Drugs and Jobs: The effect of unemployment on drug overdose deaths in America

Evan Arnold and Caleb Ren

12/06/2019

Introduction and Motivation

Overdose deaths in the US have increased dramatically since 1999

Exploratory Data Anslysis Methods

Data Summary

As a next step in our project, we collected data from the CDC in the form of the Vital Statistics Rapid Release dataset (VSRR). The VSRR data contains provisional counts of drug overdose deaths in the US as reported by agencies from all 50 states and the District of Columbia. The data is collected in on a monthly basis.

The data of import to this project is the number of deaths in each state as a result of drug overdose. Drug overdoses are counted by state agencies in acordance to World Health Organization standards, which lay out the basic guides for reporting agencies to code and classify causes of death. Drug categories that are represented in this dataset include the major drivers of the opioid epidemic like heroin (coded by T40.1), natural opioid analgesics (morphine and codeine), synthetic opioids (oxycodone, hydrocodone, oxymorphone; T40.2), methadone (T40.3), other synthetics (fentanyl, tramadol; T40.4) and other drugs like cocaine, methamphetamine, etc.

There were over 26052 data points from the VSRR dataset. Of those data points, many are individual observation of different coded deaths from different drugs; after reshaping and data cleaning, there are now 2652 individual observations. The data ranges from 2015 to 2019, with each state reporting 52 observations (once per month). Overdose deaths range from 55 deaths in the month of May 2015 in South Dakota to a high of 5697 in Pennsylvania in September of 2017.

Unemployment data was sourced from the Bureau of Labor Statistics. Unemployment data is published in monthly increments from the Bureau of Labor Statistics by state. Data is published beginning in 1976 and is published on the first of each month describing the previous month's unemployment rate.

There is a very specific definition of who in the labor force is considered *unemployed*. According to the BLS, those who are currently unemployed are those who are "jobless, looking for a job, and avaiable for work." People who are incarcerated, in a nursing home, or in a mental health care facility are not considered unemployed as they are not fit for work.

Using this definition, data was scraped from the BLS website and aggregated by each state and the District of Columbia. The unemployment rate in percent is given by the `unemployment` column. The lowest unemployment rate in a given state and month is Vermont in 2019 with a 2.1% unemployment rate. The highest rate is DC in 2015 with a 7.4% unemployment rate. The data itself is roughly Normally distributed with a mean of 4.2% and a median of 4.31%.

St. Louis Datasets permits: housing units authorized by building permits (raw count). This is a proxy for housing development. imports: imports in millions of dollars. This is a proxy for in-state manufactoring. income: annual income per capita

Census Bureau Dataset population: raw population

NOTE: For each state, we use the population estimate for the previous year for the entire year. This is due to a lack of available data as well as slow population growth across states.

