**Predicting State and Local Taxes based on Economy and Bureaucracy**

By DJ Cumberbatch and Graham I. Gordon

**Introduction:**

In the U.S., each state (and district) is responsible for imposing and collecting state and local taxes. The gross state and local tax per capita of a particular region of the U.S. is the total amount of state and local taxes a region collects divided by the number of people in each state in that region. Tax per capita for a given region incorporates both the amount of money a region produces, and the burden its population poses on the national economy in one statistic. As such, it is very useful in determining the economic security in said region. Taxes change regularly based on political whims, especially on the local levels. A model to help predict the amount of tax per capita a region is likely to generate would be very useful in making predictions regarding region prosperity and economic capabilities. This could also be indirectly considered a measure of a citizen's potential quality of life. The changes in economy and culture have been drastically altered by technological advances, and a pandemic. Because of the chaotic nature of the data we collected, we decided not to create a chronological model. Rather, we chose to fit a model based on economic success, quantity of government bureaucracy, and health insurance. More specifically, we used data from every state to construct a model adjusted for population which predicted the gross state and local tax collected by using the GDP, the population, the number of unemployed, the number of forms filed, the number of government workers, and the number of people who have health insurance.

**Methods:**

This data set was collected from six sources, mostly data sets built up by the federal government. [One data set](https://data.census.gov/table?q=employment%20by%20state%202021&g=010XX00US_040XX00US16)[[1]](#footnote-0) was part of the American Community Survey (ACS), and disseminated by the Census Bureau’s Population Estimates Program in 2021. The ACS is an annual demographics survey directed by the U.S. Census Bureau which collects the data samples from populations of greater than 65,000 people and extrapolates to create population estimates (making up only 26% of all counties)[[2]](#footnote-1). This should be suitable for our model because we are only looking at a singular year, so demographic shifts in those counties over time will not affect our model. This data set describes eight economic characteristics of individuals divided by state (which were then divided into many subtypes). This includes employment status, occupation, and health insurance coverage. From this, we pulled the number of people in each state who were unemployed, who had any kind of health insurance, and the number of government workers.

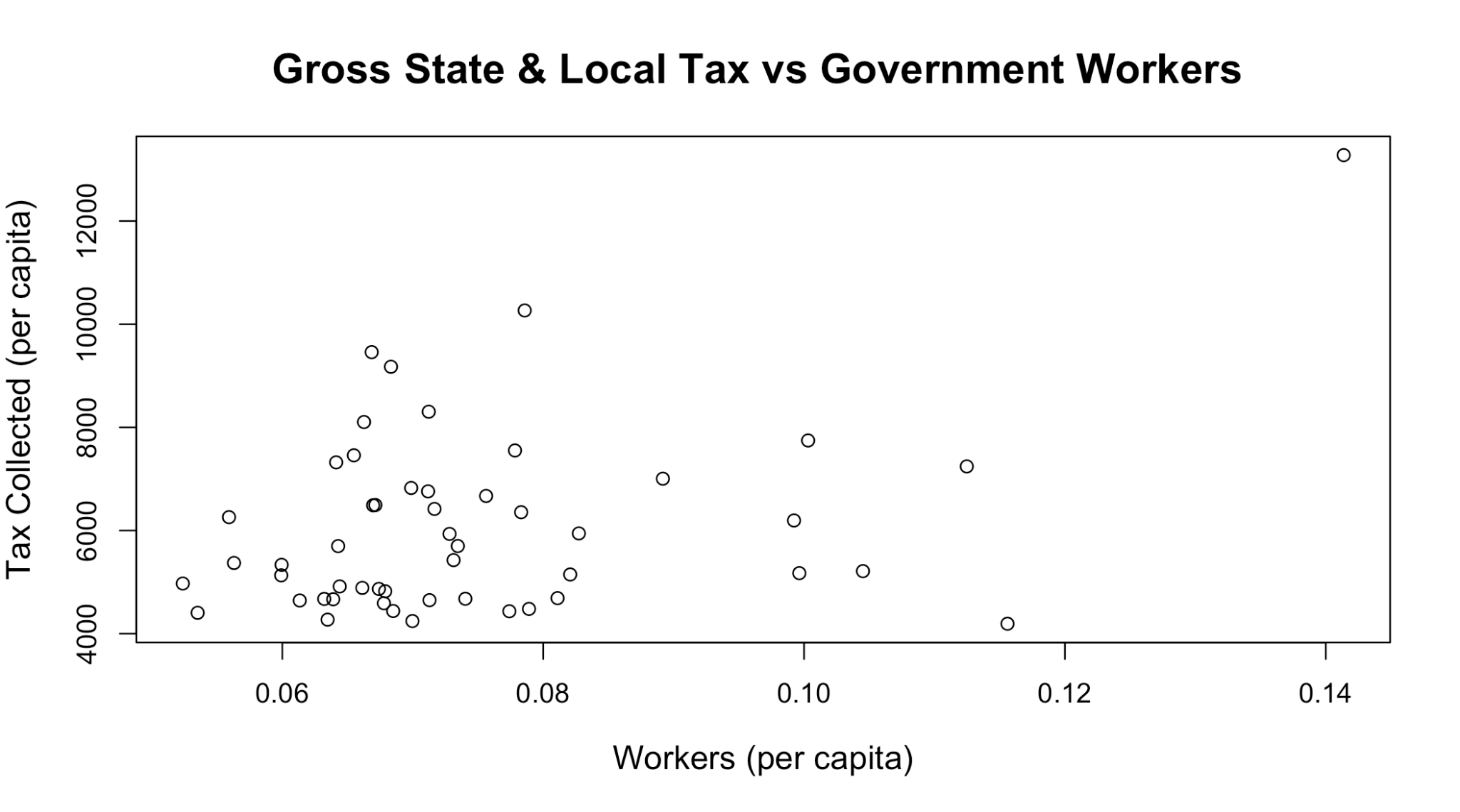
A second data set, also collected by the U.S. Census Bureau, this time through the Population Estimates Program. This program estimates the [population](https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html#v2022)[[3]](#footnote-2) based on the decennial census (last taken in 2020) and on measures of population change. Unfortunately the 2021 data was not available, so we had to use the 2022 estimates. The demographic shifts within a single year in each state is unlikely to be substantially different from those shifts in other states. As such it is reasonable to use this data in order to measure changes in other states (because the results should be marginally, and proportionally similar to a model using 2021 data). A third dataset contained the [tax per capita by state](https://taxfoundation.org/data/all/state/state-local-tax-collections-per-capita-fy-2021/) which was calculated by the Tax foundation with information obtained by the Local Government Finance Survey which is run by the U.S. Census Bureau. The fourth data set was the regions and divisions of the United States the Census Bureau uses.

