Capstone - IDV

Ken Gustafson

Contents

1 Introduction

The data chosen for this project comes from the R package https://rdrr.io/github/jdmR-packages/czerlinski1999/ . It was collected by Christopher Bingham for the American Almanac of 1974 with the intent to study fuel consumption. The data set contains 48 observations with 9 variables. The goal for this project is to predict fuel consumption. Model assumptions will be checked to see if they are being violated or not in order to assess the validity of the model.

2 Data Exploration

The response variable will be $fuel_{pc}$ which is the number of gallons consumed per person. Since there are two variables measuring fuel, the one measuring the total fuel consumption per state will be removed.

The variables under consideration for analysis with their description is the following:

- $state$ = State Name
- pop = Population
- tax = Motor fuel tax rate, in cents per gallon
- lic_n = Number of licensed drivers in thousands
- $income$ = Per capita income in thousands of dollars

- $road$ = Thousands of miles of federal highways
- $prop_{lic}$ = Proportion of population that are licensed
- $fuel_{pc}$ = Number of gallons of fuel consumed per person

Since the number of state categories are unique for each data point, it cannot be used in the analysis. After plotting each variable against one another it was found that the variables that give the population and the number of registered drivers are highly correlated ($r > .99$). The population variable will be dropped during modeling since the number of registered drivers will mostly affect fuel consumption.

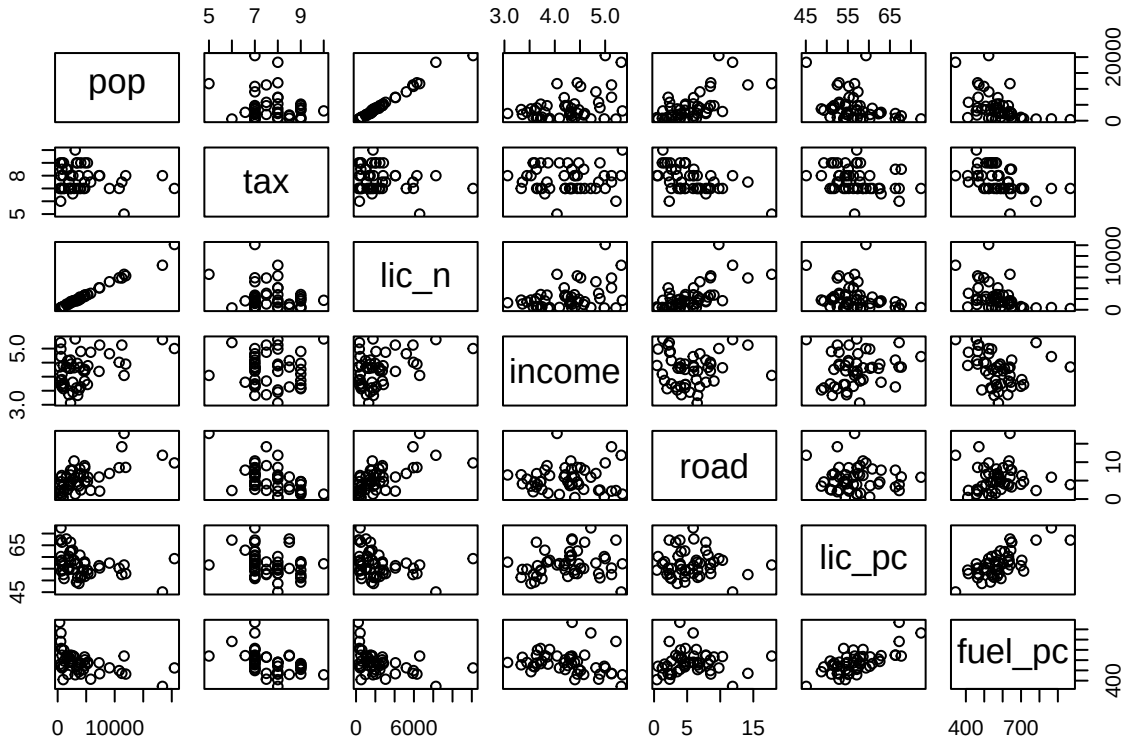
```
# Obtaining the Data
#remotes::install_github("jdmR-packages/czerlinski1999")
if (!require(remotes)) install.packages('remotes')
library(remotes)
if (!require(czerlinski1999)) install.packages('czerlinski1999')
library("czerlinski1999")
if (!require(glmnet)) install.packages('glmnet')
library(tidyverse)
library(caret)
library(data.table)
library(stringr)
library(dplyr)
library(tidyr)
library(lubridate)
library(tinytex)

data("fuel")
head(fuel)
```

```
## # A tibble: 6 x 9
##   state   pop   tax lic_n income  road fuel_n lic_pc fuel_pc
##   <chr> <int> <dbl> <int>  <dbl> <dbl>  <int>  <dbl>   <int>
## 1 ME     1029   9     540   3.57 1.98     557   52.5     541
## 2 NH      771   9     441   4.09 1.25     404   57.2     524
## 3 VT      462   9     268   3.86 1.59     259   58       561
## 4 MA     5787   7.5  3060   4.87 2.35    2396   52.9     414
## 5 RI      968   8     527   4.40 0.431    397   54.4     410
## 6 CN     3082  10    1760   5.34 1.33    1408   57.1     457
```

```
# Dropping redundant response variable
fuel <- fuel[,-7]

# Creating Pairs Plot
pairs(fuel[,c(2,3,4,5,6,7,8)])
```



