Final Project

Sean Palac

December 6, 2016

Introduction

The "War on Drugs" is a term popularized by former president Ronald Reagan. Recently, the main focus of said war has been the legalization of marijuana. But one drug that has lurked in the shadows and continues to become an ever-present threat to the American people, is heroin. According to the Centers for Disease Control and Prevention, heroin use has more than doubled in the past decade amongst American citizens. In addition, the US Department of Health and Human Services has claimed that deaths related to heroin have more than tripled since 2010. Before we can begin to address solutions to this problem, we must first discover where the problem is manifesting. So the question remains: where is the recent spike in heroin use coming from? Is the stereotypical belief that urban regions are the main contributor to drug use true, or has the issue become the less populated, rural regions?

In order to answer the previously stated question, data tables were gathered with the intention to gain information on the age adjusted death rates due to heroin as well as on population, land area, and personal income statistics. All data sets were used to find data at the county level in order to paint a more exact picture of the issue. From here, we use the population and land area statistics to find the population densities of each county. The larger densities most likely represent more populated, urban regions and the smaller densities most likely represent the rural and suburban counties. Thus, we use these age adjusted death rates due to heroin use, as predicted by the population density of counties and controlled for by economic factors such as working age male participation rate, personal income, or unemployment rate to hypothesize that the recent surge of heroin use is manifesting in the less dense, rural regions of America rather than the more dense, urban counties.

```
county_choropleth(mapper, title = "Age Adjusted Death Rates 2014")
```

Methods

The data used, begins with an evaluation of the Age Adjusted Death Rates for Drug Poisoning in counties across the US. The data was gathered by the National Center for Health Statistics and describes the estimated rate at which people in each county have died due to drug poisoning causes. According to the NCHS, "Drug-poisoning deaths are defined as having ICD-10 underlying cause-of-death codes X40-X44 (unintentional), X60-X64 (suicide), X85 (homicide), or Y10-Y14 (undetermined intent)." The recorded data spans from the year 1999 to 2014 and describes the estimated range of age adjusted death rates rather than crude rates. Thus, in order to "clean" the raw data uploaded, we must isolate that death rate data column, splitting the string value where a "-" occurs. This is done in order to separate the minimum estimated rate of death recorded and the maximum estimated rate of death per county. From here, we take the average of the two values provided by the split in order to obtain the best estimation of the true age adjusted death rate per county. We chose to record the average due to the assumption that choosing the minimum or maximum rate estimations for each county would most likely tend to provide an underestimated or overestimated prediction, respectively. In addition, choosing a randomized value within the range would not be able to tell us much information, as the randomization might not be truly representative of the county. Thus the "safest bet" is to use the average of the two extremes. After replacing this ranged data with numerical averages for each county across all years, we create a new data frame specifying the year of data desired. In this case, the most recent year, 2014, was chosen as it would be most representative of current trends in drug poisoning across the US.

Next, we upload the Bureau of Labor Statistics' (BLS) data on employment and unemployment rates per county as well as the American Community Survey's (ACS) data available through R's packages. From here, we are able to create a data frame of the working age population of specific demographics as well as the fips

