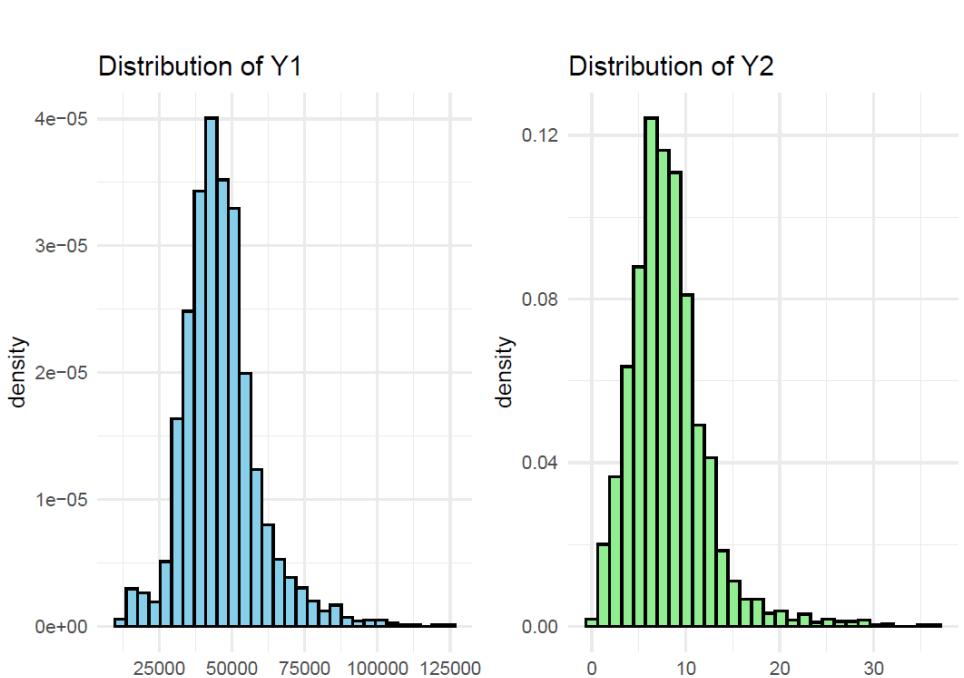
We performed several exploratory data analysis (EDA) steps to better understand the dataset. This included examining correlations between variables, analysing the distribution of the response variables, reviewing the distribution of predictor variables, and identifying potential outliers.

2.3.1 Correlation



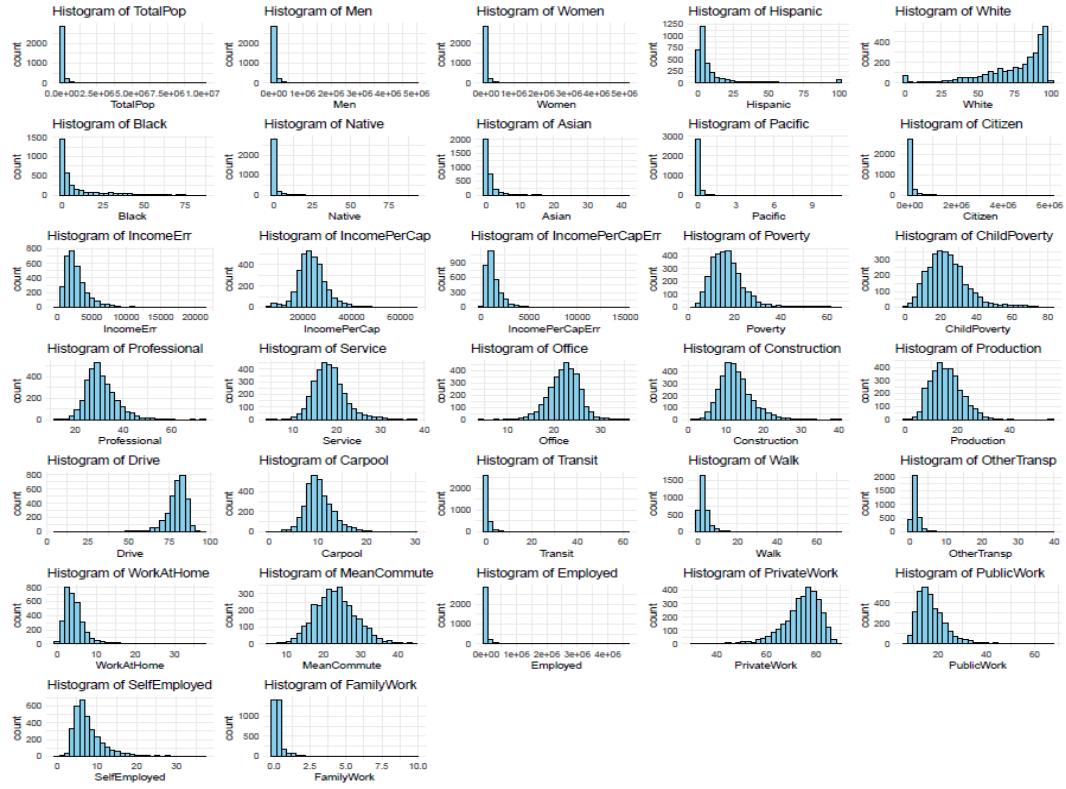
It was observed that many predictors are highly correlated, which indicates a potential violation of the multicollinearity assumption in linear regression.

2.3.2 Distribution (dependent variable)



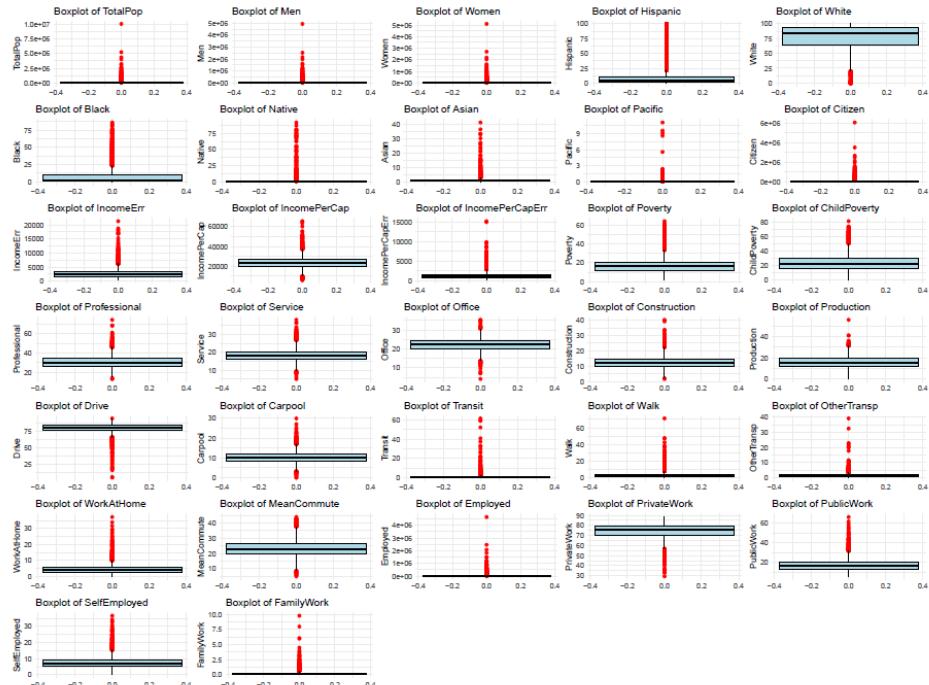
Response variables Y1 and Y2 are positively skewed, meaning transformation is probably needed.

2.3.3 Distribution (Predictors)



Except for a few predictors remaining ones are highly skewed; again, transformation is needed.

2.3.4 Outliers



Leverage plot is essential, as many outliers are present in the predictors.

2.4 Feature selection

In a correlation plot, many features are highly correlated, so among 34 features few of them are not useful. So, stepwise regression in both directions was implemented, which gave 24 predictors based on AIC.

For Y1, these predictors are selected after stepwise regression : TotalPop, Men, White, Black, Native, Asian, Pacific, Citizen, IncomeErr, IncomePerCap, IncomePerCapErr, Poverty, ChildPoverty, Professional, Office, Construction, Production, Drive, Carpool, Walk, WorkAtHome, MeanCommute, PrivateWork, PublicWork

For Y2, these predictors are selected after stepwise regression : TotalPop, Men, Hispanic, Black, Native, Citizen, IncomeErr, IncomePerCapErr, Poverty, ChildPoverty, Service, Office, Production, Drive, Carpool, Transit, Walk, WorkAtHome, MeanCommute, Employed, PrivateWork, SelfEmployed

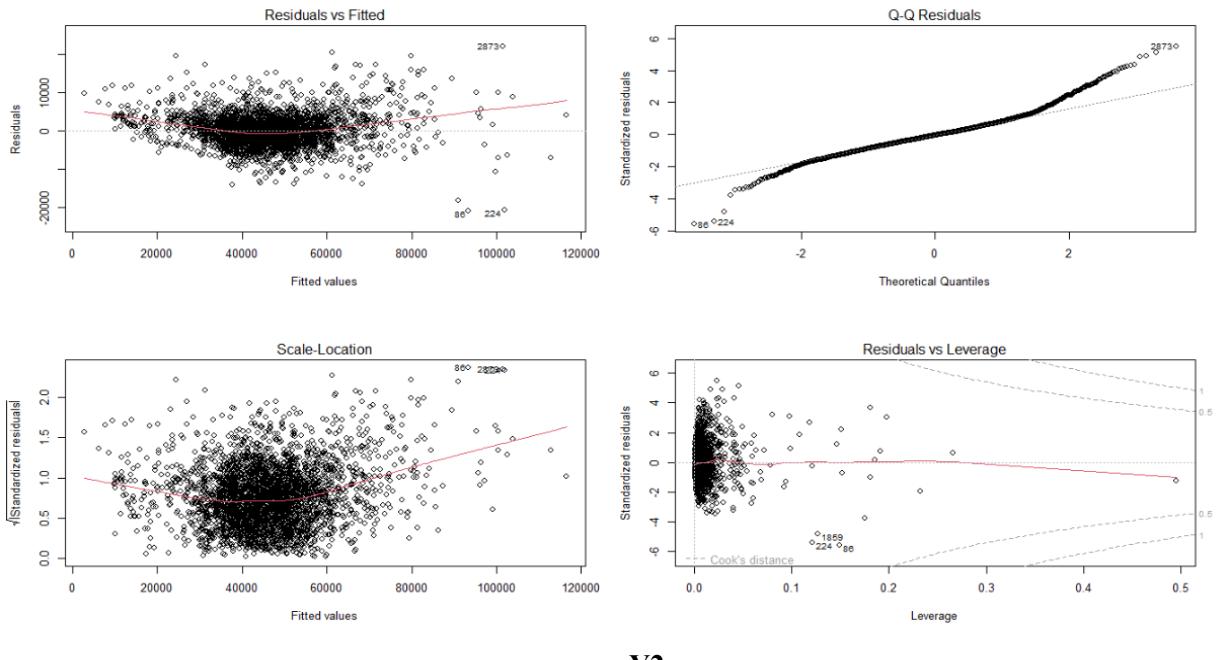
2.5 Base Model with main effect (before diagnosis)

In the base model, multiple linear regression was implemented for both Y1 and Y2 separately, with predictors obtained from the feature selection step.

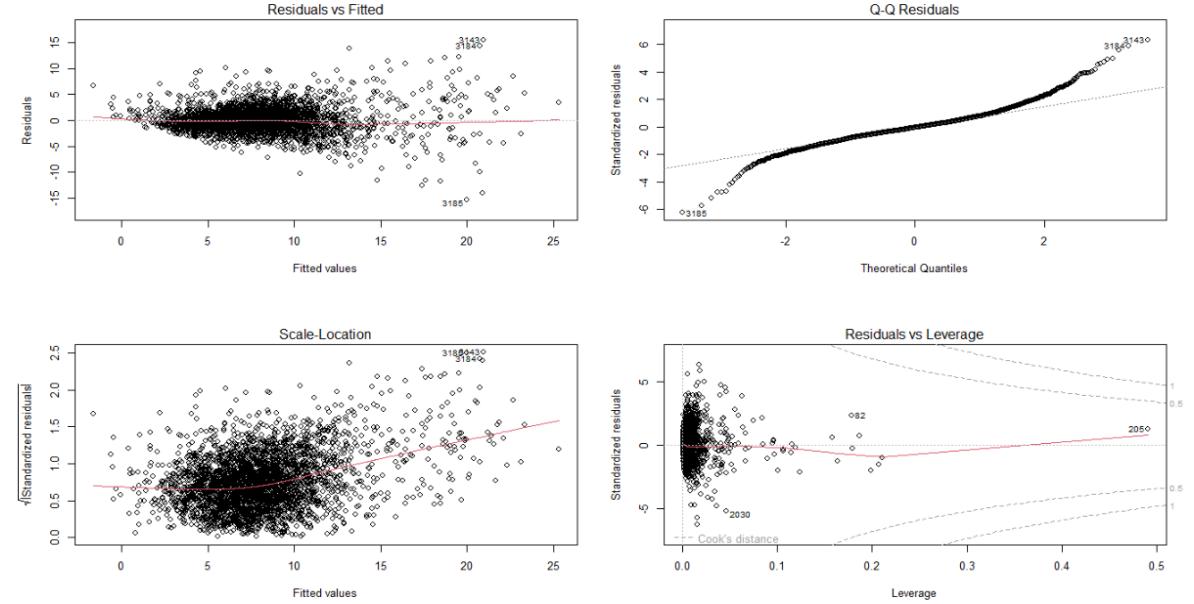
Metrics/Test	Y1	Y2
R-squared	0.9022807	0.6357207
Adjusted R-squared	0.9015461	0.6332124
AIC	62620.16	14999.82
BIC	62778.15	15145.65
Shapiro-Wilk normality test (p-value)	2.2e-16	2.2e-16
Durbin-Watson test (DW)	1.7407	1.7425
VIF(<5)	14	11

Although R-squared and Adjusted R-squared look impressive but this model fails in diagnosis as the Shapiro-Wilk normality test (p-value) is less than 0.05 and the Durbin-Watson test (DW) is less than 2, which indicates mild autocorrelation. Also, many predictors have a VIF greater than 5, which shows multicollinearity.

Y1



Y2

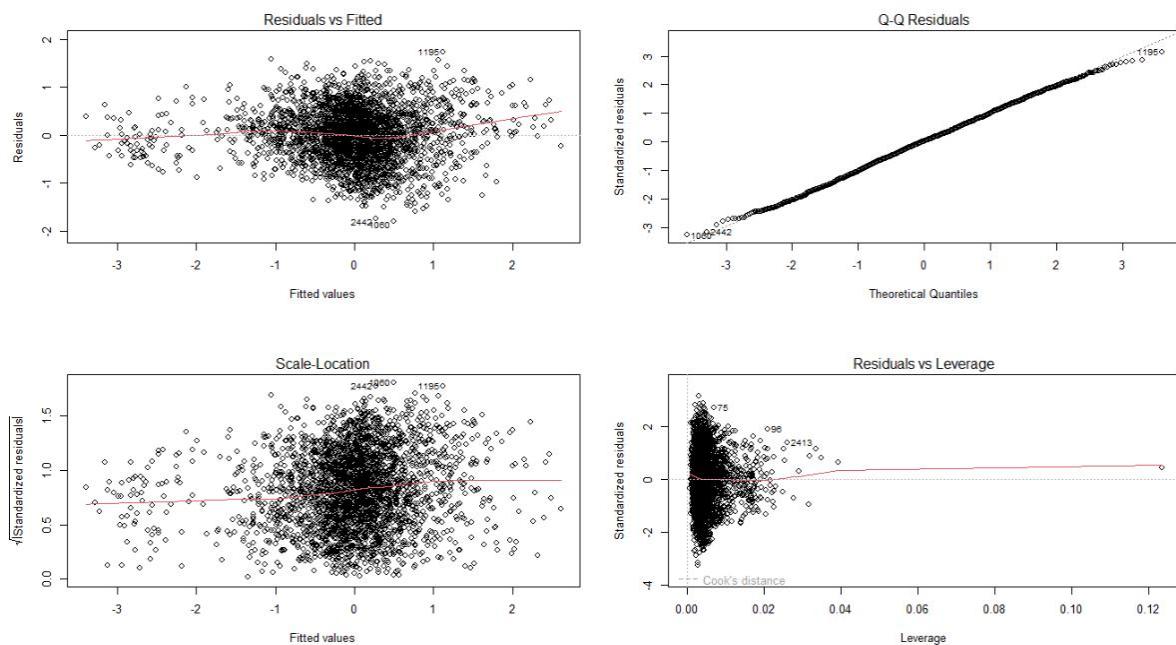


2.5.1 Base Model with main effect (After diagnosis)

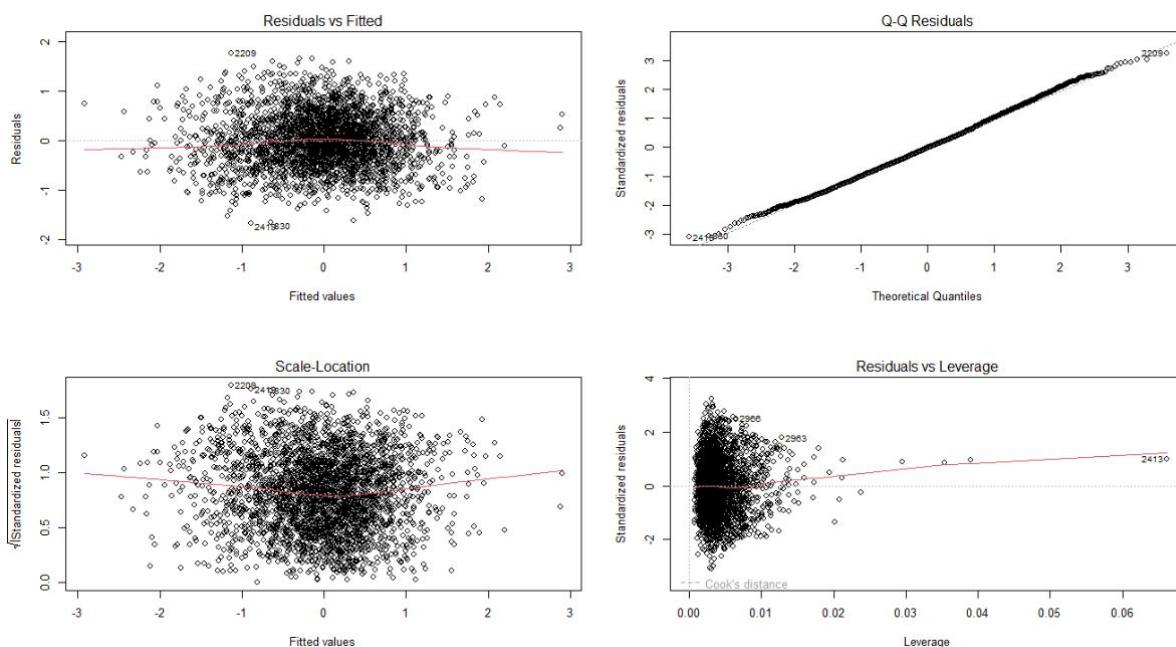
After the failure of the previous model, predictors with VIF greater than 5 were dropped. The Yeo-Johnson transformation is implemented for response variables Y1 and Y2, which is similar to the Box-Cox transformation. The Yeo-Johnson transformation is good for variables having zeros and negative values. Also, log transformation implemented for predictors. For outliers, handled using Cook's distance, for Y1 total is 223, and for Y2 total is 251. These data points were removed from the data.

Metrics/Test	Y1	Y2
R-squared	0.6333641	0.6018305
Adjusted R-squared	0.6316416	0.6003483
AIC	4965.617	4812.527
BIC	5061.692	4890.466
Shapiro-Wilk normality test (p-value)	0.05253	0.001644
Durbin-Watson test (DW)	1.6749	1.5639
VIF(<5)	14	11

Y1



Y2



2.6 Two main effects

In the last step, the base model with the main effect is combined with the two main effect models. Again, stepwise regression on both sides was implemented for all possible combinations of the two main effect models; later best combination of models was later selected, and multiple linear regression was implemented. Also, in this step another set of outliers dropped.

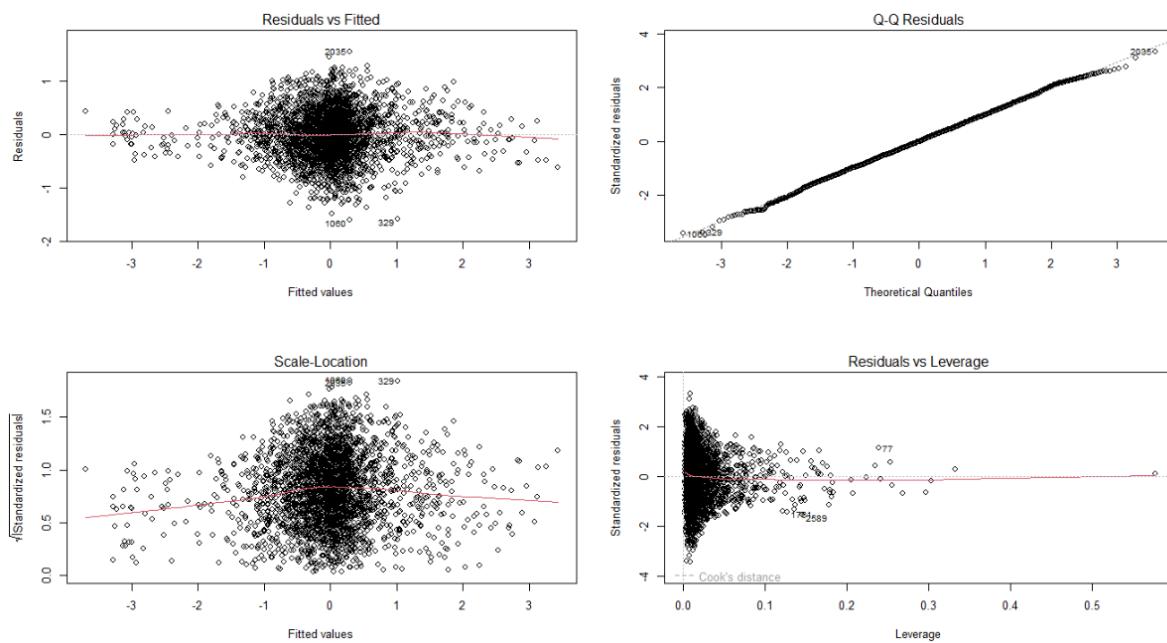
Metrics/Test	Y1	Y2
R-squared	0.7121303	0.6880194
Adjusted R-squared	0.705694	0.6835918
AIC	3820.55	3642.337
BIC	4201.359	3885.593
Shapiro-Wilk normality test (p-value)	0.3903	0.0707
Durbin-Watson test (DW)	1.7113	1.7045

3 Summary and final model

With two main effects model has improved a lot. There is mild autocorrelation in both Y1 and Y2, but considering the sample size, this mild autocorrelation can be accepted.

