Opioid Death and Prescriptions

Miguel Briones

December 14, 2017

Overview

We will analyze the relationship between death in the U.S. resulting from opioid overdose with the number of opioid prescriptions written over the span of 15 years.

First, we will load the data from a publically available dataset found on data.world.

Which state has the most overall deaths?

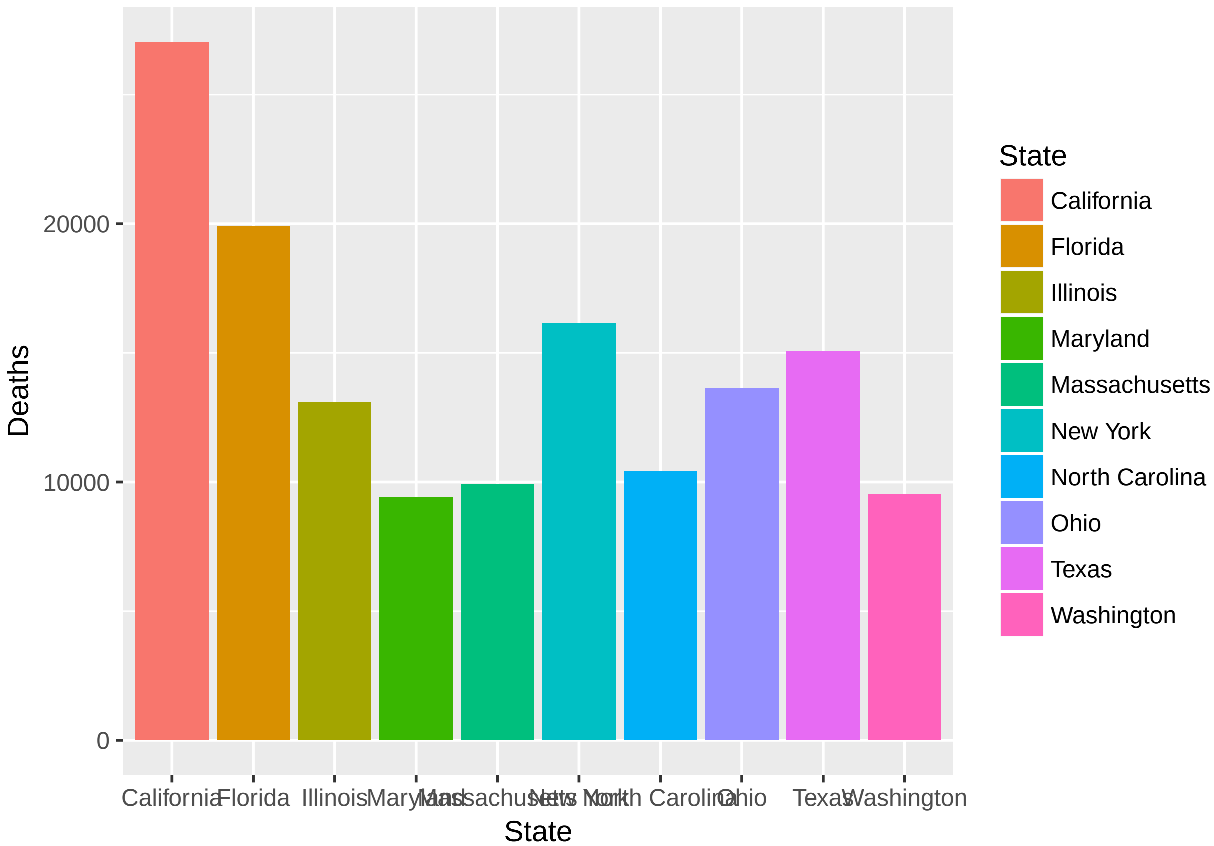
We first want to ask which state had to most overall deaths due to opioid overdosing.

```
StateDeath <- data_frame(OpioidDeath$State, OpioidDeath$Deaths)
colnames(StateDeath) <- c("State", "Deaths")
StateDeath <- subset(StateDeath, Deaths != "Suppressed")
StateDeath$Deaths <- as.numeric(as.character(StateDeath$Deaths))
StateDeath <- aggregate(. ~  State, data = StateDeath, sum)
StateDeath <- arrange(StateDeath, desc(Deaths))
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.2
```

```
StateDeath.Top <- StateDeath[1:10,]

ggplot(StateDeath.Top, aes(x=State, y=Deaths, fill = State)) + geom_bar(stat = "Identity")
```




```
StateDeath.Top
```

```
##             State Deaths
## 1      California  27044
## 2         Florida  19919
## 3        New York  16156
## 4           Texas  15050
## 5            Ohio  13623
## 6        Illinois  13072
## 7  North Carolina  10413
## 8   Massachusetts   9923
## 9      Washington   9528
## 10       Maryland   9403
```