Looking at the pairs plot, which plots each variable against one another, there is evidence that there are outliers present in the data set that could potentially be dropped.

A new variable we will look at is fuel_tax_income, which represents the percentage of income that is spent on fuel taxes per person. We will examine how this varies by state.

```
# Defining the new variable
fuel$fuel_tax_income <- fuel$fuel_pc * (fuel$tax/100) / (1000*fuel$income)

head(arrange(fuel, fuel$fuel_tax_income))
```

```
## # A tibble: 6 x 9
##   state   pop   tax lic_n income   road lic_pc fuel_pc fuel_tax_income
##   <chr> <int> <dbl> <int>  <dbl>  <dbl>  <dbl>   <int>           <dbl>
## 1 NY    18366   8    8278   5.32 11.9     45.1     344         0.00517
## 2 MA     5787   7.5  3060   4.87  2.35    52.9     414         0.00638
## 3 IL    11251   7.5  5903   5.13 14.2     52.5     471         0.00689
## 4 NJ     7367   8    4074   5.13  2.14    55.3     467         0.00729
## 5 CA    20468   7   12130   5.00  9.79    59.3     524         0.00733
## 6 RI      968   8     527   4.40  0.431   54.4     410         0.00746
```

```
tail(arrange(fuel, fuel$fuel_tax_income))
```

```
## # A tibble: 6 x 9
##   state   pop   tax lic_n income  road lic_pc fuel_pc fuel_tax_income
##   <chr> <int> <dbl> <int>  <dbl> <dbl>  <dbl>   <int>           <dbl>
```