Y1

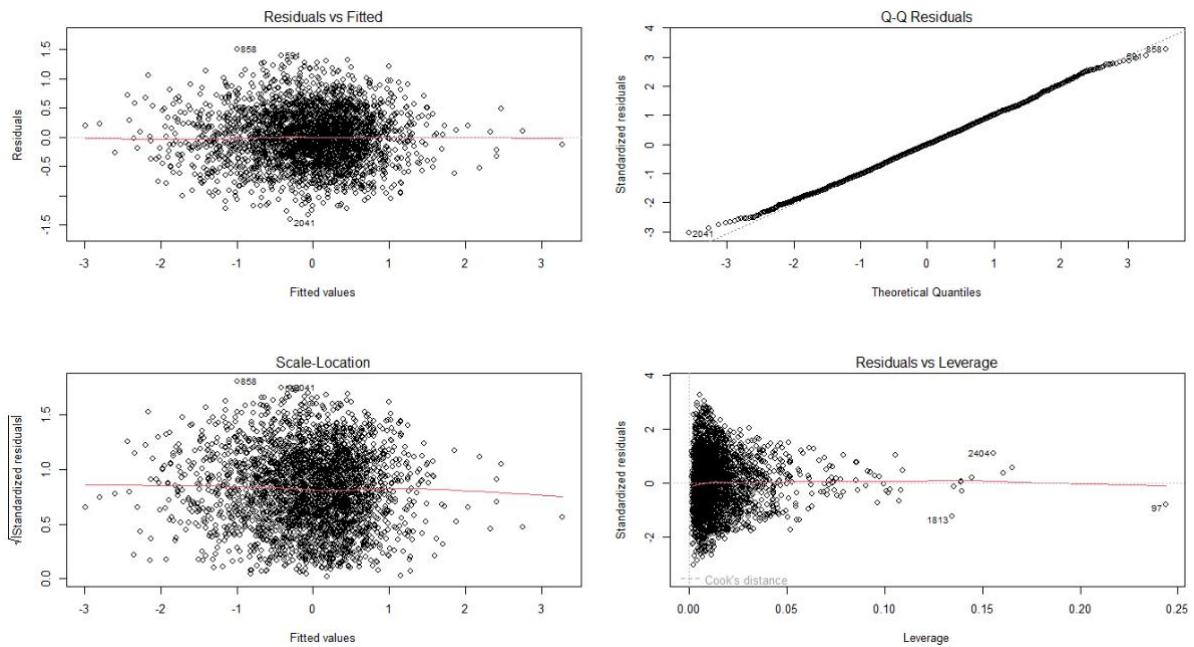
Metrics/Test	One main effect	Two main effects
R-squared	0.6333641	0.7121303
Adjusted R-squared	0.6316416	0.705694
AIC	4965.617	3820.55
BIC	5061.692	4201.359
Shapiro-Wilk normality test (p-value)	0.05253	0.3903
Durbin-Watson test (DW)	1.6749	1.7113



$$\begin{aligned}
Y \sim & White + Black + Native + Asian + Pacific + IncomeErr + IncomePerCapErr + \\
& Office + Construction + Production + Carpool + Walk + WorkAtHome + \\
& MeanCommute + White:Black + White:Native + White:Asian + \\
& White:IncomePerCapErr + White:Office + White:Construction + \\
& White:Production + White:Carpool + White:Walk + White:WorkAtHome + \\
& Black:Asian + Black:Pacific + Black:IncomeErr + Black:IncomePerCapErr + \\
& Black:Production + Black:WorkAtHome + Black:MeanCommute + \\
& Native:IncomePerCapErr + Native:Construction + Native:WorkAtHome + \\
& Native:MeanCommute + Asian:IncomeErr + Asian:Office + Asian:Production + \\
& Asian:Carpool + Asian:Walk + Asian:WorkAtHome + Asian:MeanCommute + \\
& Pacific:Production + Pacific:Carpool + Pacific:MeanCommute + \\
& IncomeErr:IncomePerCapErr + IncomeErr:Office + IncomeErr:Walk + \\
& IncomePerCapErr:Construction + IncomePerCapErr:Walk + Construction:Production + \\
& Construction:Carpool + Construction:Walk + Construction:WorkAtHome + \\
& Construction:MeanCommute + Production:Carpool + Production:Walk + \\
& Production:WorkAtHome + Production:MeanCommute + Carpool:Walk + \\
& Walk:MeanCommute + WorkAtHome:MeanCommute
\end{aligned}$$

Y2

Metrics/Test	One main effect	Two main effects
R-squared	0.6018305	0.6880194
Adjusted R-squared	0.6003483	0.6835918
AIC	4812.527	3642.337
BIC	4890.466	3885.593
Shapiro-Wilk normality test (p-value)	0.001644	0.0707
Durbin-Watson test (DW)	1.5639	1.7045



$$\begin{aligned}
Y \sim & Hispanic + Black + Native + IncomeErr + IncomePerCapErr + \\
& Service + Office + Production + MeanCommute + PrivateWork + \\
& SelfEmployed + Hispanic:Black + Hispanic:Native + Hispanic:Service + \\
& Hispanic:Production + Black:Native + Black:IncomePerCapErr + \\
& Black:Service + Black:Office + Black:MeanCommute + Native:Service + \\
& Native:PrivateWork + IncomeErr:IncomePerCapErr + IncomeErr:Service + \\
& IncomeErr:MeanCommute + IncomeErr:PrivateWork + IncomeErr:SelfEmployed + \\
& IncomePerCapErr:Office + IncomePerCapErr:Production + IncomePerCapErr:PrivateWork \\
& + Service:Office + Service:Production + Office:Production + \\
& Office:SelfEmployed + Production:MeanCommute + Production:PrivateWork + \\
& Production:SelfEmployed + MeanCommute:PrivateWork + MeanCommute:SelfEmployed
\end{aligned}$$

4 Conclusion

The best model results are given above. Mild autocorrelation is accepted because of the sample size. Overall plots look better from the earlier model, and can be improved with other transformation methods.

5 Appendix

The R Markdown has been attached. Below code section can be used to run the model for selected Y1/Y2.

```

```{r}
Y1 <- df$Income
Y2 <- df$Unemployment
Y=Y2 #Y=Y1
X <- df[, !(names(df) %in% c("Income", "Unemployment", "State", "County"))]
```

```

AMS578_project_116125547

Ajeetkumar Rai (Id : 116125547)

2025-04-30

```
#Library
```

```
required_packages <- c("tidyverse", "caret", "neuralnet", "ggplot2", "glmnet", "rpart", "rattle", "facto

for (pkg in required_packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)
    library(pkg, character.only = TRUE)}
```



```
## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3

## Warning: package 'dplyr' was built under R version 4.3.3

## Warning: package 'stringr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## vforcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyverse 1.3.1
## v purrr     1.0.4
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

## Warning: package 'caret' was built under R version 4.3.3
```

```

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##     lift

## Warning: package 'neuralnet' was built under R version 4.3.3

##
## Attaching package: 'neuralnet'
##
## The following object is masked from 'package:dplyr':
##     compute

## Warning: package 'glmnet' was built under R version 4.3.3

## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8

## Warning: package 'rpart' was built under R version 4.3.3

## Warning: package 'rattle' was built under R version 4.3.3

## Loading required package: bitops

## Warning: package 'bitops' was built under R version 4.3.3

##
## Attaching package: 'bitops'
##
## The following object is masked from 'package:Matrix':
##     %&%
##
## Rattle: A free graphical interface for data science with R.
## Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

## Warning: package 'factoextra' was built under R version 4.3.3

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

```

```

## Warning: package 'gridExtra' was built under R version 4.3.3

##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##   combine

## Warning: package 'corrplot' was built under R version 4.3.3

## corrplot 0.95 loaded

## Warning: package 'lmtest' was built under R version 4.3.3

## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##   as.Date, as.Date.numeric

## Warning: package 'car' was built under R version 4.3.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.3

##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##   recode
##
## The following object is masked from 'package:purrr':
##   some
##
## Registered S3 method overwritten by 'quantmod':
##   method           from
##   as.zoo.data.frame zoo

## Warning: package 'forecast' was built under R version 4.3.3

## Warning: package 'e1071' was built under R version 4.3.3

```

Data

```

df <- read.csv("C:/Users/Ajeet Rai/OneDrive/Desktop/SBU/Academics/Sem III/AMS 578 Regression Theory/Pro
head(df)

##      State County TotalPop   Men Women Hispanic White Black Native Asian
## 1 Alabama Autauga    55221 26745 28476      2.6 75.8 18.5  0.4  1.0
## 2 Alabama Baldwin    195121 95314 99807      4.5 83.1  9.5  0.6  0.7
## 3 Alabama Barbour   26932 14497 12435      4.6 46.2 46.7  0.2  0.4
## 4 Alabama Bibb       22604 12073 10531      2.2 74.5 21.4  0.4  0.1
## 5 Alabama Blount    57710 28512 29198      8.6 87.9  1.5  0.3  0.1
## 6 Alabama Bullock   10678  5660  5018      4.4 22.2 70.7  1.2  0.2
##      Pacific Citizen Income IncomeErr IncomePerCap IncomePerCapErr Poverty
## 1          0 40725 51281     2391     24974           1080      12.9
## 2          0 147695 50254     1263     27317            711      13.4
## 3          0 20714 32964     2973     16824            798      26.7
## 4          0 17495 38678     3995     18431           1618      16.8
## 5          0 42345 45813     3141     20532            708      16.7
## 6          0 8057 31938     5884     17580           2055      24.6
##      ChildPoverty Professional Service Office Construction Production Drive
## 1          18.6        33.2    17.0   24.2        8.6     17.1  87.5
## 2          19.2        33.1    17.7   27.1       10.8     11.2 84.7
## 3          45.3        26.8    16.1   23.1       10.8     23.1 83.8
## 4          27.9        21.5    17.9   17.8       19.0     23.7 83.2
## 5          27.2        28.5    14.1   23.9       13.5     19.9 84.9
## 6          38.4        18.8    15.0   19.7       20.1     26.4 74.9
##      Carpool Transit Walk OtherTransp WorkAtHome MeanCommute Employed PrivateWork
## 1          8.8       0.1  0.5      1.3      1.8     26.5 23986    73.6
## 2          8.8       0.1  1.0      1.4      3.9     26.4 85953    81.5
## 3         10.9       0.4  1.8      1.5      1.6     24.1 8597    71.8
## 4         13.5       0.5  0.6      1.5      0.7     28.8 8294    76.8
## 5         11.2       0.4  0.9      0.4      2.3     34.9 22189    82.0
## 6         14.9       0.7  5.0      1.7      2.8     27.5 3865    79.5
##      PublicWork SelfEmployed FamilyWork Unemployment
## 1          20.9       5.5  0.0      7.6
## 2          12.3       5.8  0.4      7.5
## 3          20.8       7.3  0.1     17.6
## 4          16.1       6.7  0.4      8.3
## 5          13.5       4.2  0.4      7.7
## 6          15.1       5.4  0.0     18.0

```

Summary

Shape

```
dim(df)
```

```
## [1] 3220 36
```

Datatypes

```
str(df)
```

```
## 'data.frame': 3220 obs. of 36 variables:
## $ State      : chr "Alabama" "Alabama" "Alabama" "Alabama" ...
## $ County     : chr "Autauga" "Baldwin" "Barbour" "Bibb" ...
## $ TotalPop   : int 55221 195121 26932 22604 57710 10678 20354 116648 34079 26008 ...
## $ Men        : int 26745 95314 14497 12073 28512 5660 9502 56274 16258 12975 ...
## $ Women      : int 28476 99807 12435 10531 29198 5018 10852 60374 17821 13033 ...
## $ Hispanic   : num 2.6 4.5 4.6 2.2 8.6 4.4 1.2 3.5 0.4 1.5 ...
## $ White       : num 75.8 83.1 46.2 74.5 87.9 22.2 53.3 73 57.3 91.7 ...
## $ Black       : num 18.5 9.5 46.7 21.4 1.5 70.7 43.8 20.3 40.3 4.8 ...
## $ Native      : num 0.4 0.6 0.2 0.4 0.3 1.2 0.1 0.2 0.2 0.6 ...
## $ Asian       : num 1 0.7 0.4 0.1 0.1 0.2 0.4 0.9 0.8 0.3 ...
## $ Pacific     : num 0 0 0 0 0 0 0 0 0 0 ...
## $ Citizen    : int 40725 147695 20714 17495 42345 8057 15581 88612 26462 20600 ...
## $ Income      : int 51281 50254 32964 38678 45813 31938 32229 41703 34177 36296 ...
## $ IncomeErr   : int 2391 1263 2973 3995 3141 5884 1793 925 2949 1710 ...
## $ IncomePerCap: int 24974 27317 16824 18431 20532 17580 18390 21374 21071 21811 ...
## $ IncomePerCapErr: int 1080 711 798 1618 708 2055 714 489 1366 1556 ...
## $ Poverty     : num 12.9 13.4 26.7 16.8 16.7 24.6 25.4 20.5 21.6 19.2 ...
## $ ChildPoverty: num 18.6 19.2 45.3 27.9 27.2 38.4 39.2 31.6 37.2 30.1 ...
## $ Professional: num 33.2 33.1 26.8 21.5 28.5 18.8 27.5 27.3 23.3 29.3 ...
## $ Service     : num 17 17.7 16.1 17.9 14.1 15 16.6 17.7 14.5 16 ...
## $ Office      : num 24.2 27.1 23.1 17.8 23.9 19.7 21.9 24.2 26.3 19.5 ...
## $ Construction: num 8.6 10.8 10.8 19 13.5 20.1 10.3 10.5 11.5 13.7 ...
## $ Production  : num 17.1 11.2 23.1 23.7 19.9 26.4 23.7 20.4 24.4 21.5 ...
## $ Drive       : num 87.5 84.7 83.8 83.2 84.9 74.9 84.5 85.3 85.1 83.9 ...
## $ Carpool     : num 8.8 8.8 10.9 13.5 11.2 14.9 12.4 9.4 11.9 12.1 ...
## $ Transit     : num 0.1 0.1 0.4 0.5 0.4 0.7 0 0.2 0.2 0.2 ...
## $ Walk        : num 0.5 1 1.8 0.6 0.9 5 0.8 1.2 0.3 0.6 ...
## $ OtherTransp : num 1.3 1.4 1.5 1.5 0.4 1.7 0.6 1.2 0.4 0.7 ...
## $ WorkAtHome  : num 1.8 3.9 1.6 0.7 2.3 2.8 1.7 2.7 2.1 2.5 ...
## $ MeanCommute : num 26.5 26.4 24.1 28.8 34.9 27.5 24.6 24.1 25.1 27.4 ...
## $ Employed    : int 23986 85953 8597 8294 22189 3865 7813 47401 13689 10155 ...
## $ PrivateWork : num 73.6 81.5 71.8 76.8 82 79.5 77.4 74.1 85.1 73.1 ...
## $ PublicWork  : num 20.9 12.3 20.8 16.1 13.5 15.1 16.2 20.8 12.1 18.5 ...
## $ SelfEmployed: num 5.5 5.8 7.3 6.7 4.2 5.4 6.2 5 2.8 7.9 ...
## $ FamilyWork  : num 0 0.4 0.1 0.4 0.4 0 0.2 0.1 0 0.5 ...
## $ Unemployment: num 7.6 7.5 17.6 8.3 7.7 18 10.9 12.3 8.9 7.9 ...
```

Descriptions

```
summary(df)
```

```
##      State          County        TotalPop         Men      
## Length:3220    Length:3220    Min.   :    85   Min.   :    42  
## Class :character Class :character  1st Qu.: 11218   1st Qu.: 5637  
## Mode  :character Mode  :character  Median : 26035  Median : 12932
```

```

##                                     Mean   : 99409   Mean   : 48897
##                                     3rd Qu.: 66430   3rd Qu.: 32993
##                                     Max.   :10038388   Max.   :4945351
##
##      Women          Hispanic        White        Black
## Min.   :    43   Min.   : 0.000   Min.   : 0.00   Min.   : 0.000
## 1st Qu.: 5572   1st Qu.: 1.900   1st Qu.:64.10  1st Qu.: 0.500
## Median : 13057  Median : 3.900   Median :84.10  Median : 1.900
## Mean   : 50512  Mean   :11.012   Mean   :75.43  Mean   : 8.665
## 3rd Qu.: 33488  3rd Qu.: 9.825   3rd Qu.:93.20  3rd Qu.: 9.600
## Max.   :5093037 Max.   :99.900   Max.   :99.80  Max.   :85.900
##
##      Native         Asian        Pacific       Citizen
## Min.   : 0.000   Min.   : 0.000   Min.   : 0.00000   Min.   :     80
## 1st Qu.: 0.100   1st Qu.: 0.200   1st Qu.: 0.00000   1st Qu.: 8450
## Median : 0.300   Median : 0.500   Median : 0.00000   Median : 19643
## Mean   : 1.724   Mean   : 1.229   Mean   : 0.08273   Mean   : 69935
## 3rd Qu.: 0.600   3rd Qu.: 1.200   3rd Qu.: 0.00000   3rd Qu.: 49920
## Max.   :92.100   Max.   :41.600   Max.   :35.30000   Max.   :6046749
##
##      Income        IncomeErr     IncomePerCap  IncomePerCapErr
## Min.   : 10499   Min.   : 270   Min.   : 5878   Min.   : 113
## 1st Qu.: 38192   1st Qu.: 1635  1st Qu.:20239  1st Qu.: 755
## Median : 44749   Median : 2406  Median :23460   Median : 1096
## Mean   : 46130   Mean   : 2850  Mean   :23982   Mean   : 1363
## 3rd Qu.: 52074   3rd Qu.: 3446  3rd Qu.:27053  3rd Qu.: 1631
## Max.   :123453   Max.   :21355  Max.   :65600   Max.   :15266
## NA's   :1
##
##      Poverty       ChildPoverty  Professional     Service
## Min.   : 1.40   Min.   : 0.00   Min.   :13.50   Min.   : 5.00
## 1st Qu.:12.10  1st Qu.:16.30  1st Qu.:26.70  1st Qu.:16.00
## Median :16.15  Median :22.70  Median :29.90  Median :18.10
## Mean   :17.49  Mean   :24.18  Mean   :30.99  Mean   :18.35
## 3rd Qu.:20.70  3rd Qu.:30.00  3rd Qu.:34.40  3rd Qu.:20.30
## Max.   :64.20   Max.   :81.60  Max.   :74.00  Max.   :38.20
## NA's   :1
##
##      Office        Construction Production     Drive
## Min.   : 4.10   Min.   : 1.70   Min.   : 0.00   Min.   : 5.20
## 1st Qu.:20.20  1st Qu.: 9.80  1st Qu.:11.50  1st Qu.:76.60
## Median :22.40  Median :12.10  Median :15.25  Median :80.70
## Mean   :22.22  Mean   :12.71  Mean   :15.73  Mean   :79.18
## 3rd Qu.:24.40  3rd Qu.:14.90  3rd Qu.:19.32  3rd Qu.:83.70
## Max.   :35.40   Max.   :40.30  Max.   :55.60  Max.   :94.60
##
##      Carpool       Transit        Walk        OtherTransp
## Min.   : 0.00   Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000
## 1st Qu.: 8.40   1st Qu.: 0.1000  1st Qu.: 1.400  1st Qu.: 0.900
## Median : 9.90   Median : 0.4000  Median : 2.400  Median : 1.300
## Mean   :10.28   Mean   : 0.9718  Mean   : 3.324  Mean   : 1.613
## 3rd Qu.:11.80   3rd Qu.: 0.8000  3rd Qu.: 4.000  3rd Qu.: 1.900
## Max.   :29.90   Max.   :61.7000  Max.   :71.200  Max.   :39.100
##
##      WorkAtHome    MeanCommute    Employed     PrivateWork
## Min.   : 0.000   Min.   : 4.90   Min.   :     62  Min.   :25.00

```

```

## 1st Qu.: 2.700 1st Qu.:19.50 1st Qu.: 4551 1st Qu.:70.50
## Median : 3.900 Median :23.00 Median : 10508 Median :75.70
## Mean   : 4.632 Mean   :23.28 Mean   : 45594 Mean   :74.22
## 3rd Qu.: 5.600 3rd Qu.:26.80 3rd Qu.: 28633 3rd Qu.:79.70
## Max.   :37.200 Max.   :44.00 Max.   :4635465 Max.   :88.30
##
##   PublicWork    SelfEmployed    FamilyWork    Unemployment
## Min.   : 5.80  Min.   :0.0000  Min.   :0.0000  Min.   : 0.000
## 1st Qu.:13.10 1st Qu.: 5.400  1st Qu.:0.1000  1st Qu.: 5.500
## Median :16.20  Median : 6.900  Median :0.2000  Median : 7.600
## Mean   :17.56  Mean   : 7.932  Mean   :0.2881  Mean   : 8.094
## 3rd Qu.:20.50 3rd Qu.: 9.400  3rd Qu.:0.3000  3rd Qu.: 9.900
## Max.   :66.20  Max.   :36.600  Max.   :9.8000  Max.   :36.500
##

