1: Introduction to Data Set

The number and types of motor vehicle crash deaths differ widely among the 50 states and the District of Columbia in the U.S. Fatality rates per capita and per vehicle miles travelled provide a way of examining motor vehicle deaths relative to the population and amount of driving. However, many factors can affect these rates, including types of vehicles driven, single or multiple vehicle collision, whether the passengers were wearing a seatbelt or not (restrained fatally injured occupants versus unrestrained fatally injured occupants), whether the accident occurred in an urban or rural area

2. Descriptive Analysis

The 2016 Car Crash dataset contains information on car accidents and fatalities in the United States. There are 51 observations for 50 states plus district of Columbia. There are 21 variables, 20 variables are numeric in nature. For each state it contains information on 21 facts (variables): State, Population, Vehicle miles travelled (millions), Fatal crashes, Deaths, Deaths per 100,000 population, Deaths per 100 million vehicle miles travelled, Car occupants, Pickup and SUV occupants, Large truck occupants, Motorcyclists, Pedestrians, Bicyclists, Unknown mode of transport, Single-vehicle, Multiple-vehicle, Unrestrained fatally injured occupants, Restrained fatally injured occupants, Unknown restraint status of fatally injured occupants, Urban and Rural. My first task was to check the dataset for missing values and or zero values. I found there to be no 'na' values and any integer value of zero made sense for that particular data set.

5 Number Summaries/ Str

All the variables in the dataset are integers with the exception of one character variable – State. The report is an analysis of fatality rates per capita so the 'Deaths per 100,000 population' variable was selected as the response variable. It is an integer variable with values ranging throughout the 51 states from min 3 in district of Columbia to max 25 in Wyoming.

Correlation

The correlation tests looking first at the response variable - 'Deaths per 100,000 population'. High Correlation with Deaths. Per. 100.million. vehicle.miles.traveled and also relatively high with Unrestrained.fatally.injured.occupants. For the variable Vehicle miles travelled (millions) I found it interesting that there is a relatively small positive correlation between this and Deaths.per.100.million.vehicle.miles.traveled .16 and Deaths.per.100.000.population .11. A preconceived idea of mine would have been that the more miles travelled would result in a strong positive correlation >.75 with the two latter variables.

I also looked at multicollinearity between the other variables with a view to looking at what variables could be dropped to build a more efficient model. I found a high degree of collinearity between Fatal Crashes and Deaths – perfectly correlated when rounded to 2 decimal places. There is also a high degree of correlation between single vehicle and multiple vehicle and pedestrians and cyclists.

Unrestrained fatally injured more highly correlated with rural deaths which tells me that road users are more likely to be unrestrained when driving in a rural area. This speaks to the urban rural divide in the United States that is evident to me when exploring the dataset.

ii. Graphical summaries:

I used the pairs function for a visual representation of correlation between the variables. This confirmed in graphs what is spoken about in ***Correlation*** above, eg please see appendix below and notice the graphs of fatal crashes and deaths are moving diagonally upward which indicates a very strong correlation between these two variables. As these two variables are highly correlated and are inferred from the response variable these were dropped from the final model.

Model Building

Firstly, the generalized linear model was fitted to all variables. The proportion deviance explained (generalization of R-squared for linear models) informs that 63% of the deaths per 100,000 population is explained by all other explanatory variables. There was an AIC rate of **269.3467.** Analysis of variance showed variables population, vehicle miles travelled, fatal crashes, Deaths.per.100.million.vehicle.miles.traveled and single vehicle to be the most statistically significant.

ii. t-tests

Running the t-test with all predictors included we see that population has a t value of -3.265 which means that as population increases deaths per 100,000 population decreases. The other variable to show a strong value here is Deaths.per.100.million.vehicle.miles.traveled. This is a binary variable that shows a positive value – this variable value increases as dependent variable increases. Conversely, urban variable decreases (-2.378) ie deaths in urban areas decrease as deaths per 100,000 population increases.

iii. F-tests ,

F – tests on the model that includes all variables show significant difference from zero of variables population, Vehicle.miles.traveled..millions. , fatal crashes, Deaths.per.100.million.vehicle.miles.traveled and Single.vehicle. These variables were dropped for the final model due to multicollinearity. Pick up and SUV vehicle always show strong F-values on the models that I ran. On model 1 Pick up and SUV had an F-value of 36.825 which is significantly different from zero and the other variables. This variable is consistently important across the models that I have run.

ANOVA (c) Perform analysis of the best regression

Population variable was dropped due to outliers and the option of using a log of the population to make it more normally distributed was decided against as the dependent variable of deaths per 100,000 population is sufficient for the population needs of this model. Fatal crashes and deaths variable was dropped as both are highly correlated with one another. Analysis of best regression ie that led to the best model fit was model 1 in R code. An R value of almost **70%(.6987)** and an AIC value of **228.5519.**

i. Coefficients and model fit

Model one show's residuals that are quite normally distributed – (Min)-4.7505; (1Q)-1.0538; (Median)0.0025; (3Q)0.9634 (Max) 5.0555. When interpreting the t values we are looking for values significantly different form zero. In terms of this the most significant variable is Deaths.per.100.million.vehicle.miles.traveled – a positive value of 2.651 ie. This variable increase as Deaths per 100,000 population increases. The other variable to show significance is the urban deaths variable – urban deaths decrease by -2.378 as deaths per 100,000 population increases.

ii. **Model Diagnostics**

The best model from my r code was model 1 which explained almost 70% of the variation in deaths per 100,000 population is explained by the variation in all other explanatory variables – per the changes that I made in this model

iii. **Outliers / Influential values**

Upon running the model with all variables and rows included, I checked the plots for outliers which showed the states Texas – which had the highest Cook's value point – and has both large urban and rural areas, Washington D.C. which is an urban area featuring zero Pickup and SUV occupants and Large truck occupants deaths. Other State outliers were North Dakota and Wyoming both rural areas with low population. States(T, DC, ND W). Model 2 was based on outliers in terms of populations.

iv. Possible Remedial Measures

On my first alternative model I dropped rows 9, 35, 44 and 51 - States(T, DC, ND W) – these were outliers on my residual graphs. I also dropped the columns 2,3,4 and 5 due to multicollinearity discussed above. This model was the best overall fit with

On model 2 I dropped the four highest and four lowest population centre rows as these were distorting the data. The Cook's distance measure show's that state's Hawaii and South Carolina are outliers.