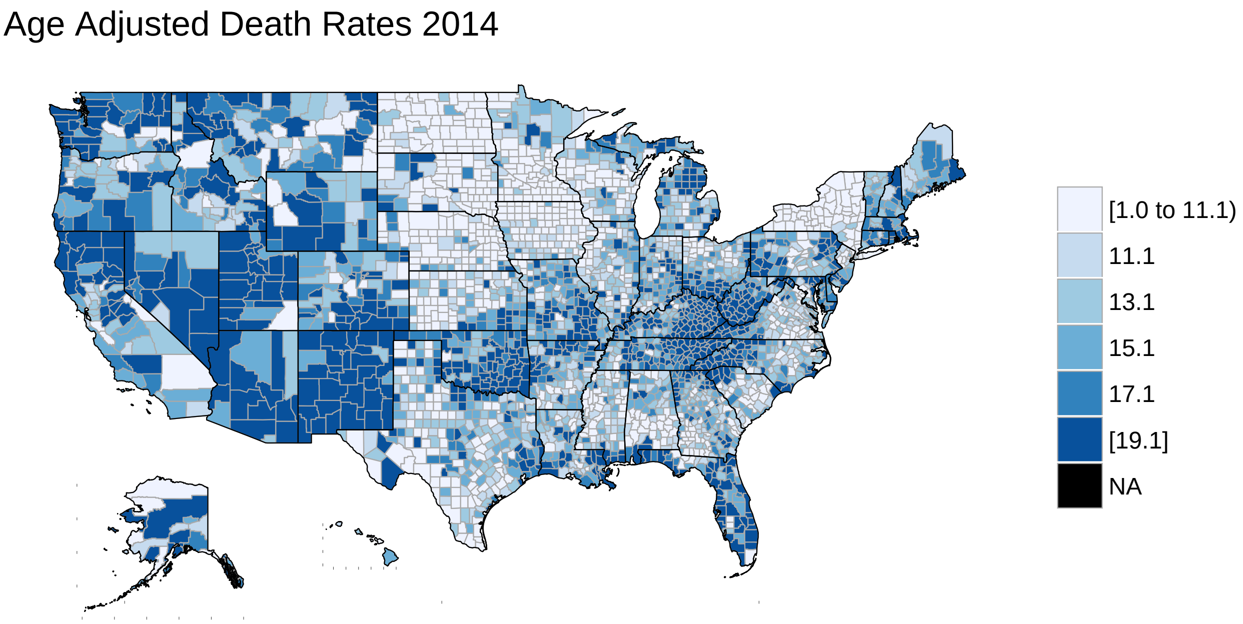

Figure 1: This maps the age adjusted rate of death per county across the US. This is to illustrate the main areas of concern in the country and to be used for comparison with later maps of residuals and fitted values of regressions. As we can see from the map, it seems that the west as well as the Appalachian Mountain area and Oklahoma are where the highest death rates seem to occur. However, the Midwest and New York have the lowest. This gives concern as to whether or not there is a regional effect or whether heroin has not penetrated the market in these areas.

codes of the counties they are associated with. This will be useful when paired with the BLS employment data as we can then combine the two data sets and calculate the labor force participation of each county. For our research, we specifically looked at the labor force participation of men aged 18 to 54 years old, as these ages represent the prime working ages of American men. It was chosen to use these data due to the revelation of the "Missing Men Phenomenon," which claims that there is a large group of working class men that are missing from the labor force unlike previous decades. Due to surveys taken by different groups, it is believed that injury and substance abuse, specifically with opioids, has led to this lack of representation. Thus, in order to control for this clearly significant influence, we add the participation rate as a control variable in our analysis.

In addition, data from the Bureau of Economic Analysis was used to acquire information about the per capita personal income of a county. This was recorded for each county in order to control for the possible confounding variable of poverty. Personal income is a measure of all the wealth gained by a specific region in a given time period, which we will use as a proxy for wealth of a county. I believe that for those with less personal income to spend, it can be assumed that it would be harder to afford prescription painkillers and over-the-counter medicine to alleviate pain. Thus, it can also be assumed that if heroin were to penetrate the market of these lower-income areas, there would be a higher rate of heroin use and thus a higher rate of heroin-related drug poisoning deaths. Therefore, in order to control for this possible confounding variable, we add a variable for per capita personal income of each county to our regression models.

Finally, we extracted data from the US Census Bureau on the land area and population statistics of each county. From here, we combined the data frame on drug poisoning statistics, which also included population counts of each county, with that of land area based on matching fips codes. This allowed us to calculate the population density of each county by dividing the total population of each county by its land area in square miles. Through this calculation, we created a new column in the new data frame representing the desired densities, which was later used to predict drug poisoning death rates. One issue that arose from this however, were outlying points that had huge leveraging effects on our regression analysis. For instance, New York County had a population density of approximately 70,000, which was well above the mean population density of 268. Thus, it was deemed necessary to remove rows of data that included population densities of 5000 or more, as they reduced the accuracy of our predictions, as well as their trustworthiness, substantially.

