Final Project PSTAT 126: Regression

Analysis

Regression Analysis of California Taxes

Sam Leathers

Vahe Barseghyan

6/01/2020

Contents

1. Introduction

2. Research Question

3 Methodology

4. Regression Analysis, Results, and Interpretation.

5. Findings & Final Model

6. Appendix

**Introduction**

This project will focus on studying the State of California’s taxes as a function of fiscal spending using the United States Census Bureau data for state governance. We will be investigating which forms of fiscal policy are having the greatest effect on the growth of states taxes over an 18 year period. We will be specifically focusing on the following forms of fiscal spending: total expenditure, education, public welfare, hospitals, police protection, prisons, highways, and direct expenditure.

**Research Question**

Using modelling methods, can we uncover statistical correlations between the forms of fiscal policy that are allocated funding by a state, and the growth of state taxes over an 18 year period?

Can the state’s Taxes be predicted by forms of spending?

**Methodology**

We will be building a linear regression model adhering to the LINE conditions for multiple predictor variables. Our multiple regression model will undergo residual analysis, visual inspection of a scatterplot matrix, and a QQ plot to ensure each of the LINE conditions are met to include: linearity, independent error terms, normal distribution of error, and equal variance of error. We will be applying transformation to our variables to ensure that linearity and error normality are maintained throughout.

**Regression Analysis, Results, and Interpretation.**

Our model building begins with defining each of our variables and then creating a base model with no predictor variables and one full model with all response variables being tested included in the model.

**Response (Y variable): Taxes**

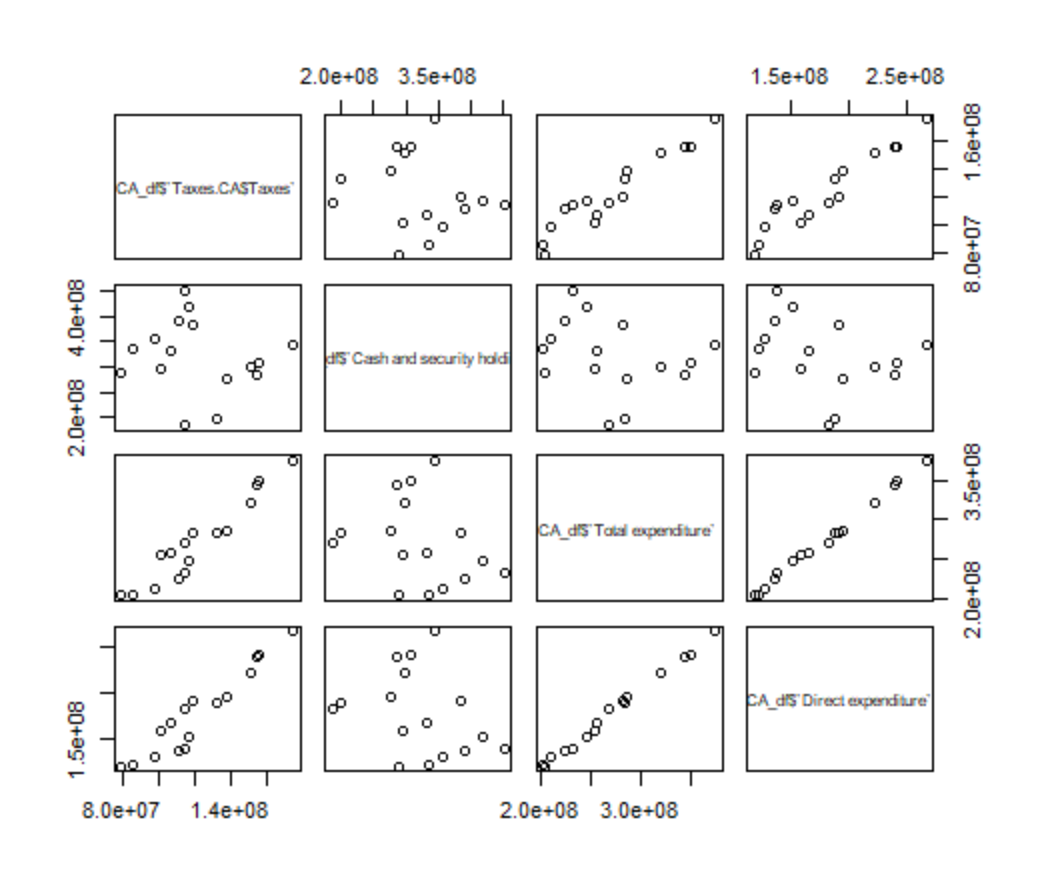
Taxes were chosen as they include all state revenue minus Federal funds from the government. We felt the interplay between federal fund flows and state fiscal policy is likely to be minimal and related to Federal policy not being analyzed.

