Final MD

Matt Burns

February 27, 2019

# Load libraries & data

library(readxl)  
library(ggplot2)

tax <- read\_excel("C:/Users/burns/OneDrive/Desktop/Matt/Grad School/DSC 520/Final Project/tax.xlsx")  
names(tax)[1]<-"State"  
names(tax)[33]<-"pop"  
tax <- tax[-c(9,52), ]   
str(tax)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 50 obs. of 33 variables:  
## $ State : chr "Alabama" "Alaska" "Arizona" "Arkansas" ...  
## $ Total Taxes : num 1.00e+07 8.97e+05 1.37e+07 9.43e+06 1.55e+08 ...  
## $ Property Taxes : num 352378 111736 943008 1097908 2513157 ...  
## $ Sales and Gross Receipts Taxes : num 5106102 260846 8315805 4590072 51602180 ...  
## $ General Sales and Gross Receipts Taxes : num 2596223 0 6300443 3314363 39189007 ...  
## $ Selective Sales and Gross Receipts Taxes : num 2509879 260846 2015362 1275709 12413173 ...  
## $ Alcoholic Beverages Sales Tax : num 210535 42430 72281 55164 368699 ...  
## $ Amusements Sales Tax : num 6 10306 0 56031 0 ...  
## $ Insurance Premiums Sales Tax : num 317657 64400 545124 192020 2561932 ...  
## $ Motor Fuels Sales Tax : num 526763 48773 898234 479879 5000539 ...  
## $ Pari-mutuels Sales Tax : num 1401 0 160 2616 14537 ...  
## $ Public Utilities Sales Tax : num 695626 4027 22337 0 714623 ...  
## $ Tobacco Products Sales Tax : num 180301 67918 317331 230527 840034 ...  
## $ Other Selective Sales and Gross Receipts Taxes: num 577590 22992 159895 259472 2912809 ...  
## $ License Taxes : num 507479 120529 482362 396891 10275132 ...  
## $ Alcoholic Beverages License : num 4224 1919 7416 4624 57406 ...  
## $ Amusements License : num 0 0 0 473 16767 ...  
## $ Corporations in General License : num 162117 0 18342 26703 75066 ...  
## $ Hunting and Fishing License : num 22931 29500 35059 26579 104698 ...  
## $ Motor Vehicle License : num 213550 38000 228970 163023 3996089 ...  
## $ Motor Vehicle Operators License : num 33964 0 31373 21825 296160 ...  
## $ Public Utilities License : num 14443 838 0 8351 674660 ...  
## $ Occupation and Business License, NEC : num 56250 46957 159454 143422 5027281 ...  
## $ Other License Taxes : num 0 3315 1748 1891 27005 ...  
## $ Income Taxes : num 3966640 67457 3906722 3231617 90655530 ...  
## $ Individual Income Taxes : num 3492904 0 3336174 2781458 80753345 ...  
## $ Corporations Net Income Taxes : num 473736 67457 570548 450159 9902185 ...  
## $ Other Taxes : num 90352 336801 32524 114345 145715 ...  
## $ Death and Gift Taxes : num 0 0 0 3 330 ...  
## $ Documentarty and Stock Transfer Taxes : num 43730 0 17328 38844 0 ...  
## $ Severance Taxes : num 46622 336801 15196 48340 68500 ...  
## $ Taxes, NEC : num 0 0 0 27158 76885 ...  
## $ pop : num 3766477 554567 5299579 2283195 30157154 ...

infra <- read\_excel("C:/Users/burns/OneDrive/Desktop/Matt/Grad School/DSC 520/Final Project/Infrastructure.xlsx")  
infra$A <- NULL  
infra$B <- NULL  
infra$C <- NULL  
colnames(infra) <- c("State","BadRoads","SpendPerDriver")  
infra[order(infra$State),]

## # A tibble: 50 x 3  
## State BadRoads SpendPerDriver  
## <chr> <dbl> <dbl>  
## 1 Alabama 0.018 405  
## 2 Alaska 0.189 2374  
## 3 Arizona 0.028 239  
## 4 Arkansas 0.102 495  
## 5 California 0.169 269  
## 6 Colorado 0.068 305  
## 7 Connecticut 0.056 446  
## 8 Delaware 0.039 671  
## 9 Florida 0.013 457  
## 10 Georgia 0.019 254  
## # ... with 40 more rows

str(infra)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 50 obs. of 3 variables:  
## $ State : chr "Alabama" "Alaska" "Arizona" "Arkansas" ...  
## $ BadRoads : num 0.018 0.189 0.028 0.102 0.169 0.068 0.056 0.039 0.013 0.019 ...  
## $ SpendPerDriver: num 405 2374 239 495 269 ...