```

Imputing missing values

```

dim(df)

## [1] 3220 36

df = na.omit(df)
dim(df)

## [1] 3218 36

```

Independent vs dependent

```

Y1 <- df$Income
Y2 <- df$Unemployment
Y=Y1
X <- df[, !(names(df) %in% c("Income", "Unemployment", "State", "County"))]

```

Exploratory data analysis

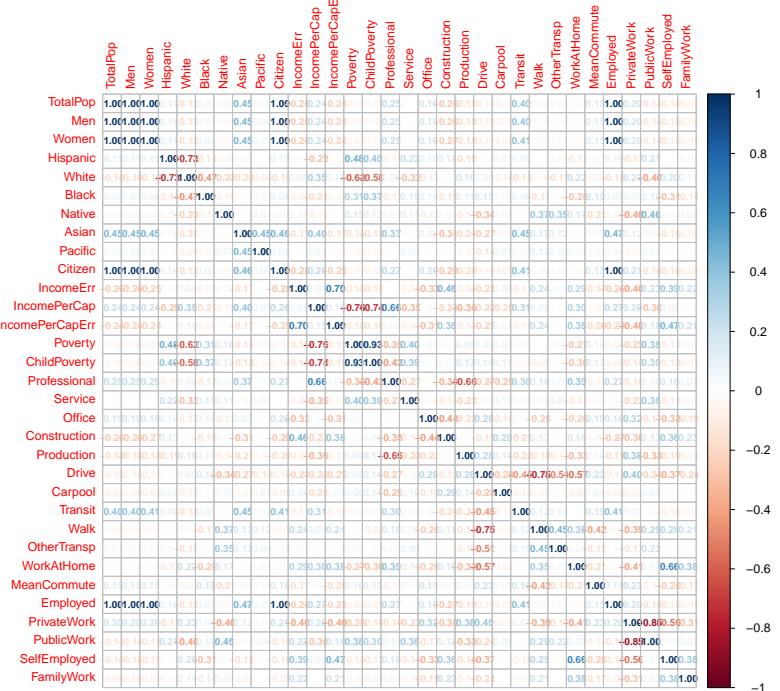
Correlation

As we can see collinearity between features, we'll need check VIF carefully.

```

correlation_matrix <- cor(X)
corrplot(correlation_matrix, method = "number", number.cex = 0.7, tl.cex = 0.8)

```



Distributin (Dependent)

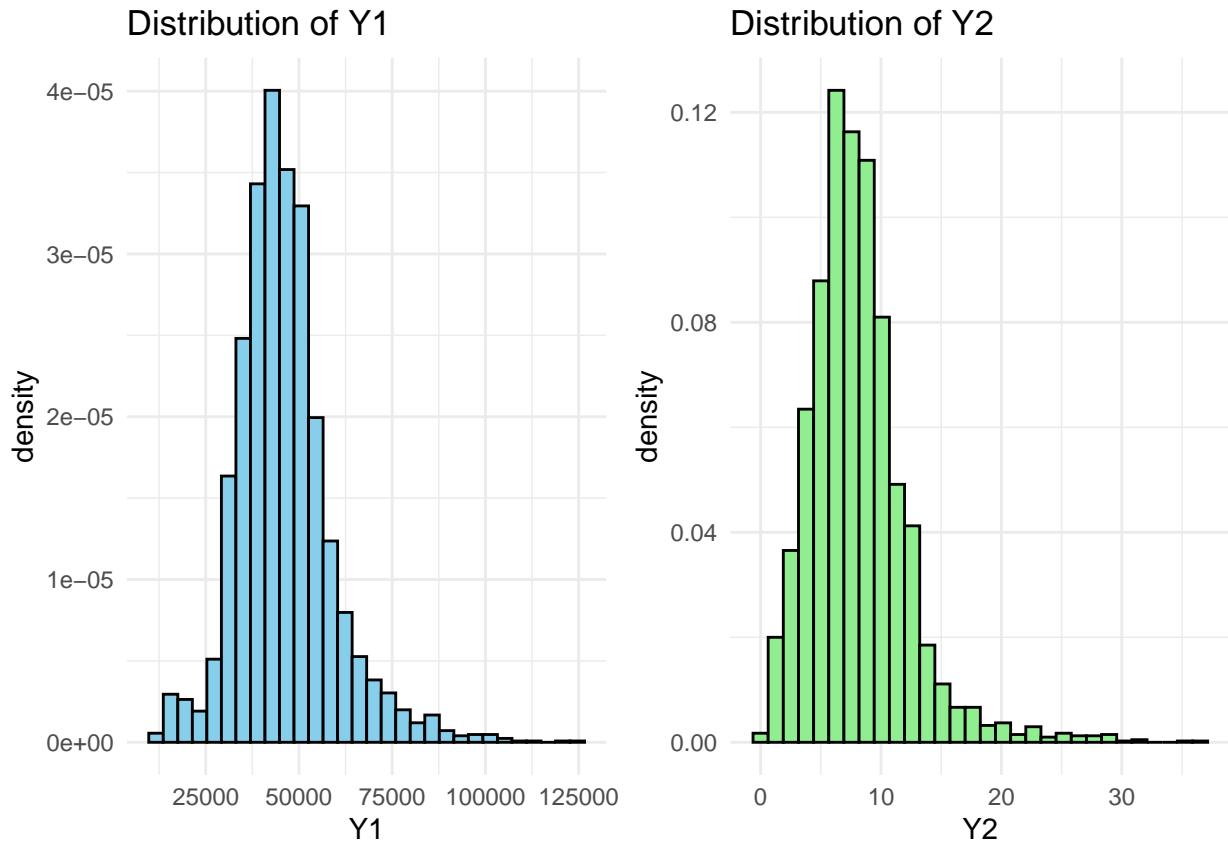
Y1 and Y2 are slightly skewed, and transformation is needed to prevent this.

```
p1 <- ggplot(data = data.frame(Y1), aes(x = Y1)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'skyblue', color = 'black') +
  ggtitle('Distribution of Y1') +
  theme_minimal()

p2 <- ggplot(data = data.frame(Y2), aes(x = Y2)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'lightgreen', color = 'black') +
  ggtitle('Distribution of Y2') +
  theme_minimal()

grid.arrange(p1, p2, ncol = 2)
```

```
## Warning: The dot-dot notation ('..density..') was deprecated in ggplot2 3.4.0.
## i Please use 'after_stat(density)' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



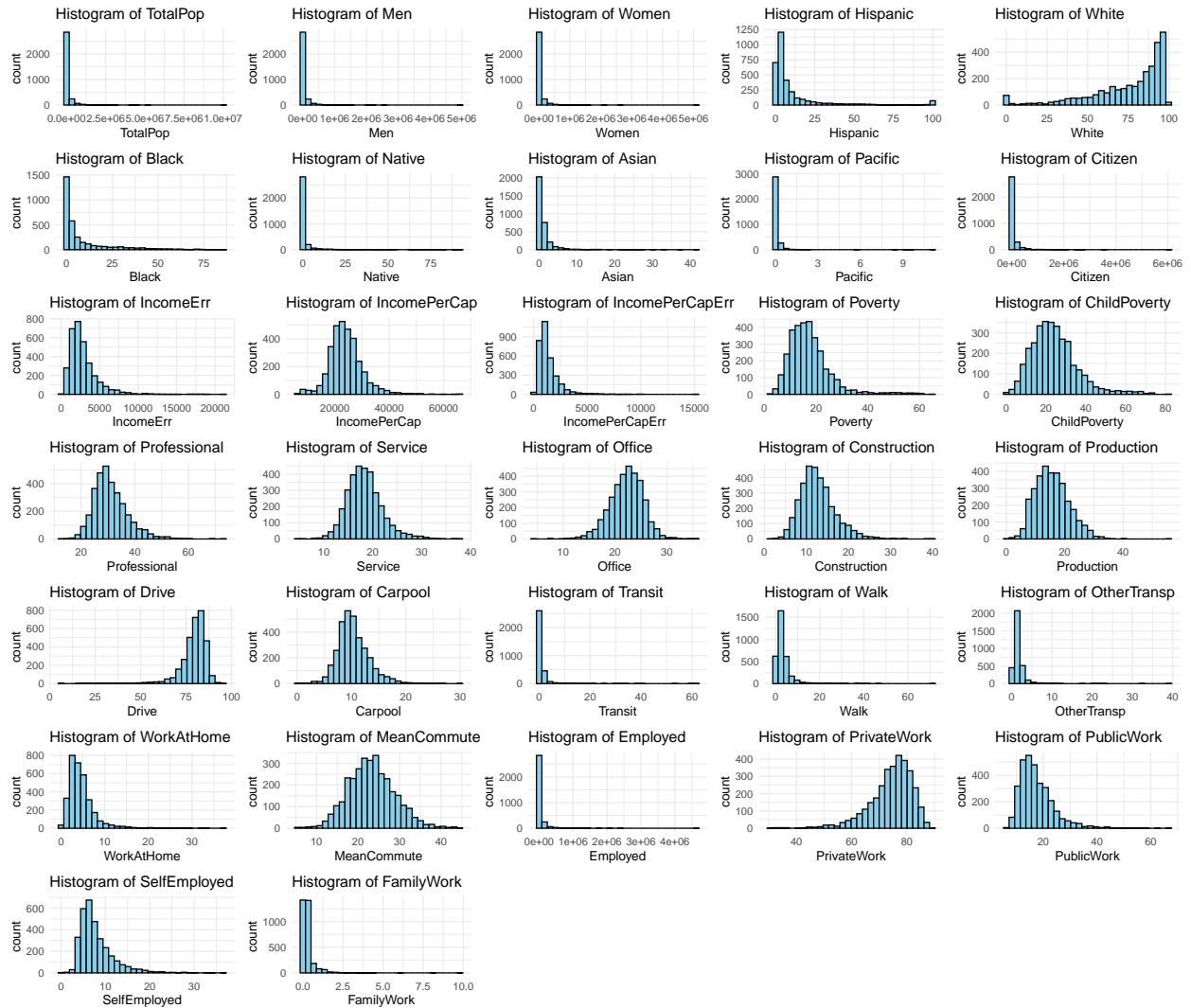
Distributin (Independent)

Independent features are skewed, and transformation is needed to prevent this.

```
plots <- lapply(names(X), function(colname) {
  if (is.numeric(X[[colname]])) {
    ggplot(X, aes_string(x = colname)) +
      geom_histogram(bins = 30, fill = "skyblue", color = "black") +
      ggttitle(paste("Histogram of", colname)) +
      theme_minimal() } else {
    ggplot(X, aes_string(x = colname)) +
      geom_bar(fill = "lightgreen", color = "black") +
      ggttitle(paste("Bar Plot of", colname)) +
      theme_minimal()}})

## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`'.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
grid.arrange(grobs = plots, ncol = 5)
```

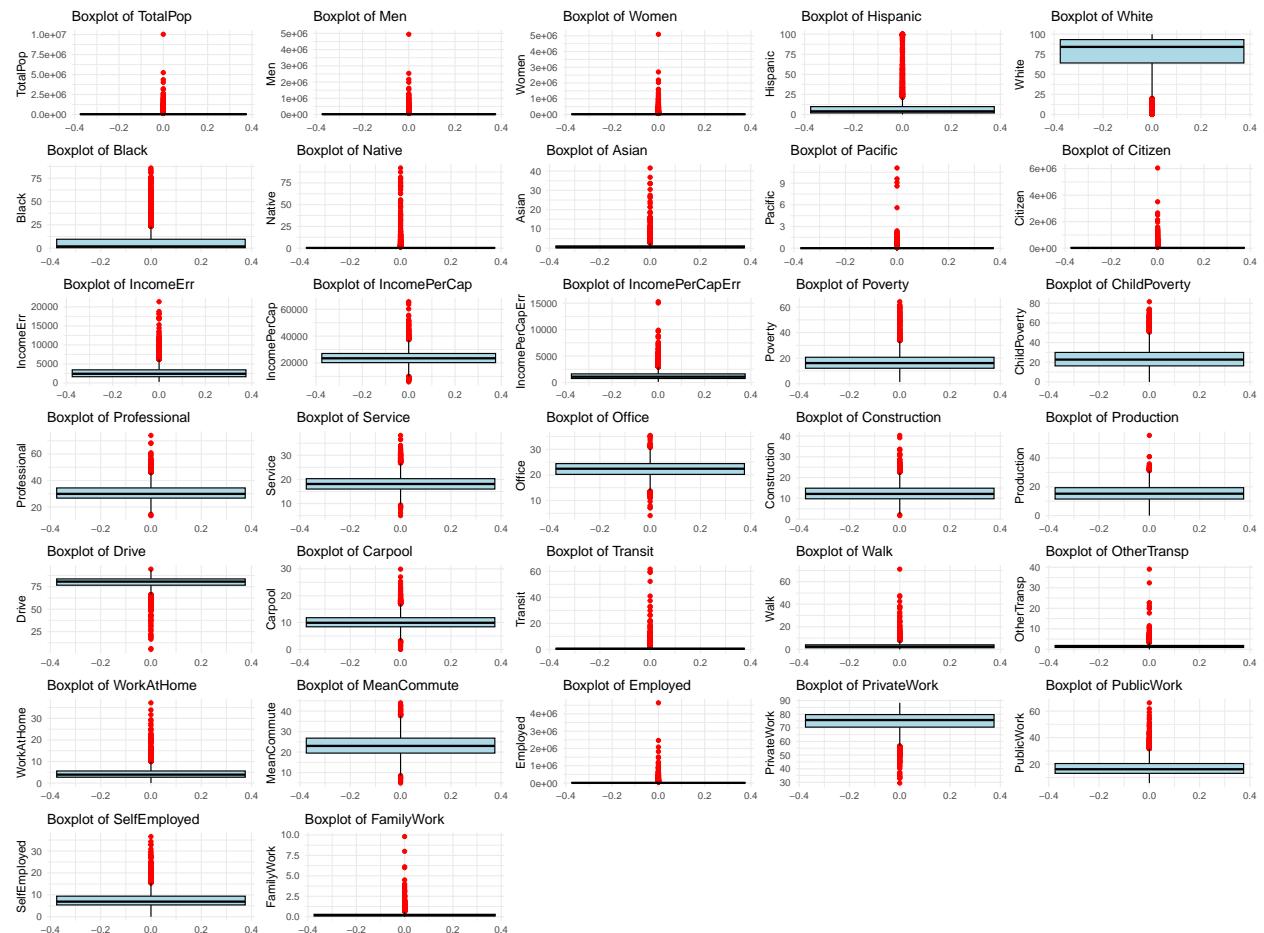


Outliers

Need careful treatment for outliers as it will effect cook's distance

```
plots <- lapply(names(X), function(colname) {
  if (is.numeric(X[[colname]])) {
    ggplot(X, aes_string(y = colname)) +
      geom_boxplot(fill = "lightblue", color = "black", outlier.color = "red") +
      ggtitle(paste("Boxplot of", colname)) +
      theme_minimal()
  } else {
    NULL
  }
})
plots <- Filter(Negate(is.null), plots)
```

```
grid.arrange(grobs = plots, ncol = 5)
```



Feature selection

As we have seen above so many features are correlated and among 34 features many of them are not contributing in Y1/Y2.

So, we will use STEP wise model in both direction to selected only meaningful features.

```
full_model <- lm(Y ~ ., data = X)
stepwise_model_both <- step(full_model, direction = "both")
```

```
## Start: AIC=53496.15
## Y ~ TotalPop + Men + Women + Hispanic + White + Black + Native +
##      Asian + Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##      Poverty + ChildPoverty + Professional + Service + Office +
##      Construction + Production + Drive + Carpool + Transit + Walk +
```

```

##      OtherTransp + WorkAtHome + MeanCommute + Employed + PrivateWork +
##      PublicWork + SelfEmployed + FamilyWork
##
##
## Step: AIC=53496.15
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##      Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##      Poverty + ChildPoverty + Professional + Service + Office +
##      Construction + Production + Drive + Carpool + Transit + Walk +
##      OtherTransp + WorkAtHome + MeanCommute + Employed + PrivateWork +
##      PublicWork + SelfEmployed + FamilyWork
##
##          Df  Sum of Sq      RSS     AIC
## - Service      1 1.9808e+05 5.2321e+10 53494
## - Office       1 9.5063e+05 5.2322e+10 53494
## - Production   1 1.2568e+06 5.2322e+10 53494
## - Transit      1 1.4236e+06 5.2322e+10 53494
## - OtherTransp   1 1.5730e+06 5.2323e+10 53494
## - Construction 1 1.8211e+06 5.2323e+10 53494
## - SelfEmployed 1 2.1011e+06 5.2323e+10 53494
## - Hispanic     1 2.2760e+06 5.2323e+10 53494
## - Professional 1 2.6270e+06 5.2324e+10 53494
## - Drive        1 2.9825e+06 5.2324e+10 53494
## - Walk         1 3.0939e+06 5.2324e+10 53494
## - Carpool      1 3.5828e+06 5.2325e+10 53494
## - WorkAtHome    1 4.3709e+06 5.2325e+10 53494
## - FamilyWork    1 4.4875e+06 5.2326e+10 53494
## - PrivateWork   1 8.5938e+06 5.2330e+10 53495
## - PublicWork    1 9.1306e+06 5.2330e+10 53495
## - Black         1 2.0583e+07 5.2342e+10 53495
## - Employed      1 3.1169e+07 5.2352e+10 53496
## <none>           5.2321e+10 53496
## - Native        1 4.6610e+07 5.2368e+10 53497
## - White         1 5.4779e+07 5.2376e+10 53498
## - Pacific        1 1.2380e+08 5.2445e+10 53502
## - TotalPop      1 3.3719e+08 5.2658e+10 53515
## - ChildPoverty   1 3.8681e+08 5.2708e+10 53518
## - Citizen        1 5.2462e+08 5.2846e+10 53526
## - Asian          1 5.5916e+08 5.2880e+10 53528
## - Men            1 6.0032e+08 5.2921e+10 53531
## - IncomeErr      1 1.1968e+09 5.3518e+10 53567
## - MeanCommute    1 3.8026e+09 5.6124e+10 53720
## - Poverty         1 3.9329e+09 5.6254e+10 53727
## - IncomePerCapErr 1 4.5101e+09 5.6831e+10 53760
## - IncomePerCap   1 3.5643e+10 8.7964e+10 55166
##
## Step: AIC=53494.16
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##      Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##      Poverty + ChildPoverty + Professional + Office + Construction +
##      Production + Drive + Carpool + Transit + Walk + OtherTransp +
##      WorkAtHome + MeanCommute + Employed + PrivateWork + PublicWork +
##      SelfEmployed + FamilyWork
##

```

```

##                                     Df Sum of Sq      RSS     AIC
## - Transit                         1 1.4335e+06 5.2323e+10 53492
## - OtherTransp                      1 1.5835e+06 5.2323e+10 53492
## - SelfEmployed                     1 2.0763e+06 5.2323e+10 53492
## - Hispanic                         1 2.2555e+06 5.2324e+10 53492
## - Drive                            1 2.9977e+06 5.2324e+10 53492
## - Walk                             1 3.1100e+06 5.2324e+10 53492
## - Carpool                          1 3.6008e+06 5.2325e+10 53492
## - WorkAtHome                       1 4.3882e+06 5.2326e+10 53492
## - FamilyWork                        1 4.4590e+06 5.2326e+10 53492
## - PrivateWork                       1 8.5465e+06 5.2330e+10 53493
## - PublicWork                        1 9.0820e+06 5.2330e+10 53493
## - Black                            1 2.0644e+07 5.2342e+10 53493
## - Employed                          1 3.1151e+07 5.2352e+10 53494
## <none>                            5.2321e+10 53494
## - Native                           1 4.6544e+07 5.2368e+10 53495
## - White                            1 5.4881e+07 5.2376e+10 53496
## + Service                          1 1.9808e+05 5.2321e+10 53496
## - Pacific                           1 1.2381e+08 5.2445e+10 53500
## - Office                            1 2.9460e+08 5.2616e+10 53510
## - TotalPop                         1 3.3759e+08 5.2659e+10 53513
## - ChildPoverty                     1 3.8676e+08 5.2708e+10 53516
## - Citizen                           1 5.2499e+08 5.2846e+10 53524
## - Asian                            1 5.5896e+08 5.2880e+10 53526
## - Men                               1 6.0104e+08 5.2922e+10 53529
## - Production                        1 8.8543e+08 5.3207e+10 53546
## - Construction                      1 1.1263e+09 5.3447e+10 53561
## - IncomeErr                         1 1.1971e+09 5.3518e+10 53565
## - Professional                      1 2.3568e+09 5.4678e+10 53634
## - MeanCommute                       1 3.8039e+09 5.6125e+10 53718
## - Poverty                           1 3.9334e+09 5.6255e+10 53725
## - IncomePerCapErr                  1 4.5102e+09 5.6831e+10 53758
## - IncomePerCap                      1 3.5663e+10 8.7984e+10 55165
##
## Step:  AIC=53492.25
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##     Poverty + ChildPoverty + Professional + Office + Construction +
##     Production + Drive + Carpool + Walk + OtherTransp + WorkAtHome +
##     MeanCommute + Employed + PrivateWork + PublicWork + SelfEmployed +
##     FamilyWork
##
##                                     Df Sum of Sq      RSS     AIC
## - OtherTransp                      1 1.0917e+06 5.2324e+10 53490
## - SelfEmployed                     1 2.1242e+06 5.2325e+10 53490
## - Hispanic                          1 2.2783e+06 5.2325e+10 53490
## - FamilyWork                        1 4.5138e+06 5.2327e+10 53491
## - PrivateWork                       1 8.6434e+06 5.2331e+10 53491
## - PublicWork                        1 9.1840e+06 5.2332e+10 53491
## - Black                            1 2.0603e+07 5.2343e+10 53492
## - Employed                          1 3.0905e+07 5.2354e+10 53492
## <none>                            5.2323e+10 53492
## - Native                           1 4.6584e+07 5.2369e+10 53493
## - White                            1 5.4788e+07 5.2377e+10 53494

```