We can see that the state of California had the most deaths across the 15 years of available data with 27,044 deaths.

What year was most fatal for California in relation to that years population size?

The next question we can ask is what year was the most fatal for California in relation to it's population size that given year?

```
CaliDeath <- data_frame(OpioidDeath$State, OpioidDeath$Year, OpioidDeath$Deaths, OpioidDeath$Population)
colnames(CaliDeath) <- c("State", "Year", "Deaths", "Population")
CaliDeath <- subset(CaliDeath, Deaths != "Suppressed" & State == "California")
```

```r
CaliDeath$Deaths <- as.numeric(as.character(CaliDeath$Deaths))
CaliDeath <- arrange(CaliDeath, desc(Deaths))
CaliDeath[,2:3]
```

```
## # A tibble: 16 x 2
##     Year Deaths
##    <int>  <dbl>
##  1  2014   2159
##  2  2009   2128
##  3  2013   2088
##  4  2010   2059
##  5  2011   2057
##  6  2008   1889
##  7  2012   1847
##  8  2007   1762
##  9  2006   1602
## 10  1999   1598
## 11  2002   1583
## 12  2004   1547
## 13  2003   1530
## 14  2005   1485
## 15  2000   1105
## 16  2001    605
```

We can see that before population normalization, 2014 was the most fatal year with 2,159 deaths.

```r
CaliDeath$Porportion <- CaliDeath$Deaths/CaliDeath$Population *exp(10)
CaliDeath <- arrange(CaliDeath, desc(Porportion))
CaliDeath[,c("Year", "Porportion")]
```

```
## # A tibble: 16 x 2
##     Year Porportion
##    <int>      <dbl>
##  1  2009  1.2681483
##  2  2014  1.2255690
##  3  2010  1.2173873
##  4  2011  1.2020733
##  5  2013  1.1997974
##  6  2008  1.1366957
##  7  2007  1.0706290
##  8  2012  1.0694362
##  9  1999  1.0507203
## 10  2002  0.9998868
## 11  2006  0.9796008
## 12  2004  0.9578454
## 13  2003  0.9559567
## 14  2005  0.9129551
## 15  2000  0.7185728
## 16  2001  0.3864913
```