Results

```
## 
## Call:
## lm(formula = final$deaths ~ final$pop.density)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -12.6578  -4.6055  -0.6094   5.3215   6.3452 
## 
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       1.365e+01  9.602e-02 142.210  < 2e-16 ***
## final$pop.density 7.146e-04  2.261e-04   3.161  0.00159 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 4.937 on 3063 degrees of freedom
## Multiple R-squared:  0.003251,   Adjusted R-squared:  0.002926 
## F-statistic: 9.991 on 1 and 3063 DF,  p-value: 0.001589
```

The first test that was run was that of a simple linear regression, where our dependent variable was the age adjusted death rates per county and our independent variable was population density. Performing this regression with a single variable did give us a significant p-value of .0016 along with a beta coefficient of .0007146, which could suggest that population density does indeed positively influence death rates due to heroin. The beta coefficient can be summed up to mean that for each unit increase in population density, the death rate increases by .0007146. Although this coefficient seems small, the amount at which population density may increase is very large; thus, for population densities over 100, of which there are many, the death rate will increase by quite a bit. However, I believe this correlation may occur due to the effects of other confounding variables playing a part in the relationship between counties with large population densities and high death rates. Thus, the variables unemployment rate, labor force, and personal income were added to the model separately. In addition, interaction terms were added to each respective model due the strong possibility of collinearity, especially in the cases of unemployment rate and labor force. After running each simple regression and gathering their respective p-values, all but two regression returned a p-value that indicated population density was a significant determinant of average heroin death rates per county. The only regressions that did not return similar findings were one that involved population density with unemployment rate as well as an interaction term as independent variables and one with population density and personal income as independent variables with an interaction term as well. Thus, for simple linear regression models, we seem to have information pointing to the idea that as population density increases, so does the age adjusted death rate due to heroin use. However, when control variables are applied, it seems that population density is not very predictive of death rates due to heroin use.

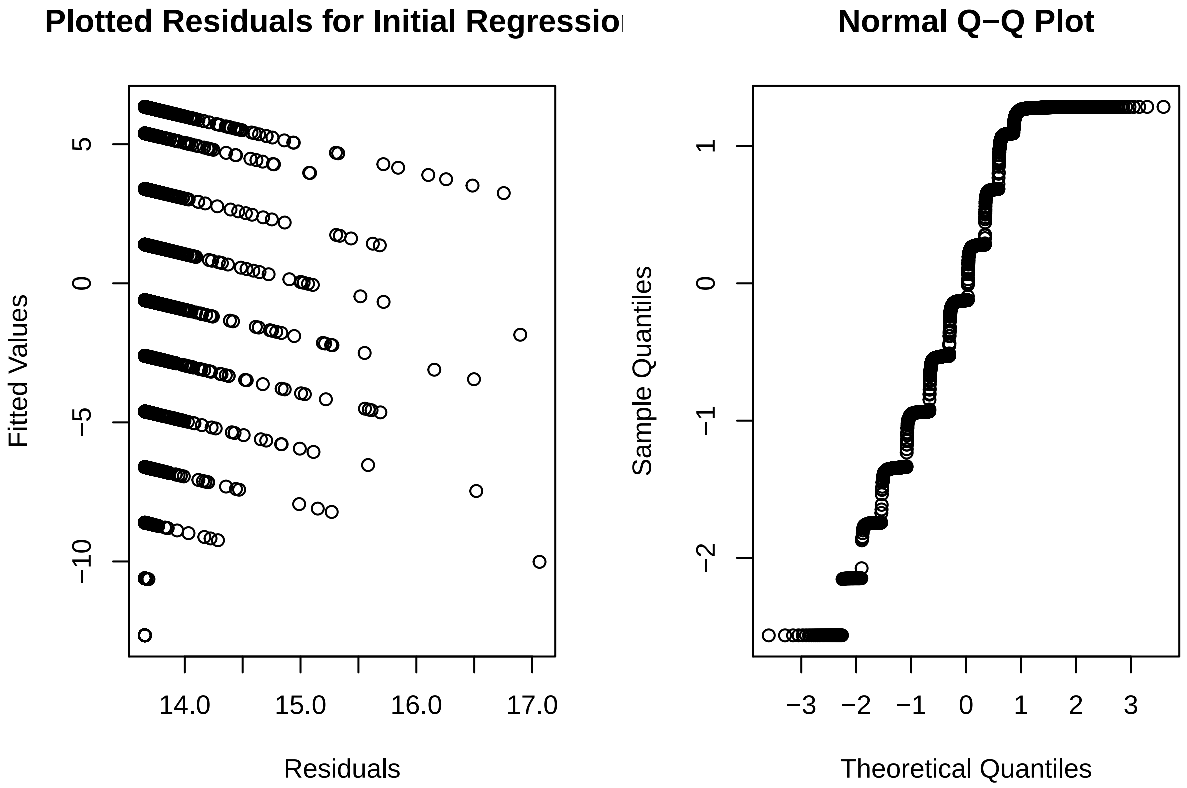

Figure 2: As seen by the fact that our residuals plot contains seemingly parallel lines and how our Normal QQ plot has a pattern to it that does not follow a straight line, our independent variables must be correlated in some fashion. The parrallel lines following a downward sloping, linear pattern indicates to us that there are multiple classes of results and possibly that Linear Discriminant Analysis would be a better test for this data.