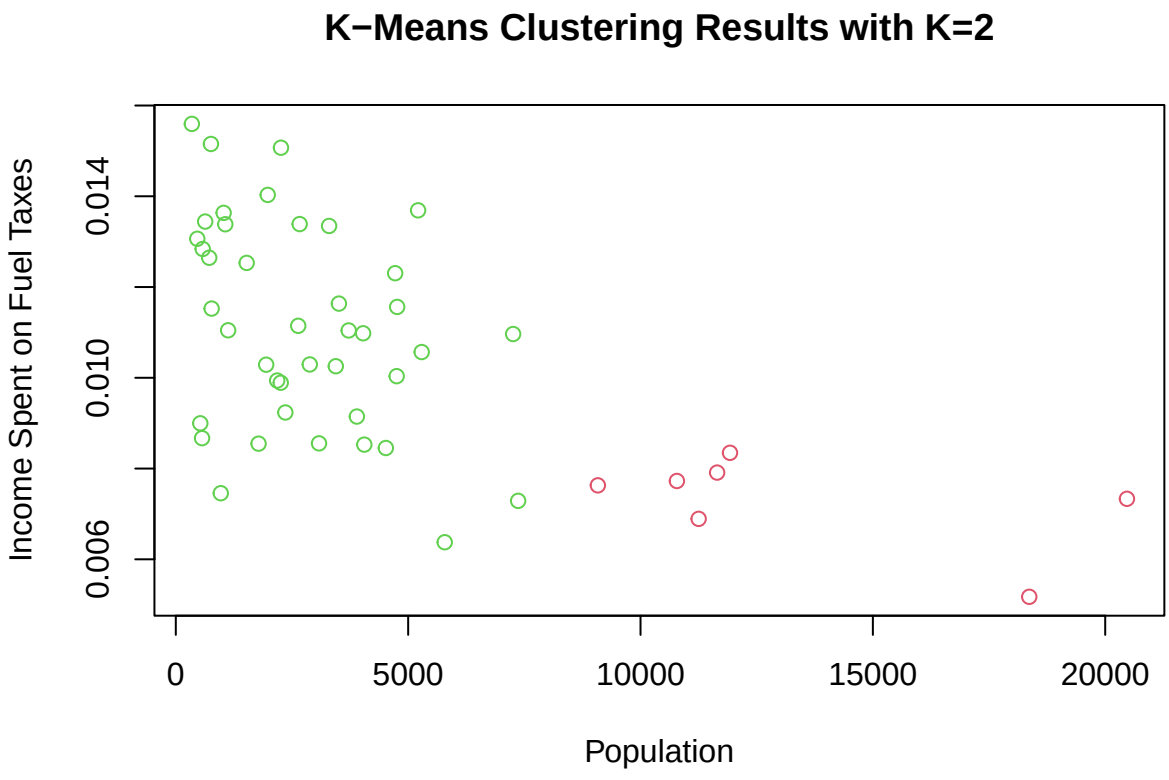
```
## 1 ME     1029   9     540   3.57  1.98   52.5     541          0.0136
## 2 NC     5214   9    2835   3.72  4.75   54.4     566          0.0137
## 3 AR     1978   7.5  1081   3.36  4.12   54.7     628          0.0140
## 4 MS     2263   8    1309   3.06  6.52   57.8     577          0.0151
## 5 ID      756   8.5   501   3.64  3.27   66.3     648          0.0152
## 6 WY      345   7     232   4.34  3.90   67.2     968          0.0156
```

Looking at the percent of income spend on fuel tax, we see that people in the northeastern states spend a lower percentage of their income on fuel taxes. This could be from the availability of mass transit as the fuel per person is less. On the other end, more rural states spend higher amounts of income on fuel tax, likely due to lower per capita incomes there.

2.1 Clustering

We will explore if any clusters are evident in the data. Below is a visual that shows two clusters where states that have large populations tend to spend less on fuel taxes, while states with lower populations spend more.

```
fuelmatrix <- as.matrix(fuel[,c(2,9)])
k <- kmeans(fuelmatrix, centers = 2, nstart = 25)
plot(fuelmatrix,col=k$cluster+1,main = "K-Means Clustering Results with K=2",xlab = "Population",ylab =
```



Outliers

Checking for outliers, Texas was identified as the biggest continental state but does not have the population to match it. Another outlier is California which has extreme values in the x direction. This can cause the state to affect the regression line much more strongly than other states. Both states were removed from the data set.

```
# Dropping Texas and California
fuel <- fuel[-c(37,48),]
fuel
```

```
## # A tibble: 46 x 9
##    state   pop   tax lic_n income   road lic_pc fuel_pc fuel_tax_income
##    <chr> <int> <dbl> <int>  <dbl>  <dbl>  <dbl>   <int>           <dbl>
##  1 ME     1029   9     540   3.57  1.98    52.5     541         0.0136
##  2 NH      771   9     441   4.09  1.25    57.2     524         0.0115
##  3 VT      462   9     268   3.86  1.59    58       561         0.0131
##  4 MA     5787   7.5  3060   4.87  2.35    52.9     414         0.00638
##  5 RI      968   8     527   4.40  0.431   54.4     410         0.00746
##  6 CN     3082  10    1760   5.34  1.33    57.1     457         0.00855
##  7 NY    18366   8    8278   5.32 11.9     45.1     344         0.00517
##  8 NJ     7367   8    4074   5.13  2.14    55.3     467         0.00729
##  9 PA    11926   8    6312   4.45  8.58    52.9     464         0.00835
## 10 OH    10783   7    5948   4.51  8.51    55.2     498         0.00773
## # ... with 36 more rows
```