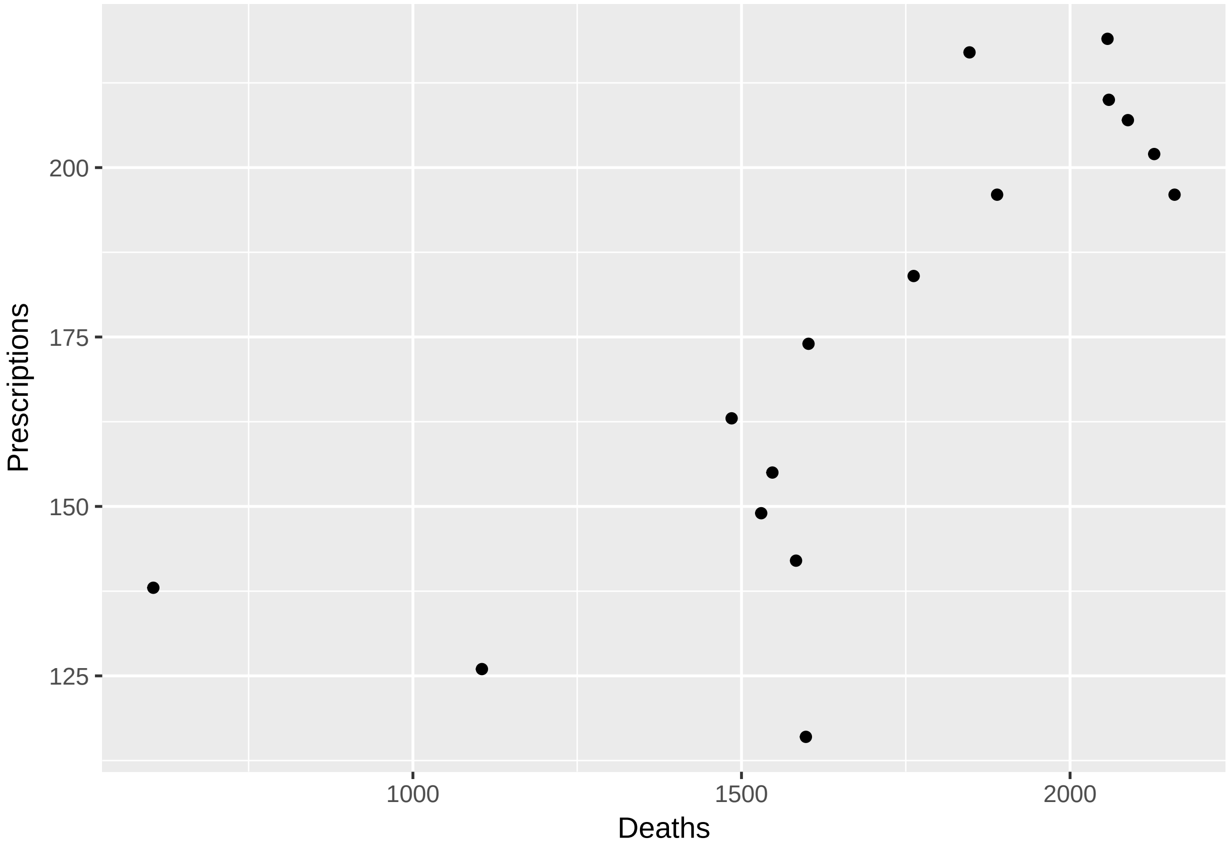
When normalized to the size of the population, 2009 was the most fatal year.

Is there a relationship between perscriptions dispensed and number of deaths for the state of California?

The next question we wanted to ask was whether there was a relationship between the number of deaths per year and the number of prescriptions for opioids filled per year.

```
CaliPrescrip <- data_frame(OpioidDeath$State, OpioidDeath$Year, OpioidDeath$Deaths, OpioidDeath$Prescri
colnames(CaliPrescrip) <- c("State", "Year", "Deaths", "Prescriptions")
CaliPrescrip <- subset(CaliPrescrip, Deaths != "Suppressed" & State == "California")
CaliPrescrip$Deaths <- as.numeric(as.character(CaliPrescrip$Deaths))

ggplot(CaliPrescrip, aes(x = Deaths, y = Prescriptions)) + geom_point()
```




```
CaliPrescripRelation <-lm(Prescriptions~Deaths, data=CaliPrescrip)
summary(CaliPrescripRelation)
```

```
##
## Call:
## lm(formula = Prescriptions ~ Deaths, data = CaliPrescrip)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.746 -10.704   3.129   9.500  32.537
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 66.90685   24.09571   2.777 0.014845 *
## Deaths       0.06373    0.01387   4.593 0.000418 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.14 on 14 degrees of freedom
## Multiple R-squared:  0.6011, Adjusted R-squared:  0.5726
## F-statistic:  21.1 on 1 and 14 DF,  p-value: 0.0004178
```

We see that our intercept is 66.9 and our slope is 0.06. We can intepret it as California having an average increase of 67 deaths per year over the course of 15 years. The slope indicates that every 1 increase in death is equal to 1/6th of a prescription. Our R squared is 60%, indicating that roughly 60% of the variance in deaths can be explained by the number ofprescriptions. Our F-statstic is 21.1, with a p-value of >0.01, indicating that there is a strong relationship Prescriptions and Number of Deaths in the state of California. However, it can be stronger.