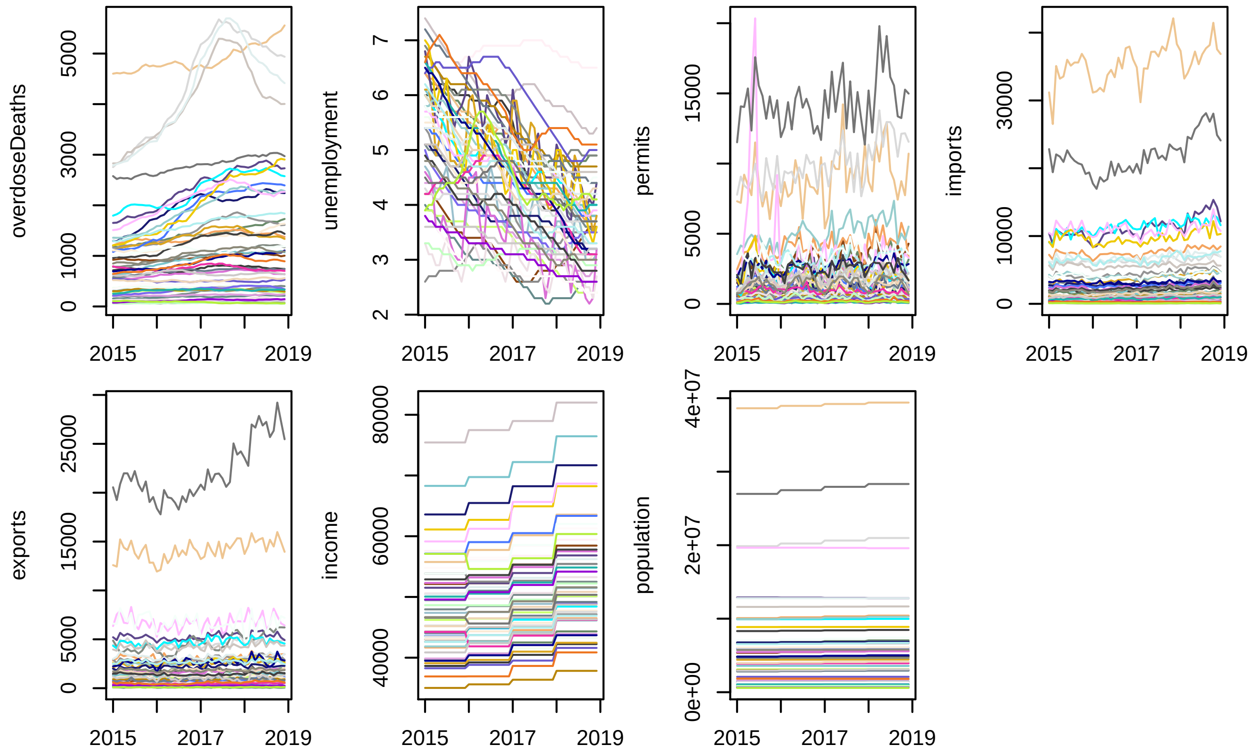
R Data region: data from R. We have to include the region for DC specifically

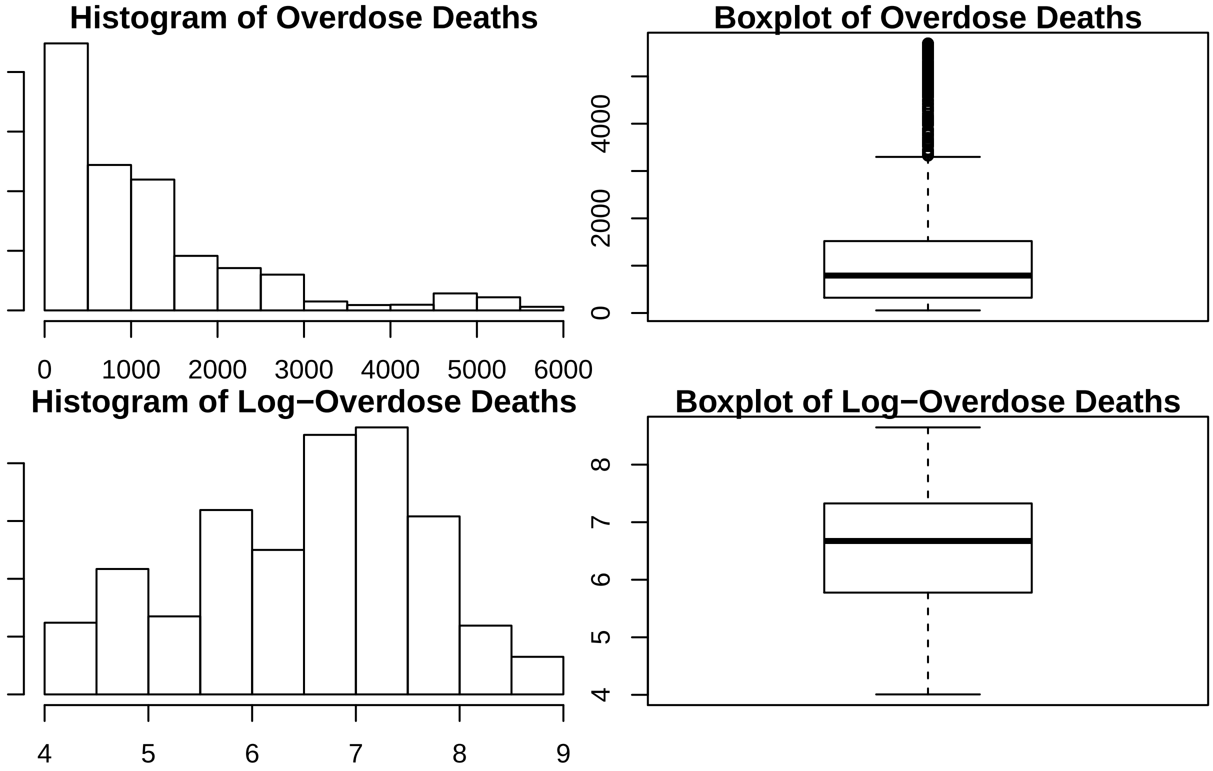
<<<<<<< HEAD Normalize overdose deaths, permits, and imports. These variables are now relative values per 100000 people.

