ST404 Assignment 1 Report

Alex Walters, 1921659 Jena Moteea, 1939940 Remus Gong, 1918934
James Keith, 1827052

Contents

1 Executive summary

- **In this report, we are analysing the cancer dataset (2013) in US counties. We have observed that the dataset provided is incomplete and we have treated missing as well as abnormal observations from the data**
- **We notice that there is an exhaustive amount of factors used to predict cancer death rate in US counties and therefore, we suggest that we can omit some predictor variables to reduce complexity of our analysis**
- **We finally acknowledge that our dataset needs to be further curated, transformed and investigated before fitting our model**

2 Findings

2.1 Missing and suspicious values

We identified 152 missing values in PctEmployed16_Over and 61 abnormal values in AvgHouseholdSize. These values were investigated and we decide to remove the observations associated with them before any further exploring of the data as we determined that they were missing completely at random.

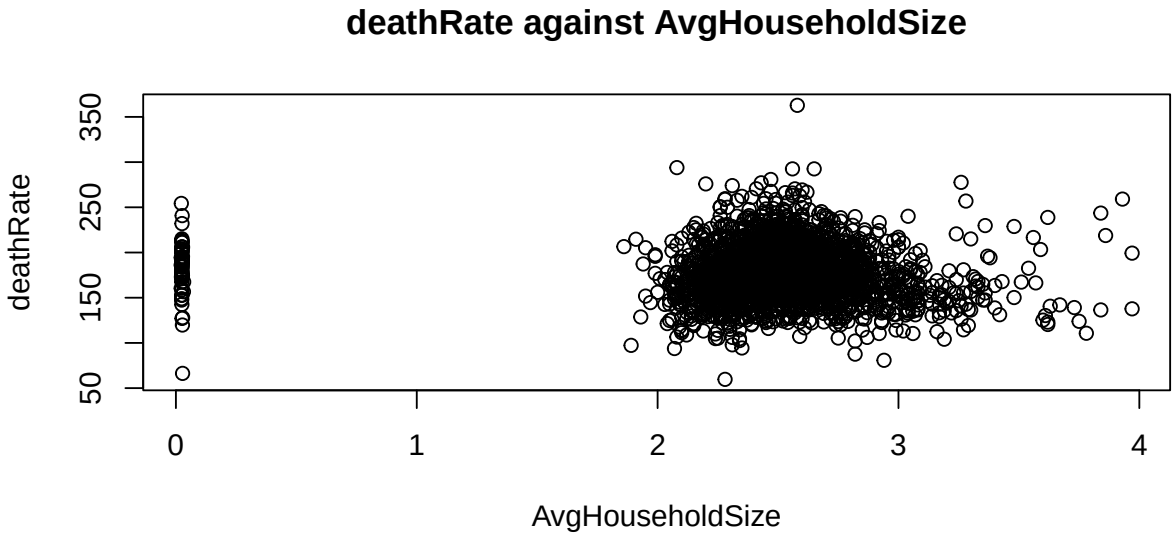

Figure 1: The abnormal values near zero were investigated and removed before further analysis

2.2 Skewness, Heteroskedasticity and Linearity

We found most predictor variables linear and no signs of heteroscedascity. However, we identified heteroscedascity in 3 variables: PctBlack, incidenceRate, medIncome. Potential outliers may

undermine the heteroscedascity. We used power transformations to fix heteroscedascity. We also found non-linearity in AvgHouseholdsize,MedianAgeFemale and MedianAgeMale. We suggest using a more complex model to improve linearity. We found that some variables are skewed and we applied simple power transfomations to fix the skewness. However, due to a massive right skew in PctBlack, we could not completely remove the skewness and so this needs to be further explored when fitting a model. We noticed that the power transformations used in fixing heteroscedascity and skewness are different and we should prioritise heteroscedascity.

2.3 Outliers

We used box plots to discover outliers. We have a considerable amount of outliers in medIncome, PctBlack and incidenceRate. The outliers in medIncome and PctBlack are reasonable after some research of the US demographics. However, there are some extreme values in incidenceRate. We need to further investigate into this before we can make decisions on how we handle or remove them.