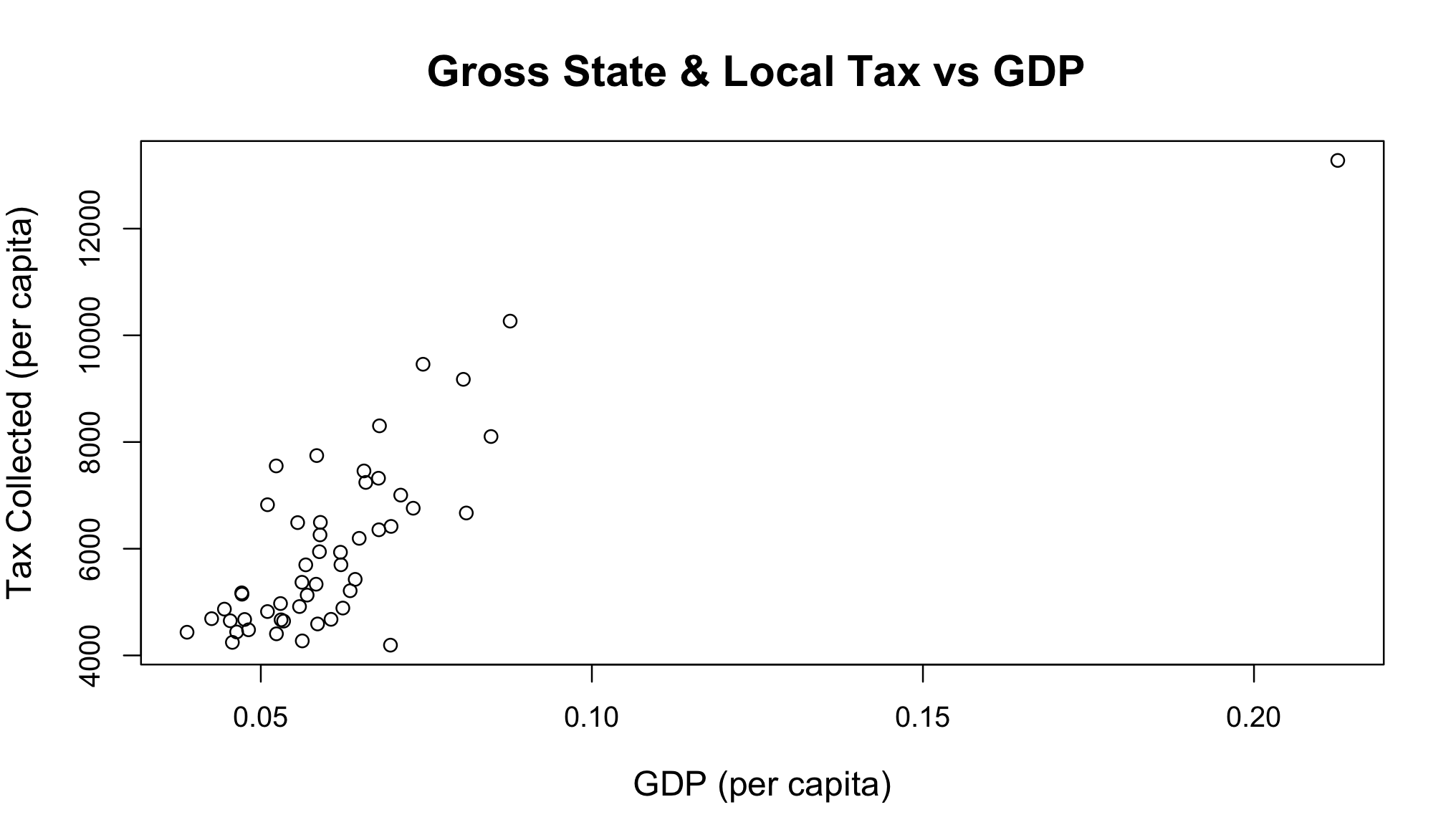
Another data set, this one collected by the IRS in the year 2021, is a record of the [number of forms filed](https://www.irs.gov/pub/irs-soi/21dbs01t03nr.xlsx) by type and state.[[4]](#footnote-3) From it we selected only the total number of forms filed per state as a possible measure of bureaucratic complexity. The sixth data set gave us the [real GDP](https://apps.bea.gov/itable/?ReqID=70&step=1#eyJhcHBpZCI6NzAsInN0ZXBzIjpbMSwyOSwyNSwzMSwyNiwyNywzMCwzMF0sImRhdGEiOltbIlRhYmxlSWQiLCI1MzEiXSxbIk1ham9yX0FyZWEiLCIwIl0sWyJTdGF0ZSIsWyIwIl1dLFsiQXJlYSIsWyJYWCJdXSxbIlN0YXRpc3RpYyIsIjEiXSxbIlVuaXRfb2ZfbWVhc3VyZSIsIkxldmVscyJdLFsiWWVhciIsWyIyMDIxIl1dLFsiWWVhckJlZ2luIiwiLTEiXSxbIlllYXJfRW5kIiwiLTEiXV19) of each state in millions of chained 2017 dollars.[[5]](#footnote-4) The GDP being locked either to 2017 dollars is a little concerning considering that the rest of our tax data is in 2021 dollars. However, our primary use of the GDP per capita per state is centered around the proportional difference between states (where each State is described by the other variables).

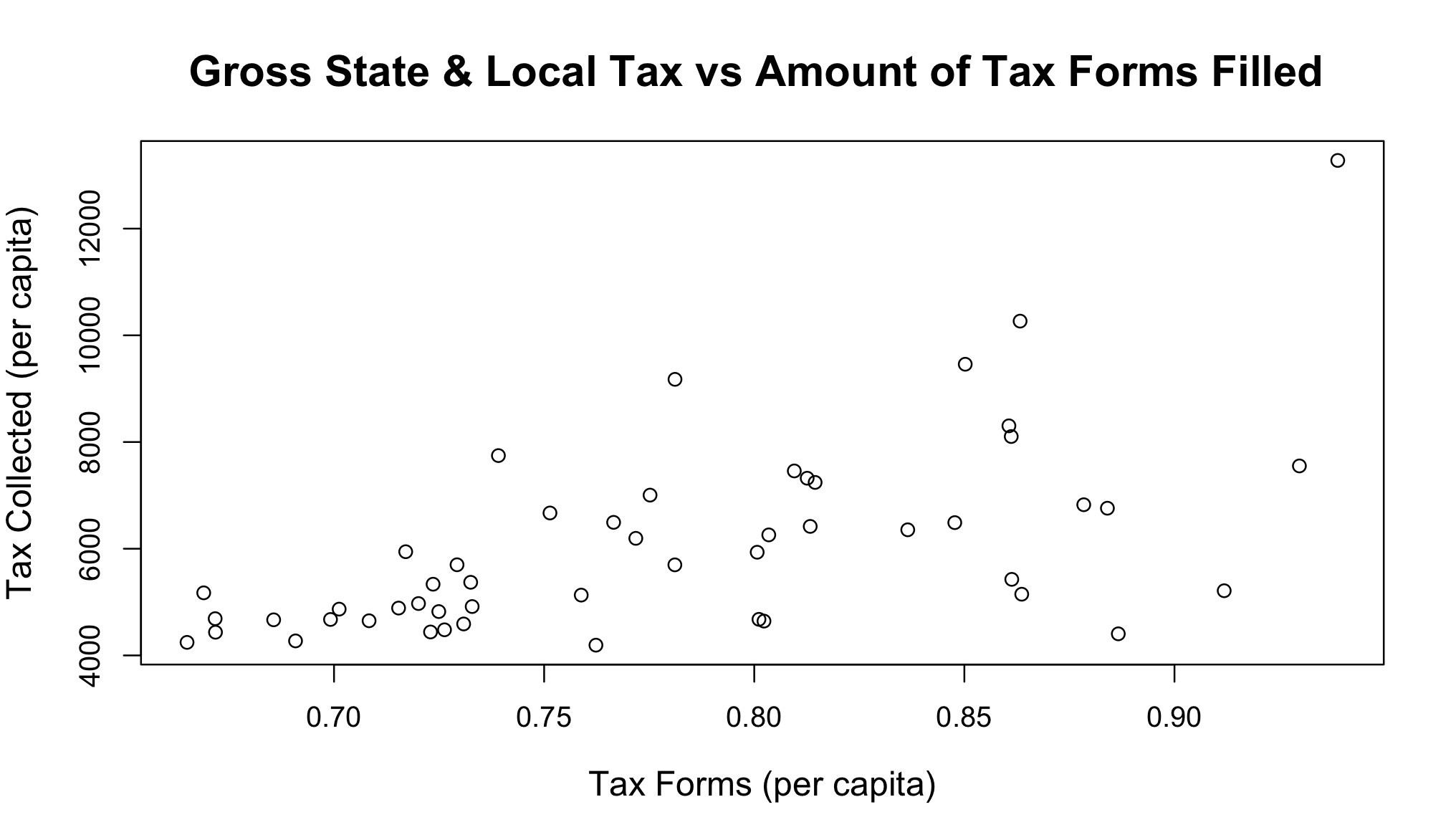
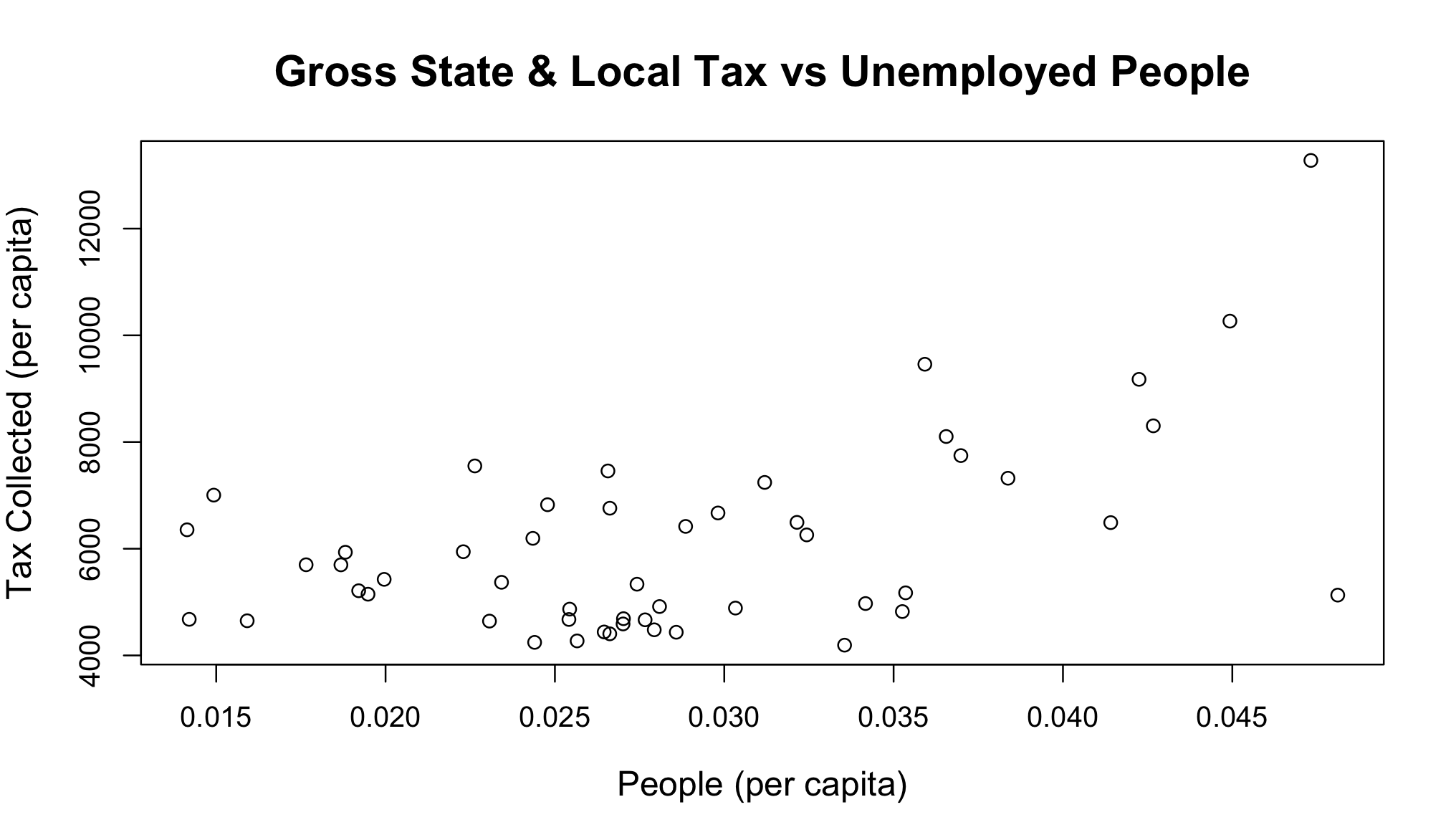
**Results**

