Large Data, Sparsity and LASSO

- Introduction
- Reading the Data and EDA
- Slide with R Output
- Slide with Plot
- County Level Effects on AQI

Introduction

The goal of this study is to examine the impact of certain variable on the climate by examining the AQI of counties across the United States of America using data collected by the EPA.

There are two smaller sub studies in this presentation: One examining the effects of the Climate Alliance legislative program, and another examining the correlation between aspects of counties and the air quality.

- Introduction
- Reading the Data and EDA
- Slide with R Output
- Slide with Plot
- County Level Effects on AQI

Heatmap

To begin we read the data in from the EPA datasets.

```
## `summarise()` has grouped output by 'state'. You can override using the `.groups` argument.
```

- Introduction
- Reading the Data and EDA
- Slide with R Output
- Slide with Plot
- County Level Effects on AQI

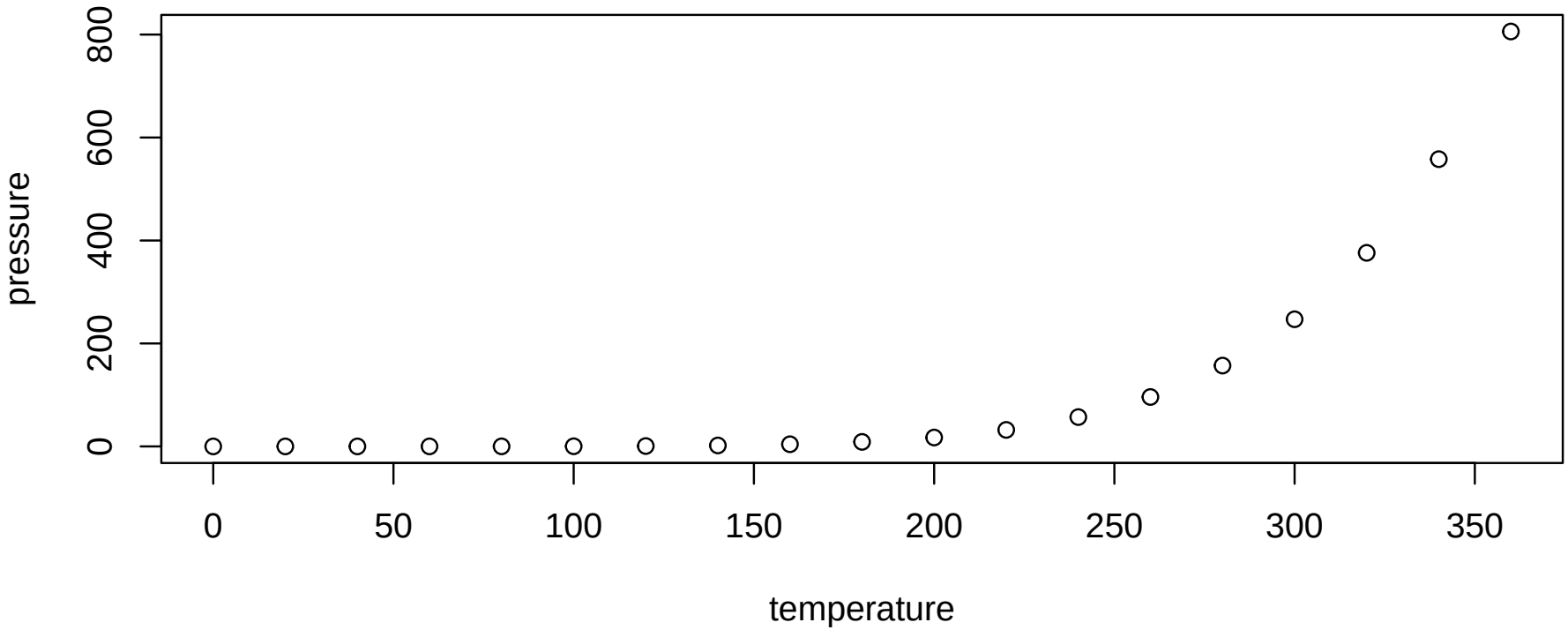
Slide with R Output

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2
##  1st Qu.:12.0   1st Qu.: 26
##  Median :15.0   Median : 36
##  Mean   :15.4   Mean   : 43
##  3rd Qu.:19.0   3rd Qu.: 56
##  Max.   :25.0   Max.   :120
```

- Introduction
- Reading the Data and EDA
- Slide with R Output
- Slide with Plot
- County Level Effects on AQI

Slide with Plot



- Introduction
- Reading the Data and EDA
- Slide with R Output
- Slide with Plot
- County Level Effects on AQI

County Level Effects on AQI

Using the data found by the USDA's Economic Research Service, we look for predictors in counties to determine air quality and find correlations.

This begins by merging the 2019 AQI with the latest USDA ERS data. We use 2019 data to avoid skewing due to the 2020 West Coast fires.

Merging AQI Data with County Data

To begin the analysis, we start by merging county data with AQI data. We start by merging all three sets of ERS county data, and then we merge by county and state.

We only take the data from year 2019 to keep it consistent. We are avoiding using 2020 data due to the fires on the West coast skewing data.

Running the LASSO Algorithm

Break the cleaned and merged dataset into X and Y for use with cv.glmnet. We use set.seed(1) for consistency.