3 Modeling

3.1 Multiple Linear Regression

First the data will be divided into a training and validation set where the validation set comprises 20% of the fuel data set. Each model will get trained on the training set. The two models under consideration will be a multiple linear regression model (along with a reduced version) and a penalized least squares model (Regularization). Specifically, a ridge regression model will be used which limits the sum of the squares of the beta coefficients.

```
# Validation set will be 20% of fuel data
set.seed(1, sample.kind="Rounding")
test_index <- createDataPartition(y = fuel$fuel_pc, times = 1, p = 0.2, list = FALSE)
trainset <- fuel[-test_index,]
temp <- fuel[test_index,]
validation <- temp

removed <- anti_join(temp, validation)
trainset <- rbind(trainset, removed)

# Cross Validation function
rmse <- function(realfuel,predictedfuel){
  dif <- sqrt(mean((realfuel - predictedfuel)^2))
  return(dif)
}
```

```
# Full Model
lm1 <- lm(fuel_pc ~ ( tax + lic_n + income + road + lic_pc),data=trainset)
summary(lm1)
```

```
##
## Call:
```

```
## lm(formula = fuel_pc ~ (tax + lic_n + income + road + lic_pc),
##     data = trainset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.255  -51.889   -4.272   32.566  236.001
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 410.22855  237.12482   1.730  0.09464 .
## tax         -31.28355   15.71184  -1.991  0.05631 .
## lic_n        -0.01902    0.01556  -1.222  0.23188
## income      -49.19829   28.88527  -1.703  0.09960 .
## road          4.99419    7.52994   0.663  0.51260
## lic_pc       11.00344    3.24244   3.394  0.00208 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.72 on 28 degrees of freedom
## Multiple R-squared:  0.6442, Adjusted R-squared:  0.5807
## F-statistic: 10.14 on 5 and 28 DF,  p-value: 1.275e-05
```

```
# Reduced Model
lmred <- lm(fuel_pc ~ ( tax + income + lic_pc),data=trainset)
summary(lmred)
```