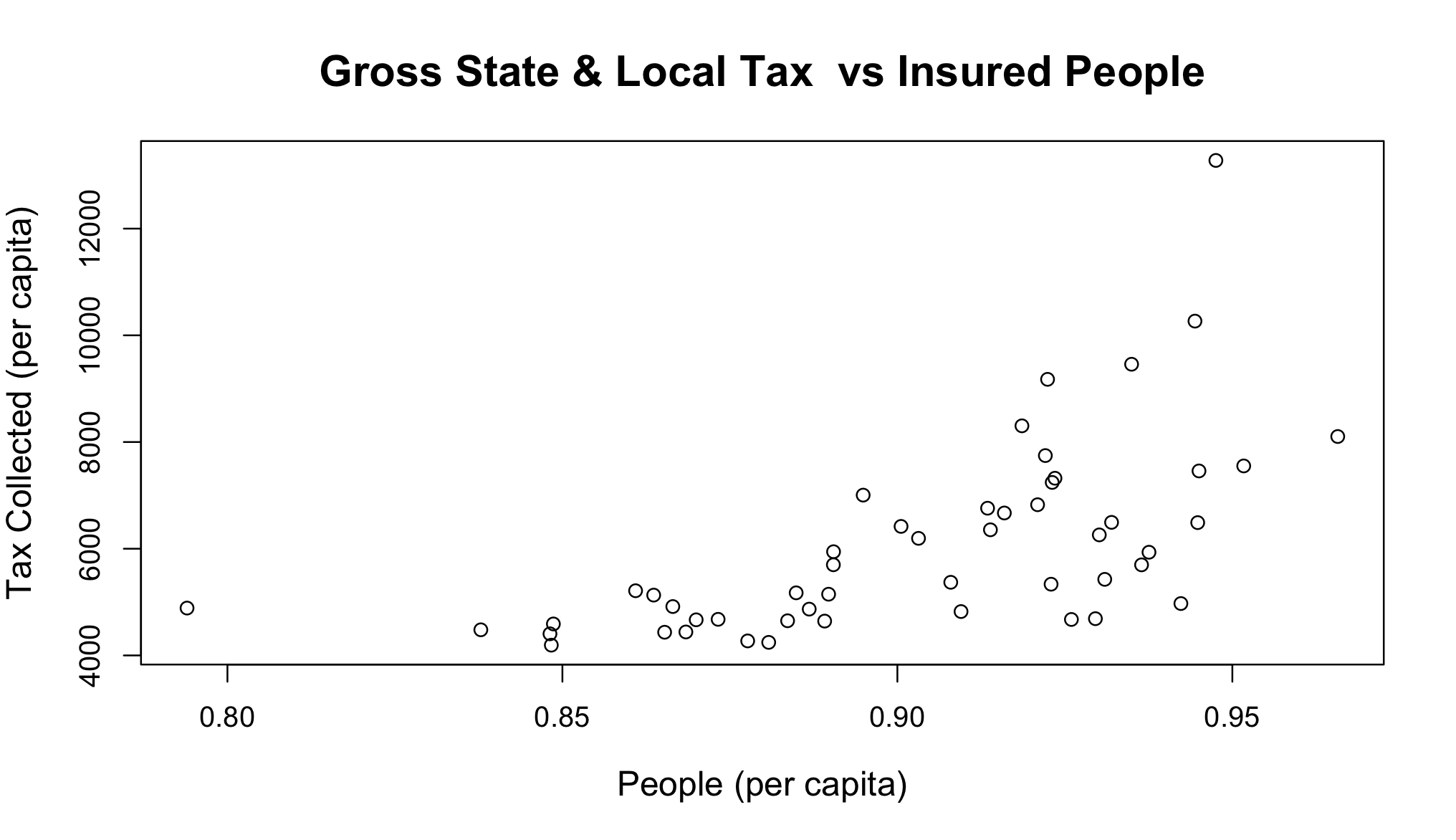
To better understand the dataset, we found the typical values of each variable on our model. Using the median values of said variables, we can describe the common U. S. state to have per capita, 5426 in gross state and local tax, GDP of 0.05857, 0.7718 filed tax forms, 0.02702 unemployed people, 0.9095 insured people and 0.06998 federal workers and the South has the most amount of states. It is important to note that the District of Columbia is used as a data point which has a considerably higher state and local tax because of how progressive the territory is. We found that it is important to keep this outlier because although not a state it represents a critical part of the U.S. population with their own government and tax collection parameters.