Edu <- read\_excel("C:/Users/burns/OneDrive/Desktop/Matt/Grad School/DSC 520/Final Project/Education.xlsx")  
Edu <- Edu[-c(9,40,46), ]   
colnames(Edu) <- c("State","HSRate","HSRank","BachRate","BachRank","AdvRate","AdvRank")  
str(Edu)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 50 obs. of 7 variables:  
## $ State : chr "Alabama" "Alaska" "Arizona" "Arkansas" ...  
## $ HSRate : num 0.843 0.921 0.86 0.848 0.818 0.907 0.899 0.884 0.869 0.854 ...  
## $ HSRank : num 43 5 34 42 50 14 18 26 31 39 ...  
## $ BachRate: num 0.235 0.28 0.275 0.211 0.314 0.381 0.376 0.3 0.273 0.288 ...  
## $ BachRank: num 44 26 30 48 14 2 4 19 31 23 ...  
## $ AdvRate : num 0.087 0.101 0.102 0.075 0.116 0.14 0.166 0.122 0.098 0.107 ...  
## $ AdvRank : num 38 27 25 49 14 7 3 12 29 20 ...

IMR <- read\_excel("C:/Users/burns/OneDrive/Desktop/Matt/Grad School/DSC 520/Final Project/INFANT\_MORTALITY\_RATES2017.xlsx")  
IMR$URL<- NULL  
str(IMR)

## Classes 'tbl\_df', 'tbl' and 'data.frame': 50 obs. of 3 variables:  
## $ STATE : chr "AL" "AK" "AZ" "AR" ...  
## $ RATE : num 7.4 5.6 5.7 8.2 4.2 4.5 4.5 6.6 6.1 7.2 ...  
## $ DEATHS: num 434 59 469 307 1973 ...

The data I pulled above came from the internet as detailed in my project summary. It is worth noting that Puerto Rico and D.C. were intentionally knocked out. Also, the dataframes will be manipulated as necessary throughout the project.

Now that the raw data is in R Studio, I need to make sure the rows (i.e. state align).

# Check that all dataframes have the states in the same order

states <- data.frame(t=tax$State,i=infra$State,e=Edu$State,h=IMR$STATE)  
print(states)

## t i e h  
## 1 Alabama Alabama Alabama AL  
## 2 Alaska Alaska Alaska AK  
## 3 Arizona Arizona Arizona AZ  
## 4 Arkansas Arkansas Arkansas AR  
## 5 California California California CA  
## 6 Colorado Colorado Colorado CO  
## 7 Connecticut Connecticut Connecticut CT  
## 8 Delaware Delaware Delaware DE  
## 9 Florida Florida Florida FL  
## 10 Georgia Georgia Georgia GA  
## 11 Hawaii Hawaii Hawaii HI  
## 12 Idaho Idaho Idaho ID  
## 13 Illinois Illinois Illinois IL  
## 14 Indiana Indiana Indiana IN  
## 15 Iowa Iowa Iowa IA  
## 16 Kansas Kansas Kansas KS  
## 17 Kentucky Kentucky Kentucky KY  
## 18 Louisiana Louisiana Louisiana LA  
## 19 Maine Maine Maine ME  
## 20 Maryland Maryland Maryland MD  
## 21 Massachusetts Massachusetts Massachusetts MA  
## 22 Michigan Michigan Michigan MI  
## 23 Minnesota Minnesota Minnesota MN  
## 24 Mississippi Mississippi Mississippi MS  
## 25 Missouri Missouri Missouri MO  
## 26 Montana Montana Montana MT  
## 27 Nebraska Nebraska Nebraska NE  
## 28 Nevada Nevada Nevada NV  
## 29 New Hampshire New Hampshire New Hampshire NH  
## 30 New Jersey New Jersey New Jersey NJ  
## 31 New Mexico New Mexico New Mexico NM  
## 32 New York New York New York NY  
## 33 North Carolina North Carolina North Carolina NC  
## 34 North Dakota North Dakota North Dakota ND  
## 35 Ohio Ohio Ohio OH  
## 36 Oklahoma Oklahoma Oklahoma OK  
## 37 Oregon Oregon Oregon OR  
## 38 Pennsylvania Pennsylvania Pennsylvania PA  
## 39 Rhode Island Rhode Island Rhode Island RI  
## 40 South Carolina South Carolina South Carolina SC  
## 41 South Dakota South Dakota South Dakota SD  
## 42 Tennessee Tennessee Tennessee TN  
## 43 Texas Texas Texas TX  
## 44 Utah Utah Utah UT  
## 45 Vermont Vermont Vermont VT  
## 46 Virginia Virginia Virginia VA  
## 47 Washington Washington Washington WA  
## 48 West Virginia West Virginia West Virginia WV  
## 49 Wisconsin Wisconsin Wisconsin WI  
## 50 Wyoming Wyoming Wyoming WY