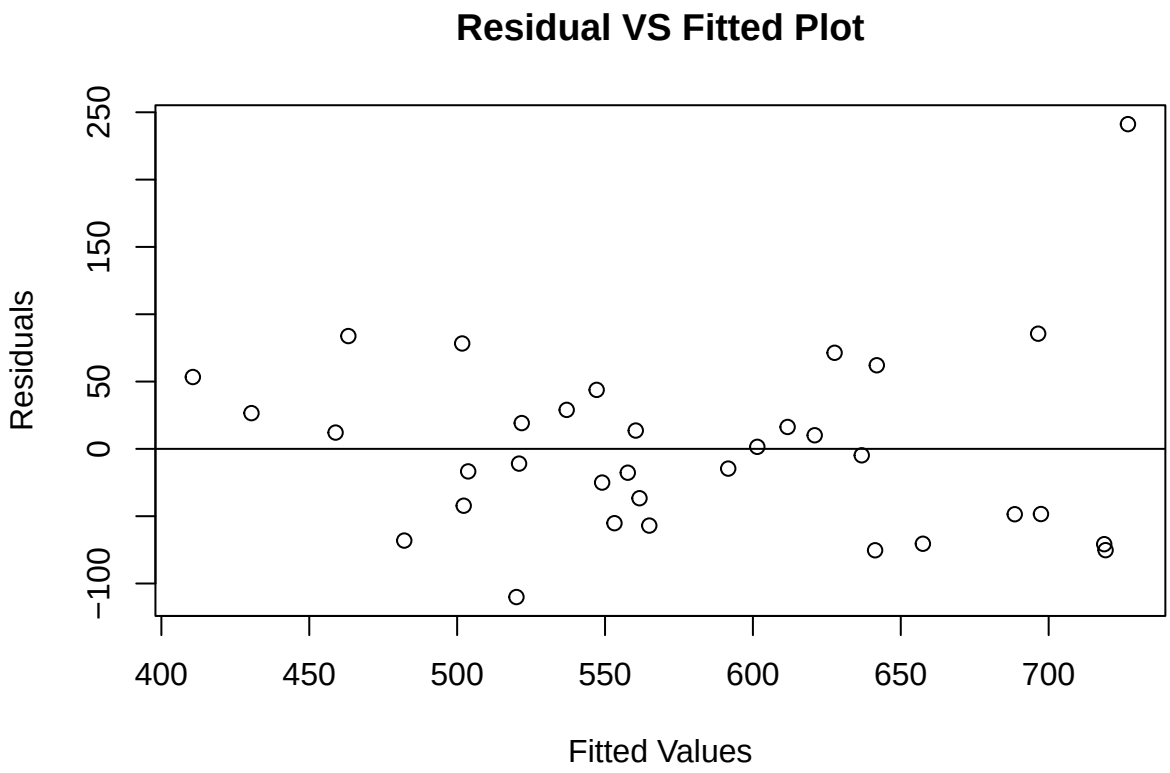
```
##
## Call:
## lm(formula = fuel_pc ~ (tax + income + lic_pc), data = trainset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.050  -48.504   -7.866   28.355  241.187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   333.46     219.15   1.522  0.13858
## tax           -30.20      14.18  -2.129  0.04155 *
## income        -69.61      23.63  -2.946  0.00617 **
## lic_pc         13.50       2.57   5.254 1.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 70.41 on 30 degrees of freedom
## Multiple R-squared:  0.6221, Adjusted R-squared:  0.5844
## F-statistic: 16.46 on 3 and 30 DF,  p-value: 1.645e-06
```

```
# Cross Validation statistic for Multiple Linear Regression Model
pred <- predict(lmred,newdata = validation)
rmse(validation$fuel_pc,pred)
```

```
## [1] 57.41359
```

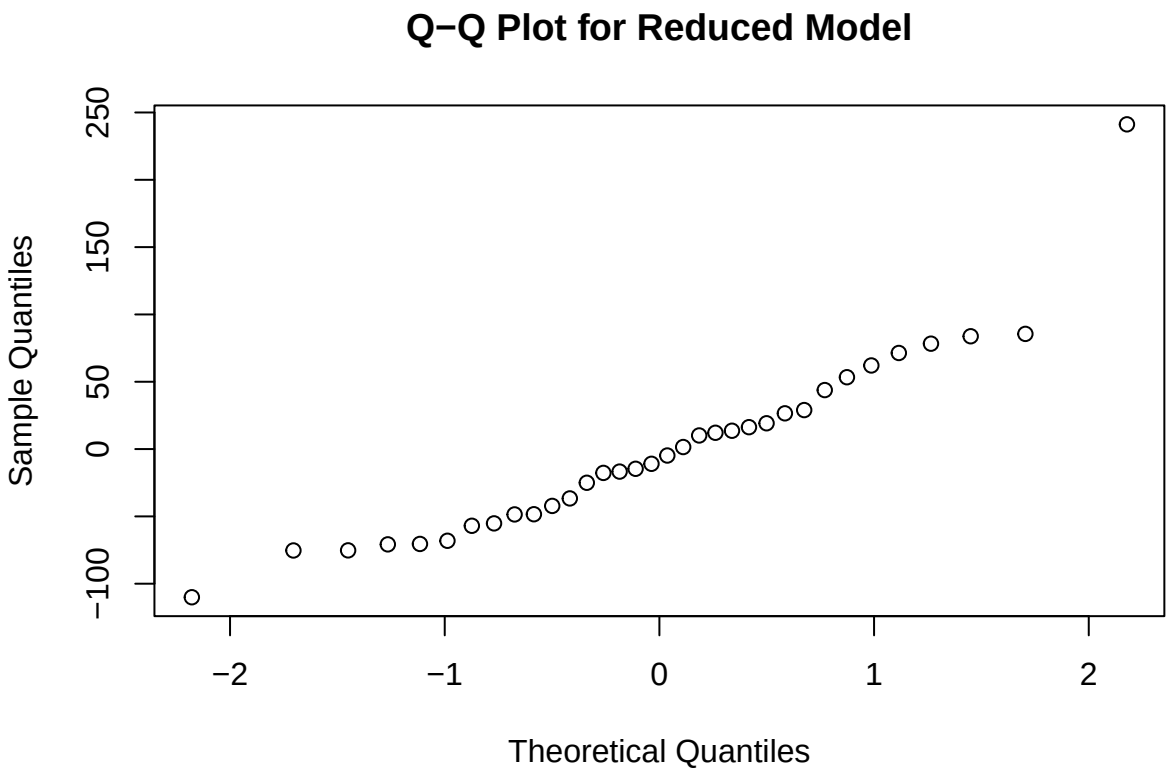
As an initial model, a first order regression model was fit using all the remaining predictor variables. The full model is $fuel_{pc} = tax + lic_n + income + road + prop_{lic}$.

```
# Residual Plot
plot(fitted(lmred),lmred$residuals,xlab = "Fitted Values",ylab = "Residuals",main = "Residual VS Fitted
abline(0,0)
```