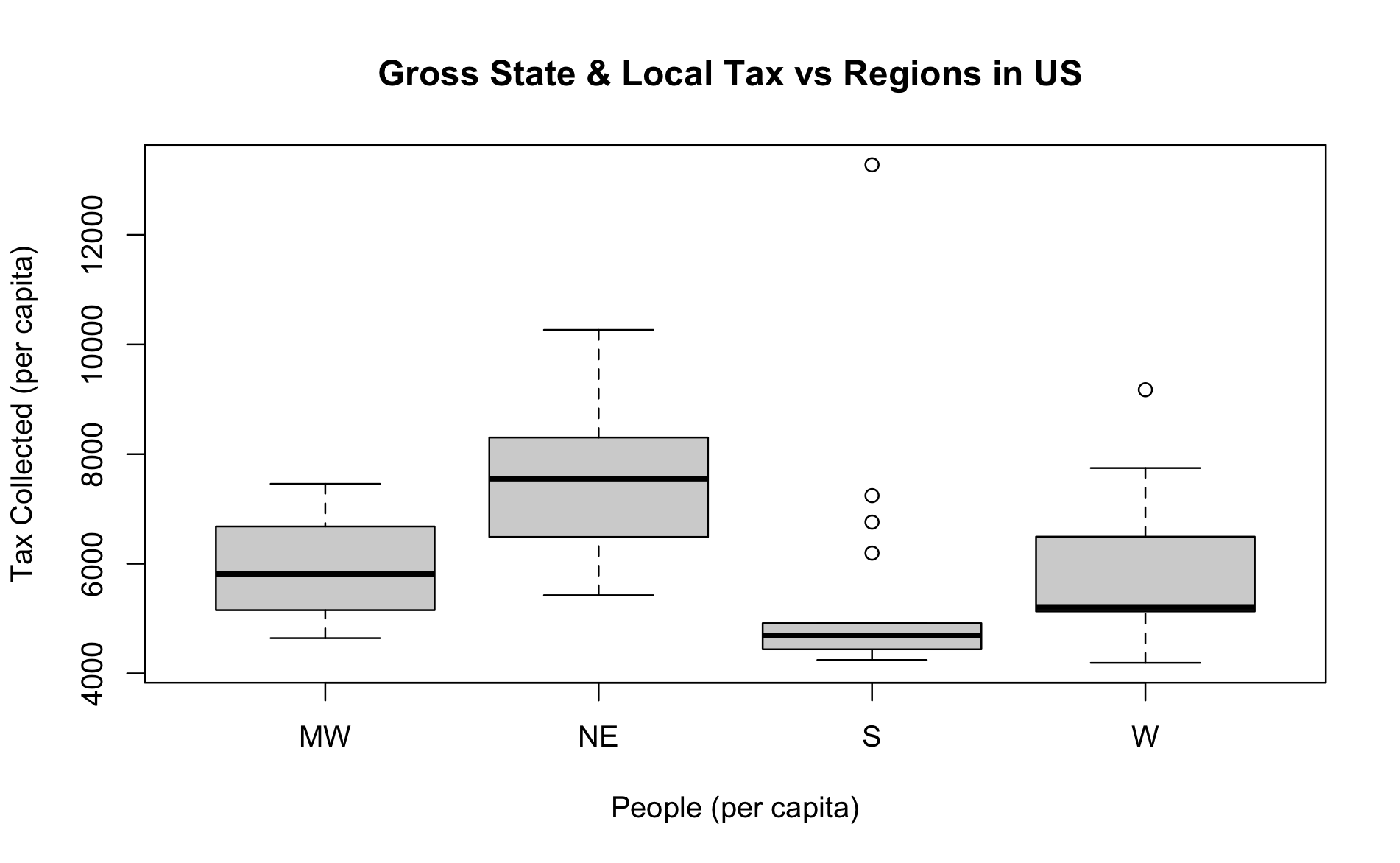
For our quantitative predictors we found our correlation coefficient to be the lowest between government workers and gross state and local tax although they have a positive relationship with r = 0.388 but it doesn't suggest any linearity. the plotted graph looks like the following:



The highest correlation coefficient is found between gross state and local tax and GDP with r = 0.7860338 making sense in practice as taxes collected directly boost the economy or GDP of a state in some way. This strong correlation makes a suggestion on linearity verified by the graph as seen below:

For the other quantitative variables their correlation coefficient falls between the two values stated before. Nothing notable happens with their values other than there being evidence in the graph's positive relationship as their correlation coefficient increases as expected. To state the correlation coefficients between the rest of variable and gross state and local tax. For unemployed people r = 0.517, tax forms filled have a r = 0.603, similarly for insured people r = 0.616. As stated before the relationship between them being positive and linear is supported by the following graphs:



For our only categorical variable we found our median federal gross tax by group by state per capita to be: For the Midwest median is 5893, Northeast is 7631 having the highest median between groups, the South has the lowest median with 5482, and the West with 5875 falls in a happy middle ground.

Having this many predictor variables poses the question of collinearity between variables. We found the lowest correlation to be r =0.005 and the highest to be r = 0.55. To see if interaction between terms was too high between in the highest of cases we ran for variation of inflation coefficients and we found the mean of the VIF to be 1.64 with our highest VIF value coming from the regions of the U.S. being 3.29 and because the values are under 5 we decided they wouldnt pose a problem for our model.

**Linear Models:**

Within some state of the U.S., let:

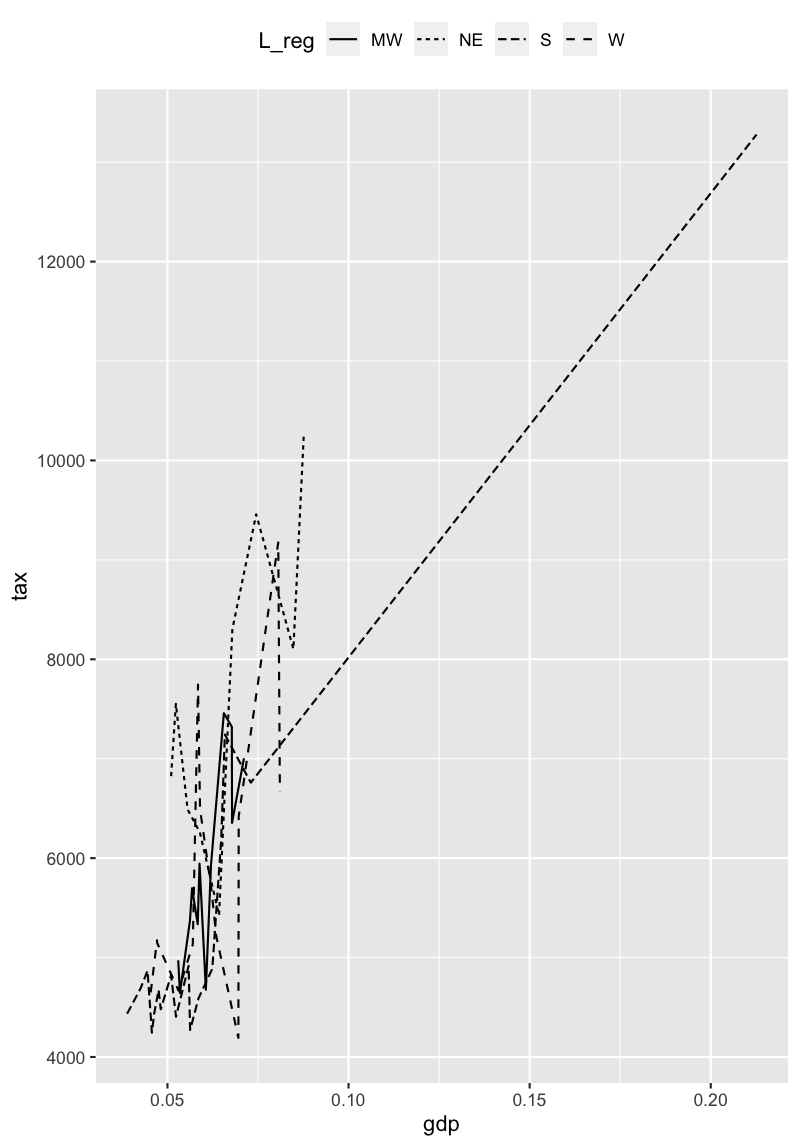
* be the gross state and local tax per capita
* be the number of forms filed per capita
* be the number of unemployed persons per capita
* the number of people with insurance per capita
* be the number of government workers per capita
* be the gross gdp per capita of this state