```

## + Transit          1 1.4335e+06 5.2321e+10 53494
## + Service         1 2.0800e+05 5.2322e+10 53494
## - Pacific          1 1.2415e+08 5.2447e+10 53498
## - Walk             1 1.7202e+08 5.2495e+10 53501
## - Drive             1 2.7629e+08 5.2599e+10 53507
## - Carpool           1 2.9082e+08 5.2614e+10 53508
## - Office            1 2.9562e+08 5.2618e+10 53508
## - TotalPop          1 3.3863e+08 5.2661e+10 53511
## - WorkAtHome        1 3.7853e+08 5.2701e+10 53513
## - ChildPoverty      1 3.8679e+08 5.2709e+10 53514
## - Citizen            1 5.2464e+08 5.2847e+10 53522
## - Asian              1 5.6018e+08 5.2883e+10 53525
## - Men                1 6.0217e+08 5.2925e+10 53527
## - Production         1 8.8666e+08 5.3209e+10 53544
## - Construction       1 1.1256e+09 5.3448e+10 53559
## - IncomeErr          1 1.1968e+09 5.3519e+10 53563
## - Professional        1 2.3562e+09 5.4679e+10 53632
## - MeanCommute         1 3.8029e+09 5.6126e+10 53716
## - Poverty             1 3.9346e+09 5.6257e+10 53724
## - IncomePerCapErr    1 4.5089e+09 5.6832e+10 53756
## - IncomePerCap        1 3.5664e+10 8.7987e+10 55163
##
## Step: AIC=53490.32
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##     Poverty + ChildPoverty + Professional + Office + Construction +
##     Production + Drive + Carpool + Walk + WorkAtHome + MeanCommute +
##     Employed + PrivateWork + PublicWork + SelfEmployed + FamilyWork
##
## Df   Sum of Sq      RSS     AIC
## - SelfEmployed      1 2.2034e+06 5.2326e+10 53488
## - Hispanic          1 2.2763e+06 5.2326e+10 53488
## - FamilyWork         1 4.6060e+06 5.2328e+10 53489
## - PrivateWork        1 8.8086e+06 5.2333e+10 53489
## - PublicWork         1 9.3541e+06 5.2333e+10 53489
## - Black              1 2.0543e+07 5.2344e+10 53490
## - Employed           1 3.1636e+07 5.2355e+10 53490
## <none>                  5.2324e+10 53490
## - Native             1 4.7236e+07 5.2371e+10 53491
## - White              1 5.4828e+07 5.2379e+10 53492
## + OtherTransp        1 1.0917e+06 5.2323e+10 53492
## + Transit            1 9.4177e+05 5.2323e+10 53492
## + Service            1 2.0842e+05 5.2324e+10 53492
## - Pacific            1 1.2329e+08 5.2447e+10 53496
## - Walk               1 1.7772e+08 5.2501e+10 53499
## - Office              1 3.0476e+08 5.2629e+10 53507
## - Carpool             1 3.3661e+08 5.2660e+10 53509
## - TotalPop            1 3.8389e+08 5.2708e+10 53512
## - Drive               1 3.8598e+08 5.2710e+10 53512
## - ChildPoverty        1 3.8676e+08 5.2711e+10 53512
## - WorkAtHome          1 4.3877e+08 5.2763e+10 53515
## - Citizen             1 5.2444e+08 5.2848e+10 53520
## - Asian               1 5.6132e+08 5.2885e+10 53523
## - Men                 1 6.8535e+08 5.3009e+10 53530

```

```

## - Production      1 8.9190e+08 5.3216e+10 53543
## - Construction   1 1.1322e+09 5.3456e+10 53557
## - IncomeErr       1 1.1965e+09 5.3520e+10 53561
## - Professional    1 2.3631e+09 5.4687e+10 53630
## - MeanCommute     1 3.8397e+09 5.6163e+10 53716
## - Poverty          1 3.9461e+09 5.6270e+10 53722
## - IncomePerCapErr 1 4.5089e+09 5.6833e+10 53754
## - IncomePerCap     1 3.5812e+10 8.8136e+10 55166
##
## Step: AIC=53488.46
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##      Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##      Poverty + ChildPoverty + Professional + Office + Construction +
##      Production + Drive + Carpool + Walk + WorkAtHome + MeanCommute +
##      Employed + PrivateWork + PublicWork + FamilyWork
##
##              Df  Sum of Sq      RSS      AIC
## - Hispanic      1 2.2615e+06 5.2328e+10 53487
## - Black         1 2.0616e+07 5.2347e+10 53488
## - FamilyWork    1 2.2413e+07 5.2348e+10 53488
## - Employed      1 3.1971e+07 5.2358e+10 53488
## <none>           5.2326e+10 53488
## - Native        1 4.7146e+07 5.2373e+10 53489
## - White         1 5.4941e+07 5.2381e+10 53490
## + SelfEmployed   1 2.2034e+06 5.2324e+10 53490
## + OtherTransp    1 1.1709e+06 5.2325e+10 53490
## + Transit        1 1.0129e+06 5.2325e+10 53490
## + Service        1 1.8240e+05 5.2326e+10 53490
## - Pacific        1 1.2444e+08 5.2450e+10 53494
## - Walk           1 1.7731e+08 5.2503e+10 53497
## - Office          1 3.0640e+08 5.2632e+10 53505
## - Carpool         1 3.3684e+08 5.2663e+10 53507
## - TotalPop        1 3.8395e+08 5.2710e+10 53510
## - Drive           1 3.8518e+08 5.2711e+10 53510
## - ChildPoverty    1 3.8641e+08 5.2712e+10 53510
## - WorkAtHome      1 4.3756e+08 5.2764e+10 53513
## - Citizen         1 5.2344e+08 5.2849e+10 53518
## - Asian            1 5.6189e+08 5.2888e+10 53521
## - Men              1 6.8555e+08 5.3012e+10 53528
## - Production       1 8.9403e+08 5.3220e+10 53541
## - Construction     1 1.1337e+09 5.3460e+10 53555
## - IncomeErr        1 1.1949e+09 5.3521e+10 53559
## - Professional      1 2.3681e+09 5.4694e+10 53629
## - MeanCommute      1 3.8414e+09 5.6167e+10 53714
## - PublicWork        1 3.9045e+09 5.6230e+10 53718
## - Poverty           1 3.9543e+09 5.6280e+10 53721
## - PrivateWork       1 4.1319e+09 5.6458e+10 53731
## - IncomePerCapErr   1 4.5067e+09 5.6833e+10 53752
## - IncomePerCap      1 3.5834e+10 8.8160e+10 55165
##
## Step: AIC=53486.59
## Y ~ TotalPop + Men + White + Black + Native + Asian + Pacific +
##      Citizen + IncomeErr + IncomePerCap + IncomePerCapErr + Poverty +
##      ChildPoverty + Professional + Office + Construction + Production +

```

```

##      Drive + Carpool + Walk + WorkAtHome + MeanCommute + Employed +
##      PrivateWork + PublicWork + FamilyWork
##
##          Df  Sum of Sq      RSS     AIC
## - FamilyWork      1 2.2334e+07 5.2351e+10 53486
## - Employed        1 3.1472e+07 5.2360e+10 53487
## <none>                  5.2328e+10 53487
## + Hispanic        1 2.2615e+06 5.2326e+10 53488
## + SelfEmployed    1 2.1887e+06 5.2326e+10 53488
## + OtherTransp     1 1.1685e+06 5.2327e+10 53489
## + Transit         1 1.0095e+06 5.2327e+10 53489
## + Service         1 1.6287e+05 5.2328e+10 53489
## - Walk            1 1.7866e+08 5.2507e+10 53496
## - Pacific          1 1.8011e+08 5.2508e+10 53496
## - Office           1 3.0625e+08 5.2634e+10 53503
## - Carpool          1 3.3557e+08 5.2664e+10 53505
## - TotalPop         1 3.8267e+08 5.2711e+10 53508
## - Drive             1 3.8662e+08 5.2715e+10 53508
## - ChildPoverty     1 3.8951e+08 5.2718e+10 53508
## - WorkAtHome        1 4.4268e+08 5.2771e+10 53512
## - Citizen           1 5.2955e+08 5.2858e+10 53517
## - Native            1 5.6386e+08 5.2892e+10 53519
## - Men               1 6.8406e+08 5.3012e+10 53526
## - Production         1 9.0010e+08 5.3228e+10 53539
## - Construction       1 1.1369e+09 5.3465e+10 53554
## - IncomeErr          1 1.1992e+09 5.3527e+10 53558
## - Asian              1 1.3174e+09 5.3646e+10 53565
## - Black              1 2.1749e+09 5.4503e+10 53616
## - Professional        1 2.3720e+09 5.4700e+10 53627
## - MeanCommute         1 3.8514e+09 5.6180e+10 53713
## - PublicWork          1 3.9022e+09 5.6230e+10 53716
## - Poverty             1 3.9583e+09 5.6287e+10 53719
## - PrivateWork         1 4.1380e+09 5.6466e+10 53730
## - IncomePerCapErr     1 4.5049e+09 5.6833e+10 53750
## - White               1 7.4249e+09 5.9753e+10 53912
## - IncomePerCap        1 3.5842e+10 8.8170e+10 55164
##
## Step:  AIC=53485.97
## Y ~ TotalPop + Men + White + Black + Native + Asian + Pacific +
##     Citizen + IncomeErr + IncomePerCap + IncomePerCapErr + Poverty +
##     ChildPoverty + Professional + Office + Construction + Production +
##     Drive + Carpool + Walk + WorkAtHome + MeanCommute + Employed +
##     PrivateWork + PublicWork
##
##          Df  Sum of Sq      RSS     AIC
## - Employed          1 3.0990e+07 5.2382e+10 53486
## <none>                  5.2351e+10 53486
## + FamilyWork         1 2.2334e+07 5.2328e+10 53487
## + SelfEmployed        1 1.9943e+07 5.2331e+10 53487
## + Hispanic            1 2.1822e+06 5.2348e+10 53488
## + OtherTransp          1 8.4910e+05 5.2350e+10 53488
## + Transit             1 7.2473e+05 5.2350e+10 53488
## + Service              1 2.8674e+05 5.2350e+10 53488
## - Pacific              1 1.7971e+08 5.2530e+10 53495

```

```

## - Walk          1 1.8257e+08 5.2533e+10 53495
## - Office        1 3.0845e+08 5.2659e+10 53503
## - Carpool       1 3.3318e+08 5.2684e+10 53504
## - TotalPop      1 3.8241e+08 5.2733e+10 53507
## - Drive          1 3.8771e+08 5.2738e+10 53508
## - ChildPoverty   1 3.9695e+08 5.2748e+10 53508
## - WorkAtHome     1 4.5945e+08 5.2810e+10 53512
## - Citizen         1 5.2627e+08 5.2877e+10 53516
## - Native          1 5.6844e+08 5.2919e+10 53519
## - Men             1 6.8260e+08 5.3033e+10 53526
## - Production      1 9.0088e+08 5.3251e+10 53539
## - Construction    1 1.1621e+09 5.3513e+10 53555
## - IncomeErr        1 1.2108e+09 5.3561e+10 53558
## - Asian            1 1.3207e+09 5.3671e+10 53564
## - Black             1 2.1566e+09 5.4507e+10 53614
## - Professional      1 2.3701e+09 5.4721e+10 53626
## - MeanCommute       1 3.8370e+09 5.6188e+10 53712
## - Poverty           1 3.9444e+09 5.6295e+10 53718
## - PublicWork         1 3.9764e+09 5.6327e+10 53720
## - PrivateWork        1 4.2727e+09 5.6623e+10 53736
## - IncomePerCapErr   1 4.5545e+09 5.6905e+10 53752
## - White              1 7.4025e+09 5.9753e+10 53910
## - IncomePerCap       1 3.5820e+10 8.8171e+10 55162
##
## Step: AIC=53485.87
## Y ~ TotalPop + Men + White + Black + Native + Asian + Pacific +
##     Citizen + IncomeErr + IncomePerCap + IncomePerCapErr + Poverty +
##     ChildPoverty + Professional + Office + Construction + Production +
##     Drive + Carpool + Walk + WorkAtHome + MeanCommute + PrivateWork +
##     PublicWork
##
##                               Df  Sum of Sq      RSS      AIC
## <none>                      5.2382e+10 53486
## + Employed        1 3.0990e+07 5.2351e+10 53486
## + FamilyWork      1 2.1852e+07 5.2360e+10 53487
## + SelfEmployed    1 1.9355e+07 5.2362e+10 53487
## + Hispanic         1 1.6973e+06 5.2380e+10 53488
## + OtherTransp     1 1.5088e+06 5.2380e+10 53488
## + Transit          1 1.3551e+06 5.2380e+10 53488
## + Service          1 2.6332e+05 5.2381e+10 53488
## - Pacific           1 1.6759e+08 5.2549e+10 53494
## - Walk              1 1.9161e+08 5.2573e+10 53496
## - Office             1 3.1372e+08 5.2695e+10 53503
## - Carpool            1 3.4274e+08 5.2724e+10 53505
## - ChildPoverty       1 3.8845e+08 5.2770e+10 53508
## - Drive               1 4.0174e+08 5.2783e+10 53508
## - TotalPop            1 4.1628e+08 5.2798e+10 53509
## - WorkAtHome          1 4.7718e+08 5.2859e+10 53513
## - Native              1 5.6097e+08 5.2943e+10 53518
## - Citizen              1 5.6475e+08 5.2946e+10 53518
## - Men                 1 6.6787e+08 5.3049e+10 53525
## - Production            1 8.7883e+08 5.3260e+10 53537
## - Construction          1 1.1431e+09 5.3525e+10 53553
## - IncomeErr             1 1.1974e+09 5.3579e+10 53557

```

```

## - Asian          1 1.2899e+09 5.3671e+10 53562
## - Black         1 2.1790e+09 5.4561e+10 53615
## - Professional   1 2.3395e+09 5.4721e+10 53624
## - MeanCommute    1 3.8441e+09 5.6226e+10 53712
## - PublicWork     1 3.9779e+09 5.6359e+10 53719
## - Poverty        1 4.0267e+09 5.6408e+10 53722
## - PrivateWork    1 4.2693e+09 5.6651e+10 53736
## - IncomePerCapErr 1 4.5413e+09 5.6923e+10 53751
## - White          1 7.5118e+09 5.9893e+10 53915
## - IncomePerCap    1 3.6404e+10 8.8786e+10 55182

summary(stepwise_model_both)

##
## Call:
## lm(formula = Y ~ TotalPop + Men + White + Black + Native + Asian +
##      Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##      Poverty + ChildPoverty + Professional + Office + Construction +
##      Production + Drive + Carpool + Walk + WorkAtHome + MeanCommute +
##      PrivateWork + PublicWork, data = X)
##
## Residuals:
##      Min       1Q     Median      3Q      Max 
## -20911.8 -2378.7   -120.8   2101.6  22020.5 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -4.567e+04  4.513e+03 -10.118 < 2e-16 ***
## TotalPop     -9.203e-02  1.827e-02  -5.037 4.98e-07 ***
## Men          2.214e-01  3.470e-02   6.381 2.02e-10 ***
## White        -1.121e+02  5.237e+00 -21.398 < 2e-16 ***
## Black        -7.532e+01  6.536e+00 -11.525 < 2e-16 ***
## Native        7.301e+01  1.249e+01   5.848 5.49e-09 ***
## Asian         3.879e+02  4.374e+01   8.867 < 2e-16 ***
## Pacific       -6.946e+02  2.173e+02  -3.196 0.00141 ** 
## Citizen       -2.879e-02  4.907e-03  -5.867 4.88e-09 *** 
## IncomeErr      4.851e-01  5.679e-02   8.543 < 2e-16 ***
## IncomePerCap    1.300e+00  2.759e-02   47.107 < 2e-16 ***
## IncomePerCapErr -1.767e+00  1.062e-01  -16.638 < 2e-16 ***
## Poverty        -4.786e+02  3.055e+01 -15.667 < 2e-16 ***
## ChildPoverty    -9.329e+01  1.917e+01  -4.866 1.19e-06 ***
## Professional     3.235e+02  2.709e+01  11.942 < 2e-16 ***
## Office          1.508e+02  3.448e+01   4.373 1.26e-05 *** 
## Construction    2.517e+02  3.016e+01   8.347 < 2e-16 ***
## Production      1.861e+02  2.542e+01   7.319 3.14e-13 *** 
## Drive           1.344e+02  2.717e+01   4.949 7.86e-07 *** 
## Carpool          1.764e+02  3.860e+01   4.571 5.04e-06 *** 
## Walk            1.474e+02  4.312e+01   3.418 0.00064 *** 
## WorkAtHome       2.333e+02  4.326e+01   5.393 7.42e-08 *** 
## MeanCommute      2.325e+02  1.519e+01  15.308 < 2e-16 ***
## PrivateWork      4.546e+02  2.818e+01  16.132 < 2e-16 *** 
## PublicWork       4.831e+02  3.102e+01  15.572 < 2e-16 *** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

## 
## Residual standard error: 4050 on 3193 degrees of freedom
## Multiple R-squared:  0.9023, Adjusted R-squared:  0.9015
## F-statistic:  1228 on 24 and 3193 DF,  p-value: < 2.2e-16

selected_features_step <- names(coef(stepwise_model_both))[-1]
selected_features_step

## [1] "TotalPop"      "Men"          "White"         "Black"
## [5] "Native"        "Asian"         "Pacific"       "Citizen"
## [9] "IncomeErr"      "IncomePerCap"   "IncomePerCapErr" "Poverty"
## [13] "ChildPoverty"   "Professional"  "Office"        "Construction"
## [17] "Production"    "Drive"         "Carpool"       "Walk"
## [21] "WorkAtHome"    "MeanCommute"   "PrivateWork"   "PublicWork"

```

Filtering out these features

```
X=X[selected_features_step]
```

Base Model with main effect before diagnosis

```

fit1 <- lm(Y ~ ., data = X)
base_model_main_effect <- summary(fit1)
r2 <- base_model_main_effect$r.squared
adj_r2 <- base_model_main_effect$adj.r.squared

```

Base Model diagnosis with main effect

```

cat("R2:", r2, "\n")

## R2: 0.9022807

cat("Adjusted R2:", adj_r2, "\n")

## Adjusted R2: 0.9015461

cat("AIC:", AIC(fit1), "\n")

## AIC: 62620.16

cat("BIC:", BIC(fit1), "\n")

## BIC: 62778.15

```

```
print(dwtest(fit1))

##
##  Durbin-Watson test
##
## data: fit1
## DW = 1.7407, p-value = 4.733e-14
## alternative hypothesis: true autocorrelation is greater than 0
```

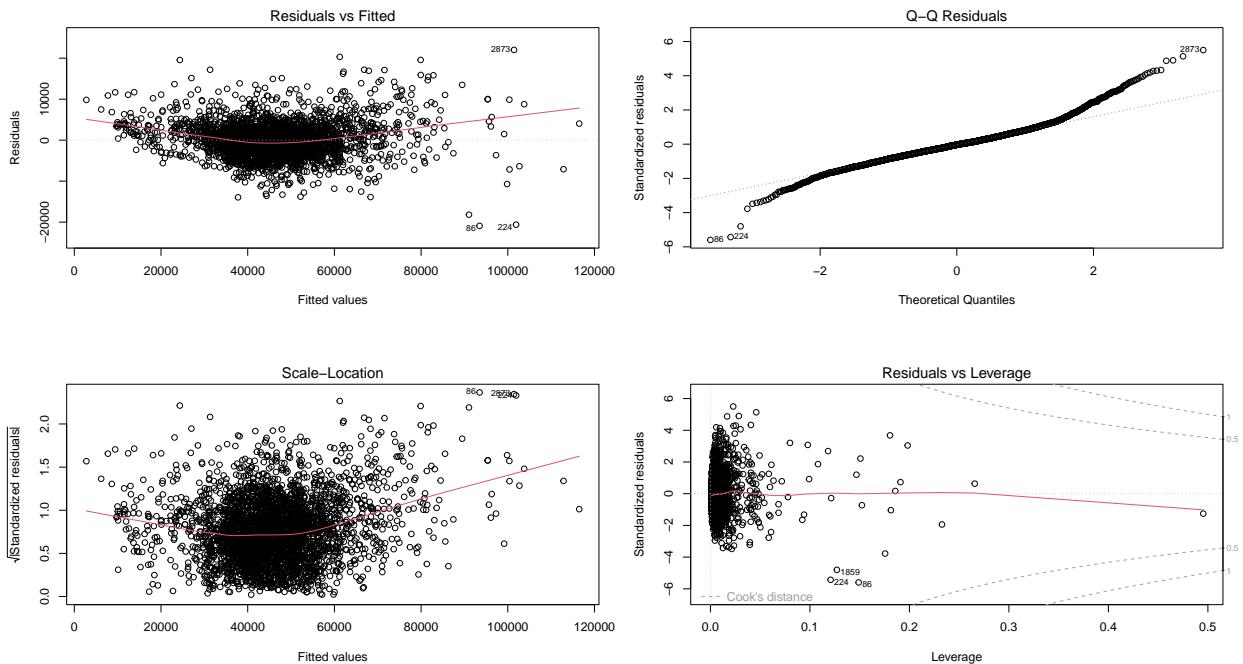
```
print(shapiro.test(residuals(fit1)))
```

```
##
##  Shapiro-Wilk normality test
##
## data: residuals(fit1)
## W = 0.96419, p-value < 2.2e-16
```

```
print(vif(fit1))
```

| | TotalPop | Men | White | Black | Native |
|----|-----------------|-------------|--------------|--------------|--------------|
| ## | 6677.425309 | 5799.286884 | 2.825381 | 1.708575 | 1.607888 |
| ## | Asian | Pacific | Citizen | IncomeErr | IncomePerCap |
| ## | 2.556547 | 1.433631 | 198.749750 | 2.319333 | 5.726323 |
| ## | IncomePerCapErr | Poverty | ChildPoverty | Professional | Office |
| ## | 2.386229 | 12.661735 | 9.854505 | 5.837453 | 2.380646 |
| ## | Construction | Production | Drive | Carpool | Walk |
| ## | 3.166153 | 4.171999 | 8.392551 | 2.471069 | 4.989173 |
| ## | WorkAtHome | MeanCommute | PrivateWork | PublicWork | |
| ## | 3.695681 | 1.416821 | 9.517031 | 7.870448 | |