I also brought in another column - Area in square miles – which enabled me to create a new column in the r dataset - square.miles_deaths.peronehundthoupop. I did this by dividing deaths per 100,000 population by area. This enabled me to identify Washington DC and Delaware as outlier states in terms of deaths per 100,000 population to area in square miles ratio.

Conclusions

The main takeaway from my analysis of the data was the rural urban divide best encapsulated by the state Texas with both big rural and urban areas

- The four outlier states from model 1 (T, DC, ND W) should have a separate analysis run on them. These are all States with particularities. Texas is a big state in terms of population but also area. It has big urban and rural areas. District of Columbia is really an administrative region – the smallest in terms of area – as such there are no rural fatalities. North Dakota and Wyoming are rural areas.

- I would recommend running a separate analysis on the four biggest and four smallest states run separate analysis for them with information on individual counties. Texas for example is a highly populous state with a huge rural area as well. A separate analysis is recommended for this State.

- Population is inversely related to death. The higher the population the lower deaths per 100,000 population. Urban areas are more tightly policed leading to improved driver behaviour. Premiums should decrease if a driver lives in an urban area.

- Regardless of the model that I ran Pick up and SUV models was a significant variable. These are normally associated with rural areas. I would recommend a higher premium on this type of vehicle especially in rural areas. Unrestrained fatally injured deaths are also more likely to occur in rural areas which featured a .**90** correlation while urban areas were less correlated **.82**

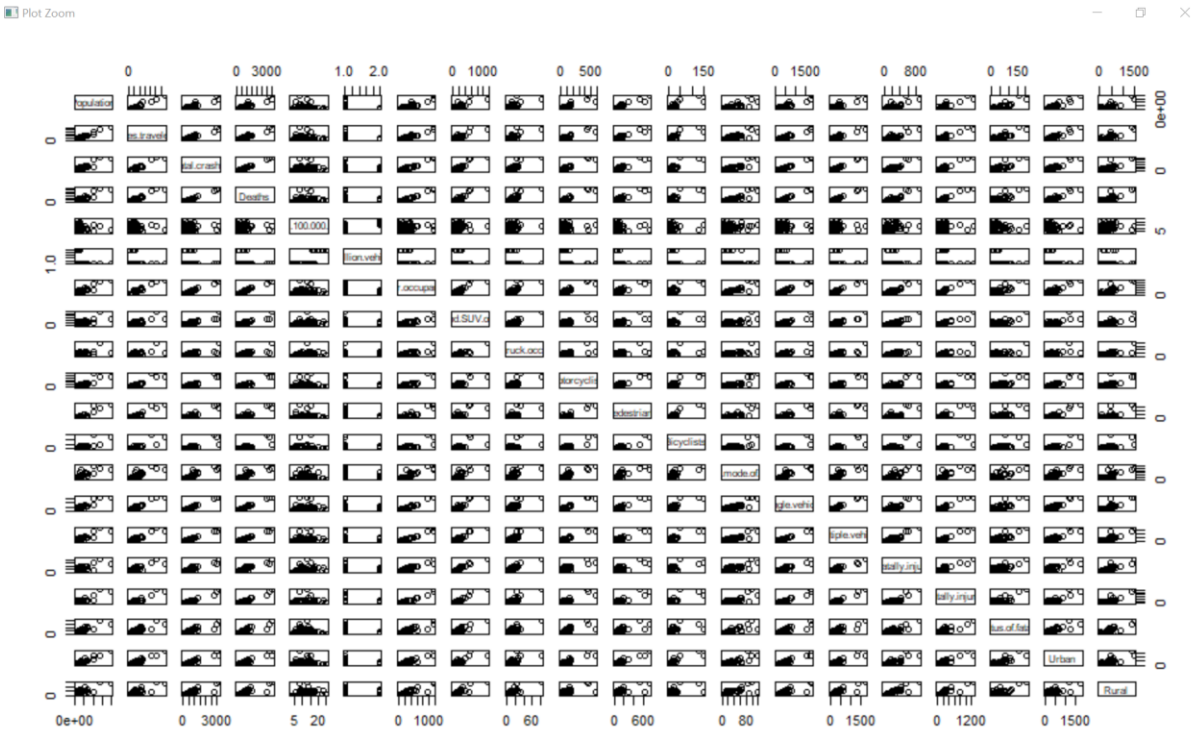
- When states are ordered in decreasing value per state area, we see that 4 of the first five states in terms of area have more pickup and suv fatalities than car fatalities. There are only 11 States where Pick up and SUV are greater than car fatalities. Another reason that I would recommend a higher premium for Pick up and SUVs

- District of Columbia is a big outlier in terms of the area size of the State and deaths per 100,000 population. I would recommend an increase in premium for people living in this State

- From my model four hypotheses test I would charge a higher premium to those motorists living in a state where the deaths per 100m miles travelled is 2.