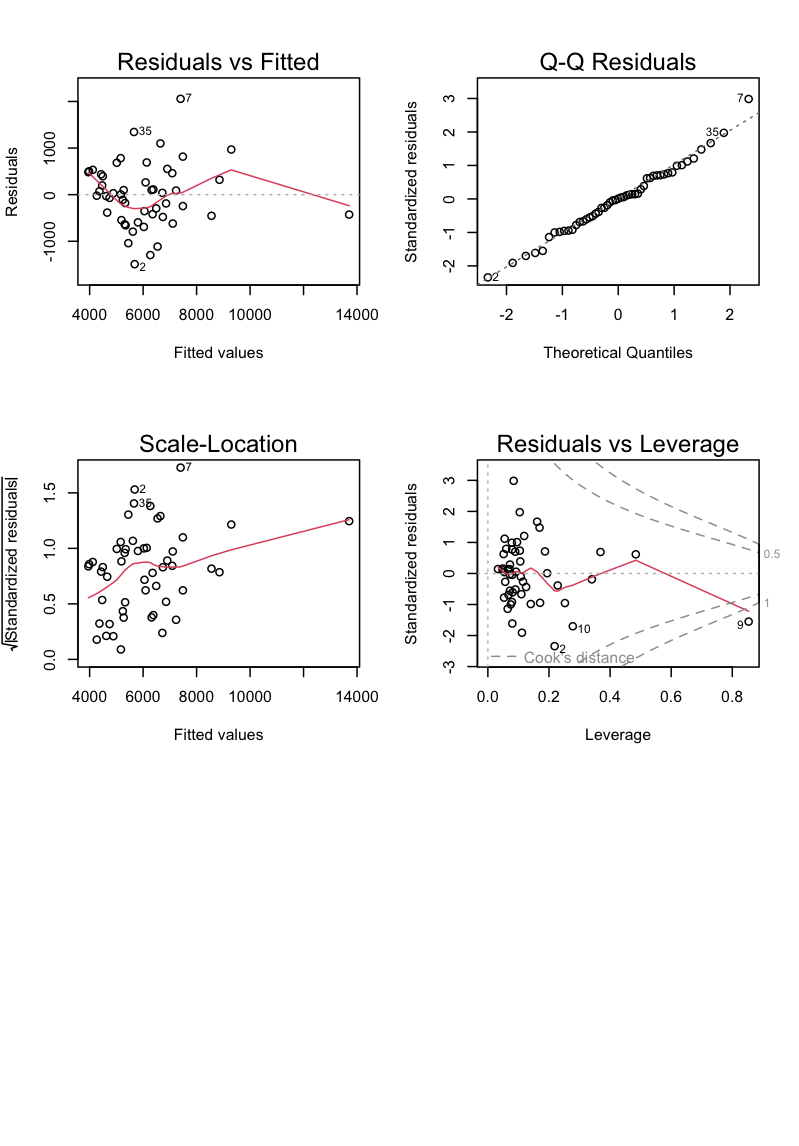
Also let be an indicator function equaling 1 if the tax is in general region . is one of four regions: the Northeast (NE), the South (S), the West (W) , or the Midwest (MW). Then the full model can be written as:

.

This is an ANCOVA linear model because it is composed of both quantitative, and categorical variables. This particular model does not have interaction terms, we will determine later in this paper if they are necessary. The model was built using dummy coding such that the linear model can be built directly from the summary statistics. For analysis regarding the diagnostics, the Pearson coefficient, and the F-test later on we used the default treatment coding.

Important to note that for our intercept , Is the mean of our gross state and local tax if all of the predictors are zero. However, our expectation was a low or even negative value as we are doing our model per capita and to account for people who don’t work the model needs to utilize extremely negative values and correct with the rest of the model. The coefficient for our GDP per capita per state is , and it is expected that for every 1 unit increase in GDP per capita there will be a increase in gross federal income. In terms of our categorical variable the model is built around the Midwest and each of the coefficients for the other regional levels have the coefficients to account for the geographical differences. In practical terms the Midwest is our default and the coefficients for the South, Northeast and West change the model for that specific region when called for.

We must check the diagnostics to ensure that all the assumptions requisite to fitting a model are within acceptable limits. The assumptions for an ANCOVA model are as follows: linearity, homogeneity, homoscedasticity, normality of residuals, independence, absence of perfect or near perfect multicollinearity. Linearity is shown if the fitted vs residual graphs show a linear relationship, and by the Pearson correlation which for this model is . This means that the Pearson correlation coefficient must be calculated via effect coding to act as an accurate representation of the linearity and correlation our response variable (gross state and local tax) has with our predictor variables. Linearity can also be seen graphically in the tax collected per capita vs. dependent variable graphs located in appendix B. Homogeneity of regression means that the slopes of the covariates should be the same across models. This can be seen in the graph to the right. It is a parallel slopes model. Although it is somewhat disrupted by the MW line’s offshoot to the right, a close observation will note that although the lines move about a lot, they always move in the same way. In other words, they always move together/parallel to each other. ANCOVA models will still function very well if the slopes are somewhat parallel, so it is not worth further investigating. In order to run regression, we desire that the error is random. It can be seen in the residual plot below that the error values appear to be randomly vertically scattered (independently, identically, normally distributed). This is enough to say that the residuals have homoscedasticity. Our data must also be normally distributed. The Q-Q plot below demonstrates that our data appears to be fairly normally distributed, as the Q-Q plot points appear to follow the line of normalcy. 



The multicollinearity of the model can be measured via a correlation matrix, or a variance inflation factor matrix, for both of which we omitted the factors. Both can also be found in appendix B. Because almost all variance inflation factors (and the mean of the VIF) are well below two, and because almost all correlation matrices are well below .8, multicollinearity should not be a problem. However for GDP, the correlation is nearing .8, and the VIF is a little above 2. As such we should be cautious of GDP, although it is still unlikely that it will be problematic in the formation of a model.

We performed a t-test for the model to determine whether or not the number of government workers per capita were significant to the full ANCOVA model. Our null hypothesis is that the number of government workers per capita are not significant to the full model. Our alternate hypothesis is that the number of government workers per capita are significant to the full model. After the t-test, we found the t-value to be 1.021, there are df = 24 degrees of freedom, and The p-value was 0.313077, which is much greater than even a .10 alpha level. Because the p-value was much greater than the alpha level, we fail to reject a null hypothesis. This means it is very unlikely that the number of government workers per capita is significant to the full model. After performing this test to all variables we found our reduced model (using effect coding) letting

* be the gross state and local tax per capita
* be the number of unemployed persons per capita
* be the number of people with insurance per capita
* gross gdp of this state to be

The model is:

We performed an F test to determine whether the new reduced model we found was a better estimate than the full model. : there is no difference between models : there is a difference between reduced model and full model *a* < 0.05. We found that when we ran the anova p = 0.07 with 5 dfs and F = 2.1738 and by our hypothesis we failed to reject the null hypothesis but there is no statistical difference between models and because we have less variables in the reduced model, we choose the simpler estimate between models.

We ran a prediction on what the new data would look like using the fake state that has all the median values from all variables remaining. Therefore we used the values of GDP as 0.05857, 0.02702 for unemployed people and 0.9095 for insured people. When the prediction was run we found that the model has a fit of 5965.65 with a range of (4299.856, 7631.443) with 95% confidence. We then added the other median values to predict the full model with filled forms = 0.7718 , government workers as 0.06998 and the state being from the South. In this case, we found our fit to be 5635.572 and with a range of (4006.909 , 7264.236) with 95% confidence.