```
par(mfrow = c(2, 2))
plot(fit1)
```



As we can see we got good R² but assumptions failed.

Treatment in main effect

Dropping features with VIF < 5

```
vif_values <- vif(fit1)
low_vif_features <- names(vif_values)[vif_values < 5]
print(low_vif_features)
```

```
## [1] "White"          "Black"          "Native"         "Asian"
## [5] "Pacific"        "IncomeErr"       "IncomePerCapErr" "Office"
## [9] "Construction"   "Production"     "Carpool"        "Walk"
## [13] "WorkAtHome"     "MeanCommute"
```

```
X <- X[low_vif_features]
dim(X)
```

```
## [1] 3218 14
```

Transformation

```
Y <- bestNormalize::yeojohnson(Y)$x.t
X[abs(apply(X, 2, e1071::skewness)) > 1] <- lapply(X[abs(apply(X, 2, e1071::skewness)) > 1], log1p) #
```

```

##      White     Black    Native    Asian Pacific IncomeErr IncomePerCapErr
## 1 4.341205 2.9704145 0.3364722 0.69314718      0 7.779885      6.985642
## 2 4.432007 2.3513753 0.4700036 0.53062825      0 7.142037      6.568078
## 3 3.854394 3.8649314 0.1823216 0.33647224      0 7.997663      6.683361
## 4 4.324133 3.1090610 0.3364722 0.09531018      0 8.293049      7.389564
## 5 4.487512 0.9162907 0.2623643 0.09531018      0 8.052615      6.563856
## 6 3.144152 4.2724907 0.7884574 0.18232156      0 8.680162      7.628518
##      Office Construction Production Carpool      Walk WorkAtHome MeanCommute
## 1    24.2          8.6       17.1      8.8 0.4054651  1.0296194      26.5
## 2    27.1         10.8       11.2      8.8 0.6931472  1.5892352      26.4
## 3    23.1         10.8       23.1     10.9 1.0296194  0.9555114      24.1
## 4    17.8         19.0       23.7     13.5 0.4700036  0.5306283      28.8
## 5    23.9         13.5       19.9     11.2 0.6418539  1.1939225      34.9
## 6    19.7         20.1       26.4     14.9 1.7917595  1.3350011      27.5

```

Base model after diagnosis(VIF filter+transformation) with main effect

```

fit2 <- lm(Y ~ ., data = X)
base_model_main_effect <- summary(fit2)
r2 <- base_model_main_effect$r.squared
adj_r2 <- base_model_main_effect$adj.r.squared

cat("R2:", r2, "\n")

## R2: 0.5565732

cat("Adjusted R2:", adj_r2, "\n")

## Adjusted R2: 0.554635

cat("AIC:", AIC(fit2), "\n")

## AIC: 6546.338

cat("BIC:", BIC(fit2), "\n")

## BIC: 6643.563

print(dwtest(fit2))

## 
## Durbin-Watson test
##
## data: fit2
## DW = 1.5695, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0

```

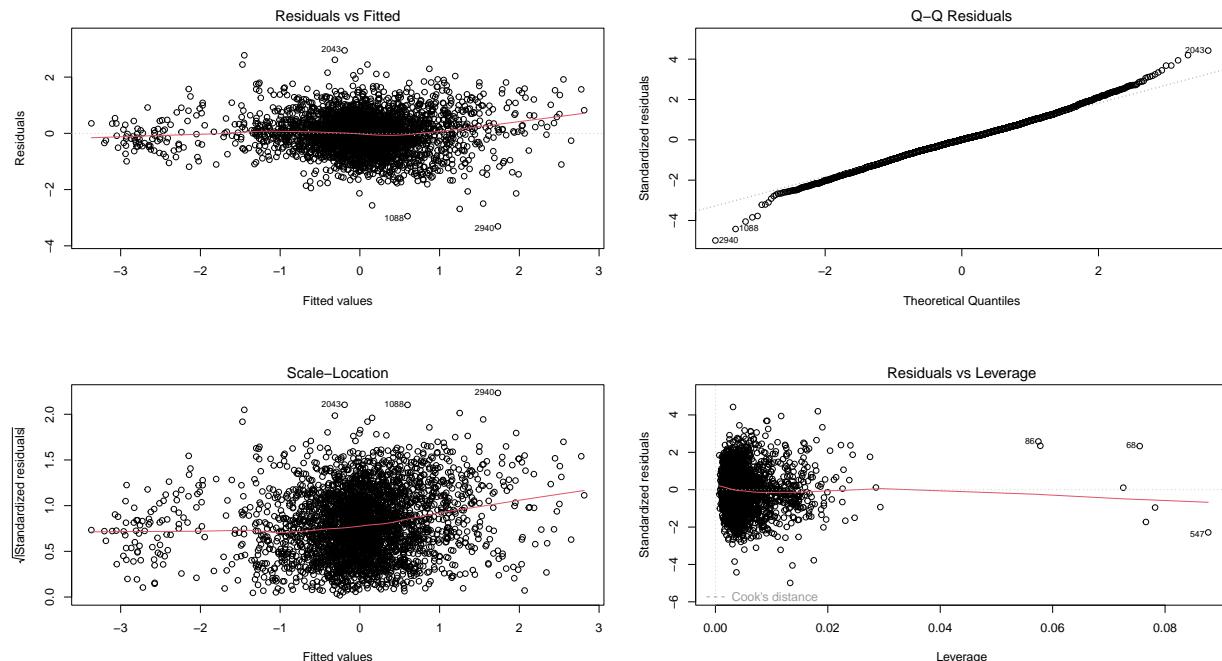
```
print(shapiro.test(residuals(fit2)))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: residuals(fit2)  
## W = 0.99474, p-value = 2.442e-09
```

```
print(vif(fit2))
```

```
##          White           Black           Native          Asian          Pacific  
## 1.263248 1.429725 1.205693 1.942369 1.190704  
## IncomeErr IncomePerCapErr Office Construction Production  
## 3.537887 3.732609 1.657800 1.889478 1.648585  
## Carpool      Walk WorkAtHome MeanCommute  
## 1.156902 1.781497 1.778833 1.423280
```

```
par(mfrow = c(2, 2))  
plot(fit2)
```



```
## Dropping outliers for cooks distance
```

```
dim(X)
```

```
## [1] 3218 14
```

```

cooks_dist <- cooks.distance(fit2)
threshold <- 4 / length(cooks_dist)
influential_points <- which(cooks_dist > threshold)
X <- X[-influential_points, ]
Y <- Y[-influential_points]

```

Base model after diagnosis(Influential points) with main effect

```

fit3 <- lm(Y ~ ., data = X)
base_model_main_effect <- summary(fit3)
r2 <- base_model_main_effect$r.squared
adj_r2 <- base_model_main_effect$adj.r.squared

cat("R2:", r2, "\n")

## R2: 0.6333641

cat("Adjusted R2:", adj_r2, "\n")

## Adjusted R2: 0.6316416

cat("AIC:", AIC(fit3), "\n")

## AIC: 4965.617

cat("BIC:", BIC(fit3), "\n")

## BIC: 5061.692

print(dwtest(fit3))

##
## Durbin-Watson test
##
## data: fit3
## DW = 1.6749, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0

print(shapiro.test(residuals(fit3)))

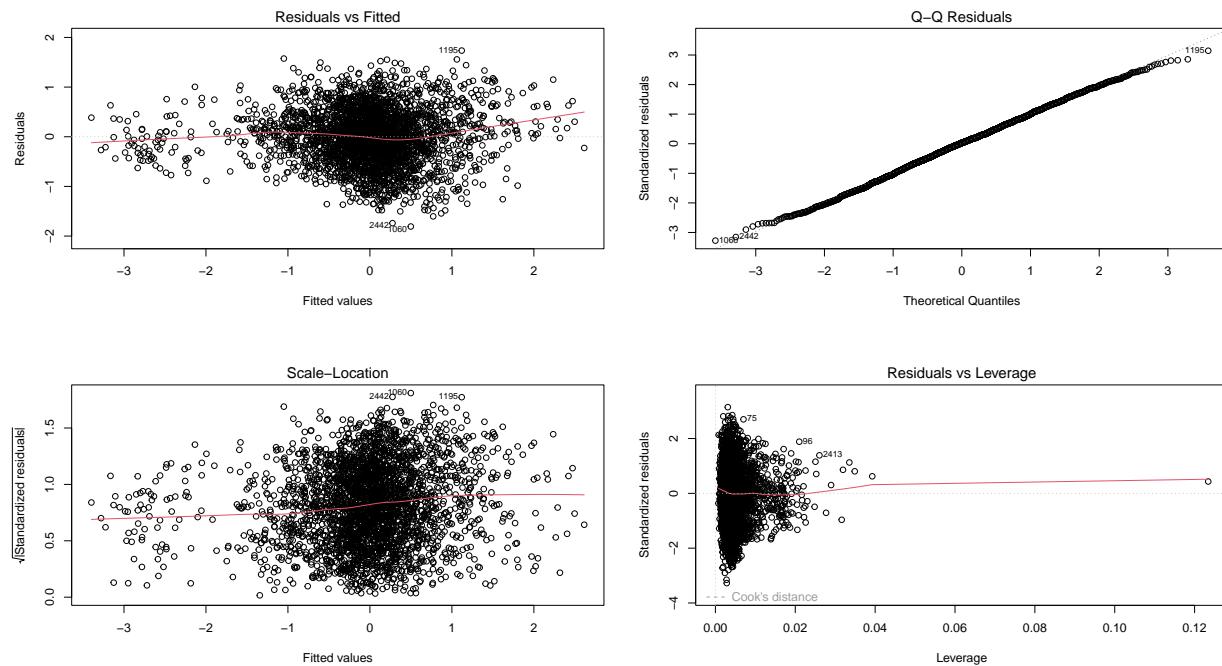
##
## Shapiro-Wilk normality test
##
## data: residuals(fit3)
## W = 0.9989, p-value = 0.05253

```

```
print(vif(fit3))
```

```
##          White           Black          Native          Asian          Pacific
## 1.298041    1.460077    1.186241    2.010300    1.127789
## IncomeErr  IncomePerCapErr          Office      Construction      Production
## 3.482675    3.742256    1.725720    1.901318    1.793523
## Carpool            Walk      WorkAtHome      MeanCommute
## 1.154905    1.820614    1.875744    1.391618
```

```
par(mfrow = c(2, 2))
plot(fit3)
```



Only multicollinearity has been passed, but plot has been improved.

two main effect

All combination of two main effect has been implemented and the best features extracted using STEP.

```
full_model <- lm(Y ~ .^2, data = X)
step_model <- step(full_model, direction = "both", trace = FALSE)
final_formula <- formula(step_model)
final_fit <- lm(final_formula, data = X)
cooks_dist <- cooks.distance(final_fit)
threshold <- 4 / length(cooks_dist)
influential_points <- which(cooks_dist > threshold)
```

```

X <- X[!influential_points, , drop = FALSE]
Y <- Y[!influential_points]
final_fit <- lm(final_formula, data = X)
model_summary <- summary(final_fit)
r2 <- model_summary$r.squared
adj_r2 <- model_summary$adj.r.squared

cat("R2:", r2, "\n")

## R2: 0.7121303

cat("Adjusted R2:", adj_r2, "\n")

## Adjusted R2: 0.705694

cat("AIC:", AIC(final_fit), "\n")

## AIC: 3820.55

cat("BIC:", BIC(final_fit), "\n")

## BIC: 4201.359

print(final_formula)

## Y ~ White + Black + Native + Asian + Pacific + IncomeErr + IncomePerCapErr +
##     Office + Construction + Production + Carpool + Walk + WorkAtHome +
##     MeanCommute + White:Black + White:Native + White:Asian +
##     White:IncomePerCapErr + White:Office + White:Construction +
##     White:Production + White:Carpool + White:Walk + White:WorkAtHome +
##     Black:Asian + Black:Pacific + Black:IncomeErr + Black:IncomePerCapErr +
##     Black:Production + Black:WorkAtHome + Black:MeanCommute +
##     Native:IncomePerCapErr + Native:Construction + Native:WorkAtHome +
##     Native:MeanCommute + Asian:IncomeErr + Asian:Office + Asian:Production +
##     Asian:Carpool + Asian:Walk + Asian:WorkAtHome + Asian:MeanCommute +
##     Pacific:Production + Pacific:Carpool + Pacific:MeanCommute +
##     IncomeErr:IncomePerCapErr + IncomeErr:Office + IncomeErr:Walk +
##     IncomePerCapErr:Construction + IncomePerCapErr:Walk + Construction:Production +
##     Construction:Carpool + Construction:Walk + Construction:WorkAtHome +
##     Construction:MeanCommute + Production:Carpool + Production:Walk +
##     Production:WorkAtHome + Production:MeanCommute + Carpool:Walk +
##     Walk:MeanCommute + WorkAtHome:MeanCommute

print(dwtest(final_fit))

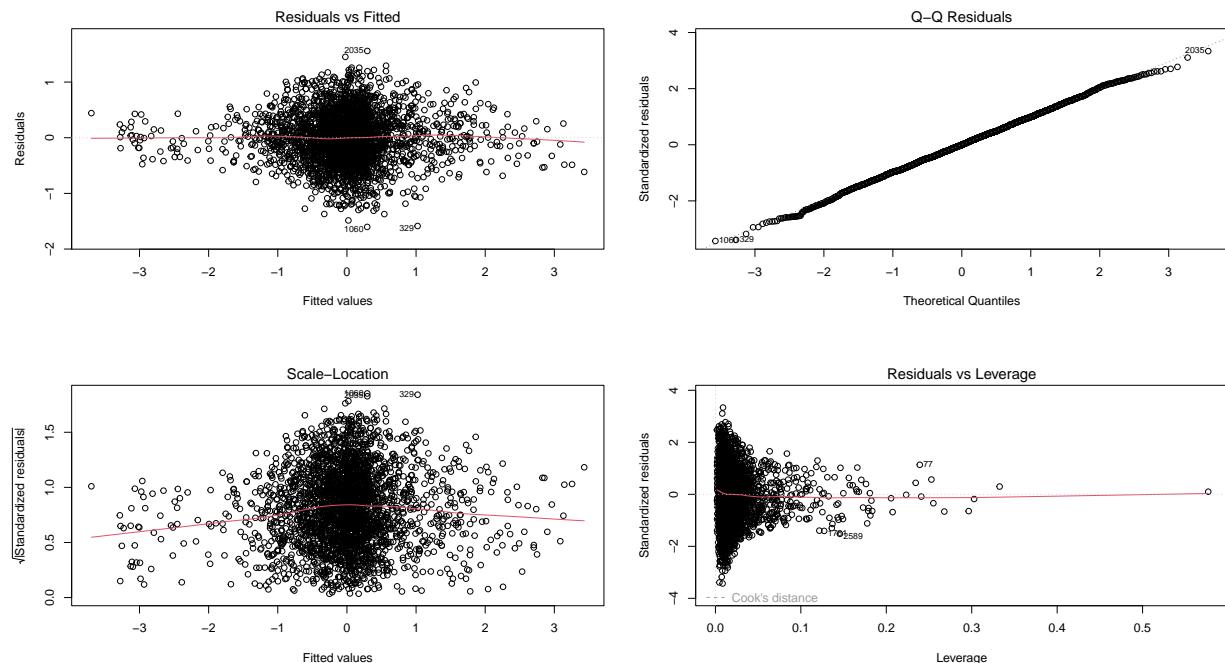
##
## Durbin-Watson test
##
## data: final_fit
## DW = 1.7113, p-value = 2.99e-15
## alternative hypothesis: true autocorrelation is greater than 0

```

```
print(shapiro.test(residuals(final_fit)))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: residuals(final_fit)  
## W = 0.99932, p-value = 0.3903
```

```
par(mfrow = c(2, 2))  
plot(final_fit)
```



AMS578_project_116125547

Ajeetkumar Rai (Id : 116125547)

2025-04-30

#Library

```
required_packages <- c("tidyverse", "caret", "neuralnet", "ggplot2", "glmnet",
  "rpart", "rattle", "factoextra", "cluster", "gridExtra", "corrplot", "lmtest", "car", "forecast", "e1071")
for (pkg in required_packages) {
  if (!requireNamespace(pkg, quietly = TRUE)) {
    install.packages(pkg)
  }
  library(pkg, character.only = TRUE)}
```



```
## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'ggplot2' was built under R version 4.3.3

## Warning: package 'tidyverse' was built under R version 4.3.3

## Warning: package 'readr' was built under R version 4.3.3

## Warning: package 'purrr' was built under R version 4.3.3

## Warning: package 'dplyr' was built under R version 4.3.3

## Warning: package 'stringr' was built under R version 4.3.3

## Warning: package 'forcats' was built under R version 4.3.3

## Warning: package 'lubridate' was built under R version 4.3.3

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.4      v readr     2.1.5
## vforcats   1.0.0      v stringr   1.5.1
## v ggplot2   3.5.1      v tibble    3.2.1
## v lubridate 1.9.4      v tidyverse 1.3.1
## v purrr     1.0.4

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

## Warning: package 'caret' was built under R version 4.3.3
```

```

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##     lift

## Warning: package 'neuralnet' was built under R version 4.3.3

##
## Attaching package: 'neuralnet'
##
## The following object is masked from 'package:dplyr':
##     compute

## Warning: package 'glmnet' was built under R version 4.3.3

## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##     expand, pack, unpack
##
## Loaded glmnet 4.1-8

## Warning: package 'rpart' was built under R version 4.3.3

## Warning: package 'rattle' was built under R version 4.3.3

## Loading required package: bitops

## Warning: package 'bitops' was built under R version 4.3.3

##
## Attaching package: 'bitops'
##
## The following object is masked from 'package:Matrix':
##     %&%
##
## Rattle: A free graphical interface for data science with R.
## Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.

## Warning: package 'factoextra' was built under R version 4.3.3

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

```

```

## Warning: package 'gridExtra' was built under R version 4.3.3

##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##   combine

## Warning: package 'corrplot' was built under R version 4.3.3

## corrplot 0.95 loaded

## Warning: package 'lmtest' was built under R version 4.3.3

## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##   as.Date, as.Date.numeric

## Warning: package 'car' was built under R version 4.3.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.3.3

##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##   recode
##
## The following object is masked from 'package:purrr':
##   some
##
## Registered S3 method overwritten by 'quantmod':
##   method           from
##   as.zoo.data.frame zoo

## Warning: package 'forecast' was built under R version 4.3.3

## Warning: package 'e1071' was built under R version 4.3.3

```

Data

```

df <- read.csv("C:/Users/Ajeet Rai/OneDrive/Desktop/SBU/Academics/Sem III/AMS 578 Regression Theory/Pro
head(df)

##      State County TotalPop   Men Women Hispanic White Black Native Asian
## 1 Alabama Autauga    55221 26745 28476       2.6 75.8 18.5  0.4  1.0
## 2 Alabama Baldwin    195121 95314 99807       4.5 83.1  9.5  0.6  0.7
## 3 Alabama Barbour   26932 14497 12435       4.6 46.2 46.7  0.2  0.4
## 4 Alabama Bibb      22604 12073 10531       2.2 74.5 21.4  0.4  0.1
## 5 Alabama Blount    57710 28512 29198       8.6 87.9  1.5  0.3  0.1
## 6 Alabama Bullock   10678  5660  5018       4.4 22.2 70.7  1.2  0.2
##      Pacific Citizen Income IncomeErr IncomePerCap IncomePerCapErr Poverty
## 1          0 40725 51281     2391     24974           1080      12.9
## 2          0 147695 50254     1263     27317            711      13.4
## 3          0 20714 32964     2973     16824            798      26.7
## 4          0 17495 38678     3995     18431           1618      16.8
## 5          0 42345 45813     3141     20532            708      16.7
## 6          0 8057 31938     5884     17580           2055      24.6
##      ChildPoverty Professional Service Office Construction Production Drive
## 1          18.6        33.2    17.0   24.2       8.6     17.1  87.5
## 2          19.2        33.1    17.7   27.1      10.8    11.2  84.7
## 3          45.3        26.8    16.1   23.1      10.8    23.1  83.8
## 4          27.9        21.5    17.9   17.8      19.0    23.7  83.2
## 5          27.2        28.5    14.1   23.9      13.5    19.9  84.9
## 6          38.4        18.8    15.0   19.7      20.1    26.4  74.9
##      Carpool Transit Walk OtherTransp WorkAtHome MeanCommute Employed PrivateWork
## 1          8.8        0.1    0.5     1.3      1.8     26.5  23986    73.6
## 2          8.8        0.1    1.0     1.4      3.9     26.4  85953    81.5
## 3         10.9        0.4    1.8     1.5      1.6     24.1  8597    71.8
## 4         13.5        0.5    0.6     1.5      0.7     28.8  8294    76.8
## 5         11.2        0.4    0.9     0.4      2.3     34.9  22189    82.0
## 6         14.9        0.7    5.0     1.7      2.8     27.5  3865    79.5
##      PublicWork SelfEmployed FamilyWork Unemployment
## 1          20.9        5.5    0.0      7.6
## 2          12.3        5.8    0.4      7.5
## 3          20.8        7.3    0.1     17.6
## 4          16.1        6.7    0.4      8.3
## 5          13.5        4.2    0.4      7.7
## 6          15.1        5.4    0.0     18.0

```

Summary

Shape