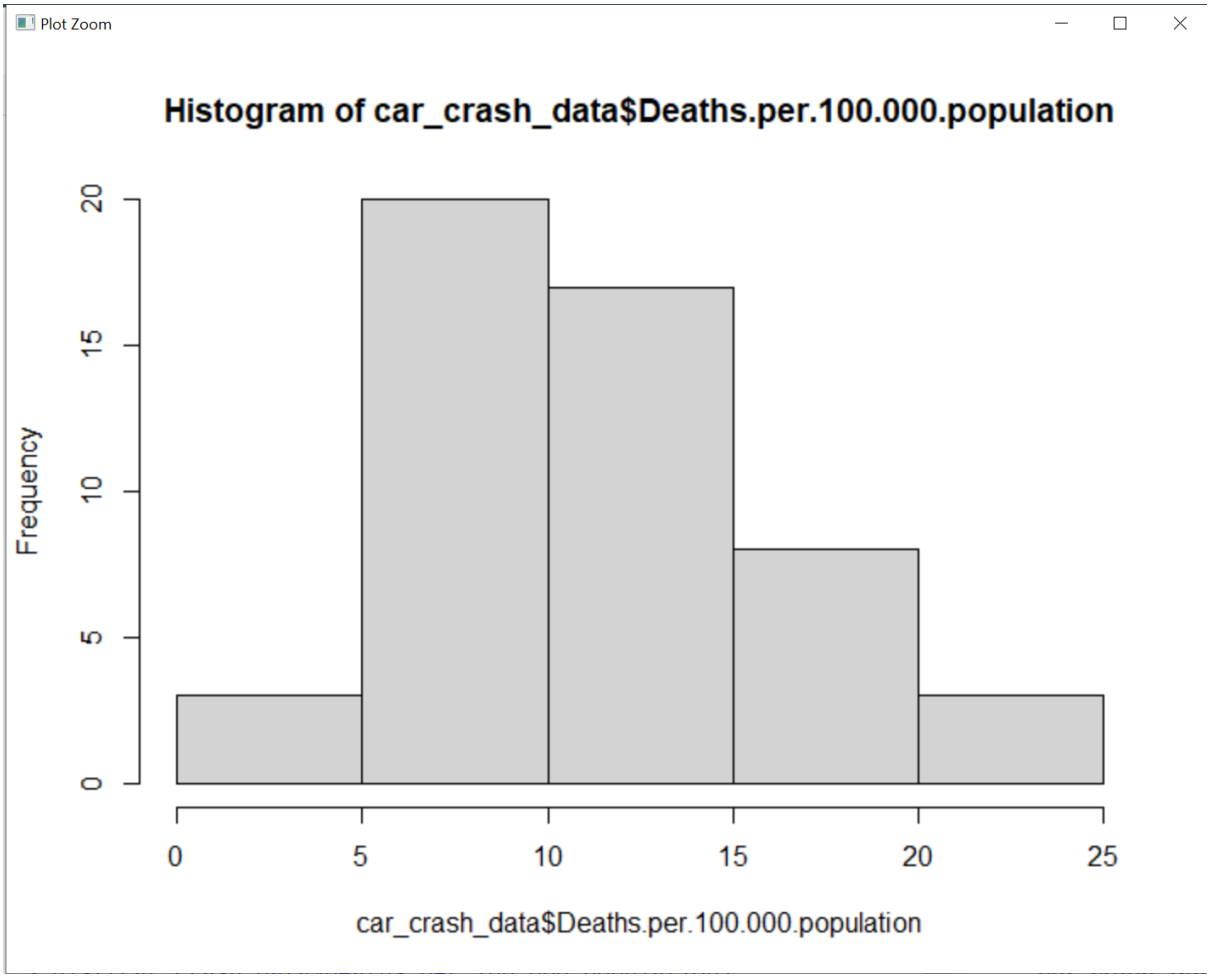
Appendix

Pairs function correlation between variables

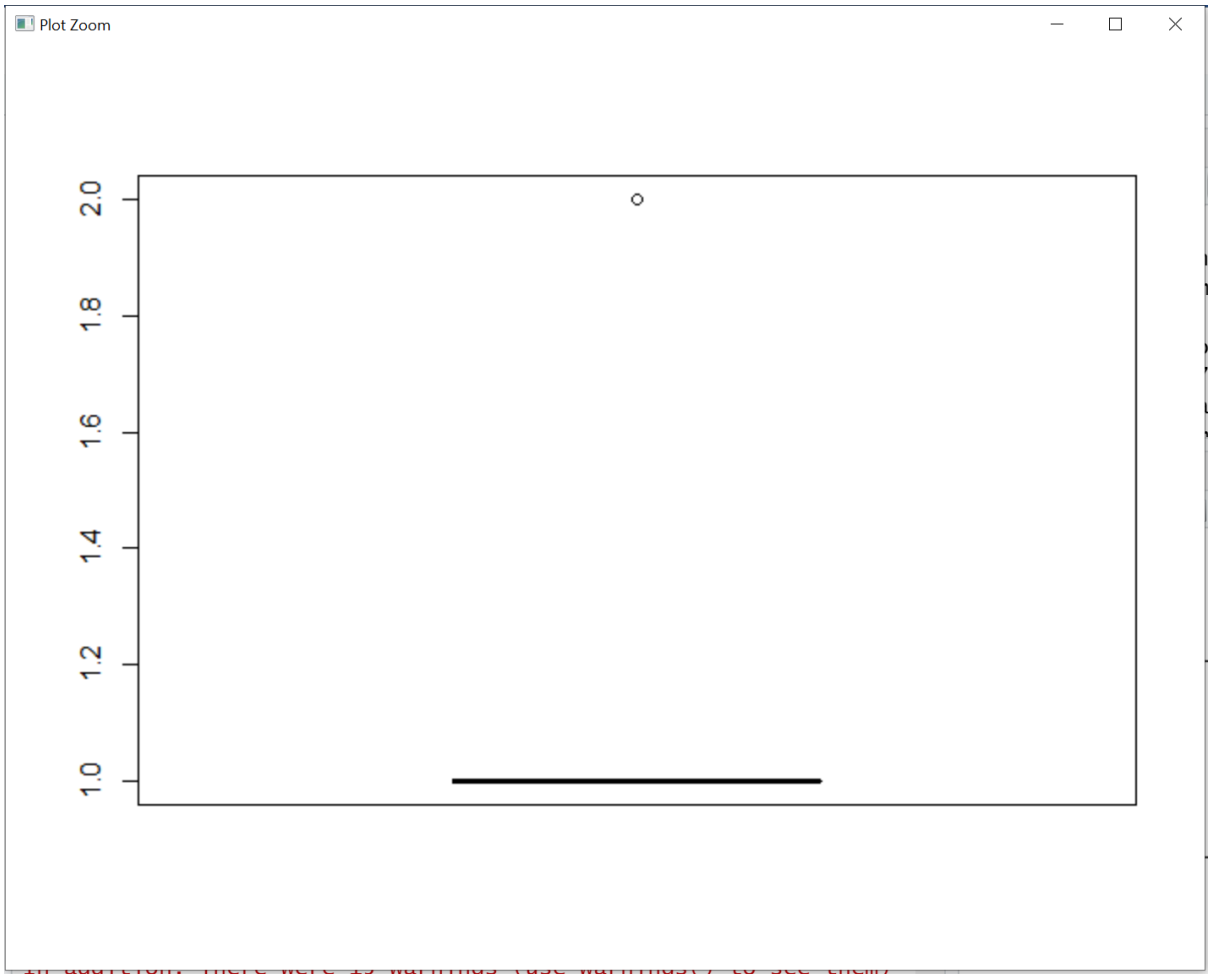


High correlation between fatal crashes and deaths, Pickup and SUV occupants & Large truck occupants, pedestrians and bicyclists