2.4 Multicollinearity

We found that incidenceRate, medIncome, povertyPercent, PctEmployed16_Over, PctUnemployed16_Over, PctPrivateCoverage and PctPublicCoverage have high absolute correlation with deathRate and should be taken into consideration as important predictors for when it comes to making a linear model. There are also some predictors that have almost no correlation at all such as AvgHouseholdSize and MedianAgeFemale which is slightly surprising as you might expect counties with older populations to incur more deaths. We'd need to investigate this further once we've made a linear model.

2.5 Correlation with deathRate

There are also a lot of variables tha suffer from the problem of multicollinearity due to rerpesenting very similar things. For example PercentMarried and PctMarriedHouseholds correlate very highly and could be considered to represent the same force. This is also true for many other variables in the dataset and focus should be placed on studying variance inflation factors and other multicollinearity diagnosis methods once a model has been fitted. If our initial suspicions about multicollinearity are proven true by further analysis then we can consider the offending variables as candidates for removal to make an eventual model less complex.

2.6 Recommendations

- Remove observations with missing values and highly abnormal values
- Apply appropriate transformations to fix cases of non-linearity, heteroskedasticty, skewness and non-normality to ensure we have a better fit
- Further investigate outliers to make sure they do not have undue influence when we come to make a model
- Perform further analysis on highly correlated predictor variables and verify that they are candidates for removal
- Take into special consideration variables with significant correlation with deathRate

3 Statistical Methodology

3.1 Checking the summary and initial EDA

Looking at the dataset we note that there is 1 character variable, 1 factor variable and 16 continuous variables. The character variable Geography is just an identifier of the observation and can hence be ignored for statistical analysis and should not be used in a linear model. It, however, can be utilised for data visualisation and analysis of geographic trends in the United States.

```
#Checking number of observations and suspiciously low values
length(cancer$Geography)
length(cancer$AvgHouseholdSize[cancer$AvgHouseholdSize<0.1])
```

```
## [1] 3047
## [1] 61
```

There are 3047 pieces of data in our dataset. That is a large amount of data but it doesn't actually equal the total amount of US counties which number 3143 in total (https://en.wikipedia.org/wiki/County_(United_States)). This means our data is not fully representative of the entire United States but the proportion of counties recorded is high enough so that the data should still be a representative sample.

There are 152 missing values in PctEmployed16_Over which need to be checked. There are 61 values in AvgHouseholdSize underneath 0.1 which should be considered suspicious and immediately investigated before further analysis.

We identify one of these points and investigate it:

Geography	AvgHouseholdSize
Berkeley County, West Virginia	0.0263

To check the validity of this data point we find an alternate source of the data at:

https://data.census.gov/cedsci/table?q=average%20household%20size&g=0500000US54003&y=2013&tid=ACSST1Y2013.S1101

We note that this data recording AvgHouseholdSize in the same year as our data lists the size at 2.61. This is completely different and this is similar for other small values in our dataset. Hence, these are very likely incorrectly inputted data points and as there is only a small proportion of them we should treat them as missing data and then test to see whether they are MCAR.

```
#Makes a dataset where the abnormally low values in AvgHouseholdSize are NA
cancer1 <- cancer
cancer1$AvgHouseholdSize[which(cancer1$AvgHouseholdSize < 0.1)] <- NA
```

3.2 Missing Value Exploration

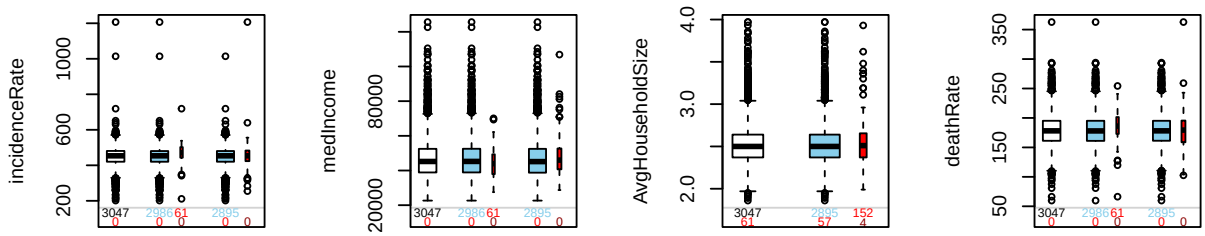

Figure 2: Box Plots showing difference between missing and non-missing data from VIM package

We use the pbox() function from the VIM package to check what these missing values represent. From the plots (Fig. 2) we note that the box plots with the missing data do not look significantly different from those without. The Box Plots for the other variables look similar to this which suggests that the data that is missing is MCAR.