```
dim(df)
```

```
## [1] 3220 36
```

Datatypes

```
str(df)
```

```
## 'data.frame': 3220 obs. of 36 variables:
## $ State      : chr "Alabama" "Alabama" "Alabama" "Alabama" ...
## $ County     : chr "Autauga" "Baldwin" "Barbour" "Bibb" ...
## $ TotalPop   : int 55221 195121 26932 22604 57710 10678 20354 116648 34079 26008 ...
## $ Men        : int 26745 95314 14497 12073 28512 5660 9502 56274 16258 12975 ...
## $ Women      : int 28476 99807 12435 10531 29198 5018 10852 60374 17821 13033 ...
## $ Hispanic   : num 2.6 4.5 4.6 2.2 8.6 4.4 1.2 3.5 0.4 1.5 ...
## $ White       : num 75.8 83.1 46.2 74.5 87.9 22.2 53.3 73 57.3 91.7 ...
## $ Black       : num 18.5 9.5 46.7 21.4 1.5 70.7 43.8 20.3 40.3 4.8 ...
## $ Native      : num 0.4 0.6 0.2 0.4 0.3 1.2 0.1 0.2 0.2 0.6 ...
## $ Asian       : num 1 0.7 0.4 0.1 0.1 0.2 0.4 0.9 0.8 0.3 ...
## $ Pacific     : num 0 0 0 0 0 0 0 0 0 0 ...
## $ Citizen    : int 40725 147695 20714 17495 42345 8057 15581 88612 26462 20600 ...
## $ Income      : int 51281 50254 32964 38678 45813 31938 32229 41703 34177 36296 ...
## $ IncomeErr   : int 2391 1263 2973 3995 3141 5884 1793 925 2949 1710 ...
## $ IncomePerCap: int 24974 27317 16824 18431 20532 17580 18390 21374 21071 21811 ...
## $ IncomePerCapErr: int 1080 711 798 1618 708 2055 714 489 1366 1556 ...
## $ Poverty     : num 12.9 13.4 26.7 16.8 16.7 24.6 25.4 20.5 21.6 19.2 ...
## $ ChildPoverty: num 18.6 19.2 45.3 27.9 27.2 38.4 39.2 31.6 37.2 30.1 ...
## $ Professional: num 33.2 33.1 26.8 21.5 28.5 18.8 27.5 27.3 23.3 29.3 ...
## $ Service     : num 17 17.7 16.1 17.9 14.1 15 16.6 17.7 14.5 16 ...
## $ Office      : num 24.2 27.1 23.1 17.8 23.9 19.7 21.9 24.2 26.3 19.5 ...
## $ Construction: num 8.6 10.8 10.8 19 13.5 20.1 10.3 10.5 11.5 13.7 ...
## $ Production  : num 17.1 11.2 23.1 23.7 19.9 26.4 23.7 20.4 24.4 21.5 ...
## $ Drive       : num 87.5 84.7 83.8 83.2 84.9 74.9 84.5 85.3 85.1 83.9 ...
## $ Carpool     : num 8.8 8.8 10.9 13.5 11.2 14.9 12.4 9.4 11.9 12.1 ...
## $ Transit     : num 0.1 0.1 0.4 0.5 0.4 0.7 0 0.2 0.2 0.2 ...
## $ Walk        : num 0.5 1 1.8 0.6 0.9 5 0.8 1.2 0.3 0.6 ...
## $ OtherTransp : num 1.3 1.4 1.5 1.5 0.4 1.7 0.6 1.2 0.4 0.7 ...
## $ WorkAtHome  : num 1.8 3.9 1.6 0.7 2.3 2.8 1.7 2.7 2.1 2.5 ...
## $ MeanCommute : num 26.5 26.4 24.1 28.8 34.9 27.5 24.6 24.1 25.1 27.4 ...
## $ Employed    : int 23986 85953 8597 8294 22189 3865 7813 47401 13689 10155 ...
## $ PrivateWork : num 73.6 81.5 71.8 76.8 82 79.5 77.4 74.1 85.1 73.1 ...
## $ PublicWork  : num 20.9 12.3 20.8 16.1 13.5 15.1 16.2 20.8 12.1 18.5 ...
## $ SelfEmployed: num 5.5 5.8 7.3 6.7 4.2 5.4 6.2 5 2.8 7.9 ...
## $ FamilyWork  : num 0 0.4 0.1 0.4 0.4 0 0.2 0.1 0 0.5 ...
## $ Unemployment: num 7.6 7.5 17.6 8.3 7.7 18 10.9 12.3 8.9 7.9 ...
```

Descriptions

```
summary(df)
```

```
##      State          County        TotalPop        Men      
## Length:3220    Length:3220    Min.   :    85   Min.   :    42  
## Class :character Class :character  1st Qu.: 11218   1st Qu.: 5637  
## Mode  :character Mode  :character  Median : 26035  Median : 12932
```

```

##                                     Mean   : 99409   Mean   : 48897
##                                     3rd Qu.: 66430   3rd Qu.: 32993
##                                     Max.   :10038388   Max.   :4945351
##
##      Women          Hispanic        White        Black
## Min.   :    43   Min.   : 0.000   Min.   : 0.00   Min.   : 0.000
## 1st Qu.: 5572   1st Qu.: 1.900   1st Qu.:64.10  1st Qu.: 0.500
## Median : 13057  Median : 3.900   Median :84.10  Median : 1.900
## Mean   : 50512  Mean   :11.012   Mean   :75.43  Mean   : 8.665
## 3rd Qu.: 33488  3rd Qu.: 9.825   3rd Qu.:93.20  3rd Qu.: 9.600
## Max.   :5093037 Max.   :99.900   Max.   :99.80  Max.   :85.900
##
##      Native         Asian        Pacific       Citizen
## Min.   : 0.000   Min.   : 0.000   Min.   : 0.00000   Min.   :     80
## 1st Qu.: 0.100   1st Qu.: 0.200   1st Qu.: 0.00000   1st Qu.: 8450
## Median : 0.300   Median : 0.500   Median : 0.00000   Median : 19643
## Mean   : 1.724   Mean   : 1.229   Mean   : 0.08273   Mean   : 69935
## 3rd Qu.: 0.600   3rd Qu.: 1.200   3rd Qu.: 0.00000   3rd Qu.: 49920
## Max.   :92.100   Max.   :41.600   Max.   :35.30000   Max.   :6046749
##
##      Income        IncomeErr     IncomePerCap  IncomePerCapErr
## Min.   : 10499   Min.   : 270   Min.   : 5878   Min.   : 113
## 1st Qu.: 38192   1st Qu.: 1635  1st Qu.:20239  1st Qu.: 755
## Median : 44749   Median : 2406  Median :23460   Median : 1096
## Mean   : 46130   Mean   : 2850  Mean   :23982   Mean   : 1363
## 3rd Qu.: 52074   3rd Qu.: 3446  3rd Qu.:27053  3rd Qu.: 1631
## Max.   :123453   Max.   :21355  Max.   :65600   Max.   :15266
## NA's   :1
##
##      Poverty       ChildPoverty  Professional     Service
## Min.   : 1.40   Min.   : 0.00   Min.   :13.50   Min.   : 5.00
## 1st Qu.:12.10  1st Qu.:16.30  1st Qu.:26.70  1st Qu.:16.00
## Median :16.15  Median :22.70  Median :29.90  Median :18.10
## Mean   :17.49  Mean   :24.18  Mean   :30.99  Mean   :18.35
## 3rd Qu.:20.70  3rd Qu.:30.00  3rd Qu.:34.40  3rd Qu.:20.30
## Max.   :64.20   Max.   :81.60  Max.   :74.00  Max.   :38.20
## NA's   :1
##
##      Office        Construction Production     Drive
## Min.   : 4.10   Min.   : 1.70   Min.   : 0.00   Min.   : 5.20
## 1st Qu.:20.20  1st Qu.: 9.80  1st Qu.:11.50  1st Qu.:76.60
## Median :22.40  Median :12.10  Median :15.25  Median :80.70
## Mean   :22.22  Mean   :12.71  Mean   :15.73  Mean   :79.18
## 3rd Qu.:24.40  3rd Qu.:14.90  3rd Qu.:19.32  3rd Qu.:83.70
## Max.   :35.40   Max.   :40.30  Max.   :55.60  Max.   :94.60
##
##      Carpool       Transit        Walk        OtherTransp
## Min.   : 0.00   Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000
## 1st Qu.: 8.40   1st Qu.: 0.1000  1st Qu.: 1.400  1st Qu.: 0.900
## Median : 9.90   Median : 0.4000  Median : 2.400  Median : 1.300
## Mean   :10.28   Mean   : 0.9718  Mean   : 3.324  Mean   : 1.613
## 3rd Qu.:11.80   3rd Qu.: 0.8000  3rd Qu.: 4.000  3rd Qu.: 1.900
## Max.   :29.90   Max.   :61.7000  Max.   :71.200  Max.   :39.100
##
##      WorkAtHome    MeanCommute    Employed     PrivateWork
## Min.   : 0.000   Min.   : 4.90   Min.   :     62  Min.   :25.00

```

```

## 1st Qu.: 2.700 1st Qu.:19.50 1st Qu.: 4551 1st Qu.:70.50
## Median : 3.900 Median :23.00 Median : 10508 Median :75.70
## Mean   : 4.632 Mean  :23.28 Mean  : 45594 Mean  :74.22
## 3rd Qu.: 5.600 3rd Qu.:26.80 3rd Qu.: 28633 3rd Qu.:79.70
## Max.   :37.200 Max.  :44.00 Max.  :4635465 Max.  :88.30
##
##   PublicWork    SelfEmployed    FamilyWork    Unemployment
## Min.   : 5.80  Min.   :0.0000  Min.   :0.0000  Min.   : 0.000
## 1st Qu.:13.10 1st Qu.: 5.400  1st Qu.:0.1000  1st Qu.: 5.500
## Median :16.20  Median : 6.900  Median :0.2000  Median : 7.600
## Mean   :17.56  Mean   : 7.932  Mean   :0.2881  Mean   : 8.094
## 3rd Qu.:20.50 3rd Qu.: 9.400  3rd Qu.:0.3000  3rd Qu.: 9.900
## Max.   :66.20  Max.   :36.600  Max.   :9.8000  Max.   :36.500
##

```

Imputing missing values

```

dim(df)

## [1] 3220 36

df = na.omit(df)
dim(df)

## [1] 3218 36

```

Independent vs dependent

```

Y1 <- df$Income
Y2 <- df$Unemployment
Y=Y2 #Y=Y1
X <- df[, !(names(df) %in% c("Income", "Unemployment", "State", "County"))]

```

Exploratory data analysis

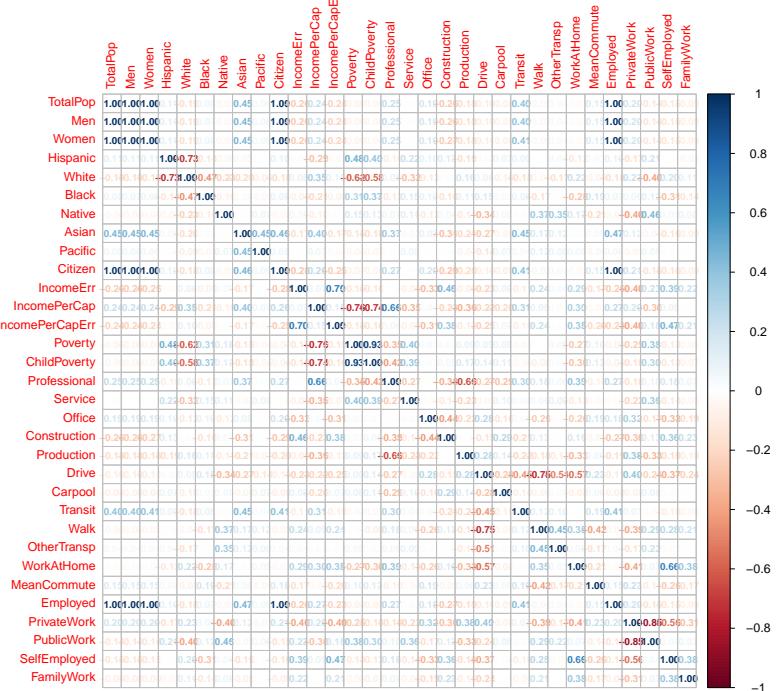
Correlation

As we can see collinearity between features, we'll need check VIF carefully.

```

correlation_matrix <- cor(X)
corrplot(correlation_matrix, method = "number", number.cex = 0.7, tl.cex = 0.8)

```



Distributin (Dependent)

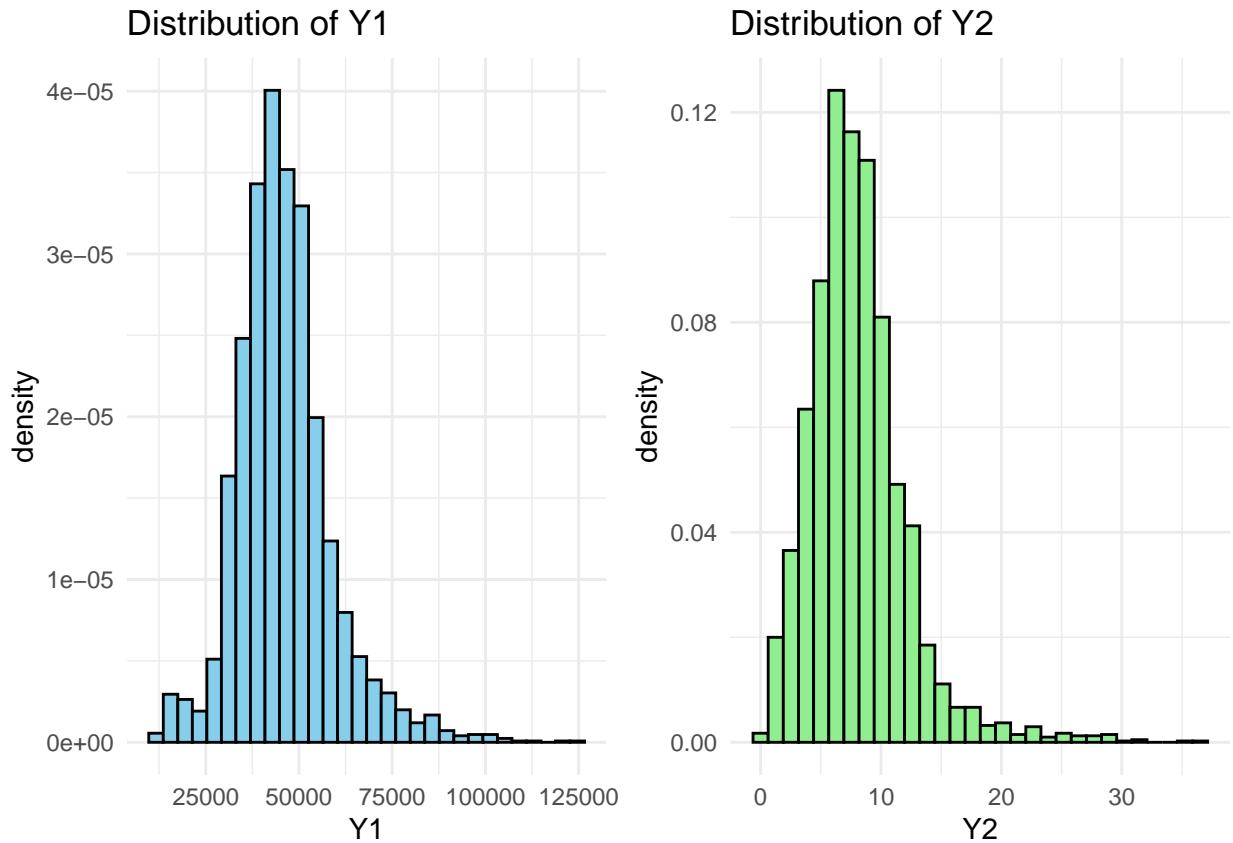
Y1 and Y2 are slightly skewed, and transformation is needed to prevent this.

```
p1 <- ggplot(data = data.frame(Y1), aes(x = Y1)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'skyblue', color = 'black') +
  ggtitle('Distribution of Y1') +
  theme_minimal()

p2 <- ggplot(data = data.frame(Y2), aes(x = Y2)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'lightgreen', color = 'black') +
  ggtitle('Distribution of Y2') +
  theme_minimal()

grid.arrange(p1, p2, ncol = 2)
```

```
## Warning: The dot-dot notation ('..density..') was deprecated in ggplot2 3.4.0.
## i Please use 'after_stat(density)' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



Distributin (Independent)

Independent features are skewed, and transformation is needed to prevent this.

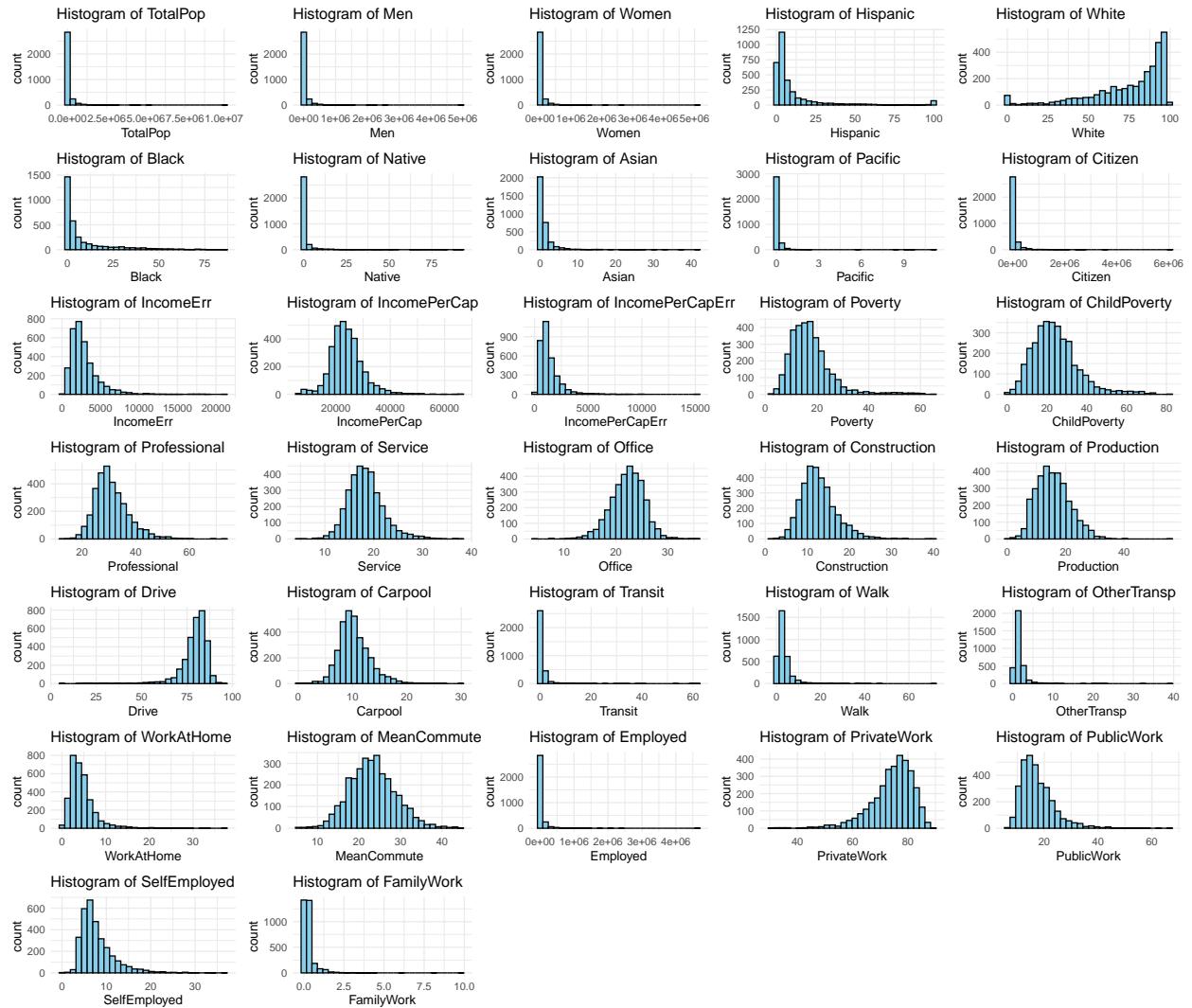
```

plots <- lapply(names(X), function(colname) {
  if (is.numeric(X[[colname]])) {
    ggplot(X, aes_string(x = colname)) +
      geom_histogram(bins = 30, fill = "skyblue", color = "black") +
      ggtitle(paste("Histogram of", colname)) +
      theme_minimal()
  } else {
    ggplot(X, aes_string(x = colname)) +
      geom_bar(fill = "lightgreen", color = "black") +
      ggtitle(paste("Bar Plot of", colname)) +
      theme_minimal()
  }
})

## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

```

```
grid.arrange(grobs = plots, ncol = 5)
```

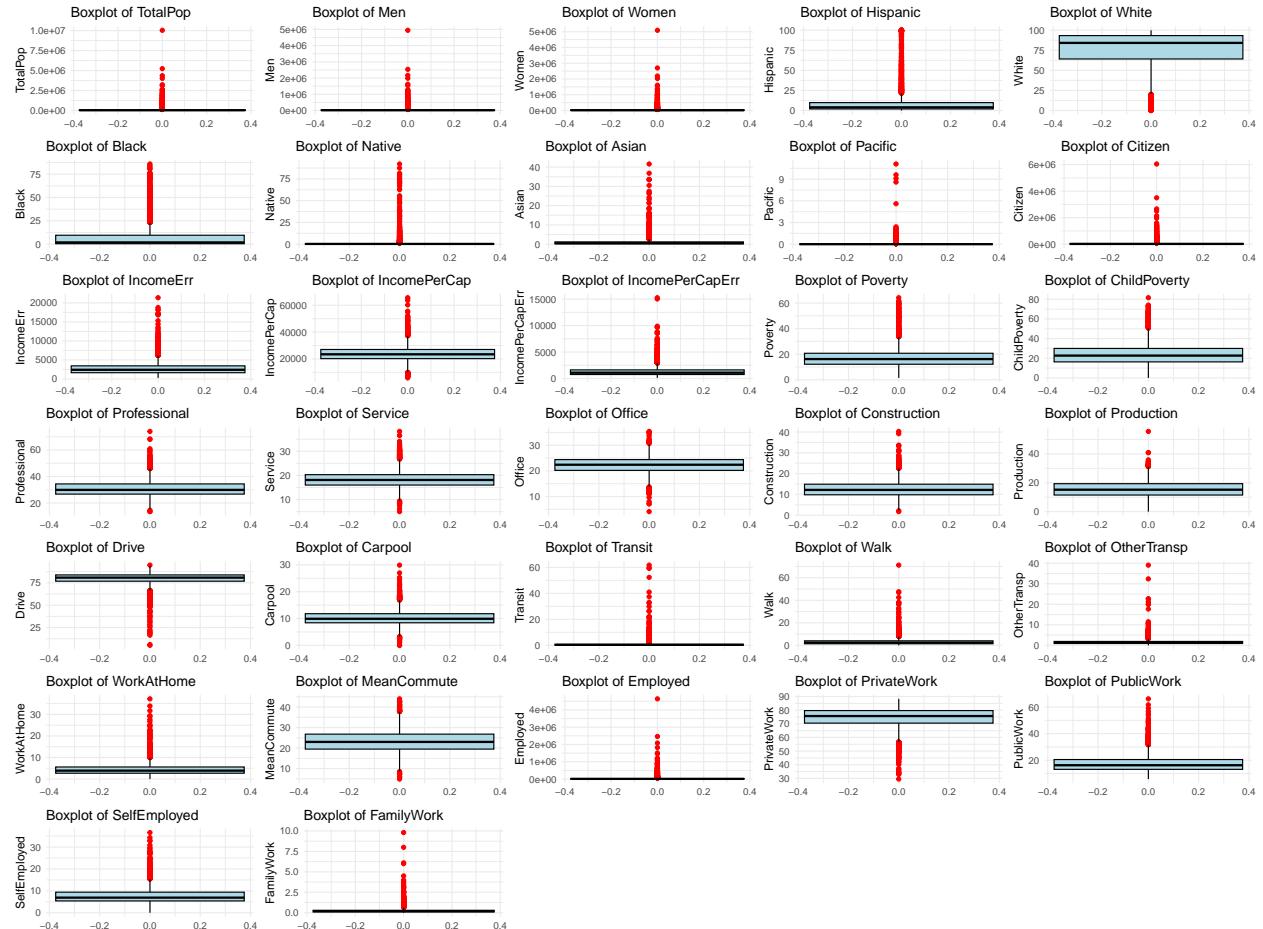


Outliers

Need careful treatment for outliers as it will effect cook's distance

```
plots <- lapply(names(X), function(colname) {
  if (is.numeric(X[[colname]])) {
    ggplot(X, aes_string(y = colname)) +
      geom_boxplot(fill = "lightblue", color = "black", outlier.color = "red") +
      ggtitle(paste("Boxplot of", colname)) +
      theme_minimal()
  } else {
    NULL
  }
})
plots <- Filter(Negate(is.null), plots)
```

```
grid.arrange(grobs = plots, ncol = 5)
```



Feature selection

As we have seen above so many features are correlated and among 34 features many of them are not contributing in Y1/Y2.