=======

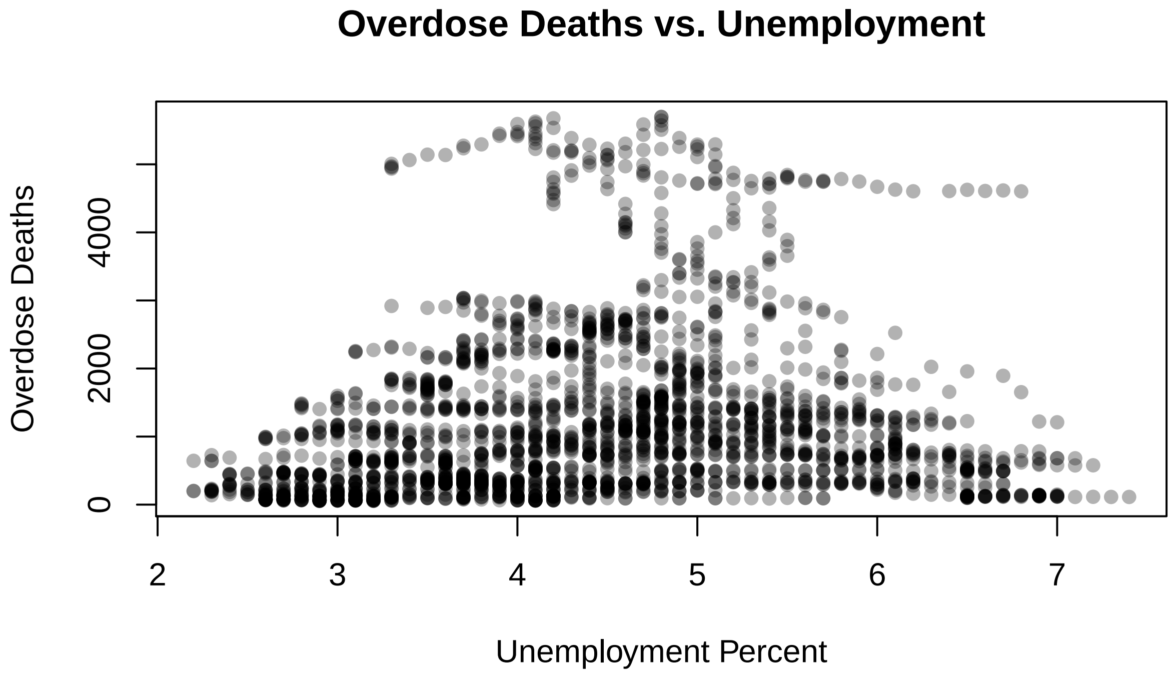
EDA

>>>>>>> 4e20851c5972817f7f45fe8698b388fa7e0bbdac





We see that the data is much closer to a Normal distribution if we apply a log transformation.