Looking at the summary of the model, the variables that were not significant at the level $\alpha = .1$ were *road* and lic_n. Therefore the variables were dropped for a reduced model. Thus the reduced model is now $fuel_{pc} = tax + income + prop_{lic}$. Checking the constant variance assumption, the residual plot above showed that the residuals have equal variance across the x values as there are no patterns and the points are sufficiently evenly spread out. Checking the normality assumption, the plot below shows that the Q-Q plot has approximately a straight line with only one observation outside the norm, so it isn't violated too badly. Thus, we conclude that the model assumptions for multiple linear regression are satisfied. The cross validation statistic for the reduced model is 57.4. We will compare this to the regularization model below.

```
# QQ-norm plot
qqnorm(lmred$residuals,main = "Q-Q Plot for Reduced Model")
```



3.2 Regularization

As an alternative model we used a penalized least squares model. Ridge regression is the method we chose to use for this modeling technique. Checking for lambda values that are between 0 and 100, the optimal lambda value was found to be approximately 12.58.

```
# Optimal Lambda Search
lambda_seq <- 10^seq(2, -2, by = -.01)
x <- data.matrix(trainset[,3:7])
y <- data.matrix(trainset[,8])

lm2 <- cv.glmnet(x, y, alpha = 0, lambda = lambda_seq)

# Optimal Value of Lambda
lm2$lambda.min
```

```
## [1] 12.58925
```

```
# Create Model With Optimal Lambda Value
lm3 <- glmnet(x, y, alpha = 0, lambda = 12.58)
```

To check how well the reduced and the regularization model perform they were run on the validation set. The calculated root mean squared error was 57.4 for the reduced multiple linear regression model and 50.6 for the regularization model. Thus the regularization model made more accurate predictions than the reduced

model. In conclusion, when using the regularization model you can expect to be off by about 50.6 gallons when predicting fuel consumption for someone in a specific state.

```
# Regularization Model Predictions
lmridgepred <- lm3$a0 + lm3$beta[1]*validation$tax + lm3$beta[2]*validation$lic_n + lm3$beta[3]*validati

# Cross Validation statistic for Regularization Model
rmse(validation$fuel_pc,lmridgepred)
```

```
## [1] 50.5715
```

4 Results

Model	RMSE Using Validation Set
Full Model	51.5
Reduced Model	57.4
Regularization Model	50.6

5 Conclusion

By looking a the signs of each beta coefficient that was estimated by the reduced model, it indicates that as the proportion of licensed drivers increases, then fuel consumption will also increase. Conversely, it indicates that as taxes increase and when income increases it will decrease fuel consumption for the state.

In conclusion, the variables that have an effect on fuel consumption per state are tax, $income$, and $prop_{lic}$. The variables $road$ and lic_n did not have a discernible effect on fuel consumption. The regularization model performed better than the reduced multiple linear regression model and on average its predictions were better by 7 gallons. A limitation of this analysis is that prediction is generally poor as there are few data points. A future direction is to use multiple years in order to predict fuel consumption for future years.