Ryan Francis

Sean Strunk

E392 Final Project Paper

**Drug Overdoses Deaths in Relation to Poverty**

As a society we value many important principles, first and foremost being the value of human life. In today’s society we are increasingly facing a threat to human life that is worthy of research directed at understanding the underlying causes of that threat. The threat to which I am referring to is the increasing epidemic of drug overdoses. Lately this issue has been increasing in both importance and magnitude as we as a nation are facing an opioid crisis that has claimed the lives of tens of thousands of people. Many factors are obviously involved in an individual ultimately dying of a drug overdose, among which poverty is commonly viewed among them. People often view poverty within the lenses of drug use and other negative societal factors. This paper seeks to find the connection between poverty with the likelihood of a person dying due to fatal drug overdoses. The focus of this paper is specifically on the state of Indiana and all the data used to analyze this topic is from Indiana.

**Literature Review**

The question of the link between poverty and drug overdoses has been covered in previous academic literature. A paper titled Neighborhood Poverty and Injection Cessation in a Sample of Injection Drug Users (2009) covers this very topic. The paper argues that the best methods to reduce the number of people who died from drug overdoses is to encourage cessation of drug use. The rest of the paper is based off of methods to measure if drug users have ceased using drugs, particularly injection drugs. The particular data set that this paper uses is neighborhood level data, data that the authors personally collected by doing interviews in their sample. He reason they use neighborhood level data is because there are certain neighborhoods that are both plagued with both poverty and common drug use. It was argued that drug use should be viewed through the lenses of job loss and other economic opportunities. This particular paper goes as far as to argue causality, i.e. that poverty causes increased injection drug use.

The other related article is titled Income Distribution and Risk of Fatal Drug Overdose and New York City Neighborhoods (2002). This paper takes a different perspective from the previous one. It focuses and on economic inequality as a causal Factor for drug use as opposed to poverty in absolute terms. The paper makes the following argument, drug use is a social activity and poorer people in economically unequal areas will feel particularly disenfranchised with the institution of that area so they will be less likely to seek out health help in the event that they require it during a drug overdose emergency. To measure the inequality they use the Gini coefficient and found that both economic inequality and poverty in absolute terms, measured by the income of the lowest 70 percent of households, were both statistically significant at the neighborhood level.

**Data and Methodology**

The data that I collected was Indiana county level data from the Indiana Department of Health and from a website titled County Health rankings. From there I got my two measures of poverty that we will be using in this paper, median income and free and reduced lunches. Free and reduced lunches is a commonly used proxy for poverty and will suffice for the purposes of this paper due to the fact that poverty is difficult to measure directly. In addition to the poverty measures, this paper will be addressing other variables that will likely have an effect on drug overdoses such as the percentage of the population that is under the age of 18, the percentage of African-Americans within a particular County, the population of a county, and the number of hospitals within a county. We will look at two possible dependent variables within this paper relevant to our research question, the total drug deaths within a county and the percentage of drug overdoses within a county that ultimately lead to a fatality.

The methodology that I used in my research was to study the changes in Indiana over time in both income and drug overdoses. To start with my project I used a fixed effects model to keep all other factors constant within Indiana counties. I have collected panel data of all the previously stated variables from the years 2014 to 2016. The interpretation of the median income coefficient is as follows; if it is negative as is expected then the lower the median income is associated with higher the drug mortality. The interpretation for the free and reduced lunch coefficient will be as follows; a positive coefficient will implies that there will be an association of higher drug deaths within the county. For the remaining control variables a positive coefficient would imply a positive relationship with increasing drug mortality and a negative coefficient would imply the opposite.