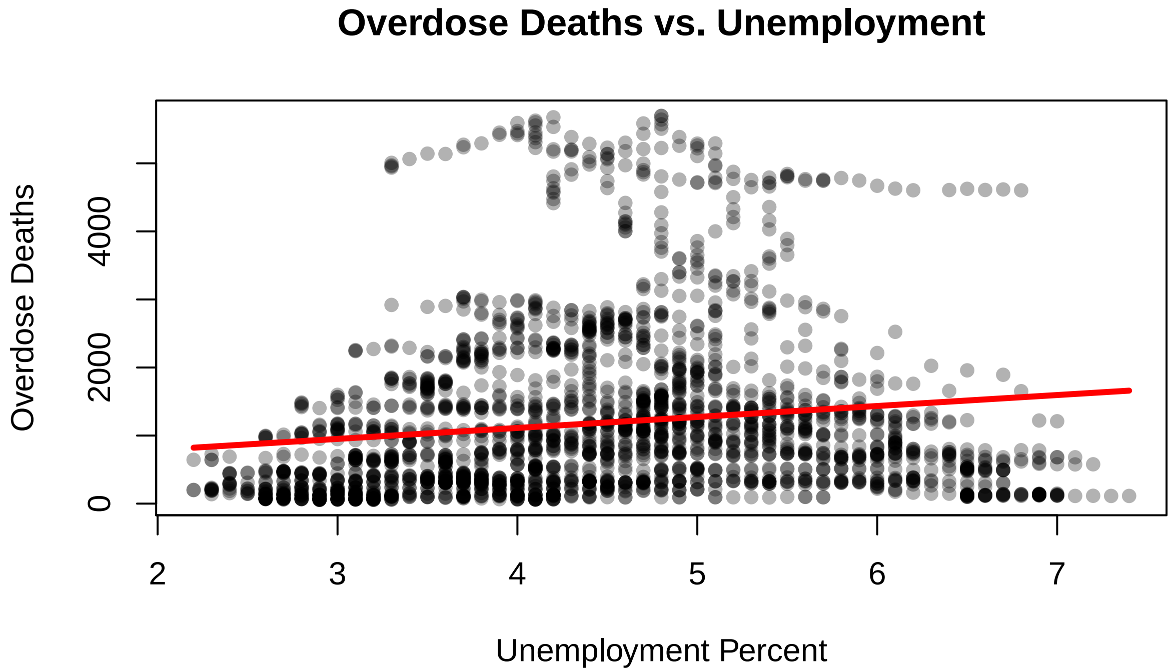


Baseline Model

```
##
## Call:
## lm(formula = overdoseDeaths ~ unemployment, data = train)
```