**Predictors (X variables):**

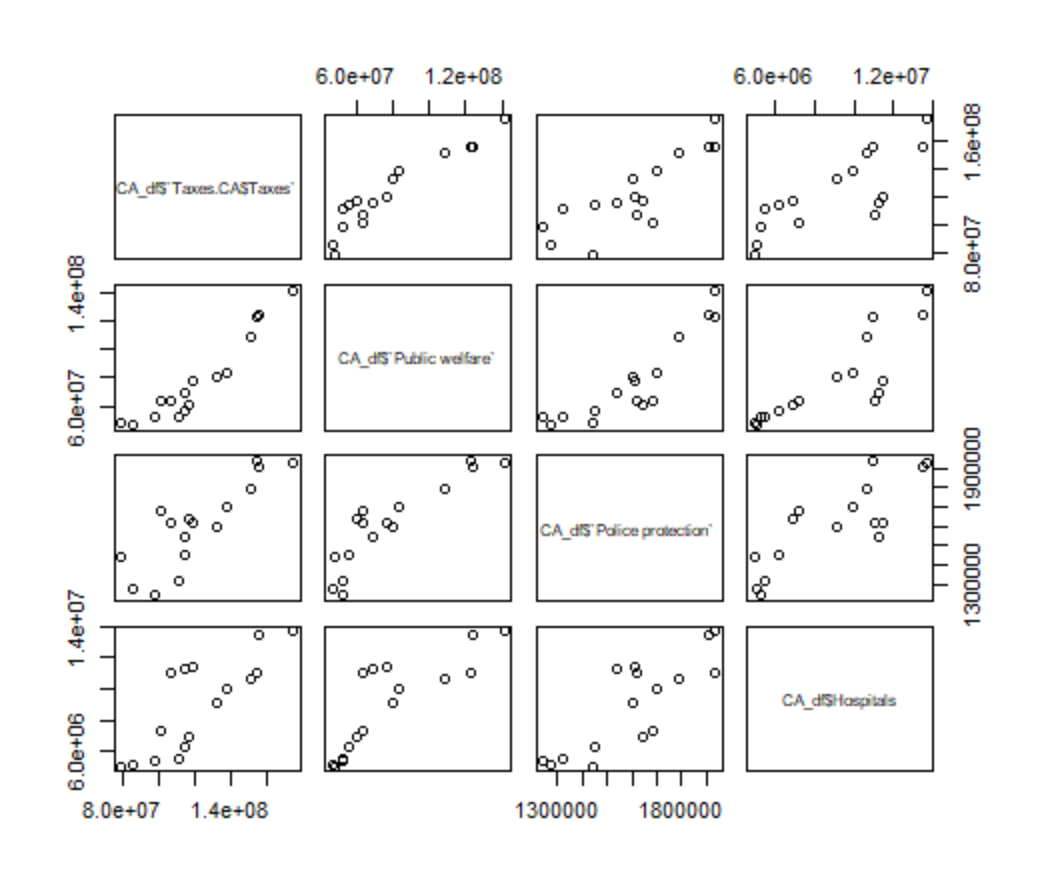
Total expenditure, direct expenditure, education, public welfare, police protection, correction, hospitals, net cash, and highways.

NOTE: Total expenditure and direct expenditure are comprised of the other predictor variables as well as other forms of spending not included in the model, please refer to the appendix for more information regarding the variables presented.

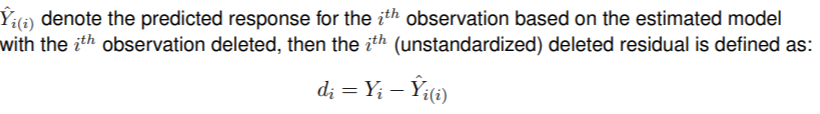
check for outliers/influential observation and remove them if any,

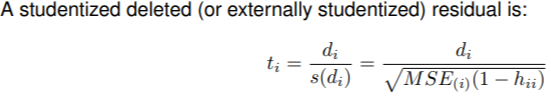
To check for linearity of the variables a scatterplot matrix of each predictor variable against the response can be created to observe their relationships. We are looking for relationships in which the variables idealistically form a line like object with no patterns or curves. Given the number of predictors we are using we will create three scatterplot matrices for legibility. 

In the figure above we can see that cash and security holdings or net cash has no linear relationship whereas both forms of expenditure are nearly linear with taxes. No obvious outlier data points are observed.