Response variable – Deaths per 10000 population – is normally distributed

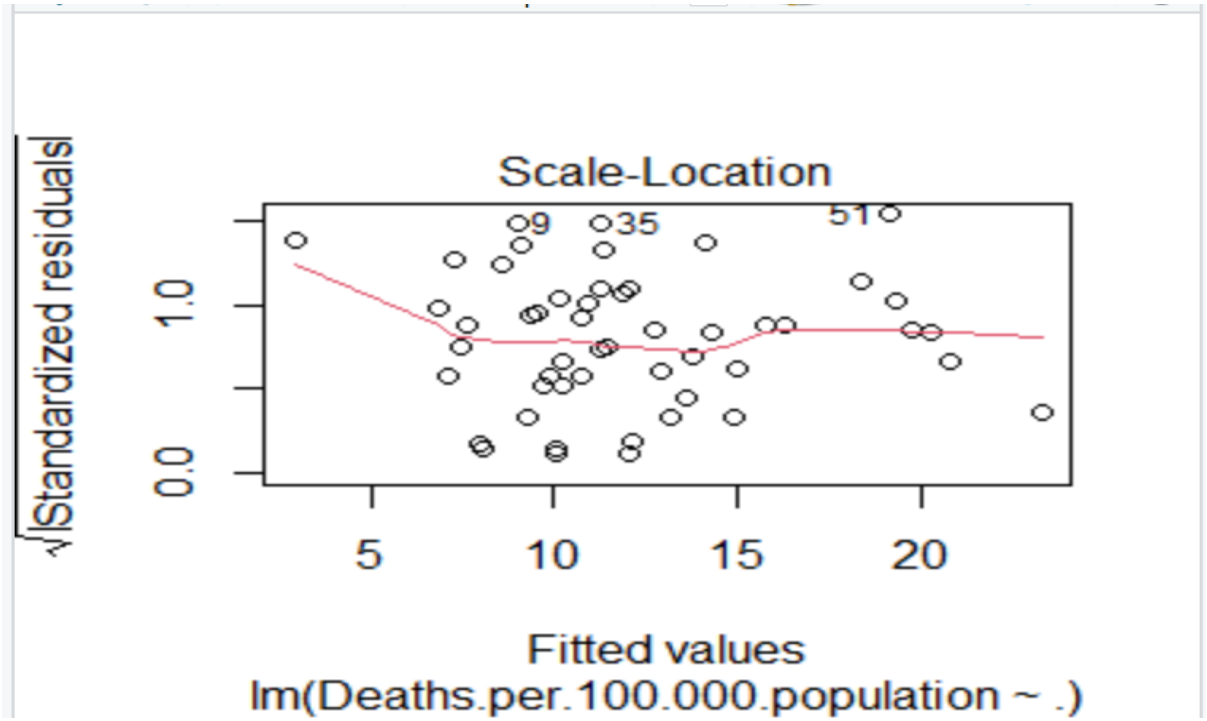


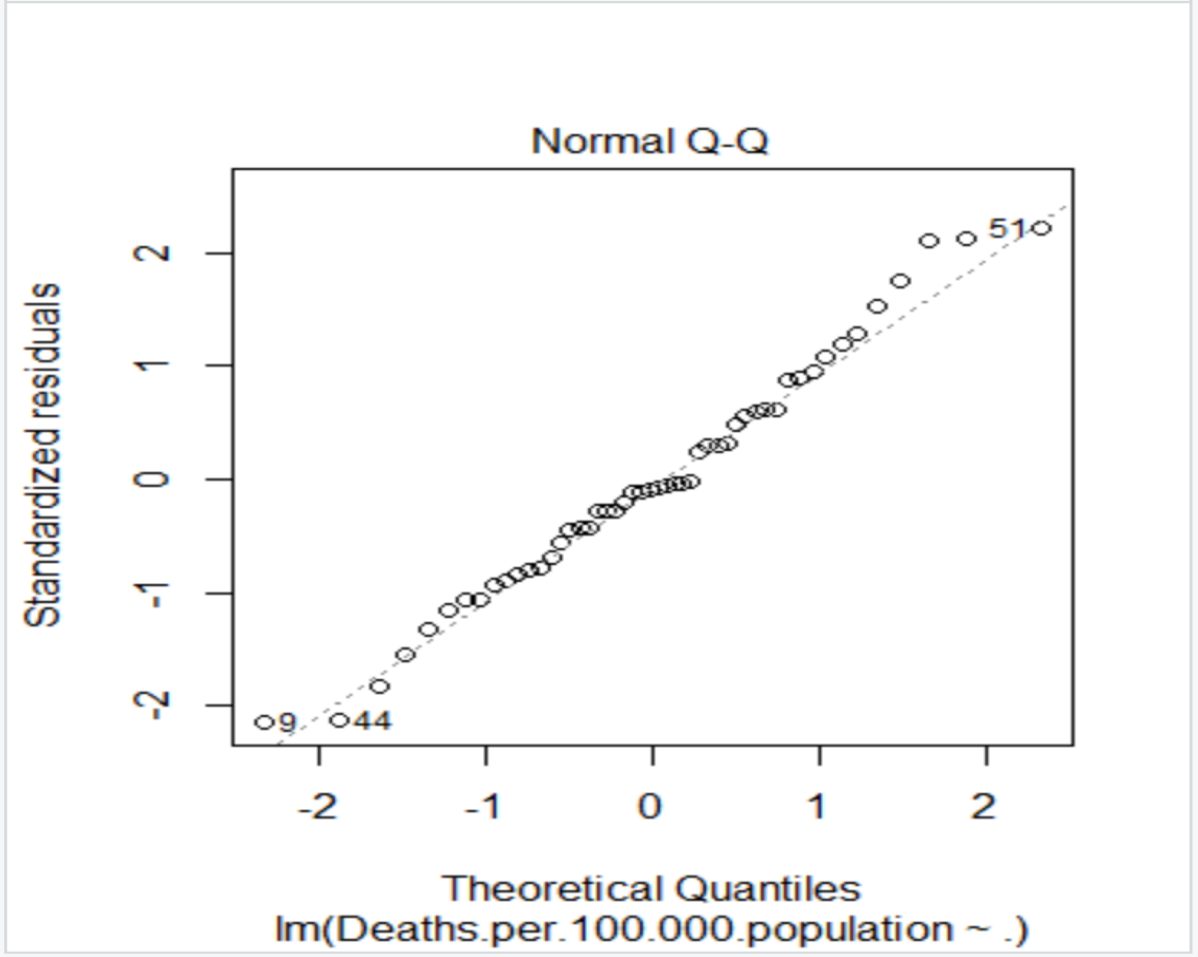
Deaths per 100m miles travelled is binary in nature.