# 

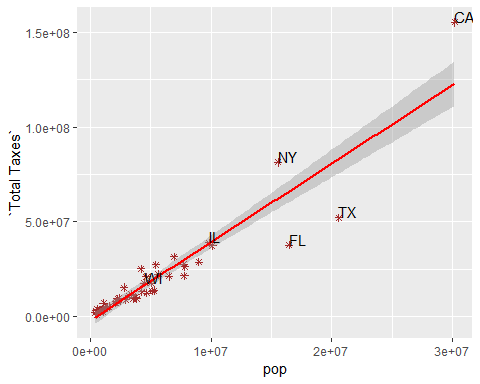
Before I make graphs, I need to add labels for state I’m specifically interested in (e.g. Illinois and Wisconsin) and also large states that don’t fall on the linear model.

# Prep labels

tax$StLab <- " "  
tax [5,34] <- "CA"  
tax [32,34] <- "NY"  
tax [43,34] <- "TX"  
tax [9,34] <- "FL"  
tax [13,34] <- "IL"  
tax [49,34] <- "WI"

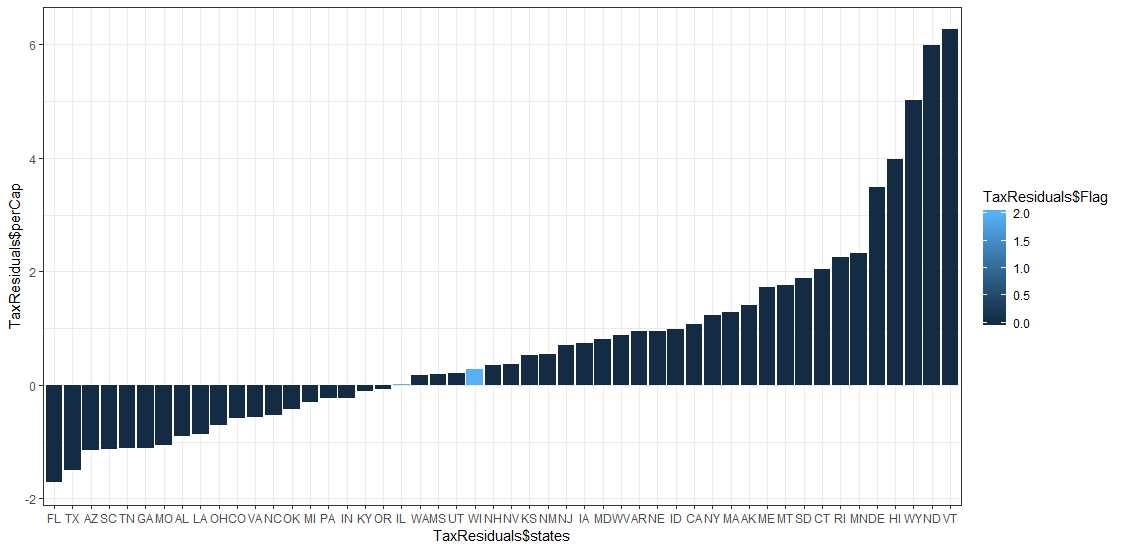
# Plot taxes & residuals for Total Taxes

ggplot(tax,aes(x=pop, y=`Total Taxes`)) + geom\_point(shape=8, color="brown") +stat\_smooth(method = "lm", col = "red") +  
 geom\_text(aes(label=StLab),hjust=0, vjust=0)



Illinois and Wisconsin are right on the linear model. That suggests they have very typical taxes for states their size. Both Florida and Texas have relatively low taxes, while NY and California are higher.