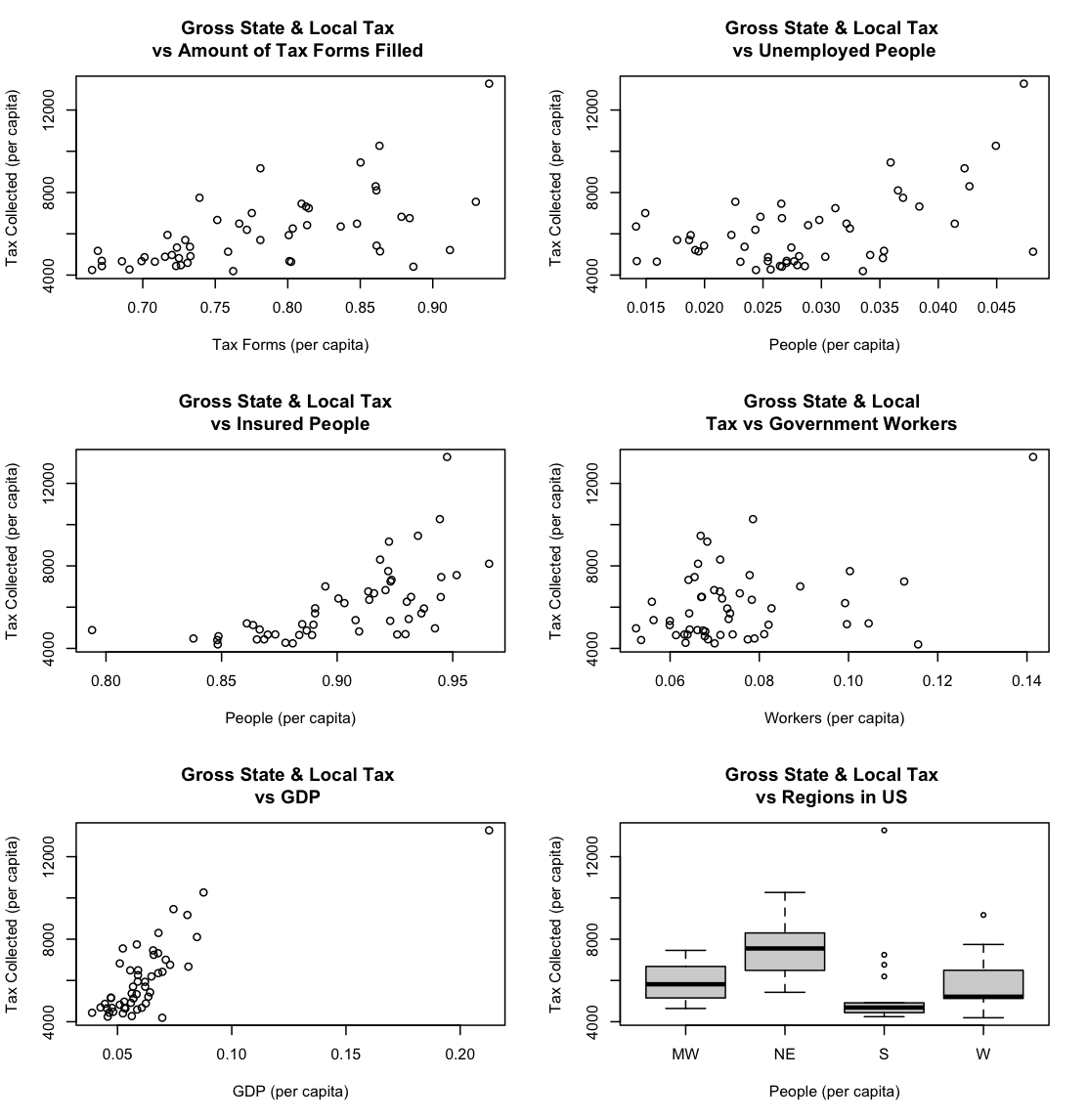
**Model Selection:**

We decided to check our model to determine if any information was lost in forming the model. In order to determine this, we ran both an AIC and a BIC using forward selection. The information criterion estimates the amount of information lost due to adding variables to the model in an attempt to improve the goodness of fit. The information criterion does not however determine the significance of a variable to the model, just how much the model benefits in comparison to the amount of information lost by adding it. As such, using the reduced model gives a more precise and accurate representation of the most capable. Both the AIC and the BIC agreed that the complete (reduced) model was the most beneficial (outputs given in Appendix D), as it gave both the lowest AIC (686.0726), and the lowest BIC (691.8681) scores. As such we have evidence to believe that the variables in the complete (reduced) model are more beneficial toward a good estimation than their absence would be.