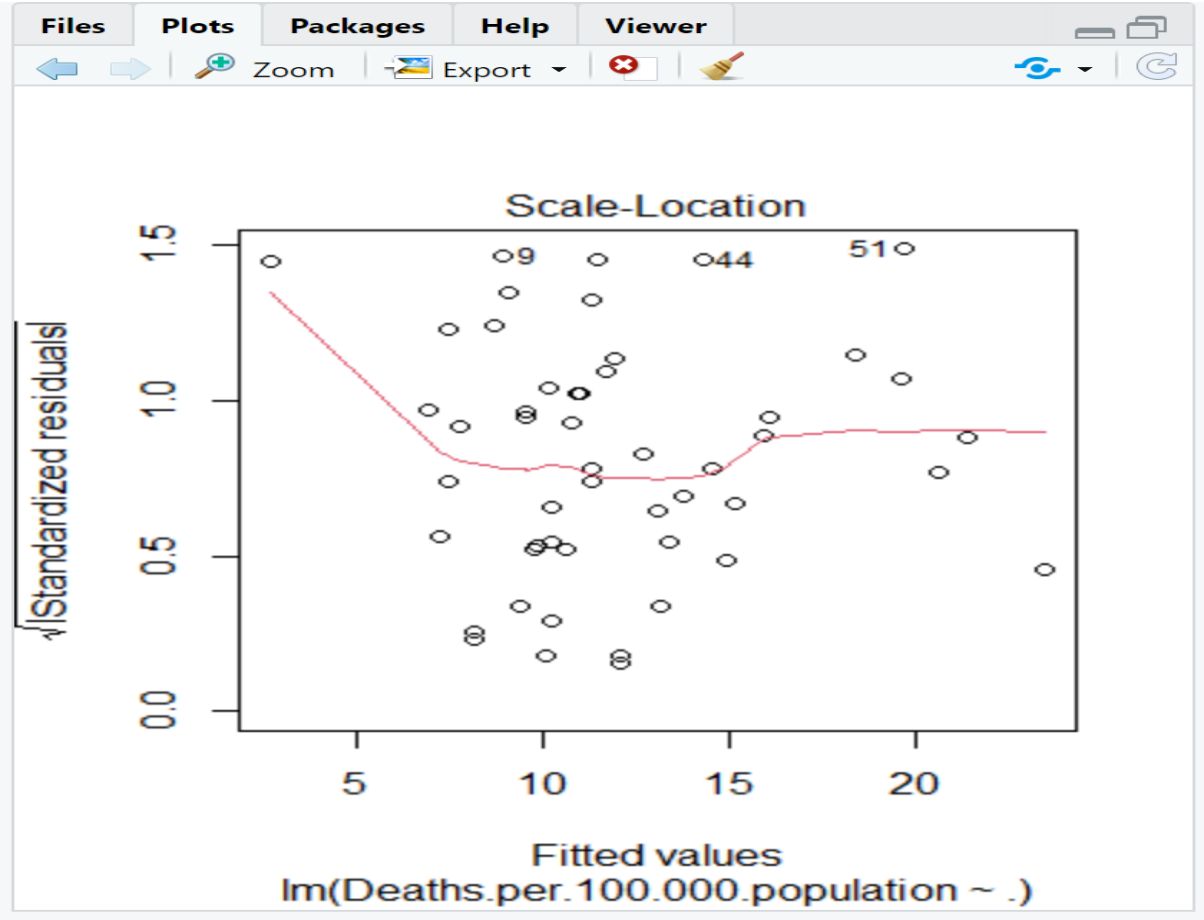


Linear model with all variables included

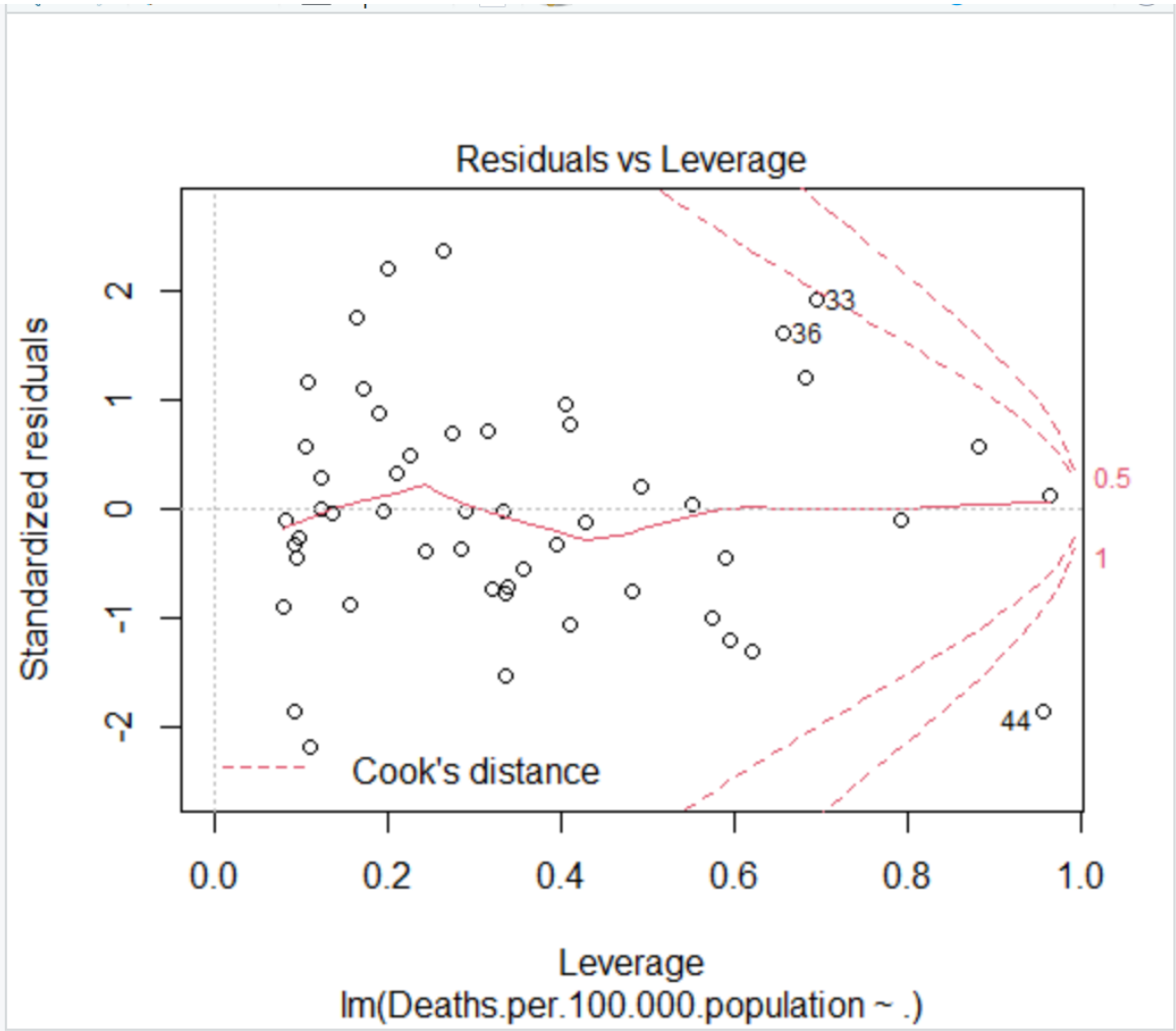
Residuals are randomly scattered around the plot, the assumption of homoscedasticity holds.