basic <- lm(tax$`Total Taxes` ~ tax$pop,tax)  
TaxResiduals <-data.frame(states=states$h,res=basic$residuals,pop=tax$pop)  
TaxResiduals$perCap <- TaxResiduals$res/TaxResiduals$pop  
TaxResiduals$Flag <- 0  
TaxResiduals [13,5] <- 2  
TaxResiduals [49,5] <- 2  
TaxResiduals$states <- factor(TaxResiduals$states, levels = TaxResiduals$states[order(TaxResiduals$perCap)])  
ggplot(TaxResiduals, aes(x = TaxResiduals$states, y = TaxResiduals$perCap, fill = TaxResiduals$Flag)) + theme\_bw() + geom\_bar(stat = "identity")



When we review residuals per capita, we see that Illinois and Wisconsin are near the middle suggestion their population pays typical taxes.

The following 5 state have the most favorable residuals per capita:

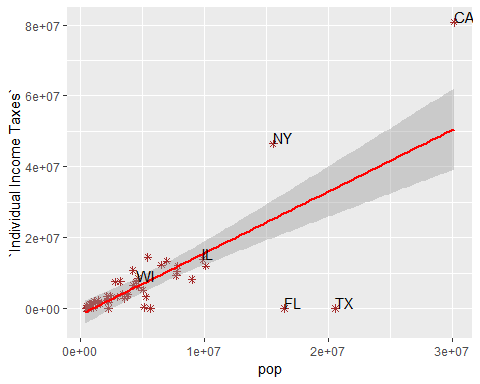
* Florida
* Texas
* Arizona
* South Carolina
* Tennessee

The following 5 state have the least favorable residuals per capita:

* Vermont
* North Dakoda
* Wyoming
* Hawaii
* Delaware

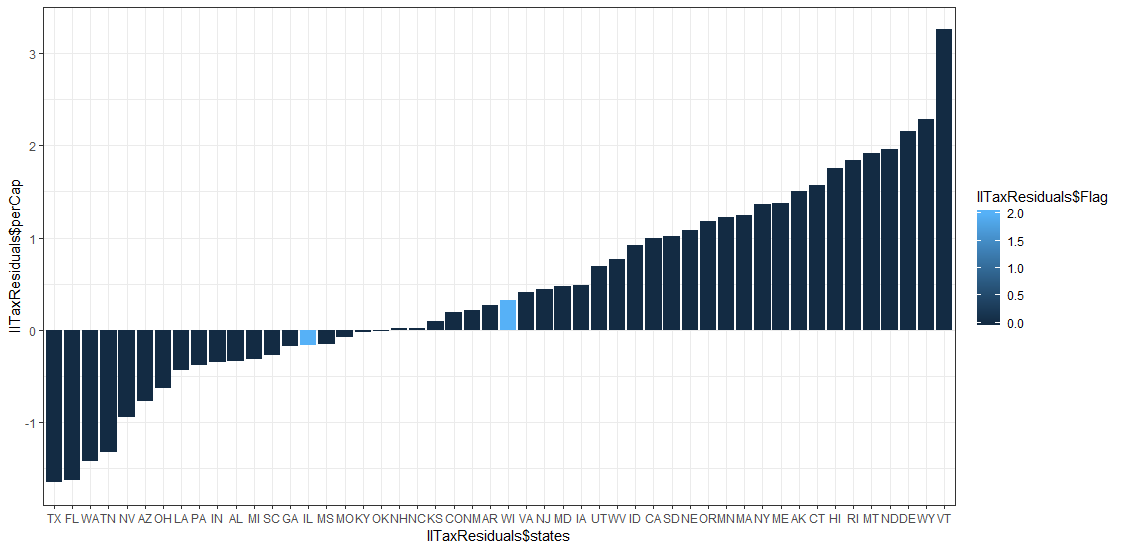
# Plot taxes & residuals for Income Taxes

ggplot(tax,aes(x=pop, y=`Individual Income Taxes`)) + geom\_point(shape=8, color="brown") +stat\_smooth(method = "lm", col = "red") +  
 geom\_text(aes(label=StLab),hjust=0, vjust=0)



The results are similar to the total taxes. It is also worth noting that Florida and Texas do not have income taxes.