**Models and Analysis**

summary(Project5)

## County FIPS Population young   
## Length:280 Length:280 Min. : 8804 Min. :15.90   
## Class :character Class :character 1st Qu.: 26446 1st Qu.:22.33   
## Mode :character Mode :character Median : 40824 Median :23.42   
## Mean : 87278 Mean :23.49   
## 3rd Qu.: 82607 3rd Qu.:24.77   
## Max. :939020 Max. :31.07   
## African DrugDeaths DrugMortalityRate MedianIncome   
## Min. : 0.1906 Min. : 10.00 Min. : 4.034 Min. :35220   
## 1st Qu.: 0.4867 1st Qu.: 16.75 1st Qu.:10.599 1st Qu.:42982   
## Median : 0.9348 Median : 31.00 Median :14.446 Median :47475   
## Mean : 3.2456 Mean : 61.79 Mean :15.985 Mean :49266   
## 3rd Qu.: 4.5020 3rd Qu.: 71.00 3rd Qu.:20.177 3rd Qu.:52495   
## Max. :27.2855 Max. :1027.00 Max. :46.635 Max. :91844   
## FreeLunch Year nHosp   
## Min. :11.56 Min. :2014 Min. : 0.000   
## 1st Qu.:34.18 1st Qu.:2014 1st Qu.: 1.000   
## Median :39.89 Median :2015 Median : 1.000   
## Mean :39.70 Mean :2015 Mean : 2.311   
## 3rd Qu.:46.40 3rd Qu.:2016 3rd Qu.: 3.000   
## Max. :67.07 Max. :2017 Max. :21.000

The summary of the data file is above. Some of the data is meaningless, for example the FIPS ID or county names. However, the rest of the summaries provide and insight into the data set. The data appears to have a lot of outliers in the max category. Further, when looking at the percent of African-American communities the median is less than a percent, but the mean is 3.3 percent. This would indicate the African population is largely clustered and segregated from other populations. This will be important later in the analysis.

**OLS**

#Multiple Linear Regression  
OLS <- lm(DrugDeaths~Population+young+African+MedianIncome+FreeLunch+Year+nHosp)  
summary(OLS)

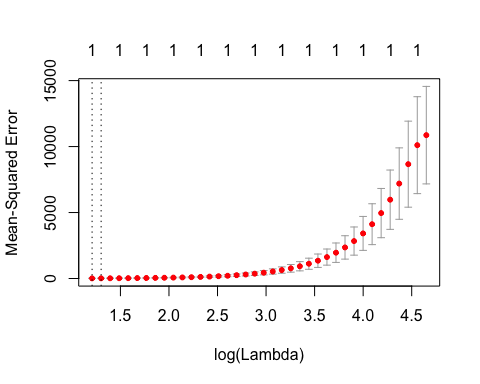
##   
## Call:  
## lm(formula = DrugDeaths ~ Population + young + African + MedianIncome +   
## FreeLunch + Year + nHosp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -157.852 -15.523 -0.563 13.104 315.476   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.373e+04 6.805e+03 3.488 0.000568 \*\*\*  
## Population 1.012e-03 7.900e-05 12.807 < 2e-16 \*\*\*  
## young -4.629e+00 1.415e+00 -3.272 0.001205 \*\*   
## African -4.870e+00 1.250e+00 -3.897 0.000123 \*\*\*  
## MedianIncome -1.254e-03 5.815e-04 -2.157 0.031853 \*   
## FreeLunch 5.517e-01 5.330e-01 1.035 0.301590   
## Year -1.170e+01 3.388e+00 -3.453 0.000642 \*\*\*  
## nHosp -5.450e+00 3.094e+00 -1.761 0.079323 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 40.89 on 272 degrees of freedom  
## Multiple R-squared: 0.8517, Adjusted R-squared: 0.8478   
## F-statistic: 223.1 on 7 and 272 DF, p-value: < 2.2e-16

The basic OLS regression analysis output is above. Both Population and Median Income indicate that while the two variables are important (p < 0.05), the estimated coeeficents are very small (~ |0.001|). The variables definitly matter, but they do not impact the variable of interest (Drug Deaths) much compared to other variables. Next, the Year variable is shown to be very significatnt (p << .05). This variable is trickier as we have not accounted for any possible changes in treatment, regualtion, enactment of new laws or enforcment of existing laws. This is a major issue of in the OLS base model and something we explore further in this section. Still, with only a three year periord, OLS is still a good first model to run to understand any large trends in the data. The last two signifcant variables, percent young(< 18) and percent African-American, both show large negative values. As a county increases 1% in youth, a decrease in drug deaths of 4.6 is predicted. As a county increases its African-American population by 1%, a decrease 4.9 drug deaths is predicted. Lastly, two variables are shown to not be very significant in the OLS model, one wosre than the other. nHosp tells one how many hospitals are in each county. At a p > .05, it does not pass the test of being staitscally significant in this model. Further in the paper, we will delve into this variable more. Free Lunch is the percent of a county that are in the Free Lunch Program. We used this as a proxy for poverty. In the base OLS model, Drug Deaths does not have a statistically significant conection to Free Lunch, or poverty, at p = 0.32. There are a few issues with this model, partly discussed above (year to year changes). Due to these issues, we will look at some other models to further probe this question.