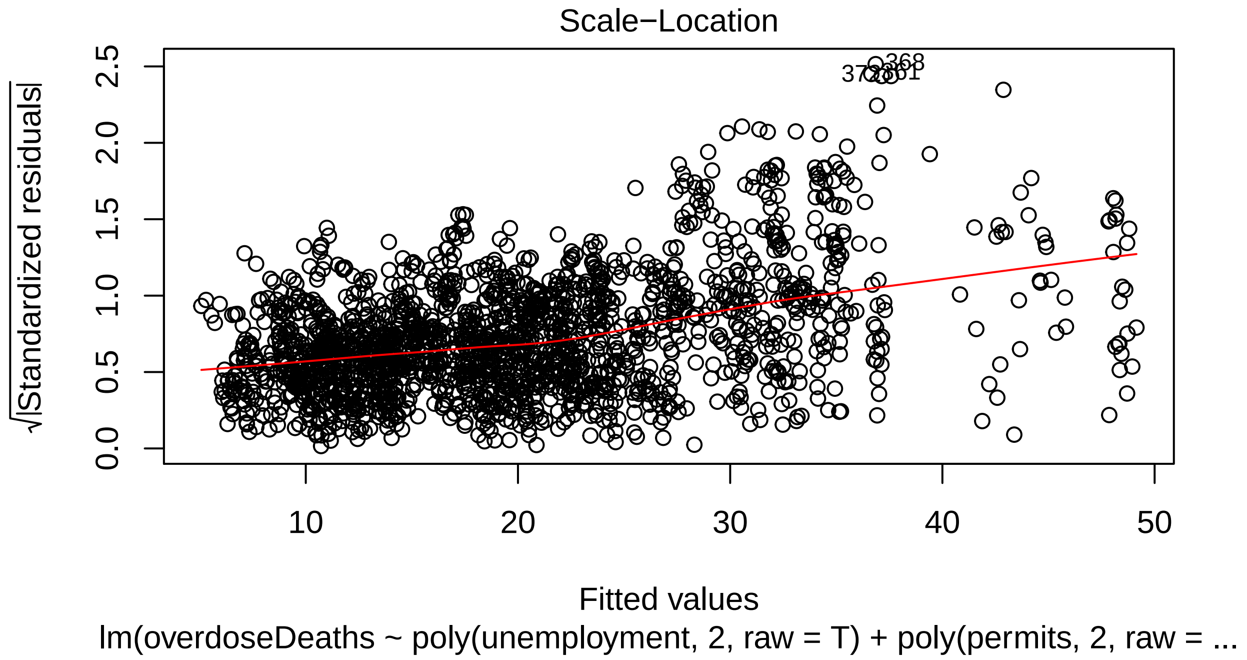
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.047  -7.266  -1.925   4.622  35.138
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   13.8459     0.8620  16.063  < 2e-16 ***
## unemployment   1.3830     0.1897   7.289 4.51e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.128 on 1956 degrees of freedom
## Multiple R-squared:  0.02645,    Adjusted R-squared:  0.02595
## F-statistic: 53.13 on 1 and 1956 DF,  p-value: 4.505e-13
```

The simple regression model has a positive coefficient for unemployment (13.164). With a t-statistic of 15.345 (p-value ≈ 0), this coefficient is very significant. The model has a positive association between unemployment and overdose deaths.



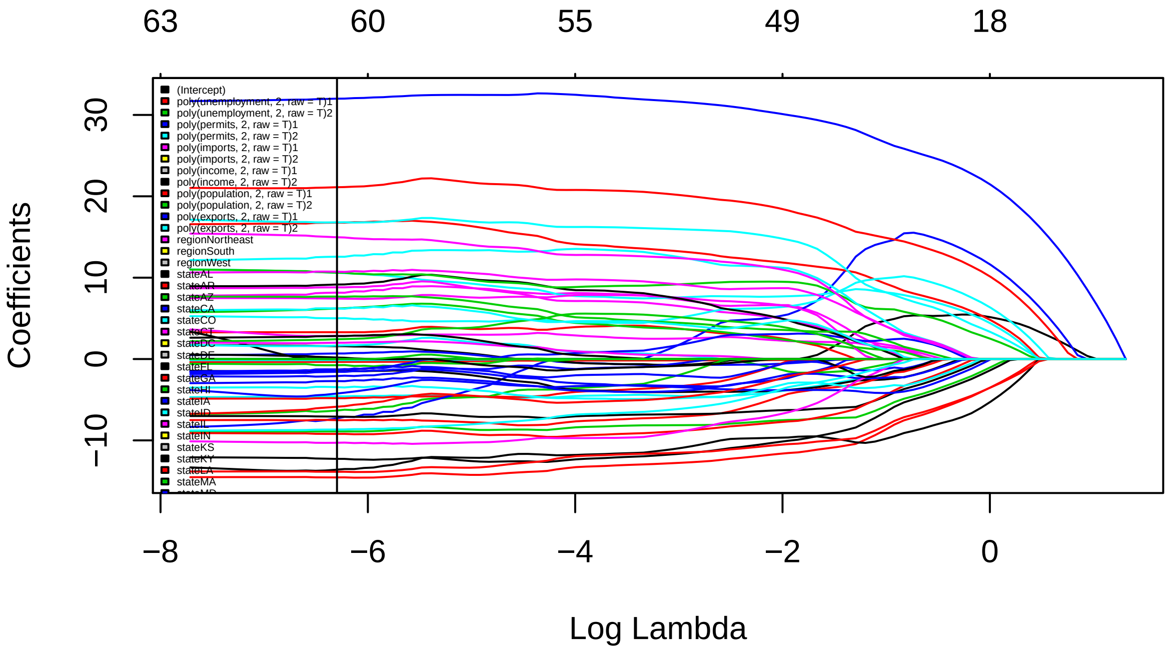
In order to better evaluate the importance of unemployment, we apply an ESS F-test. First, we fit two models: the first model contains all main effects and their respective quadratic versions besides unemployment. The second model contains the same predictors, except that it includes unemployment and its quadratic effect. Here we note that we are ignoring any time effect as of now. Additionally, we again verify the assumptions of a linear model via a plot of residuals vs. fitten values.