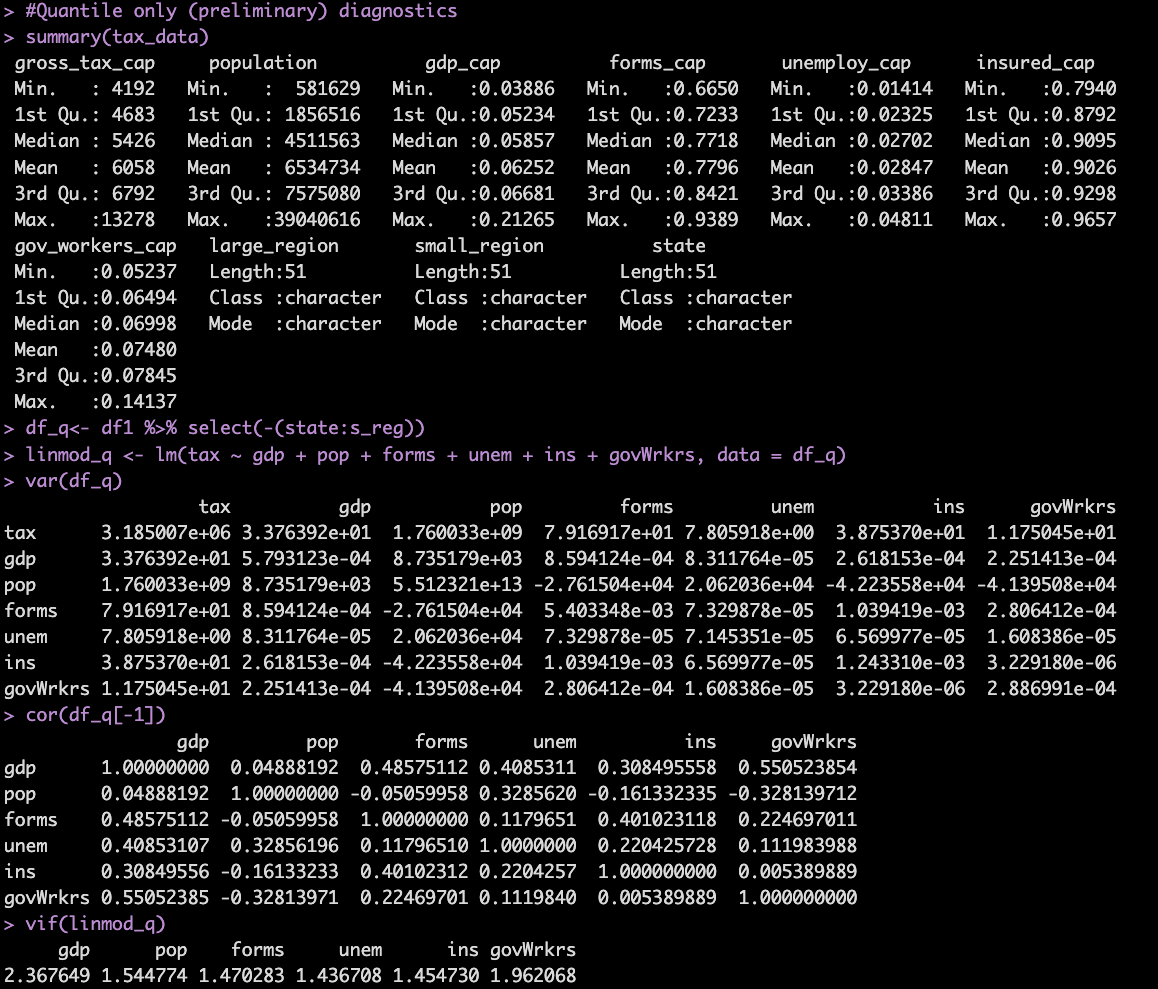
**Conclusion:**

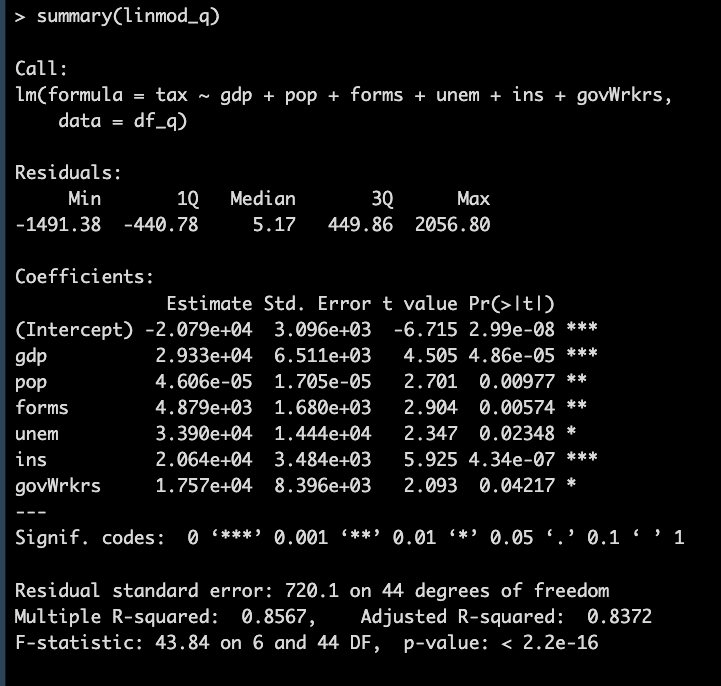
We concluded overall that there is a relationship between the predictor variables GDP, the amount of insured people per capita, unemployed persons per capita, amount of tax forms filed per capita, state government workers per capita, and the region of the U.S. the state is in, and the response variable gross state tax per capita which was the purpose of this project. We found that there is a reduced model using hypothesis testing that has similar accuracy with the full model using the data collected per capita of unemployed people per capita, insured people per capita, and GDP per capita for each state as predictor variables. This model works as a proof of feasibility, and in the future we would expect this model’s predictive ability to be tested by comparing its predictions and the actual results of years outside of 2021. Future research should also test more predictors for significance. Advisable characteristics of state specific predictor variables are measures of wealth disparity and racial make up. Exploring datasets from other years is a worthy endeavor as it might let us predict future gross tax income and missing values of tax information before the 1940’s.

**Appendix A:** Additional Graphs

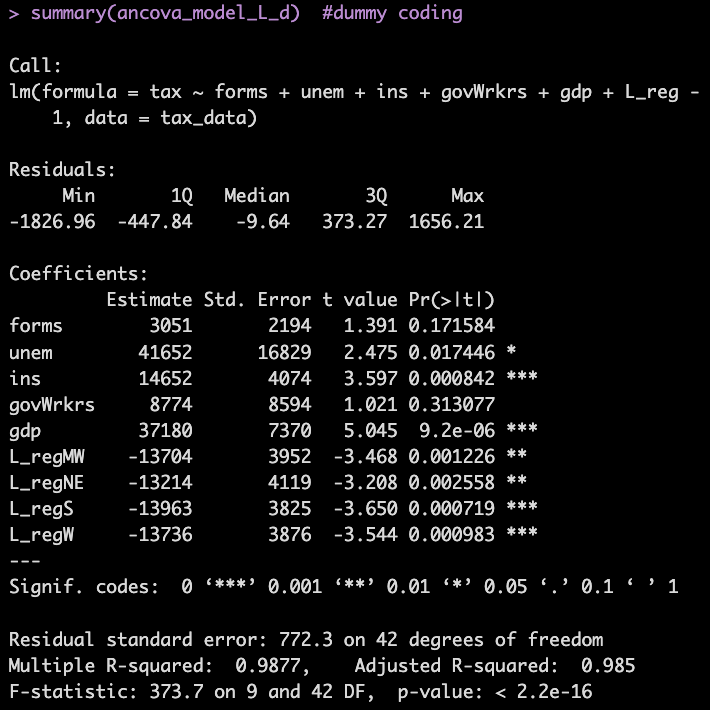
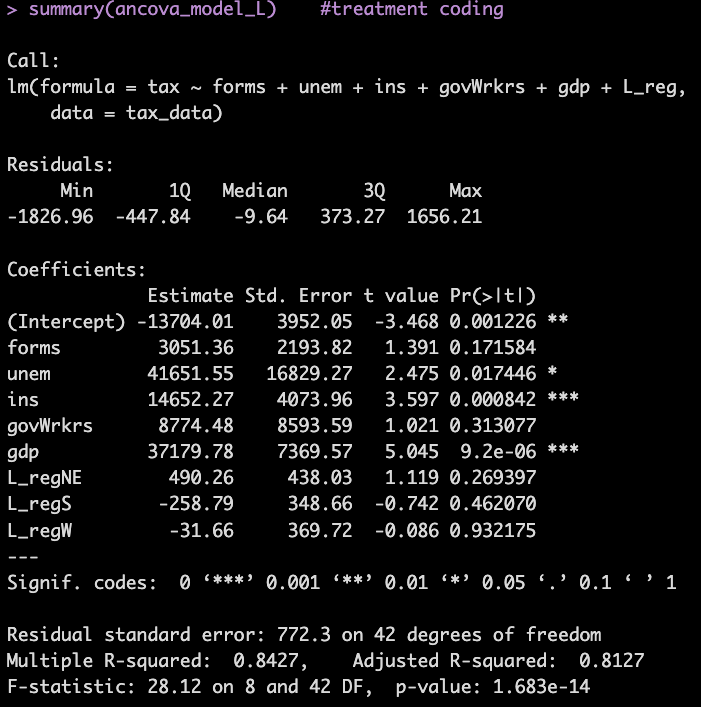


**Appendix B:** Quantitative model and diagnostics

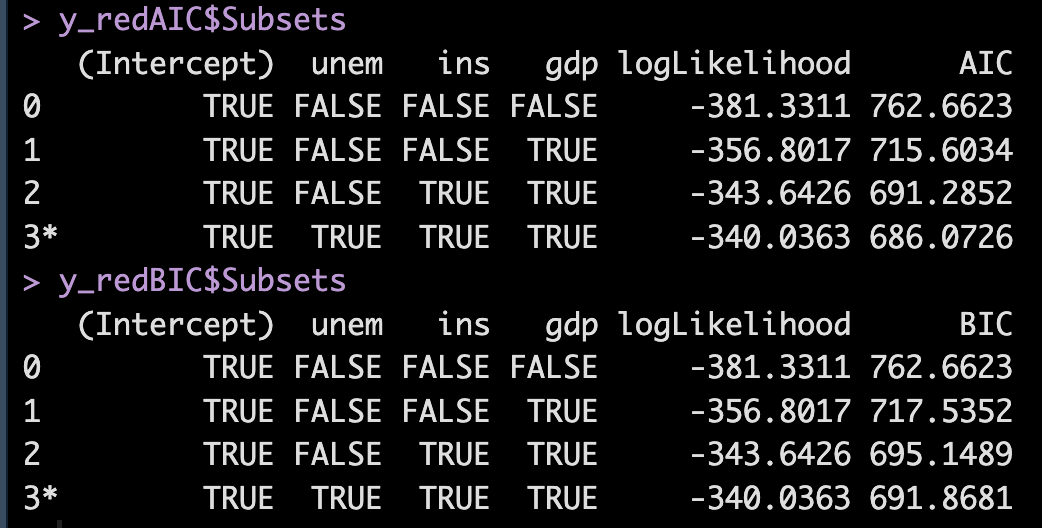




**Appendix C:** ANCOVA Models:



**Appendix D:** Model Selection

****

**Appendix E:** Code (written in R)

tax\_data <- read.csv()

library(car)

library(dplyr)

library(bestglm)

par(mfrow = c(2,2))

df1 <- data.frame(tax = tax\_data$gross\_tax\_cap, gdp = tax\_data$gdp\_cap,

pop = tax\_data$population, forms = tax\_data$forms\_cap,

unem = tax\_data$unemploy\_cap, ins = tax\_data$insured\_cap,

govWrkrs = tax\_data$gov\_workers\_cap, state = tax\_data$state,

L\_reg = tax\_data$large\_region, s\_reg =tax\_data$small\_region)

list2env(df1, envir = .GlobalEnv) #small project, shouldn't cause problems. If up-scaled, would need to remove