**Lasso**

x=model.matrix(Population~.,Project5)  
y=Project5$DrugDeaths

#lasso Regression  
lasso.cv <- cv.glmnet(x,y,alpha=1)  
plot(lasso.cv)



#lasso.cv  
fit <- lasso.cv$glmnet.fit  
lamda.min1 = lasso.cv$lambda.min  
lamda.min1

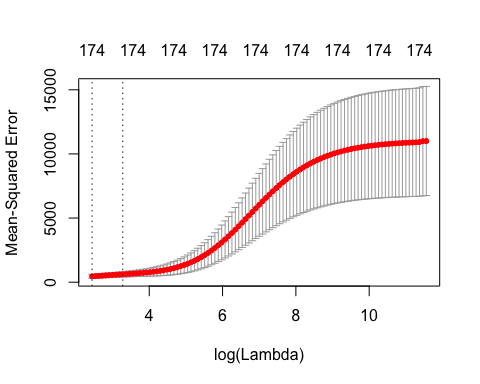
## [1] 3.3475

coef <- coef(lasso.cv)  
#coef ##Nothing interesting is displayed and it takes up a lot of space

The lasso model is the next model tried. It retruns a value of lamda min = 0.237. From the plot of MSE vs log(Lamda), the model is essentaily giving OLS as the end result. It also says that only 1 variable is important. This largely means that the model was either coded incorrectly(very possible), or that the lasso model is giving OLS. Regardless, it is not a good model for this dataset. This is becsuse lasso is best for large-p data sets where only a few variables matter.

**Ridge**

#Ridge Regression  
ridge.cv <- cv.glmnet(x,y,alpha=0)  
plot(ridge.cv)



#ridge.cv  
fit1 <- ridge.cv$glmnet.fit  
lamda.min2 = ridge.cv$lambda.min  
lamda.min2

## [1] 11.4835

coef.ridge <- coef(ridge.cv)  
#coef.ridge ##Nothing interesting is displayed and it takes up a lot of space

The ridge model is very similar in application to the lasso. However, in this case the model chooses a lamda min =77.75. This would indicate that rather than OLS, the model is heavily penlized. The coeeficents in this case are estimated for all 92 (each county is being counted seperately), but they are all approximitly 0.

Both of the advanced machine learning techniques do not do well to fit the data. This is not terribly surpising at they are optimized for different kinds of data. This means that the first OLS model is still optimal. The OLS model had it own issues, for example, it did not account for differnt years or different counties. The next model is a fixed effets OLS model. This means that the model is able to control for certain variables. In this case, the model is told to run OLS on each county, for each year and to estimate the parameters off this.

**Fixed Effects OLS**

#OLS Fixed Effects  
fixed\_effects <- plm(DrugDeaths~Population+MedianIncome+FreeLunch+young+African+nHosp+young\*FreeLunch+FreeLunch\*nHosp+FreeLunch\*African,data=Project5,model = "within",effect = "time", index = c("County","Year"))  
summary(fixed\_effects)