So, we will use STEP wise model in both direction to selected only meaningful features.

```
full_model <- lm(Y ~ ., data = X)
stepwise_model_both <- step(full_model, direction = "both")
```

```
## Start: AIC=5881.99
## Y ~ TotalPop + Men + Women + Hispanic + White + Black + Native +
##      Asian + Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##      Poverty + ChildPoverty + Professional + Service + Office +
##      Construction + Production + Drive + Carpool + Transit + Walk +
```

```

##      OtherTransp + WorkAtHome + MeanCommute + Employed + PrivateWork +
##      PublicWork + SelfEmployed + FamilyWork
##
##
## Step: AIC=5881.99
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##      Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##      Poverty + ChildPoverty + Professional + Service + Office +
##      Construction + Production + Drive + Carpool + Transit + Walk +
##      OtherTransp + WorkAtHome + MeanCommute + Employed + PrivateWork +
##      PublicWork + SelfEmployed + FamilyWork
##
##              Df Sum of Sq   RSS   AIC
## - Professional     1    0.07 19623 5880.0
## - Construction     1    0.07 19623 5880.0
## - IncomePerCap     1    0.09 19623 5880.0
## - Production       1    0.28 19624 5880.0
## - OtherTransp       1    0.45 19624 5880.1
## - Service          1    0.46 19624 5880.1
## - Office           1    0.48 19624 5880.1
## - Pacific          1    0.52 19624 5880.1
## - FamilyWork        1    1.18 19624 5880.2
## - White            1    1.22 19624 5880.2
## - Carpool          1    1.50 19625 5880.2
## - WorkAtHome        1    1.54 19625 5880.2
## - Asian             1    1.57 19625 5880.2
## - Walk              1    1.65 19625 5880.3
## - Drive             1    1.68 19625 5880.3
## - Transit           1    1.71 19625 5880.3
## - PublicWork         1    1.87 19625 5880.3
## - PrivateWork        1    2.28 19626 5880.4
## - SelfEmployed       1    2.86 19626 5880.5
## - Hispanic           1    4.31 19628 5880.7
## <none>                  19623 5882.0
## - IncomeErr          1   12.21 19635 5882.0
## - TotalPop           1   12.97 19636 5882.1
## - Black              1   14.23 19637 5882.3
## - ChildPoverty        1   17.45 19641 5882.9
## - Men                1   18.15 19641 5883.0
## - IncomePerCapErr    1   31.10 19654 5885.1
## - Native              1   41.28 19665 5886.8
## - Citizen             1  133.62 19757 5901.8
## - Employed            1  136.90 19760 5902.4
## - Poverty             1 1304.86 20928 6087.2
## - MeanCommute         1 1808.70 21432 6163.7
##
## Step: AIC=5880
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##      Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##      Poverty + ChildPoverty + Service + Office + Construction +
##      Production + Drive + Carpool + Transit + Walk + OtherTransp +
##      WorkAtHome + MeanCommute + Employed + PrivateWork + PublicWork +
##      SelfEmployed + FamilyWork
##

```

```

##                                     Df Sum of Sq   RSS   AIC
## - Construction                  1   0.00 19623 5878.0
## - IncomePerCap                 1   0.09 19623 5878.0
## - OtherTransp                  1   0.45 19624 5878.1
## - Pacific                      1   0.53 19624 5878.1
## - FamilyWork                   1   1.19 19625 5878.2
## - White                        1   1.21 19625 5878.2
## - Carpool                      1   1.49 19625 5878.2
## - WorkAtHome                   1   1.53 19625 5878.3
## - Asian                        1   1.56 19625 5878.3
## - Walk                         1   1.64 19625 5878.3
## - Drive                        1   1.68 19625 5878.3
## - Transit                      1   1.71 19625 5878.3
## - PublicWork                   1   1.88 19625 5878.3
## - PrivateWork                  1   2.30 19626 5878.4
## - SelfEmployed                 1   2.88 19626 5878.5
## - Hispanic                     1   4.30 19628 5878.7
## <none>                         19623 5880.0
## - IncomeErr                    1   12.28 19636 5880.0
## - TotalPop                     1   13.01 19636 5880.1
## - Black                        1   14.20 19638 5880.3
## - ChildPoverty                 1   17.45 19641 5880.9
## - Men                          1   18.20 19642 5881.0
## + Professional                  1   0.07 19623 5882.0
## - IncomePerCapErr              1   31.11 19654 5883.1
## - Native                       1   41.25 19665 5884.8
## - Citizen                      1   133.56 19757 5899.8
## - Employed                     1   136.87 19760 5900.4
## - Production                   1   186.73 19810 5908.5
## - Office                       1   226.95 19850 5915.0
## - Service                      1   303.83 19927 5927.4
## - Poverty                      1   1306.11 20929 6085.4
## - MeanCommute                  1   1809.29 21433 6161.8
##
## Step: AIC=5878
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCap + IncomePerCapErr +
##     Poverty + ChildPoverty + Service + Office + Production +
##     Drive + Carpool + Transit + Walk + OtherTransp + WorkAtHome +
##     MeanCommute + Employed + PrivateWork + PublicWork + SelfEmployed +
##     FamilyWork
##
##                                     Df Sum of Sq   RSS   AIC
## - IncomePerCap                 1   0.09 19623 5876.0
## - OtherTransp                  1   0.45 19624 5876.1
## - Pacific                      1   0.52 19624 5876.1
## - FamilyWork                   1   1.19 19625 5876.2
## - White                        1   1.21 19625 5876.2
## - Carpool                      1   1.49 19625 5876.2
## - WorkAtHome                   1   1.53 19625 5876.3
## - Asian                        1   1.56 19625 5876.3
## - Walk                         1   1.64 19625 5876.3
## - Drive                        1   1.68 19625 5876.3
## - Transit                      1   1.71 19625 5876.3

```

```

## - PublicWork      1    1.89 19625 5876.3
## - PrivateWork    1    2.30 19626 5876.4
## - SelfEmployed   1    2.88 19626 5876.5
## - Hispanic        1    4.30 19628 5876.7
## <none>            19623 5878.0
## - IncomeErr       1    12.66 19636 5878.1
## - TotalPop        1    13.04 19636 5878.1
## - Black           1    14.20 19638 5878.3
## - ChildPoverty    1    17.98 19641 5878.9
## - Men              1    18.31 19642 5879.0
## + Construction    1    0.00 19623 5880.0
## + Professional    1    0.00 19623 5880.0
## - IncomePerCapErr 1    31.34 19655 5881.1
## - Native           1    41.25 19665 5882.8
## - Citizen          1    134.74 19758 5898.0
## - Employed         1    137.06 19760 5898.4
## - Production       1    225.90 19849 5912.8
## - Office            1    291.07 19914 5923.4
## - Service           1    353.57 19977 5933.5
## - Poverty           1    1401.59 21025 6098.0
## - MeanCommute       1    1858.90 21482 6167.3
##
## Step: AIC=5876.02
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCapErr + Poverty +
##     ChildPoverty + Service + Office + Production + Drive + Carpool +
##     Transit + Walk + OtherTransp + WorkAtHome + MeanCommute +
##     Employed + PrivateWork + PublicWork + SelfEmployed + FamilyWork
##
##                                     Df Sum of Sq   RSS   AIC
## - OtherTransp      1    0.46 19624 5874.1
## - Pacific          1    0.56 19624 5874.1
## - White             1    1.19 19625 5874.2
## - FamilyWork        1    1.22 19625 5874.2
## - Carpool           1    1.51 19625 5874.3
## - WorkAtHome        1    1.55 19625 5874.3
## - Walk              1    1.66 19625 5874.3
## - Asian              1    1.67 19625 5874.3
## - Drive              1    1.69 19625 5874.3
## - Transit            1    1.72 19625 5874.3
## - PublicWork         1    1.91 19625 5874.3
## - PrivateWork        1    2.33 19626 5874.4
## - SelfEmployed       1    2.92 19626 5874.5
## - Hispanic           1    4.27 19628 5874.7
## <none>                19623 5876.0
## - IncomeErr          1    12.89 19636 5876.1
## - TotalPop           1    13.34 19637 5876.2
## - Black              1    14.15 19638 5876.3
## - ChildPoverty        1    17.90 19641 5877.0
## - Men                1    18.42 19642 5877.0
## + IncomePerCap        1    0.09 19623 5878.0
## + Professional         1    0.00 19623 5878.0
## + Construction         1    0.00 19623 5878.0
## - IncomePerCapErr     1    32.24 19656 5879.3

```

```

## - Native          1   41.20 19665 5880.8
## - Citizen         1   138.56 19762 5896.7
## - Employed        1   140.32 19764 5896.9
## - Production      1   267.23 19891 5917.5
## - Office           1   293.82 19917 5921.8
## - Service          1   371.26 19995 5934.3
## - Poverty          1   1650.51 21274 6133.9
## - MeanCommute      1   1860.60 21484 6165.5
##
## Step: AIC=5874.09
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Pacific + Citizen + IncomeErr + IncomePerCapErr + Poverty +
##     ChildPoverty + Service + Office + Production + Drive + Carpool +
##     Transit + Walk + WorkAtHome + MeanCommute + Employed + PrivateWork +
##     PublicWork + SelfEmployed + FamilyWork
##
##              Df Sum of Sq   RSS   AIC
## - Pacific       1    0.55 19624 5872.2
## - White          1    1.18 19625 5872.3
## - FamilyWork     1    1.23 19625 5872.3
## - Asian           1    1.65 19626 5872.4
## - PublicWork      1    1.95 19626 5872.4
## - PrivateWork     1    2.36 19626 5872.5
## - SelfEmployed    1    2.95 19627 5872.6
## - Hispanic         1    4.25 19628 5872.8
## <none>                  19624 5874.1
## - IncomeErr       1   12.87 19637 5874.2
## - TotalPop         1   13.18 19637 5874.3
## - Black            1   14.12 19638 5874.4
## - ChildPoverty     1   17.89 19642 5875.0
## - Men              1   18.24 19642 5875.1
## + OtherTransp      1    0.46 19623 5876.0
## + IncomePerCap     1    0.10 19624 5876.1
## + Professional      1    0.00 19624 5876.1
## + Construction      1    0.00 19624 5876.1
## - IncomePerCapErr   1   32.41 19656 5877.4
## - Native            1   41.16 19665 5878.8
## - Walk              1   91.49 19715 5887.1
## - WorkAtHome        1   94.98 19719 5887.6
## - Carpool           1   96.20 19720 5887.8
## - Transit            1  111.45 19735 5890.3
## - Citizen           1  138.42 19762 5894.7
## - Employed          1  140.55 19764 5895.1
## - Drive              1  161.01 19785 5898.4
## - Production         1  266.77 19891 5915.5
## - Office             1  293.37 19917 5919.8
## - Service            1  371.02 19995 5932.4
## - Poverty            1  1650.57 21274 6132.0
## - MeanCommute        1  1862.17 21486 6163.8
##
## Step: AIC=5872.18
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Citizen + IncomeErr + IncomePerCapErr + Poverty + ChildPoverty +
##     Service + Office + Production + Drive + Carpool + Transit +

```

```

##      Walk + WorkAtHome + MeanCommute + Employed + PrivateWork +
##      PublicWork + SelfEmployed + FamilyWork
##
##              Df Sum of Sq   RSS   AIC
## - FamilyWork     1     1.19 19626 5870.4
## - PublicWork     1     1.88 19626 5870.5
## - Asian          1     1.90 19626 5870.5
## - PrivateWork    1     2.29 19627 5870.6
## - White          1     2.63 19627 5870.6
## - SelfEmployed   1     2.88 19627 5870.7
## - Hispanic        1     7.44 19632 5871.4
## <none>           19624 5872.2
## - TotalPop       1    13.29 19638 5872.4
## - IncomeErr      1    13.29 19638 5872.4
## - ChildPoverty   1    17.92 19642 5873.1
## - Men             1    18.27 19643 5873.2
## - Black           1    21.70 19646 5873.7
## + Pacific          1     0.55 19624 5874.1
## + OtherTransp     1     0.45 19624 5874.1
## + IncomePerCap    1     0.14 19624 5874.2
## + Professional     1     0.02 19624 5874.2
## + Construction     1     0.02 19624 5874.2
## - IncomePerCapErr 1    32.44 19657 5875.5
## - Native           1    57.20 19682 5879.5
## - Walk             1    91.40 19716 5885.1
## - WorkAtHome       1    95.45 19720 5885.8
## - Carpool          1    96.39 19721 5885.9
## - Transit          1   110.90 19735 5888.3
## - Citizen          1   138.40 19763 5892.8
## - Employed         1   140.06 19764 5893.1
## - Drive            1   160.94 19785 5896.5
## - Production       1   266.57 19891 5913.6
## - Office           1   292.83 19917 5917.8
## - Service          1   371.30 19996 5930.5
## - Poverty          1   1650.19 21275 6130.0
## - MeanCommute      1   1862.05 21486 6161.9
##
## Step:  AIC=5870.38
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Asian +
##     Citizen + IncomeErr + IncomePerCapErr + Poverty + ChildPoverty +
##     Service + Office + Production + Drive + Carpool + Transit +
##     Walk + WorkAtHome + MeanCommute + Employed + PrivateWork +
##     PublicWork + SelfEmployed
##
##              Df Sum of Sq   RSS   AIC
## - Asian          1     1.88 19627 5868.7
## - White          1     2.59 19628 5868.8
## - PublicWork     1     4.44 19630 5869.1
## - Hispanic        1     7.37 19633 5869.6
## - PrivateWork    1     9.84 19635 5870.0
## <none>           19626 5870.4
## - IncomeErr      1    13.17 19639 5870.5
## - TotalPop       1    13.35 19639 5870.6
## - ChildPoverty   1    17.92 19644 5871.3

```

```

## - Men          1   18.33 19644 5871.4
## - SelfEmployed 1   18.76 19644 5871.5
## - Black        1   21.59 19647 5871.9
## + FamilyWork   1    1.19 19624 5872.2
## + Pacific       1    0.51 19625 5872.3
## + OtherTransp  1    0.47 19625 5872.3
## + IncomePerCap 1    0.17 19625 5872.3
## + Professional  1    0.03 19626 5872.4
## + Construction 1    0.03 19626 5872.4
## - IncomePerCapErr 1   32.52 19658 5873.7
## - Native        1   57.05 19683 5877.7
## - Walk          1   91.05 19717 5883.3
## - WorkAtHome    1   94.99 19721 5883.9
## - Carpool        1   96.17 19722 5884.1
## - Transit        1  110.47 19736 5886.4
## - Citizen        1  138.12 19764 5890.9
## - Employed       1  139.57 19765 5891.2
## - Drive          1  160.39 19786 5894.6
## - Production     1  266.31 19892 5911.8
## - Office          1  292.39 19918 5916.0
## - Service         1  372.25 19998 5928.8
## - Poverty         1 1651.10 21277 6128.3
## - MeanCommute    1 1861.65 21487 6160.0
##
## Step: AIC=5868.69
## Y ~ TotalPop + Men + Hispanic + White + Black + Native + Citizen +
##      IncomeErr + IncomePerCapErr + Poverty + ChildPoverty + Service +
##      Office + Production + Drive + Carpool + Transit + Walk +
##      WorkAtHome + MeanCommute + Employed + PrivateWork + PublicWork +
##      SelfEmployed
##
##              Df Sum of Sq   RSS   AIC
## - White          1     0.72 19628 5866.8
## - PublicWork     1     4.29 19632 5867.4
## - Hispanic        1     8.46 19636 5868.1
## - PrivateWork    1     9.57 19637 5868.3
## <none>           19627 5868.7
## - IncomeErr      1    13.46 19641 5868.9
## - TotalPop       1    15.03 19643 5869.1
## - SelfEmployed   1    18.52 19646 5869.7
## - ChildPoverty   1    18.78 19646 5869.8
## - Men            1    20.06 19648 5870.0
## + Asian          1     1.88 19626 5870.4
## + FamilyWork     1     1.16 19626 5870.5
## + Pacific         1     0.74 19627 5870.6
## + OtherTransp    1     0.45 19627 5870.6
## + IncomePerCap   1     0.34 19627 5870.6
## + Professional    1     0.15 19627 5870.7
## + Construction   1     0.14 19627 5870.7
## - IncomePerCapErr 1   32.69 19660 5872.0
## - Black          1   41.23 19669 5873.4
## - Walk           1   89.48 19717 5881.3
## - WorkAtHome     1   93.80 19721 5882.0
## - Carpool        1   96.01 19723 5882.4

```

```

## - Transit      1  108.59 19736 5884.4
## - Native       1  130.55 19758 5888.0
## - Employed    1  137.81 19765 5889.2
## - Citizen      1  137.98 19765 5889.2
## - Drive        1  159.33 19787 5892.7
## - Production   1  266.28 19894 5910.0
## - Office        1  290.81 19918 5914.0
## - Service       1  370.38 19998 5926.8
## - Poverty       1  1658.34 21286 6127.7
## - MeanCommute   1  1867.38 21495 6159.1
##
## Step: AIC=5866.8
## Y ~ TotalPop + Men + Hispanic + Black + Native + Citizen + IncomeErr +
##     IncomePerCapErr + Poverty + ChildPoverty + Service + Office +
##     Production + Drive + Carpool + Transit + Walk + WorkAtHome +
##     MeanCommute + Employed + PrivateWork + PublicWork + SelfEmployed
##
##                                     Df Sum of Sq   RSS   AIC
## - PublicWork      1      4.41 19633 5865.5
## - PrivateWork     1      9.73 19638 5866.4
## <none>                  19628 5866.8
## - IncomeErr       1     13.23 19641 5867.0
## - TotalPop        1     14.34 19643 5867.2
## - ChildPoverty    1     18.64 19647 5867.9
## - SelfEmployed    1     18.66 19647 5867.9
## - Men              1     19.41 19648 5868.0
## + Pacific          1      1.45 19627 5868.6
## + FamilyWork       1      1.15 19627 5868.6
## + White            1      0.72 19627 5868.7
## + OtherTransp      1      0.44 19628 5868.7
## + IncomePerCap     1      0.17 19628 5868.8
## + Professional     1      0.07 19628 5868.8
## + Construction     1      0.07 19628 5868.8
## + Asian             1      0.00 19628 5868.8
## - IncomePerCapErr 1     32.21 19660 5870.1
## - Walk              1     91.18 19719 5879.7
## - WorkAtHome        1     93.99 19722 5880.2
## - Hispanic          1     94.18 19722 5880.2
## - Carpool           1     97.14 19725 5880.7
## - Transit            1    113.83 19742 5883.4
## - Citizen            1    137.63 19766 5887.3
## - Employed          1    140.27 19768 5887.7
## - Drive              1    159.72 19788 5890.9
## - Production         1    268.68 19897 5908.6
## - Office              1    290.40 19919 5912.1
## - Service             1    370.61 19999 5925.0
## - Black              1    444.38 20073 5936.8
## - Native              1    528.53 20157 5950.3
## - Poverty             1   1663.64 21292 6126.6
## - MeanCommute         1   1871.44 21500 6157.9
##
## Step: AIC=5865.53
## Y ~ TotalPop + Men + Hispanic + Black + Native + Citizen + IncomeErr +
##     IncomePerCapErr + Poverty + ChildPoverty + Service + Office +