In addition to simple and multiple linear regression models, generalized linear modeling fit with a Poisson distribution was attempted to find a truly influential independent variable. This was a chosen method as our response variable only has a finite number of values it can return based on our data cleaning. Thus, when we get our data, the average death rate due to heroin of a county can be viewed as count data in a sense.

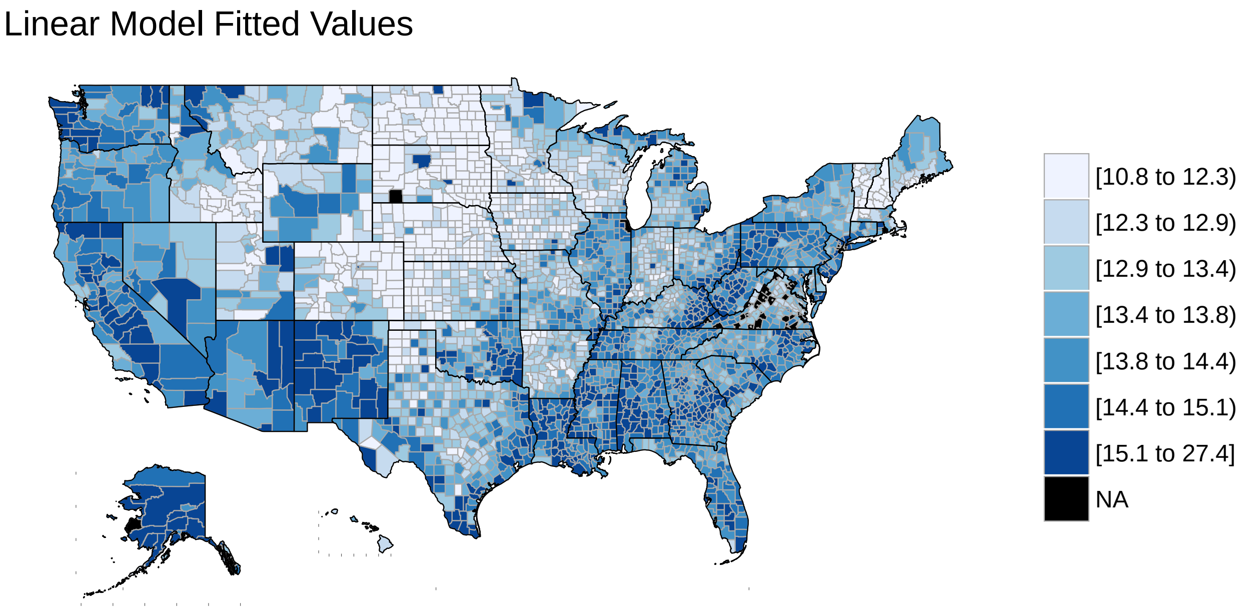

Figure 3: Here we have mapped the fitted values of our most accurate simple linear regression model. Based on how large these fitted values are in the given areas, we see that the model does not do the best at predicting death rates. In fact it overestimates nearly every region aside from those where heroin is a major issue, such as the Appalachian Mountain Range as well as the Southwest.

Therefore, it was assumed that a regression fit to a Poisson distribution would be more representative of the data collected than that of a simple or multiple linear regression model. Thus, we found that the most representative model for our data was that which regressed population density with unemployment rate and an interaction term between the two. Once again, as was the case with our multiple regression, unemployment rate seemed to be the significant variable, but population density was measured as more significant in this case compared to the multiple linear regression test. Therefore, from our fit model, we can determine that due to a large p-value of .152 associated to the population density coefficient, population density is not a good predictor of the age adjusted death rate due to heroin in a given county.