IIT <- lm(tax$`Individual Income Taxes` ~ tax$pop,tax)  
IITaxResiduals <-data.frame(states=states$h,res=IIT$residuals,pop=tax$pop)  
IITaxResiduals$perCap <- IITaxResiduals$res/IITaxResiduals$pop  
IITaxResiduals$Flag <- 0  
IITaxResiduals [13,5] <- 2  
IITaxResiduals [49,5] <- 2  
IITaxResiduals$states <- factor(IITaxResiduals$states, levels = IITaxResiduals$states[order(IITaxResiduals$perCap)])  
ggplot(IITaxResiduals, aes(x = IITaxResiduals$states, y = IITaxResiduals$perCap, fill = IITaxResiduals$Flag)) + theme\_bw() + geom\_bar(stat = "identity")



When we review residuals per capita, we see that Illinois and Wisconsin are near the middle suggestion their population pays typical income taxes.

The following 5 state have the most favorable residuals per capita:

* Texas
* Florida
* Washington
* Tennessee
* Nevada

The following 5 state have the least favorable residuals per capita:

* Vermont
* Wyoming
* Delaware
* North Dakoda
* Montana

# Check Education Impacts

Edu$pop <- tax$pop  
Edu$tax <- tax$'Total Taxes'  
Education <- lm(BachRate ~ tax + pop,Edu)  
summary(Education)

##   
## Call:  
## lm(formula = BachRate ~ tax + pop, data = Edu)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.095770 -0.027637 -0.005449 0.024863 0.104054   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.870e-01 9.301e-03 30.860 <2e-16 \*\*\*  
## tax 1.372e-09 7.825e-10 1.753 0.0861 .   
## pop -4.327e-09 3.467e-09 -1.248 0.2182   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04805 on 47 degrees of freedom  
## Multiple R-squared: 0.08367, Adjusted R-squared: 0.04467   
## F-statistic: 2.146 on 2 and 47 DF, p-value: 0.1283

I built a simpler regression to check the impacts on education. Taxes and population weren’t that significant and had small impacts on education. What I can say is that controlling for population, taxes explain only 4.4% of the bachelor degree attainment. Plus we don’t know if there’s causality here. Did the higher education levels drive a higher tax base or did the higher taxes push educational achievement?

# Check Health Impacts

IMR$pop <- tax$pop  
IMR$tax <- tax$'Total Taxes'  
Health <- lm(DEATHS ~ tax + pop,IMR)  
summary(Health)

##   
## Call:  
## lm(formula = DEATHS ~ tax + pop, data = IMR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -349.23 -36.53 -11.33 64.84 173.96   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.627e+01 1.770e+01 1.484 0.144   
## tax -1.271e-05 1.489e-06 -8.537 4.07e-11 \*\*\*  
## pop 1.312e-04 6.597e-06 19.893 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 91.43 on 47 degrees of freedom  
## Multiple R-squared: 0.9624, Adjusted R-squared: 0.9608   
## F-statistic: 601.2 on 2 and 47 DF, p-value: < 2.2e-16

I built a regression that looks at deaths during child birth and controls for population. The R2 is 96% which suggests that controlling for population, taxes can explain 96% of deaths during childbirth. This doesn’t pass the sniff test. When I get the opportunity, I’d like to dive deeper into this to see what else might be impacting these numbers.

# Check Infrastructure Impacts

infra$pop <- tax$pop  
infra$tax <- tax$'Total Taxes'  
Infrastructure <- lm(BadRoads ~ tax + SpendPerDriver + pop, infra)  
summary(Infrastructure)

##   
## Call:  
## lm(formula = BadRoads ~ tax + SpendPerDriver + pop, data = infra)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.06606 -0.02745 -0.01605 0.01334 0.18639   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.078e-02 1.627e-02 3.121 0.00311 \*\*  
## tax 2.305e-09 7.847e-10 2.938 0.00515 \*\*  
## SpendPerDriver 1.900e-05 1.925e-05 0.987 0.32866   
## pop -7.554e-09 3.526e-09 -2.142 0.03749 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04807 on 46 degrees of freedom  
## Multiple R-squared: 0.2036, Adjusted R-squared: 0.1517   
## F-statistic: 3.921 on 3 and 46 DF, p-value: 0.01419

I wanted to see if taxes had an impact on the percentage of poor roads. When controlling for population and traffic density (spend per driver), taxes only explain 15% of the incidence of bad roads. I’d like to obtain heating degree days as a proxy for winter severity. Maybe that would be tighten up this regression.