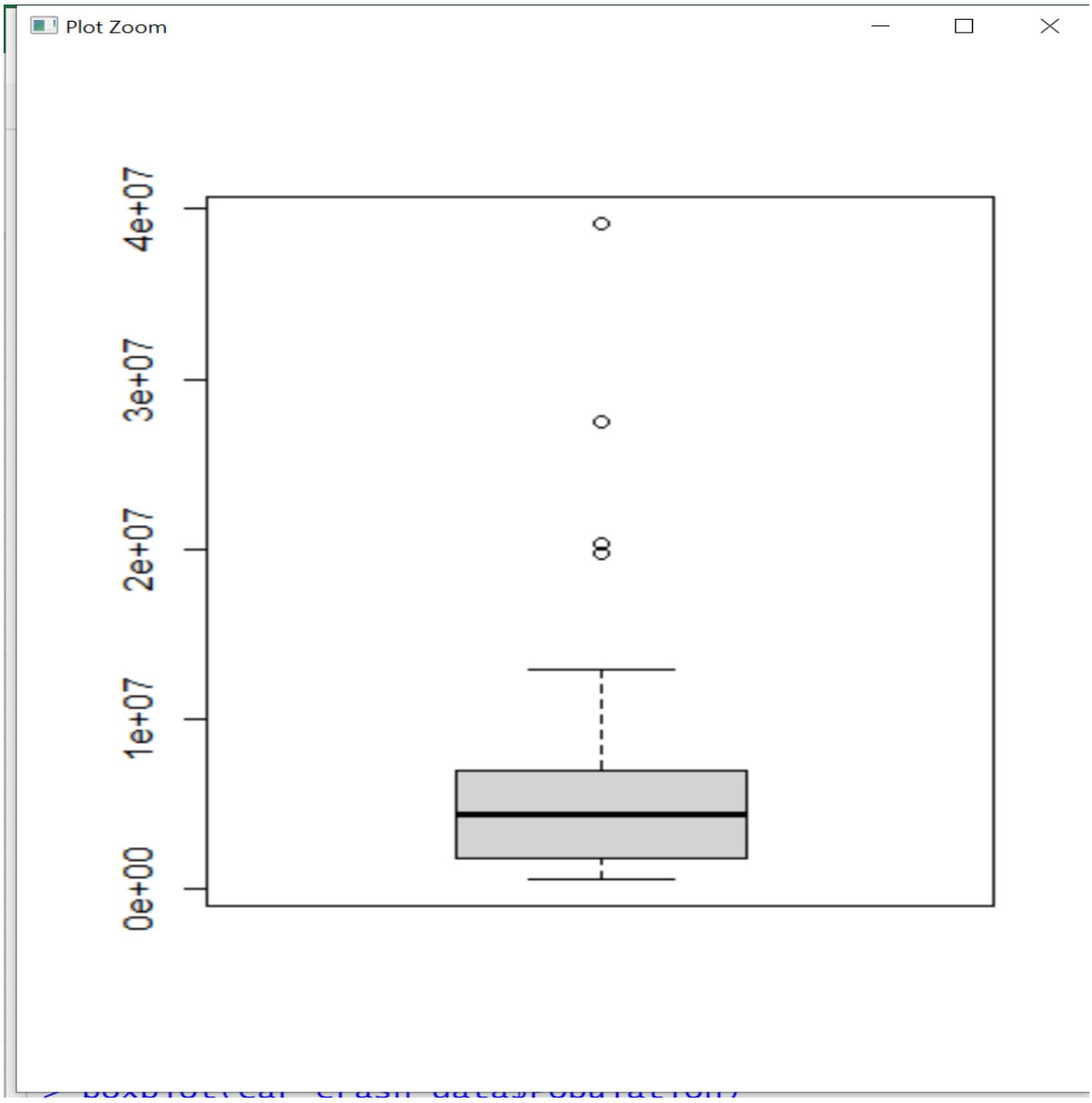

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Deaths.per.100.000.population ~ .)
9
44
51
-2
-1
0
1
2


Files
Plots
Packages
Help
Viewer
Zoom
Export
Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
9
44
51
5
10
15
20
Fitted values
lm(Deaths.per.100.000.population ~ .)