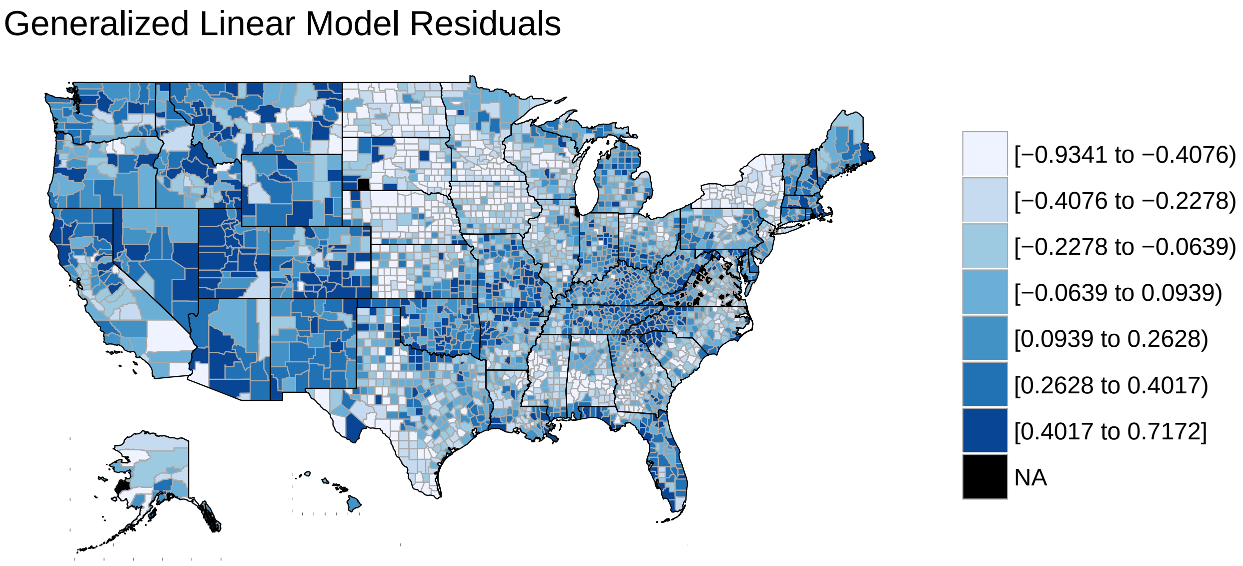

Figure 4: Here, we have mapped the residuals to our Generalized linear regression model. We can see by the layout of the graph that there still seems to be a regional influence on the death rates estimated by our model. This can be seen by the consistency of certain residuals in each region, which implies correlated residuals. Ideally, we would expect the residuals to be more random throughout the nation and have more of a checkerboard pattern.

However, upon further inspection of the residual map, it was found that our residuals seem to be very highly correlated. In particular, when checking the residual plot, there were distinct, parallel lines with slopes of ~-1, seemingly based on the values associated with our average age adjusted death rates data. In total, there were 11 different possibilities of death rates and 11 distinct lines in the residuals plot. Thus, after making this realization, Linear Discriminant Analysis tests were run. These tests are useful in determining whether our independent variables are reasonable for predicting the dependent variable, especially in when distinct cases are present, such as the 11 we see in our data. However, LDA does a better job of estimating predictions for fewer cases, so for our second LDA, we designated three specific cases as described later in the paragraph. First, was a basic LDA test with all four independent variables, population density, unemployment rate, personal income, and labor participation. From this, we found that ~92% of explained "between-group" variance comes from the population density of a county, thus making it the variable that best explains variation in LDA. After running a standard LDA regression with all variables and 11 classes to predict, we found that the LDA coefficients classify the death rate of a county correctly ~23% of the time. This value is found by summing all values along the diagonal of the table (i.e. the values predicted correctly) and dividing by the total number of predictions. In addition, we ran an LDA test where each death rate level was classified into a group: low, medium, and high. The bottom four values were classified as "low", the middle three as "medium", and the final four as "high". From here, another LDA test was run where the variables were to predict which category, low, medium, or high, a county would fall in in terms of death rate. This resulted in a correct classification ~53% of the time. Thus, we conclude that our independent variables do suffice in order to accurately predict the death rates of each county as they correctly classify the death rates of a county quite often considering how many possibilities there are.