## Oneway (time) effect Within Model  
##   
## Call:  
## plm(formula = DrugDeaths ~ Population + MedianIncome + FreeLunch +   
## young + African + nHosp + young \* FreeLunch + FreeLunch \*   
## nHosp + FreeLunch \* African, data = Project5, effect = "time",   
## model = "within", index = c("County", "Year"))  
##   
## Unbalanced Panel: n = 84, T = 1-4, N = 280  
##   
## Residuals:  
## Min. 1st Qu. Median 3rd Qu. Max.   
## -132.086 -18.039 -1.796 15.059 276.688   
##   
## Coefficients:  
## Estimate Std. Error t-value Pr(>|t|)   
## Population 8.7155e-04 8.0603e-05 10.8128 < 2.2e-16 \*\*\*  
## MedianIncome -5.7308e-04 6.5476e-04 -0.8752 0.382228   
## FreeLunch 1.2579e+01 2.8996e+00 4.3381 2.040e-05 \*\*\*  
## young 1.3909e+01 4.7362e+00 2.9367 0.003607 \*\*   
## African 1.4521e+01 4.7463e+00 3.0594 0.002443 \*\*   
## nHosp -3.2967e+01 5.1656e+00 -6.3821 7.687e-10 \*\*\*  
## FreeLunch:young -5.0750e-01 1.2036e-01 -4.2166 3.398e-05 \*\*\*  
## FreeLunch:nHosp 8.0629e-01 1.2827e-01 6.2861 1.321e-09 \*\*\*  
## FreeLunch:African -4.7757e-01 1.0874e-01 -4.3916 1.623e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Total Sum of Squares: 3031800  
## Residual Sum of Squares: 366920  
## R-Squared: 0.87898  
## Adj. R-Squared: 0.87354  
## F-statistic: 215.469 on 9 and 267 DF, p-value: < 2.22e-16

Within the fixed effects model are several new variables. These are interactions between two of the variables. Immedietly, this model shows a better fit to the data. Median Income in the only not significant variable. In the base OLS model, Median Income is signicant but not very strong. In the fixed effects OLS model (feOLS), Free Lunch takes Median Income’s place in the model. As discussed prior, this is our proxy for poverty. When one controls for county and year data, poverty becomes a strong predictor of increased drug deaths as p < .001. Further, the percent of the counties that are young or African-American show predictions for higher drug deaths, both p < .01. The coeeficents for Free Lunch, Young and African-American are all of the same order indicating similar strengths. All three are positive meaning that as they increase, so do the amount of drug deaths. The last non-interacted variable is nHosp, or the number of hospitals in each county. In OLS, we showed that this variable was not significant (p > .05), however, in the feOLS model nHosp beomes a very strong predictor of a decrease in drug deaths. In fact, nHosp has a coeeficent of about 3 times the size as Free Lunch, Young or African-American. This makes a lot of intuitive sense as the easier accsess one has to help should refelct a decrease in deadly overdoses, even for the same overdose rate.

In the feOLS model, we also ran some interaction varibales. These provide a way to study how counties high in poverty and very young, counties high in poverty and how many hospitals there are, and poverty and the percent of African-Americans impact the amount of Drug Deaths. All of the variables are p < .001 meaning they all are statistally significant. In impovershied counties with a large amount of young people the drug deaths fell, which is the oppisote compared to the individual effects of the two variables. This may mean that these areas are prone to Drug Deaths, but they have not reached that point yet. Poverty and the number of Hospitals interacted togeather produce an increase in drug deaths. When inspecting the data, the counties with the highest poverty do not have too many Hospitals, except for Marion. This interaction is difficult to interpret and may be effected by outliers. The last interaction is poverty and percent African-American. While African-American provides an increase in drug deaths, the impovershied counties with high African-American populations provides a small decrease. An important thing to note for all of the interactions is that they are two orders of magnitude smaller than the individual coeeficents. This means that they are much weaker, but as they are interacted we have more difficult interpreting them. They may offer paths to asking new questions and probing deeper questions within the data set.

**Further Questions**

The results of this paper, while interesting in by its own merits, are by no means comprehensive of all the work that could be done on this topic. Further research and inquiry into this topic could include measuring the total hospital beds available to a particular county instead of just using the aggregate number of hospitals to get a better measure into particular counties medical capacities. This paper also did not include rural measurements that could be significant in the quest to understanding drug deaths due to more rural counties possibly not having hospitals in close proximity to the residential areas within the counties. Further demographic information would provide a more detailed analysis on race or age’s impact on drug deaths. More detailed information about what kind of overdoses cause death, or occur at all, in each county would help when one tries to form public policy.

**Conclusion**

Using the Fixed Effects OLS model, we conclude that the demographics of youth, racial makeup, poverty and access to care are all statistically significant on a county by county, year by year basis. This helps us have a better understanding of the drug epidemic that is currently plaguing our state of Indiana and will help us inform any potential opinions on how we as a society should proceed to help alleviate this issue.