Next, we wanted to a correlation test between number of deaths and number of prescriptions filled.

```
cor.test(x=CaliPrescrip$Deaths, y=CaliPrescrip$Prescriptions, method = 'pearson')
```

```
##
##  Pearson's product-moment correlation
##
## data:  CaliPrescrip$Deaths and CaliPrescrip$Prescriptions
## t = 4.5932, df = 14, p-value = 0.0004178
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4541534 0.9181498
## sample estimates:
##      cor
## 0.775315
```

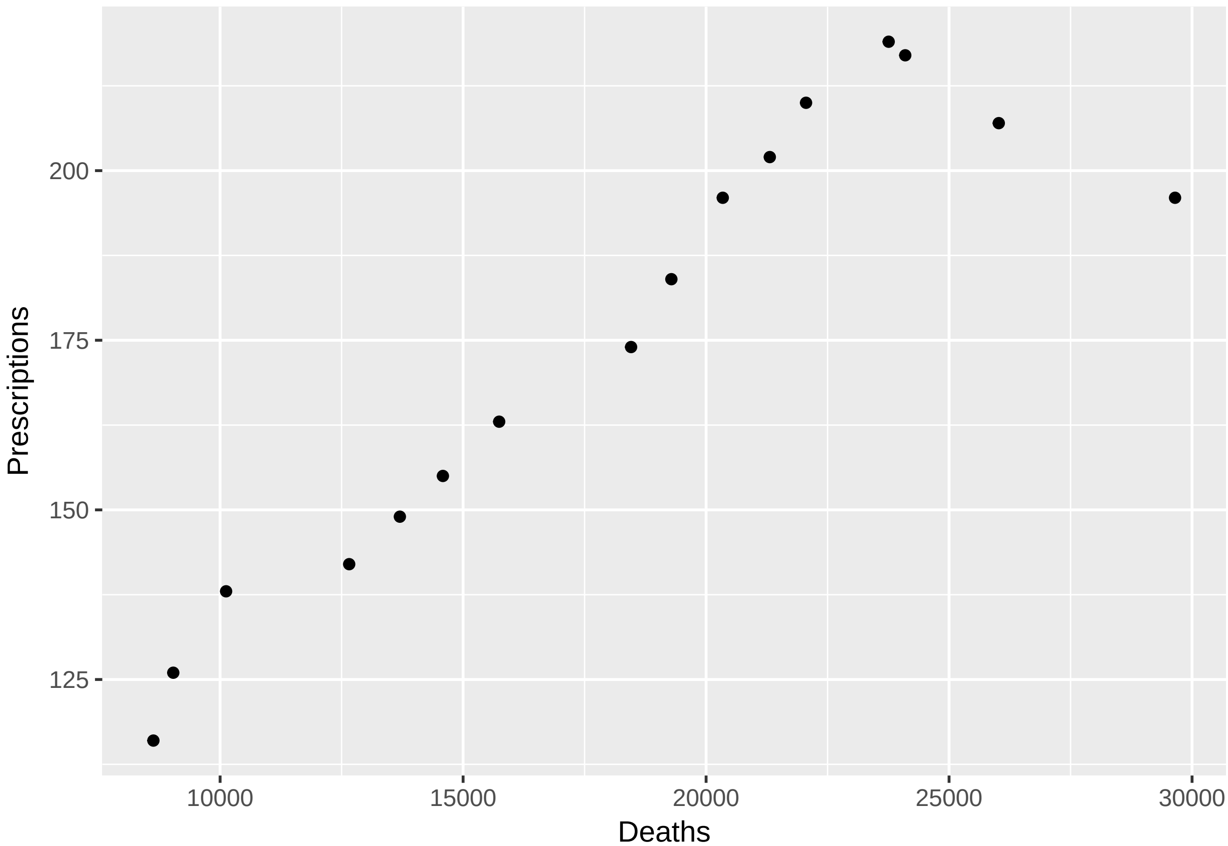
We see that our t value is equal to 4.5932 and our r is equal to 0.775315. Therefore, there is a strong, positive correlation between number of deaths and number of prescriptions given in California. Our p-value is >0.01, indicating that our correlation is statistically significant. We can reject the null hypothesis of no correlation between the two variables.