#Quantile only (preliminary) diagnostics

summary(tax\_data)

df\_q<- df1 %>% select(-(state:s\_reg))

linmod\_q <- lm(tax ~ gdp + pop + forms + unem + ins + govWrkrs, data = df\_q)

var(df\_q)

cor(df\_q[-1])

vif(linmod\_q)

summary(linmod\_q)

plot(linmod\_q)

par(mfrow = c(2,3))

plot(tax ~ forms , data = tax\_data, main ="Gross State & Local Tax \nvs Amount of Tax Forms Filled ", xlab ='Tax Forms (per capita)', ylab = 'Tax Collected (per capita)' )

plot(tax ~ unem , data = tax\_data, main ="Gross State & Local Tax \nvs Unemployed People", xlab ='People (per capita)', ylab = 'Tax Collected (per capita)')

plot(tax ~ ins , data = tax\_data, main ="Gross State & Local Tax \nvs Insured People", xlab ='People (per capita)', ylab = 'Tax Collected (per capita)')

plot(tax ~ govWrkrs , data = tax\_data, main ="Gross State & Local \nTax vs Government Workers ", xlab ='Workers (per capita)', ylab = 'Tax Collected (per capita)')

plot(tax ~ gdp , data = tax\_data, main ="Gross State & Local Tax \nvs GDP ", xlab ='GDP (per capita)', ylab = 'Tax Collected (per capita)')

boxplot(tax ~ L\_reg , tax\_data = tax\_data, main ="Gross State & Local Tax \nvs Regions in US ", xlab ='People (per capita)', ylab = 'Tax Collected (per capita)')

par(mfrow = c(2,2))

require(ggplot2)

ggplot(tax\_data,aes(x=forms,y=tax))+stat\_summary(fun.y="mean",geom="line",aes(group=L\_reg,linetype=L\_reg))+theme(legend.position = "top", legend.direction = "horizontal")

ggplot(tax\_data,aes(x=unem,y=tax))+stat\_summary(fun.y="mean",geom="line",aes(group=L\_reg,linetype=L\_reg))+theme(legend.position = "top", legend.direction = "horizontal")

ggplot(tax\_data,aes(x=ins,y=tax))+stat\_summary(fun.y="mean",geom="line",aes(group=L\_reg,linetype=L\_reg))+theme(legend.position = "top", legend.direction = "horizontal")

ggplot(tax\_data,aes(x=govWrkrs,y=tax))+stat\_summary(fun.y="mean",geom="line",aes(group=L\_reg,linetype=L\_reg))+theme(legend.position = "top", legend.direction = "horizontal")

ggplot(tax\_data,aes(x=gdp,y=tax))+stat\_summary(fun.y="mean",geom="line",aes(group=L\_reg,linetype=L\_reg))+theme(legend.position = "top", legend.direction = "horizontal")

par(mfrow = c(2,2))

# outlier

head(sort(gdp,decreasing = TRUE))

which(gdp >= .2) %>% state[.]

# fits with trend of data => kept in. then again:

df2 <- subset(df\_q, state != "District\_of\_Columbia")

linmod\_q\_reduced <- lm(df2$tax ~ df2$forms + df2$unem + df2$ins + df2$govWrkrs, data = df2)

summary(linmod\_q) #r^2 = .8567

summary(linmod\_q\_reduced)# r^2 = .6578

df3\_q <- subset(df\_q, state != "District\_of\_Columbia")

#ancova modeling

ancova\_model\_L <- lm(tax ~ forms + unem + ins + govWrkrs + gdp + L\_reg, data = tax\_data) #treatment coding

ancova\_model\_L\_d <- lm(tax ~ forms + unem + ins + govWrkrs + gdp + L\_reg-1, data = tax\_data) #dummy coding

Anova(ancova\_model\_L, type = 'III')

ancova\_model\_s <- lm(tax ~ forms + unem + ins + govWrkrs + gdp + s\_reg, data = tax\_data,)

Anova(ancova\_model\_s, type = 'III')

lmod\_e<-lm(tax ~ forms + unem + ins + govWrkrs + gdp + L\_reg, data = tax\_data,

contrasts=list(L\_reg="contr.sum"))

plot(ancova\_model\_L)

plot(ancova\_model\_L\_d)

plot(ancova\_model\_s)

ancova\_model\_Ls <- lm(tax ~ forms + unem + ins + govWrkrs + gdp + L\_reg + s\_reg, data = tax\_data)

Anova(ancova\_model\_L, type = 'III')

summary(ancova\_model\_Ls)

#model comparisons

summary(ancova\_model\_L)

summary(ancova\_model\_s) #categorical is significant

summary(lmod\_e) #effect coding = no significant improvement

anova(ancova\_model\_L, ancova\_model\_s)

tci\_L <- TukeyHSD(aov(tax ~ L\_reg))

tci\_L

tci\_s <- TukeyHSD(aov(tax ~ s\_reg))

tci\_s

plot(tci\_L)

plot(tci\_s) #while almost all cross 0, all ranges are to big to use (lowwww p-val)

plot(ancova\_model\_L)

par(mfrow=c(1,1))

t.test(gdp ~ L\_reg[2:3], df2)

#IC Model Selection

df\_IC <- df1[,c(2:7,1)]

#df\_IC[7] <- factor(df1[,7])

names(df\_IC)[7] <- 'y'

y\_AIC <- bestglm(df\_IC, IC='AIC')

y\_BIC <- bestglm(df\_IC, IC='BIC')

y\_AIC$Subsets

y\_BIC$Subsets

df\_redIC <- df1[,c("unem","ins","gdp",'tax')]

names(df\_redIC)[4] <- 'y'

y\_redAIC <- bestglm(df\_redIC, IC='AIC')

y\_redBIC <- bestglm(df\_redIC, IC='BIC')

y\_redAIC$Subsets

y\_redBIC$Subsets

1. <https://data.census.gov/table?q=employment%20by%20state%202021&g=010XX00US_040XX00US16> [↑](#footnote-ref-0)
2. Source: <https://www.census.gov/programs-surveys/acs/geography-acs/areas-published.html> [↑](#footnote-ref-1)
3. <https://www.census.gov/data/tables/time-series/demo/popest/2020s-state-total.html#v2022> [↑](#footnote-ref-2)
4. This data set: [2021](https://www.irs.gov/pub/irs-soi/21dbs01t03nr.xlsx) can be found on this page of the IRS’s website: <https://www.irs.gov/statistics/soi-tax-stats-number-of-returns-filed-by-type-of-return-and-state-and-fiscal-year-irs-data-book-table-3> [↑](#footnote-ref-3)
5. Unfortunately, this data set does not have any direct link, so it requires a link to the site its on: <https://apps.bea.gov/itable/?ReqID=70&step=1#eyJhcHBpZCI6NzAsInN0ZXBzIjpbMSwyOSwyNSwzMSwyNiwyNywzMCwzMF0sImRhdGEiOltbIlRhYmxlSWQiLCI1MzEiXSxbIk1ham9yX0FyZWEiLCIwIl0sWyJTdGF0ZSIsWyIwIl1dLFsiQXJlYSIsWyJYWCJdXSxbIlN0YXRpc3RpYyIsIjEiXSxbIlVuaXRfb2ZfbWVhc3VyZSIsIkxldmVscyJdLFsiWWVhciIsWyIyMDIxIl1dLFsiWWVhckJlZ2luIiwiLTEiXSxbIlllYXJfRW5kIiwiLTEiXV19> [↑](#footnote-ref-4)