We could replace all the data with an alternate source but as the proportion of missing data points is so small and likely MCAR, it should be safe to remove the rows with missing data from our data set. This won't make the data much less representative and shouldn't affect our statistical analysis that much when we come to build a linear model, other than slightly increasing the standard error.

```
cancer2 <- na.omit(cancer1)
```

3.3 Univariate Plots

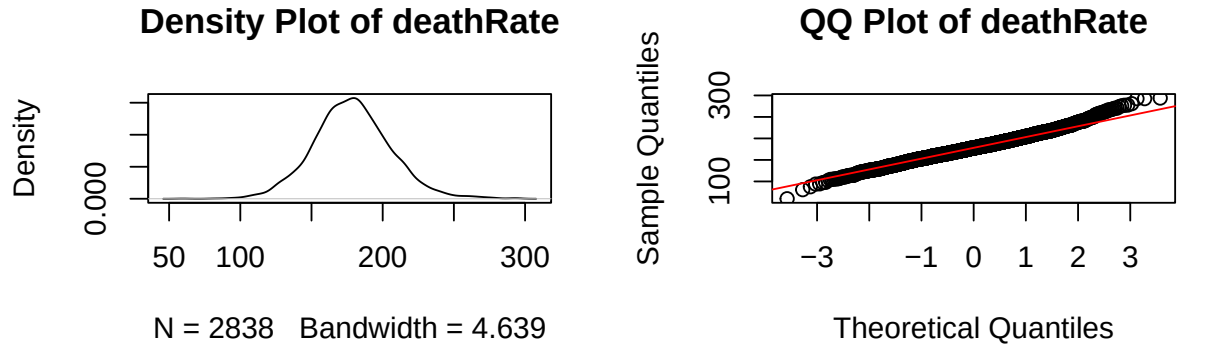

Figure 3: Density and QQ-Plots of deathRate

The assignment brief tells us we should investigate deathRate as a repsonse variable when it comes to our investigation. So we first make sure that a normal linear model is appropriate by making sure that deathRate is normally distributed. From deathRate's density and QQ Plots (Fig. 3) we can see that the variable deathRate is normally distributed so a normal linear model is appropriate to use.

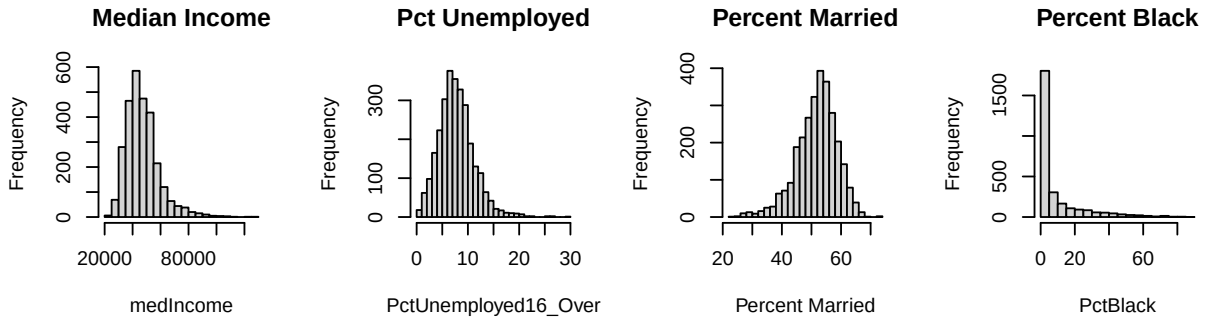

Figure 4: Histograms of 4 skewed variables

The 4 plots above (Fig. 4) give a good representation of some of our worst offenders of skew and hence non-normality. Most of our predictor variables look normally distributed from their density plots and histograms (see appendix) but medIncome, PctUnemployed16_Over, PercentMarried, PctMarriedHouseholds, povertyPercent and PctBlack all have skew. The above plots represent the amount of skew present in these other variables as well and the same respective transformations work to fix similar types of skew.