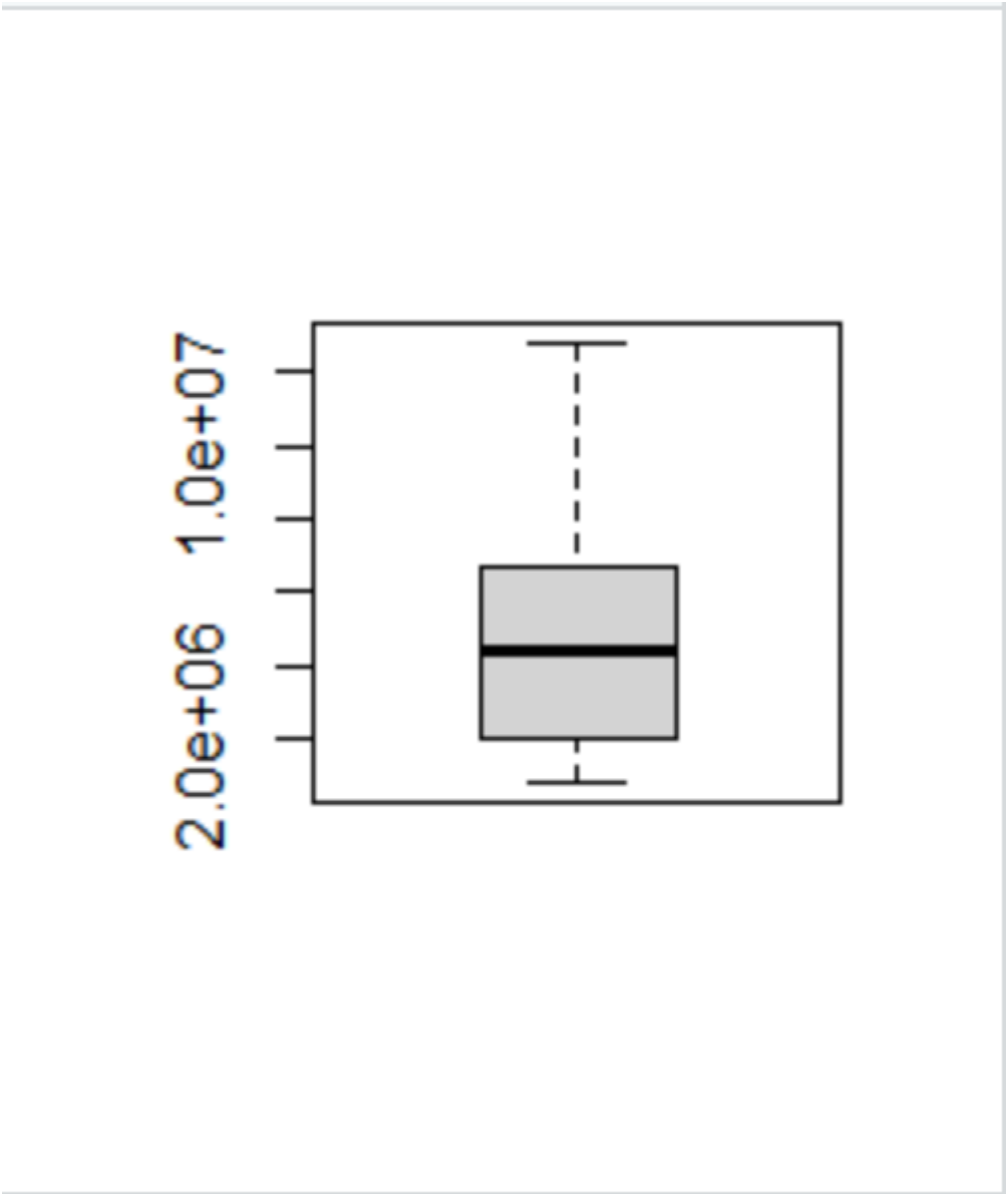
State with largest Cook's distance below is Texas(44) dropped for my initial model of best fit based on highly correlated variables and



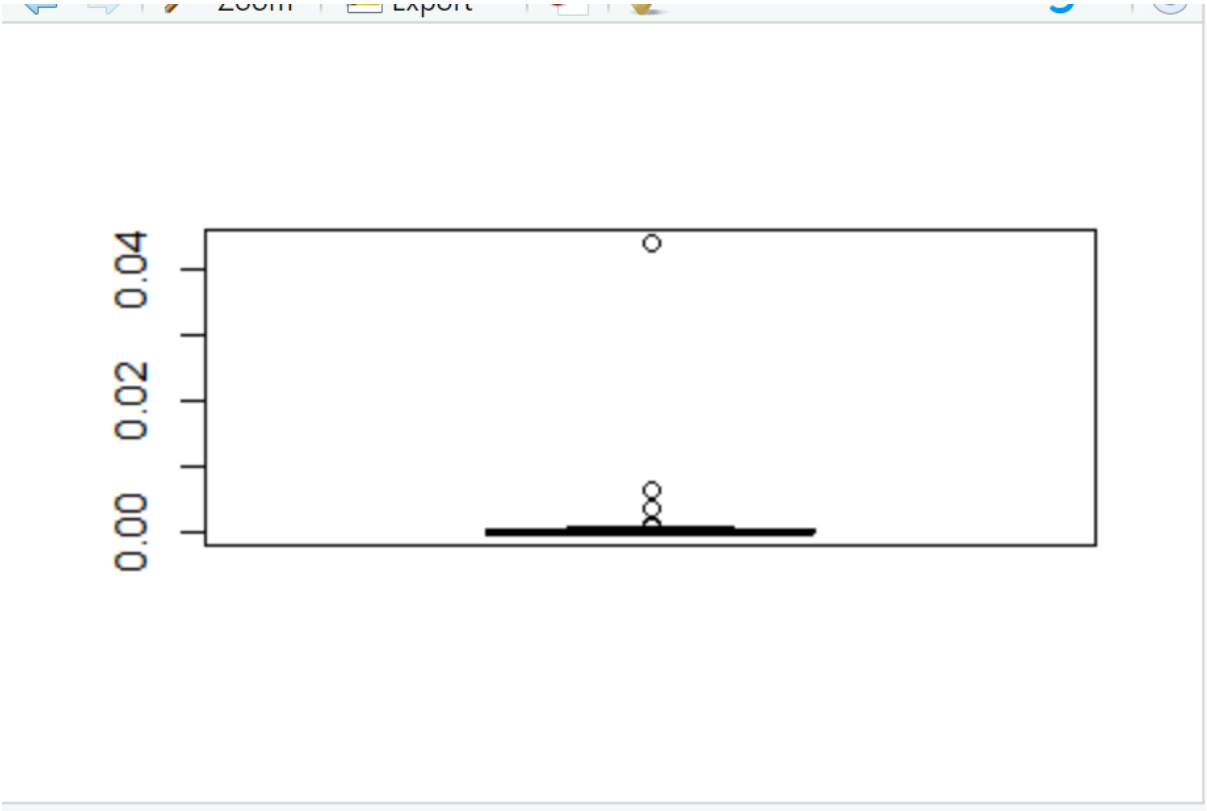
Big outliers in the population boxplot – as response variable is deaths per 100,000 population I have decided to remove this from the dataset – however row values remain – possible remedial measure would be to drop the 3 highest and lowest population centres.



After dropping the four biggest outliers and the four smallest state's in terms of population the boxplot



Boxplot(car_crash_data$Area.in.square.miles_deaths.peronehundthoupop)boxplot for this variable shows big outliers – DC and Delaware


Zoom
Export
0.04
0.02
0.00