Table 1: LDA Prediction of Individual Death Rates

	1	3.05	5.05	7.05	9.05	11.05	13.05	15.05	17.05	19.05	20
1	15	8	8	9	7	7	9	5	4	4	6
3.05	3	2	4	3	3	0	0	0	0	0	0
5.05	0	0	0	0	0	0	0	0	0	0	0
7.05	0	0	0	0	0	0	0	0	0	0	0
9.05	12	34	45	63	65	47	38	30	28	11	18
11.05	1	3	2	6	17	26	15	17	11	7	11
13.05	2	1	17	43	63	70	61	40	26	32	34
15.05	0	0	0	0	5	1	3	8	0	2	5
17.05	0	0	0	0	0	0	0	0	1	0	0
19.05	0	0	0	0	0	0	0	0	0	0	0
20	4	2	24	114	184	240	281	275	200	184	539

```
sum(diag(tab))/sum(tab)
```

```
## [1] 0.2339315
```

```
#classifies correctly ~23%
```

Table 2: LDA Prediction of Grouped Death Rates

	high	low	medium
high	1309	171	836
low	22	70	36
medium	167	184	270

```
sum(diag(tab2))/sum(tab2)
```

```
## [1] 0.5380098
```

```
#classifies data correctly ~54%
```

Conclusions

Basing our conclusions on simple linear regression and multiple regression, if we were to ignore the correlation between residuals, we would be able to conclude that population density does in fact have a strong correlation with predicting the death rates due to heroin of each county. However, these results indicated that as population density grew for a county, so did the average rates of death. Therefore, if we relied solely on this data, it could be claimed that more dense, urban areas are in fact the problem areas for the heroin epidemic, ultimately rejecting our original hypothesis that rural regions were the main problem areas. However, once other variables were added to our model, specifically unemployment, population density did very little to determine these death rates. Therefore, it cannot be justifiably claimed that population density explains an increase or decrease in age adjusted heroin death rates based off of linear regression methods.

One limitation of running the initial regressions, however, was the fact that some of the explanatory variables seemed to be correlated to the other. This could be due to the fact that each can be explained or calculated by the total population of the county. Thus, finding less correlated variables could turn up new results and provide a stabilizing factor during simple linear regression.

Furthermore, when performing Forward Stepwise Selection, it was discovered that unemployment rate seems to be more predictive than population density when death rates are regressed on either. This could be because there is no true epicenter of the heroin epidemic in terms of rural vs. urban locale but there is evidence to support the claim that being out of work influences drug abuse. One possible explanation could be that those who are unemployed turn to substance abuse to nurse the depression of being out of work, and the fact that they are out of work means they must look for less expensive substances to consume.

Moreover, based on our more robust findings from LDA testing, which hampers the effect of having a discrete variable as our dependent variable, due to the independent variables' high percentage of correct classifications, we can determine that these variables do in fact predict heroin death rates in a given county. In addition, since much of the variability is due to the population of a given county, we can assume that population density is in fact a good predictor of heroin death rates.

Therefore, based off of our findings, it is not necessarily true that rural areas, represented by population density, are more likely to have heroin related deaths. Thus we reject our null hypothesis and claim that further research must be done in order to determine what factors to look for when deciding which areas will more likely be effected by heroin use.

Bibliograph/Works Cited

The Atlantic. Atlantic Media Company, n.d. Web. 07 Dec. 2016.

"Bureau of Economic Analysis." US Department of Commerce, BEA, Bureau of Economic Analysis. N.p., n.d. Web. 07 Dec. 2016.

Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, 30 Mar. 2016. Web. 07 Dec. 2016.

"Drug Overdose Deaths*." County Health Rankings & Roadmaps. N.p., n.d. Web. 07 Dec. 2016.

"Drug-poisoning Deaths Involving Heroin: United States, 2000-2013." Centers for Disease Control and Prevention. Centers for Disease Control and Prevention, 04 Mar. 2015. Web. 07 Dec. 2016.

"Local Area Unemployment Statistics Home Page." U.S. Bureau of Labor Statistics. U.S. Bureau of Labor Statistics, n.d. Web. 07 Dec. 2016.

Secretary, HHS Office of the. "Opioids: The Prescription Drug & Heroin Overdose Epidemic." HHS.gov. N.p., 24 Mar. 2016. Web. 07 Dec. 2016.