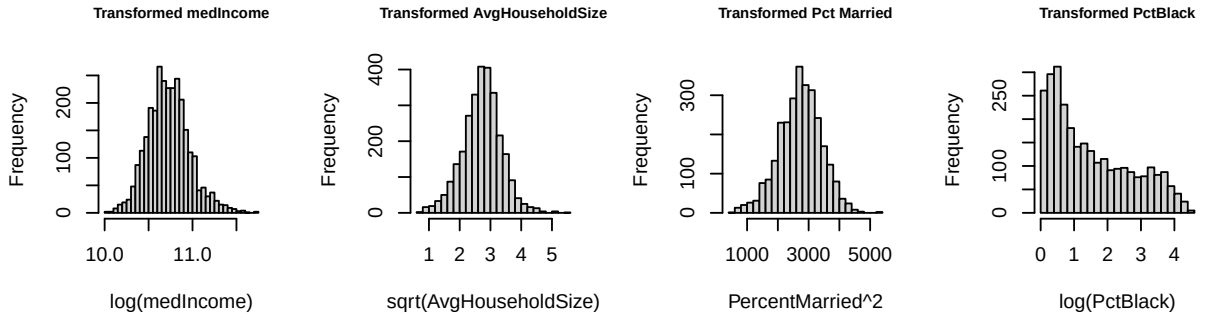

Figure 5: Histograms of 4 transformed pieces of skewed data

From the above (Fig. 5) we note that some simple transformations can be applied to fix most of these variables (log transform for median income for large right skew, square root for AvgHouseholdSize for slight right skew, square transform for PercentMarried for slight left skew). For those 3 variables the skew and normality is mostly fixed. However, for PctBlack a log transform was not sufficient for its large right skew. We made a shift in pctBlack data before log transforming it as there were some zeroes in the pctBlack data which cannot be log transformed. This indicates that the data may not even be normally distributed and would need to be handled differently when it comes to our statistical model.

3.4 Bivariate Plots

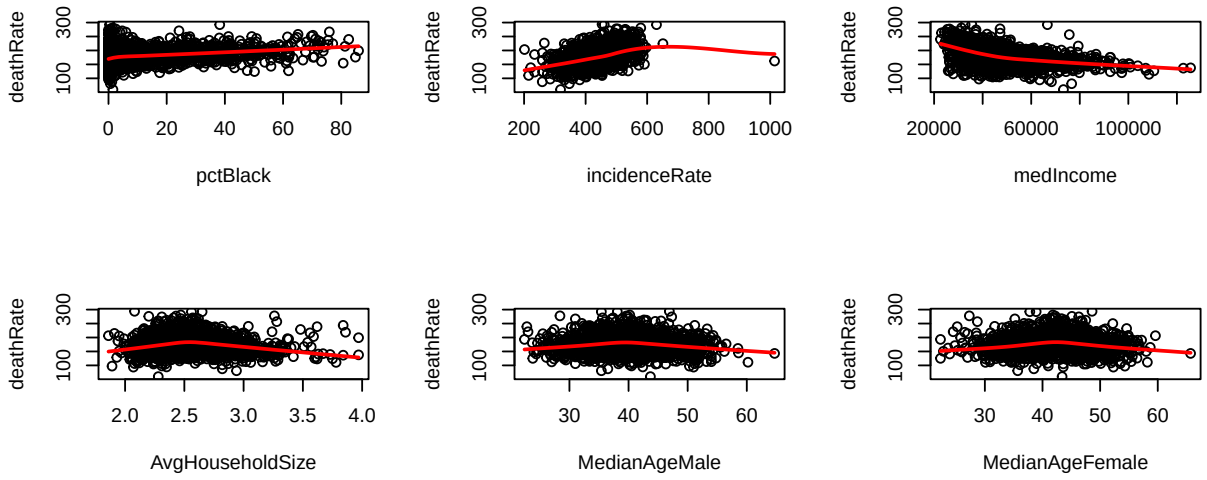

Figure 6: Plots showing deathRate against other variables

Most predictor variables in US cancer dataset show signs of linearity and no heteroscedasticity. However, from the bivariate plots above, we can observe definite heteroscedasticity in PctBlack, incidenceRate and medIncome. We might need to perform further investigation after fitting a model and we can use spreadLevelPlot() to find an appropriate power transformation to fix heteroscedasticity. We were also able to observe that the outliers of incidence rate might have a high influence underlying its heteroscedasticity and non-linearity. The transformations that have been used to fix the skew of the data may not be the same as those that would fix heteroskedasticity or non-linearity. In this case we should prioritise heteroskedasticity and linearity as those are more important to fitting a good model than normality.