Is there a relationship between perscriptions dispensed and number of deaths nationwide between 1999-2014?

Next, we wanted to scale back our analysis and ask if there was a relationship between the number of perscriptions given and the number of deaths due to opioids across the U.S.

```
UsPrescripYear <- data_frame(OpioidDeath$Year, OpioidDeath$Prescriptions.Dispensed.by.US.Retailers.in.th
colnames(UsPrescripYear) <- c("Year", "Prescriptions")
UsPrescripYear <- UsPrescripYear[1:16,]
UsYearDeaths <- data_frame(OpioidDeath$Year, OpioidDeath$Deaths)
colnames(UsYearDeaths) <- c("Year", "Deaths")
UsYearDeaths <- subset(UsYearDeaths, Deaths != "Suppressed")
UsYearDeaths$Deaths <- as.numeric(as.character(UsYearDeaths$Deaths))
UsYearDeaths <- aggregate(. ~  Year, data = UsYearDeaths, sum)
```

```r
UsPrescripDeath <- merge(UsPrescripYear, UsYearDeaths, by = 'Year')
ggplot(UsPrescripDeath, aes(x = Deaths, y = Prescriptions)) + geom_point()
```




```r
UsPrescripDeathRelation <-lm(Prescriptions~Deaths, data=UsPrescripDeath)
summary(UsPrescripDeathRelation)
```

```
##
## Call:
## lm(formula = Prescriptions ~ Deaths, data = UsPrescripDeath)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.933  -4.310  -1.111  10.508  16.283
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.494e+01  1.037e+01   8.190 1.04e-06 ***
## Deaths      4.958e-03  5.431e-04   9.129 2.85e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.29 on 14 degrees of freedom
## Multiple R-squared:  0.8562, Adjusted R-squared:  0.8459
## F-statistic: 83.34 on 1 and 14 DF,  p-value: 2.852e-07
```

We find that our intercept is 84.94 and our slope is 0.0049. We can intepret it as the US having an average increase of 85 deaths per year over the course of 15 years.The slope indicates that for every 1 increase in death is equal to 1/400th of a prescription. Our r squared is 86%, indicating that roughly 86% of the variance in deaths can be explained by the number ofprescriptions. Our F-statstic is 83.34, with a p-value of >0.01, indicating that there is a strong relationship Prescriptions and Number of Deaths in the US.

Next, we wanted to run a correlation between the number of deaths and the number of prescriptions filled nationwide.

```
cor.test(x=UsPrescripDeath$Deaths, y=UsPrescripDeath$Prescriptions, method = 'pearson')
```

```
##
##  Pearson's product-moment correlation
##
## data:  UsPrescripDeath$Deaths and UsPrescripDeath$Prescriptions
## t = 9.1293, df = 14, p-value = 2.852e-07
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7935987 0.9741745
## sample estimates:
##       cor
## 0.9252994
```