```
## Analysis of Variance Table
##
## Model 1: overdoseDeaths ~ poly(permits, 2, raw = T) + poly(imports, 2,
##     raw = T) + poly(income, 2, raw = T) + poly(population, 2,
##     raw = T) + poly(exports, 2, raw = T) + region + state
## Model 2: overdoseDeaths ~ poly(unemployment, 2, raw = T) + poly(permits,
##     2, raw = T) + poly(imports, 2, raw = T) + poly(income, 2,
##     raw = T) + poly(population, 2, raw = T) + poly(exports, 2,
##     raw = T) + region + state
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   1897 18806
## 2   1895 17554  2    1252.1 67.584 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

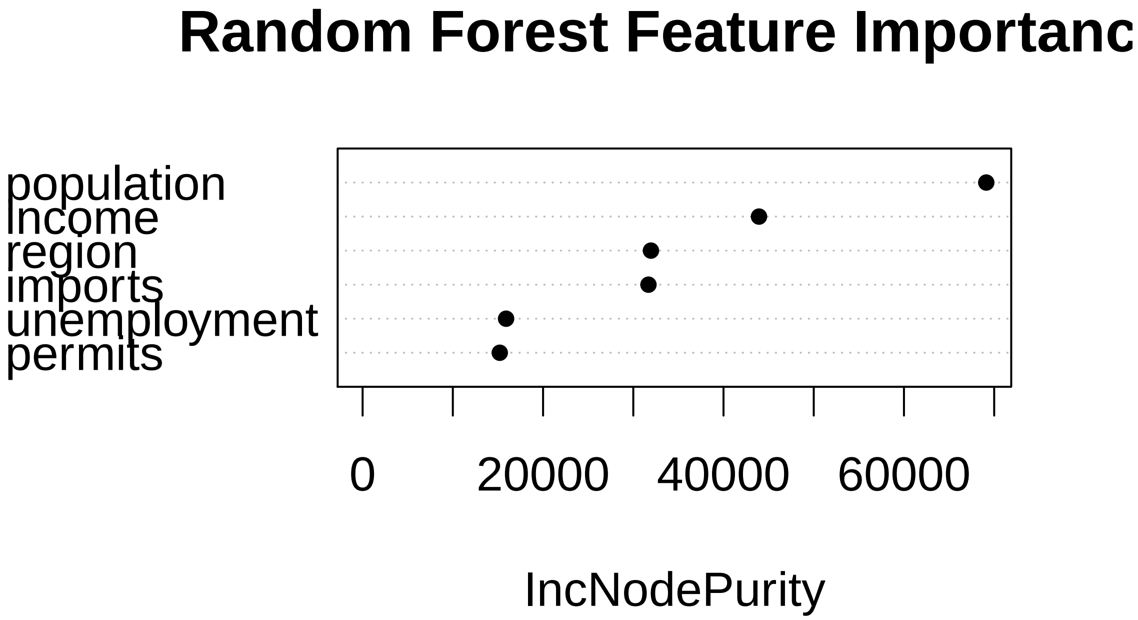


With an F-statistic of 12.208 (with 2434, 2 degrees of freedom) and a corresponding p-value << 0.001, unemployment provides significant predictive ability. In order to be thorough, we examine the importance of unemployment in two additional ways: lasso regression, and random forrest regression. Additionally, there seems to be no underlying structure to the residuals, the residuals appear to have reasonably constant variance, and are reasonably normally distributed: the assumptions of our models (linearity, homoscedasticity of residuals, and normality) appear to be reasonably satisfied. As mentions above, we are purposefully ignoring a time effect so samples are likely correlated and thus not completely independent. We shall investigate this effect further in the analysis.

We begin with lasso regression. Due to l1 penalty of lasso regression, less important variables quickly converge to zero as the regularization parameter increases. Below, we fit a lasso model (with the same dedign matrix as that of the full quadratic kitchen sink model) with a series of possible regularization parameters. We then consider the order in which the variable coefficients shrink to zero.



Lasso feature selection agrees with the results of the ESS F-test. Unemployment (the non-quadratic effect) is one of the last coefficients to shrink to zero along with population and imports. We further corroborate this conclusion with a random forest model. Below, we fit a random forest model with all main effects and then examine it's relative feature importance.



Unemployment is relatively not important in the random forest model.

Results

Conclusions and Decisions