Moreover, we can see non-linearity in AvgHouseholdsize,MedianAgeFemale and MedianAgeMale. We notice a concave shape for incidence rate and AvgHouseholdsize so we advise having a more complex model, perhaps with a quadratic term might be improve linearity as the data is not monotonic.

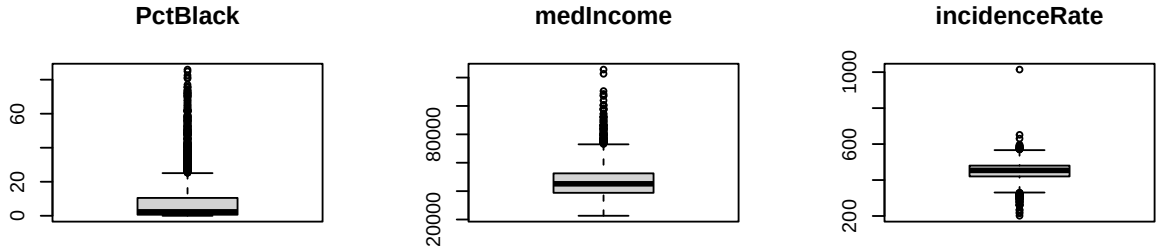

Figure 7: BoxPlots of our variables

Our box plots (Fig. 7 and see Appendix) show that we have quite a number of what we would consider outliers across all our variables apart from binnedInc, which would be impossible due to the bins intervals. We have a severe amount of outliers in medIncome. This is most likely due to natural causes such as a CEO of a large company or a doctor (see Reference about high paid jobs). We also observe significant outliers in PctBlack and this is illustrated by the very long tail as shown in the histogram above (See Fig. 4). This might be due to PctBlack being an unstable predictor variable. We observe significantly high percentages of over 50% in south and southeast region of the US, in particular, in Mississippi, Georgia, Alabama and North and South Carolina. They indeed form part of the top 10 US state with the highest percentage of Black residents. (See Reference about Black population in US)

The boxplot for Incidence rate shows the existence of extreme high values which is also illustrated in our bivariate plots in Fig. 6 There are potential outliers in PctPrivateCoverage and povertyPercent.We might want to further investigate into these and decide how we might want to treat them before fitting the model. Possible options might include deleting the outliers or imputing them.

We suggest that the incidence rate in Williamsburg city, Virginia can be considered as a candidate for removal. (See reference for Williamsburg city, Virginia)

3.5 Multicollinearity

We can see that there is potential multicollinearity between: PercentMarried and PctMarried Households (correlation 0.87), PctUnemployed16_over and PctEmployed16_Over (correlation -0.65), MedianAgeFemale and MedianAgeMale (correlation 0.94) (See Appendix). We further used Pearson correlation test (See Appendix) to check for multicollinearity between percentPoverty and PctEmployed_Over16 (correlation -0.74), PctPrivateCoverage (correlation -0.82), PctEmpPrivCoverage (correlation -0.68), and PctPublicCoverage (correlation 0.65).Therefore, we might consider discarding some of the predictor variables due to high multicollinearity to improve accuracy when fitting a model.

We can observe that for the first 9 bins medIncome and binnedInc show very similar results. We will therefore consider only using medIncome in our model.

3.6 Correlation with deathRate

We note that none of the variables highly correlate with deathRate but there are a number with medium correlation with deathRate that should be noted when it comes to building a model. These are incidenceRate, medIncome, povertyPercent, PctEmployed16_Over, PctUnemployed16_Over, PctPrivateCoverage and PctPublicCoverage. However, a lot of these also correlate highly with each other so the correlation with deathRate will be down to these variables measuring the same thing. It should also be noted that MedianAgeMale, MedianAgeFemale and AvgHouseholdSize have almost no correlation with deathRate so are candidates for removal if we were to go on to make an explanatory linear model. It makes sense that both MedianAgeMale and MedianAgeFemale both correlate so little with deathRate as they measure very similar things as explained in section 1.5.

incidenceRate	medIncome	povertyPercent	PctEmployed16_Over	PctPrivateCoverage	PctPublicCoverage	MedianAgeFemale	AvgHouseholdSize
0.42	-0.43	0.43	-0.42	-0.39	0.4	0.01	-0.04