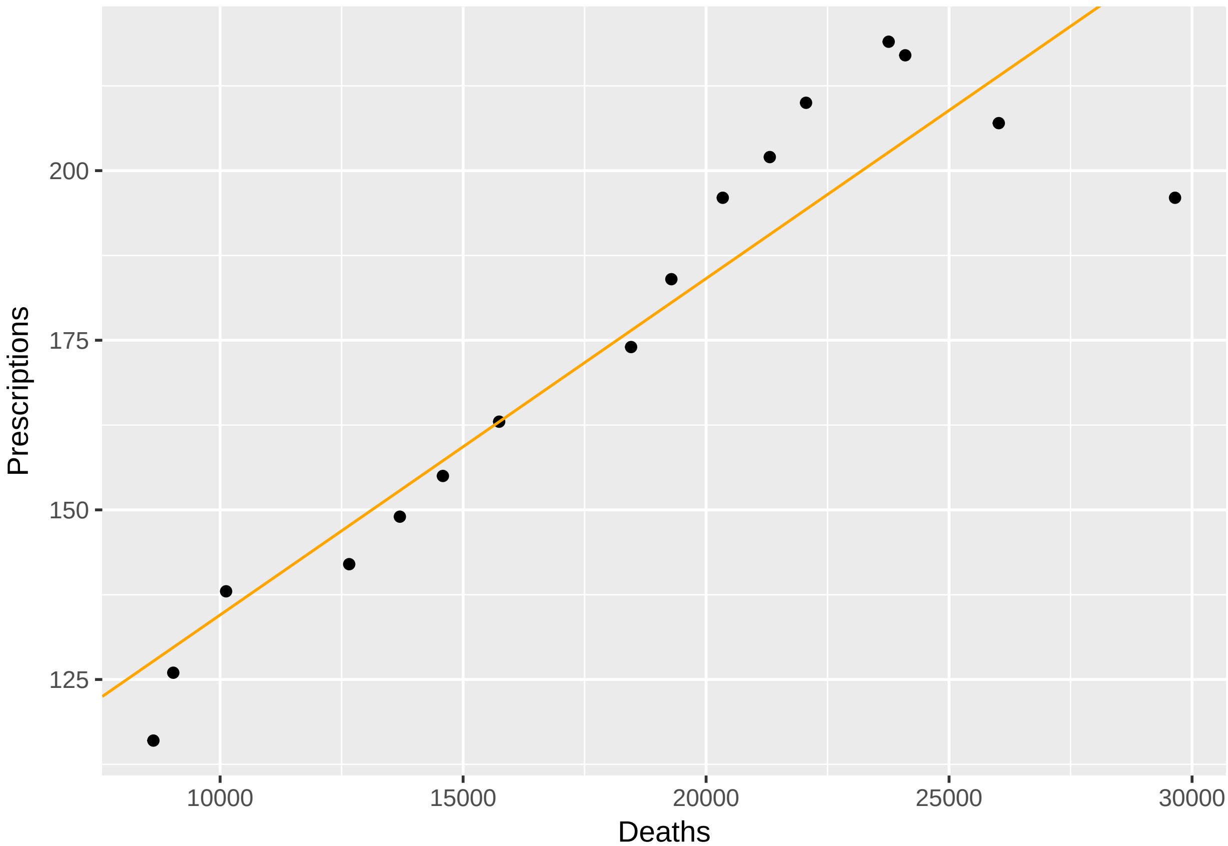
We find that our t value is equal to 9.1293 and our r is equal to 0.9252994, indicating that there is a strong, positive correlation between the number of deaths and the number of prescriptions given in the US. Our p-value is >0.01, indicating that our correlation is statistically significant. We can reject the null hypothesis of no correlation between the two variables.

Using our lm model for US deaths related to prescriptions, can we predict furture deaths due to prescriptions?

```
USPrescripModel <- train(Prescriptions ~ Deaths, data = UsPrescripDeath, method = "lm")

USPrescripcoef.icept <- coef(USPrescripModel$finalModel)[1]
USPrescripcoef.slope <- coef(USPrescripModel$finalModel)[2]

ggplot(data = UsPrescripDeath, aes(x = Deaths, y = Prescriptions)) +
  geom_point() +
  geom_abline(slope = USPrescripcoef.slope, intercept = USPrescripcoef.icept, color = "orange")
```



Although limited in it's sample size, we can predict that as prescriptions increase across the U.S., so will deaths. One of the interesting takeaways is that more factors may be necessary to investigate, such as income level of state and age, to make a more robust predictive model.