ST404 Assignment 2

Frank Or, Remos Gong, Sam Glanfield, Thomas Broadbent

2022-02-23

Contents

1 Findings

2 Statistical Methodology

We first combined our results and findings from our preliminary EDA which is mainly discussed in Outliers and Transformations.

2.1 Outliers

Counties with high Incidence Rates, namely 'Union County, Florida' and 'Williamsburg City, Virginia' were removed from the model building process as they had high influence over the model and were not reflective of the rest of the data.

For counties with Average Household Size less than one we took the decision to scale the transformations by 100 and keep them in the dataset.

2.2 Transformations

2.3 Modelling Approach and Variable Selection

2.3.1 AIC Forward and Backward Variable Selection

2.3.2 RIDGE Regression

2.4 Statistical Interpretation and Validation

3 References

4 Appendix

```
library(dplyr)

## Included Libraries
# For pipe operator and general mutation
load('cancer.rdata')

# Removing outlier incidence rates 'Union County, Florida' and 'Williamsburg City, Virginia'
cancer <- filter(cancer, incidenceRate <= 700)

# Scale average household sizes that are less than 1 by 100
cancer[which(cancer$AvgHouseholdSize < 1), ]$AvgHouseholdSize <-
  100*cancer[which(cancer$AvgHouseholdSize < 1), ]$AvgHouseholdSize
```