```

```

##      Production + Drive + Carpool + Transit + Walk + WorkAtHome +
##      MeanCommute + Employed + PrivateWork + SelfEmployed
##
##              Df Sum of Sq   RSS   AIC
## <none>                   19633 5865.5
## - IncomeErr      1     12.47 19645 5865.6
## - TotalPop       1     13.95 19647 5865.8
## - ChildPoverty    1     18.59 19651 5866.6
## - Men            1     19.03 19652 5866.6
## + PublicWork      1      4.41 19628 5866.8
## + FamilyWork      1      3.73 19629 5866.9
## + Pacific          1      1.45 19631 5867.3
## + White           1      0.84 19632 5867.4
## + OtherTransp      1      0.52 19632 5867.4
## + IncomePerCap     1      0.07 19633 5867.5
## + Asian            1      0.01 19633 5867.5
## + Professional      1      0.00 19633 5867.5
## + Construction      1      0.00 19633 5867.5
## - IncomePerCapErr   1     33.27 19666 5869.0
## - Walk             1     89.39 19722 5878.1
## - WorkAtHome        1     90.40 19723 5878.3
## - Hispanic          1     92.16 19725 5878.6
## - Carpool           1     95.81 19728 5879.2
## - Transit            1    112.82 19745 5882.0
## - PrivateWork        1    117.71 19750 5882.8
## - Citizen            1    136.74 19769 5885.9
## - Employed           1    140.90 19774 5886.5
## - Drive              1    157.91 19791 5889.3
## - SelfEmployed        1    165.57 19798 5890.6
## - Production          1    269.31 19902 5907.4
## - Office              1    289.11 19922 5910.6
## - Service             1    367.91 20001 5923.3
## - Black               1    441.73 20074 5935.1
## - Native              1    524.98 20158 5948.4
## - Poverty             1   1667.40 21300 6125.8
## - MeanCommute          1   1867.03 21500 6155.9

```

```
summary(stepwise_model_both)
```

```

##
## Call:
## lm(formula = Y ~ TotalPop + Men + Hispanic + Black + Native +
##     Citizen + IncomeErr + IncomePerCapErr + Poverty + ChildPoverty +
##     Service + Office + Production + Drive + Carpool + Transit +
##     Walk + WorkAtHome + MeanCommute + Employed + PrivateWork +
##     SelfEmployed, data = X)
##
## Residuals:
##      Min        1Q        Median         3Q        Max 
## -15.3955  -1.3416  -0.0866   1.2015  15.5527 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.317e+01  3.309e+00  3.980 7.04e-05 ***
```

```

## TotalPop      -1.691e-05  1.123e-05 -1.507  0.131986
## Men          3.783e-05  2.150e-05  1.760  0.078569 .
## Hispanic     1.205e-02  3.111e-03  3.873  0.000110 ***
## Black         3.354e-02  3.955e-03  8.479  < 2e-16 ***
## Native        7.231e-02  7.823e-03  9.243  < 2e-16 ***
## Citizen       1.402e-05  2.972e-06  4.717  2.49e-06 ***
## IncomeErr     -4.828e-05  3.390e-05 -1.424  0.154429
## IncomePerCapErr -1.465e-04  6.296e-05 -2.327  0.020033 *
## Poverty       2.743e-01  1.665e-02 16.473  < 2e-16 ***
## ChildPoverty   -2.003e-02  1.152e-02 -1.739  0.082085 .
## Service        1.153e-01  1.490e-02  7.738  1.35e-14 ***
## Office          1.199e-01  1.749e-02  6.859  8.28e-12 ***
## Production     7.540e-02  1.139e-02  6.620  4.19e-11 ***
## Drive           -1.588e-01  3.133e-02 -5.069  4.22e-07 ***
## Carpool         -1.418e-01  3.590e-02 -3.949  8.03e-05 ***
## Transit          -1.605e-01  3.745e-02 -4.285  1.88e-05 ***
## Walk            -1.542e-01  4.042e-02 -3.814  0.000139 ***
## WorkAtHome      -1.411e-01  3.679e-02 -3.836  0.000128 ***
## MeanCommute     1.607e-01  9.222e-03 17.431  < 2e-16 ***
## Employed        -2.246e-05  4.690e-06 -4.789  1.76e-06 ***
## PrivateWork     -4.424e-02  1.011e-02 -4.377  1.24e-05 ***
## SelfEmployed    -1.000e-01  1.927e-02 -5.191  2.22e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.479 on 3195 degrees of freedom
## Multiple R-squared:  0.6357 , Adjusted R-squared:  0.6332
## F-statistic: 253.4 on 22 and 3195 DF, p-value: < 2.2e-16

```

```

selected_features_step <- names(coef(stepwise_model_both))[-1]
selected_features_step

```

```

## [1] "TotalPop"      "Men"          "Hispanic"      "Black"
## [5] "Native"        "Citizen"       "IncomeErr"     "IncomePerCapErr"
## [9] "Poverty"        "ChildPoverty"  "Service"       "Office"
## [13] "Production"    "Drive"         "Carpool"       "Transit"
## [17] "Walk"          "WorkAtHome"   "MeanCommute"  "Employed"
## [21] "PrivateWork"   "SelfEmployed"

```

Filtering out these features

```
X=X[selected_features_step]
```

Base Model with main effect before diagnosis

```

fit1 <- lm(Y ~ ., data = X)
base_model_main_effect <- summary(fit1)
r2 <- base_model_main_effect$r.squared
adj_r2 <- base_model_main_effect$adj.r.squared

```

Base Model diagnosis with main effect

```
cat("R2:", r2, "\n")

## R2: 0.6357207

cat("Adjusted R2:", adj_r2, "\n")

## Adjusted R2: 0.6332124

cat("AIC:", AIC(fit1), "\n")

## AIC: 14999.82

cat("BIC:", BIC(fit1), "\n")

## BIC: 15145.65

print(dwtest(fit1))

## 
## Durbin-Watson test
##
## data: fit1
## DW = 1.7425, p-value = 7.563e-14
## alternative hypothesis: true autocorrelation is greater than 0

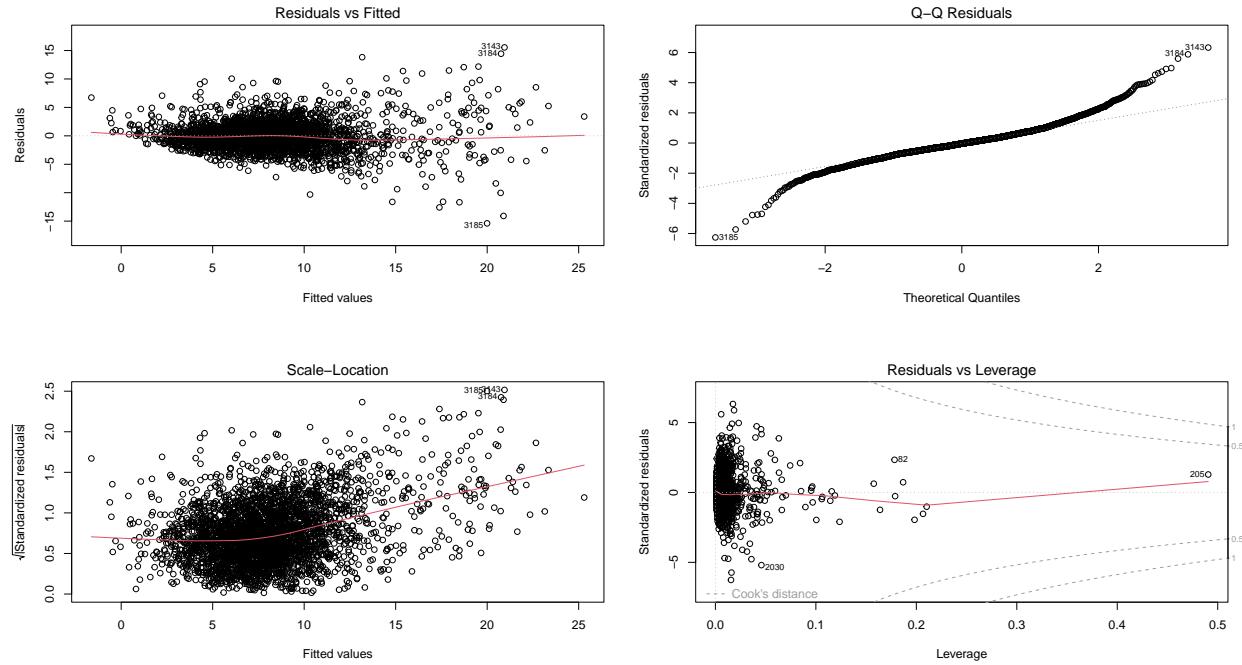
print(shapiro.test(residuals(fit1)))

## 
## Shapiro-Wilk normality test
##
## data: residuals(fit1)
## W = 0.94976, p-value < 2.2e-16

print(vif(fit1))

##          TotalPop        Men      Hispanic       Black       Native
## 6729.858380 5943.820288 1.876346 1.670589 1.685241
##      Citizen IncomeErr IncomePerCapErr     Poverty ChildPoverty
## 194.670501  2.206393   2.239038 10.044622 9.494422
##      Service    Office   Production      Drive    Carpool
## 1.536147   1.634983   2.235852 29.793033 5.707091
##      Transit      Walk WorkAtHome MeanCommute Employed
## 6.876556   11.706473   7.134674 1.394586 258.166916
## PrivateWork SelfEmployed
## 3.269309   2.977632
```

```
par(mfrow = c(2, 2))
plot(fit1)
```



As we can see we got good R² but assumptions failed.

Treatment in main effect

Dropping features with VIF<5

```
vif_values <- vif(fit1)
low_vif_features <- names(vif_values)[vif_values < 5]
print(low_vif_features)

## [1] "Hispanic"          "Black"           "Native"          "IncomeErr"
## [5] "IncomePerCapErr"   "Service"         "Office"          "Production"
## [9] "MeanCommute"       "PrivateWork"     "SelfEmployed"

X <- X[low_vif_features]
dim(X)

## [1] 3218    11
```

Transformation

```

Y <- bestNormalize::yeojohnson(Y)$x.t
X[abs(apply(X, 2, e1071::skewness)) > 1] <- lapply(X[abs(apply(X, 2, e1071::skewness)) > 1], log1p) # .
head(X)

##   Hispanic     Black    Native IncomeErr IncomePerCapErr Service Office
## 1 1.280934 2.9704145 0.3364722 7.779885      6.985642 17.0 24.2
## 2 1.704748 2.3513753 0.4700036 7.142037      6.568078 17.7 27.1
## 3 1.722767 3.8649314 0.1823216 7.997663      6.683361 16.1 23.1
## 4 1.163151 3.1090610 0.3364722 8.293049      7.389564 17.9 17.8
## 5 2.261763 0.9162907 0.2623643 8.052615      6.563856 14.1 23.9
## 6 1.686399 4.2724907 0.7884574 8.680162      7.628518 15.0 19.7
##   Production MeanCommute PrivateWork SelfEmployed
## 1          17.1        26.5   4.312141  1.871802
## 2          11.2        26.4   4.412798  1.916923
## 3          23.1        24.1   4.287716  2.116256
## 4          23.7        28.8   4.354141  2.041220
## 5          19.9        34.9   4.418841  1.648659
## 6          26.4        27.5   4.388257  1.856298

```

Base model after diagnosis(VIF filter+transformation) with main effect

```

fit2 <- lm(Y ~ ., data = X)
base_model_main_effect <- summary(fit2)
r2 <- base_model_main_effect$r.squared
adj_r2 <- base_model_main_effect$adj.r.squared

cat("R2:", r2, "\n")

## R2: 0.4901547

cat("Adjusted R2:", adj_r2, "\n")

## Adjusted R2: 0.4884054

cat("AIC:", AIC(fit2), "\n")

## AIC: 6989.489

cat("BIC:", BIC(fit2), "\n")

## BIC: 7068.484

print(dwtest(fit2))

##
## Durbin-Watson test
##
## data: fit2
## DW = 1.574, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0

```

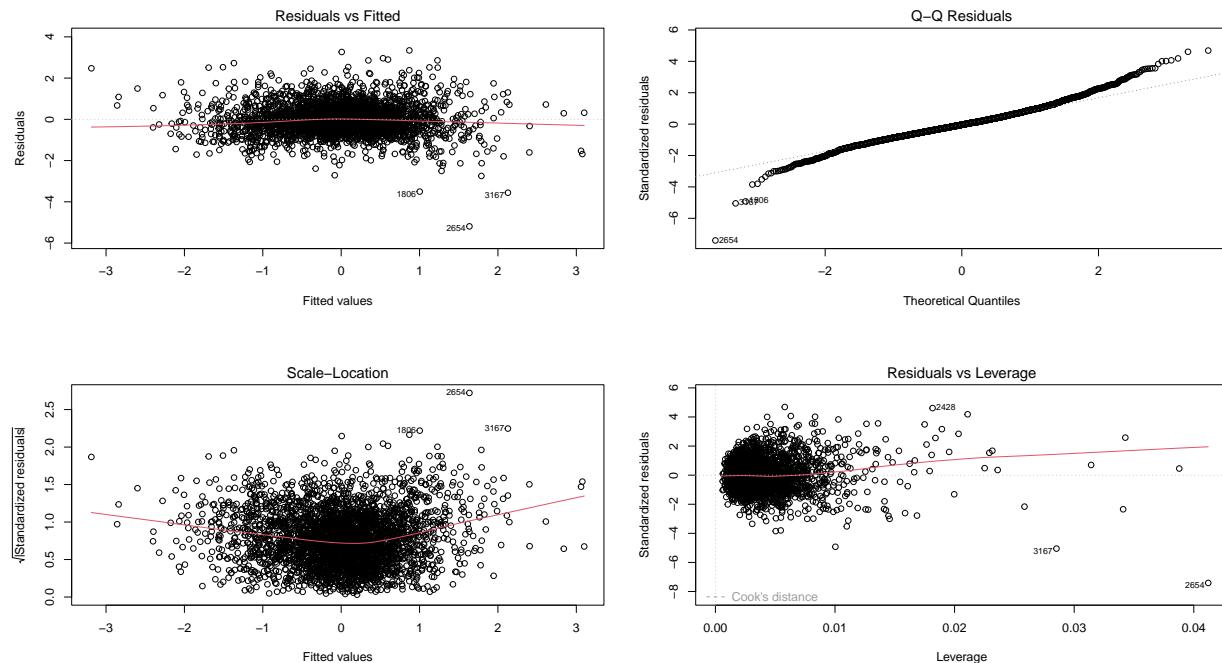
```
print(shapiro.test(residuals(fit2)))
```

```
##  
## Shapiro-Wilk normality test  
##  
## data: residuals(fit2)  
## W = 0.97761, p-value < 2.2e-16
```

```
print(vif(fit2))
```

```
##          Hispanic            Black           Native      IncomeErr IncomePerCapErr  
## 1.137995 1.366685 1.351824 3.453582 3.722128  
## Service        Office Production MeanCommute PrivateWork  
## 1.203293 1.460795 1.462111 1.190094 2.205046  
## SelfEmployed  
## 1.884212
```

```
par(mfrow = c(2, 2))  
plot(fit2)
```



```
## Dropping outliers for cooks distance
```

```
dim(X)
```

```
## [1] 3218 11
```

```

cooks_dist <- cooks.distance(fit2)
threshold <- 4 / length(cooks_dist)
influential_points <- which(cooks_dist > threshold)
X <- X[-influential_points, ]
Y <- Y[-influential_points]

```

Base model after diagnosis(Influential points) with main effect

```

fit3 <- lm(Y ~ ., data = X)
base_model_main_effect <- summary(fit3)
r2 <- base_model_main_effect$r.squared
adj_r2 <- base_model_main_effect$adj.r.squared

cat("R2:", r2, "\n")

## R2: 0.6018305

cat("Adjusted R2:", adj_r2, "\n")

## Adjusted R2: 0.6003483

cat("AIC:", AIC(fit3), "\n")

## AIC: 4812.527

cat("BIC:", BIC(fit3), "\n")

## BIC: 4890.466

print(dwtest(fit3))

##
## Durbin-Watson test
##
## data: fit3
## DW = 1.5639, p-value < 2.2e-16
## alternative hypothesis: true autocorrelation is greater than 0

print(shapiro.test(residuals(fit3)))

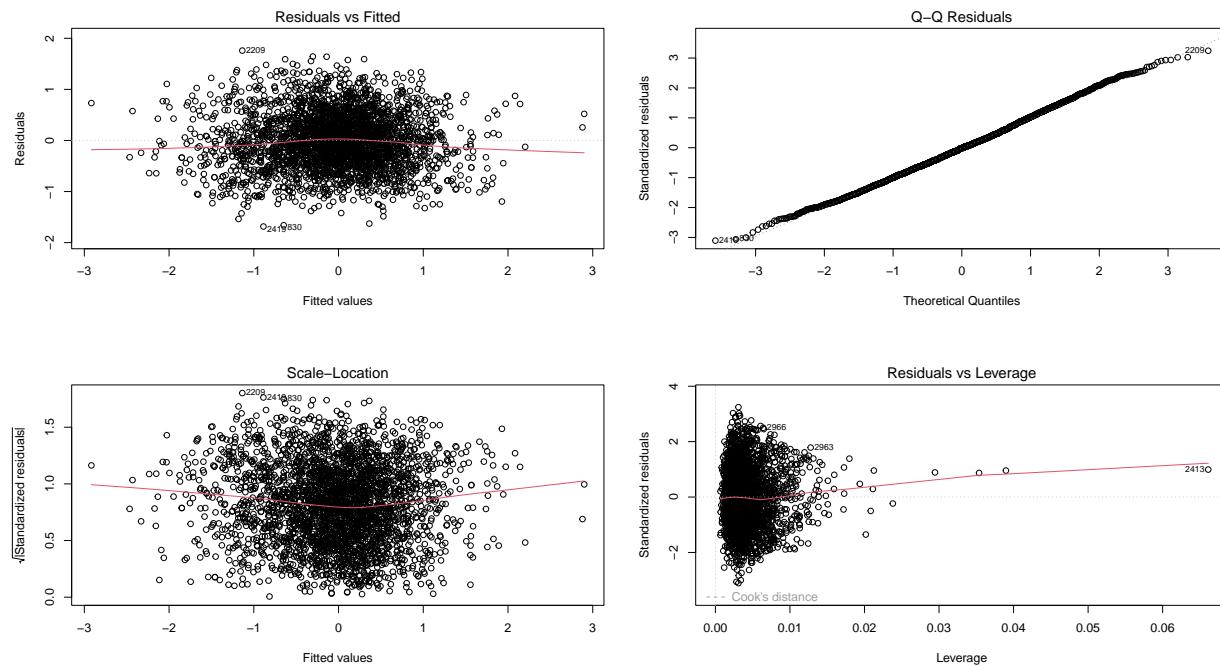
##
## Shapiro-Wilk normality test
##
## data: residuals(fit3)
## W = 0.99816, p-value = 0.001644

```

```
print(vif(fit3))
```

```
##          Hispanic           Black          Native      IncomeErr IncomePerCapErr
## 1.120317    1.418119    1.233576    3.443550     3.723739
##       Service         Office Production MeanCommute PrivateWork
## 1.194446    1.507933    1.512994    1.157602     2.264456
## SelfEmployed
## 1.998758
```

```
par(mfrow = c(2, 2))
plot(fit3)
```



Only multicollinearity has been passed, but plot has been improved.

two main effect

All combination of two main effect has been implemented and the best features extracted using STEP.

```
full_model <- lm(Y ~ .^2, data = X)
step_model <- step(full_model, direction = "both", trace = FALSE)
final_formula <- formula(step_model)
final_fit <- lm(final_formula, data = X)
cooks_dist <- cooks.distance(final_fit)
threshold <- 4 / length(cooks_dist)
influential_points <- which(cooks_dist > threshold)
```

```

X <- X[!influential_points, , drop = FALSE]
Y <- Y[!influential_points]
final_fit <- lm(final_formula, data = X)
model_summary <- summary(final_fit)
r2 <- model_summary$r.squared
adj_r2 <- model_summary$adj.r.squared
cat("R2:", r2, "\n")

## R2: 0.6880194

cat("Adjusted R2:", adj_r2, "\n")

## Adjusted R2: 0.6835918

cat("AIC:", AIC(final_fit), "\n")

## AIC: 3642.337

cat("BIC:", BIC(final_fit), "\n")

## BIC: 3885.593

print(final_formula)

## Y ~ Hispanic + Black + Native + IncomeErr + IncomePerCapErr +
##     Service + Office + Production + MeanCommute + PrivateWork +
##     SelfEmployed + Hispanic:Black + Hispanic:Native + Hispanic:Service +
##     Hispanic:Production + Black:Native + Black:IncomePerCapErr +
##     Black:Service + Black:Office + Black:MeanCommute + Native:Service +
##     Native:PrivateWork + IncomeErr:IncomePerCapErr + IncomeErr:Service +
##     IncomeErr:MeanCommute + IncomeErr:PrivateWork + IncomeErr:SelfEmployed +
##     IncomePerCapErr:Office + IncomePerCapErr:Production + IncomePerCapErr:PrivateWork +
##     Service:Office + Service:Production + Office:Production +
##     Office:SelfEmployed + Production:MeanCommute + Production:PrivateWork +
##     Production:SelfEmployed + MeanCommute:PrivateWork + MeanCommute:SelfEmployed

print(dwtest(final_fit))

##
## Durbin-Watson test
##
## data: final_fit
## DW = 1.7045, p-value = 1.321e-15
## alternative hypothesis: true autocorrelation is greater than 0

print(shapiro.test(residuals(final_fit)))

```

```

##  

## Shapiro-Wilk normality test  

##  

## data: residuals(final_fit)  

## W = 0.9989, p-value = 0.0707

```

```

par(mfrow = c(2, 2))
plot(final_fit)

```