```
## Note: Using an external vector in selections is ambiguous.
## i Use ‘all_of(select_cols)‘ instead of ‘select_cols‘ to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.


## Anova Table (Type II tests)
##
## Response: med.aqi
##                       Sum Sq  Df F value  Pr(>F)
## state                  15670  48    3.44 2.5e-13 ***
## PctEmpAgriculture        109   1    1.15  0.2848
## PctEmpConstruction       174   1    1.83  0.1761
## PctEmpFIRE               734   1    7.73  0.0055 **
## Age65AndOlderPct2010      50   1    0.53  0.4676
## Ed4AssocDegreePct        774   1    8.16  0.0044 **
## FemaleHHPct             1681   1   17.71 2.8e-05 ***
## HH65PlusAlonePct         578   1    6.09  0.0138 *
## Ed3SomeCollegeNum        737   1    7.77  0.0054 **
## ForeignBornMexNum        610   1    6.43  0.0114 *
## NetMigrationNum0010     1698   1   17.90 2.6e-05 ***
## Residuals              89962 948
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Backwards Selection with Anova

From the Anova call above, we see that Age65AndOlderPct2010 is the least relevant, so we remove it.

```
## Anova Table (Type II tests)
##
## Response: med.aqi
##                       Sum Sq  Df F value  Pr(>F)
## state                  15623  48    3.43 2.9e-13 ***
## PctEmpAgriculture         92   1    0.97  0.3246
## PctEmpConstruction       143   1    1.50  0.2205
## PctEmpFIRE               723   1    7.62  0.0059 **
## Ed4AssocDegreePct        744   1    7.84  0.0052 **
## FemaleHHPct             1652   1   17.41 3.3e-05 ***
## HH65PlusAlonePct         950   1   10.01  0.0016 **
## Ed3SomeCollegeNum        732   1    7.72  0.0056 **
## ForeignBornMexNum        618   1    6.52  0.0108 *
## NetMigrationNum0010     1683   1   17.74 2.8e-05 ***
## Residuals              90012 949
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Backwards Selection with Anova

From the Anova call above, we see that PctEmpAgriculture is the least relevant, so we remove it.

```
## Anova Table (Type II tests)
##
## Response: med.aqi
##                      Sum Sq  Df F value  Pr(>F)
## state                 16002  48    3.51 8.3e-14 ***
## PctEmpConstruction      124   1    1.31 0.25270
## PctEmpFIRE             1037   1   10.93 0.00098 ***
## Ed4AssocDegreePct       685   1    7.22 0.00732 **
## FemaleHHPct            1667   1   17.58 3.0e-05 ***
## HH65PlusAlonePct       1046   1   11.03 0.00093 ***
## Ed3SomeCollegeNum       786   1    8.29 0.00408 **
## ForeignBornMexNum       614   1    6.47 0.01112 *
## NetMigrationNum0010    1704   1   17.96 2.5e-05 ***
## Residuals             90104 950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Backwards Selection with Anova

From the Anova call above, we see that PctEmpConstruction is the least relevant, so we remove it.

```
## Anova Table (Type II tests)
##
## Response: med.aqi
##                      Sum Sq  Df F value  Pr(>F)
## state                 16606  48    3.65 1.1e-14 ***
## PctEmpFIRE             1127   1   11.88 0.00059 ***
## Ed4AssocDegreePct       733   1    7.73 0.00555 **
## FemaleHHPct            1974   1   20.81 5.7e-06 ***
## HH65PlusAlonePct       1139   1   12.01 0.00055 ***
## Ed3SomeCollegeNum       814   1    8.58 0.00348 **
## ForeignBornMexNum       582   1    6.13 0.01347 *
## NetMigrationNum0010    1679   1   17.69 2.8e-05 ***
## Residuals             90228 951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Examining the Final Fit - Do the Assumptions of the Linear Model Hold Up?

```
##                       Estimate Std. Error t value Pr(>|t|)
## PctEmpFIRE            6.13e-01   1.78e-01    3.45 5.91e-04
## Ed4AssocDegreePct    -5.29e-01   1.90e-01   -2.78 5.55e-03
## FemaleHHPct           5.35e-01   1.17e-01    4.56 5.73e-06
## HH65PlusAlonePct     -4.73e-01   1.36e-01   -3.47 5.53e-04
## Ed3SomeCollegeNum     1.42e-05   4.86e-06    2.93 3.48e-03
## ForeignBornMexNum     2.15e-05   8.67e-06    2.48 1.35e-02
## NetMigrationNum0010   2.78e-05   6.61e-06    4.21 2.84e-05
```

From the final model, we see that most of the impact on AQI is geographical. For example, the increase from ForeignBornMexNum and NetMigrationNum could signal that states closer to the Mexican border tend to have worse AQIs due to their location. However, the most clear predictors are the states themselves.

The assumptions for linearity appear to hold up until about 1 standard deviation below the mean.

Diagnostic Plots