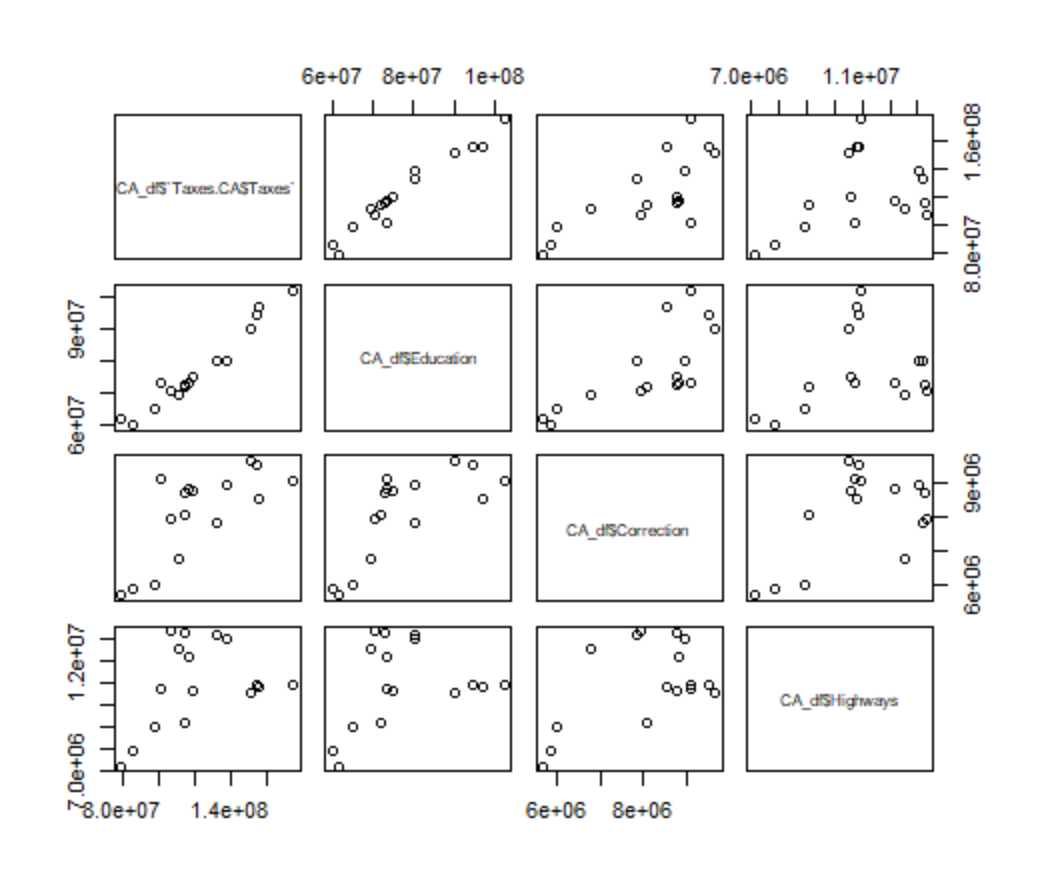
Among these variables we can note that public welfare is near linear whereas police protection and hospitals appear to be nearly linear. Police protection may have an outlier. We will be investigating further into this by using externally studentized residuals, as the value appears to be outlying on the y axis.





hii: Leverage – quantifies x value outliers, assists with y outliers.

Source: lecture slides.



Education is the only variable which appears linear to begin with, correction and highways appear to have a less defined linear trend which may influence their performance on the model.

> CA.upper <- lm( CA\_df$`Taxes.CA$Taxes` ~ + CA\_df$`Cash and security holdings` + CA\_df$`Total expenditure`

+ + CA\_df$`Direct expenditure` + CA\_df$Education + CA\_df$Correction

+ + CA\_df$`Public welfare` + CA\_df$`Police protection`

+ + CA\_df$Hospitals + CA\_df$Highways )

> rstudent(CA.upper)

1 2 3 4 5 6

0.12322653 0.69424481 -0.92391152 -1.29102588 3.88531183 0.14474219

7 8 9 10 11 12

-2.74337325 0.72822118 -1.15431070 -0.10779395 1.48444747 0.77213563

13 14 15 16

-0.06156611 0.59448230 -0.94810710 0.00173102

With the studentized deleted results above we can see that data point 5, associated with 2007 is an outlier as it is beyond the threshold value of 3. To maintain continuity in the data set, all data for 2007 will be removed. We decided this by creating both models with and without data from 2007, and cross-compared the descriptiveness of the results.

 perform stepwise/best-subsets regression to narrow down the number of predictors,

We chose to search for the best model by going through three separate model building processes in order to validate our chosen model. First we went through step-wise regression utilizing the F-test statistic for the most significant variable to be added next, iterating until only significant variables were in the model. We then chose to use Akaike’s criterion which produced the same model we had arrived at with our own step-wise regression.

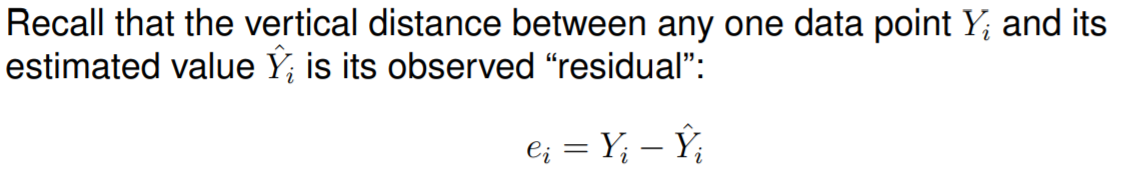
Finally we used regsubsets() function within the leaps library in R in order to iterate through the process again, yet this time optimizing the models for R^2 and adjusted R^2 ( add definitions/ formula for R2 adjusted R2, F-test etc) . This last method produces the best model with each possible number of configurations, ( 2, 3,4… predictors etc. for all available predictors)) When optimized for R2 and adjusted R2 the last significant jump while adding predictors is for a model with 3 predictors. Both of those models corroborate the findings with the first two methodologies in model building.

Each of the models indicated that the most significant forms of fiscal policy on the States Taxes were: education, highway, and police protection.

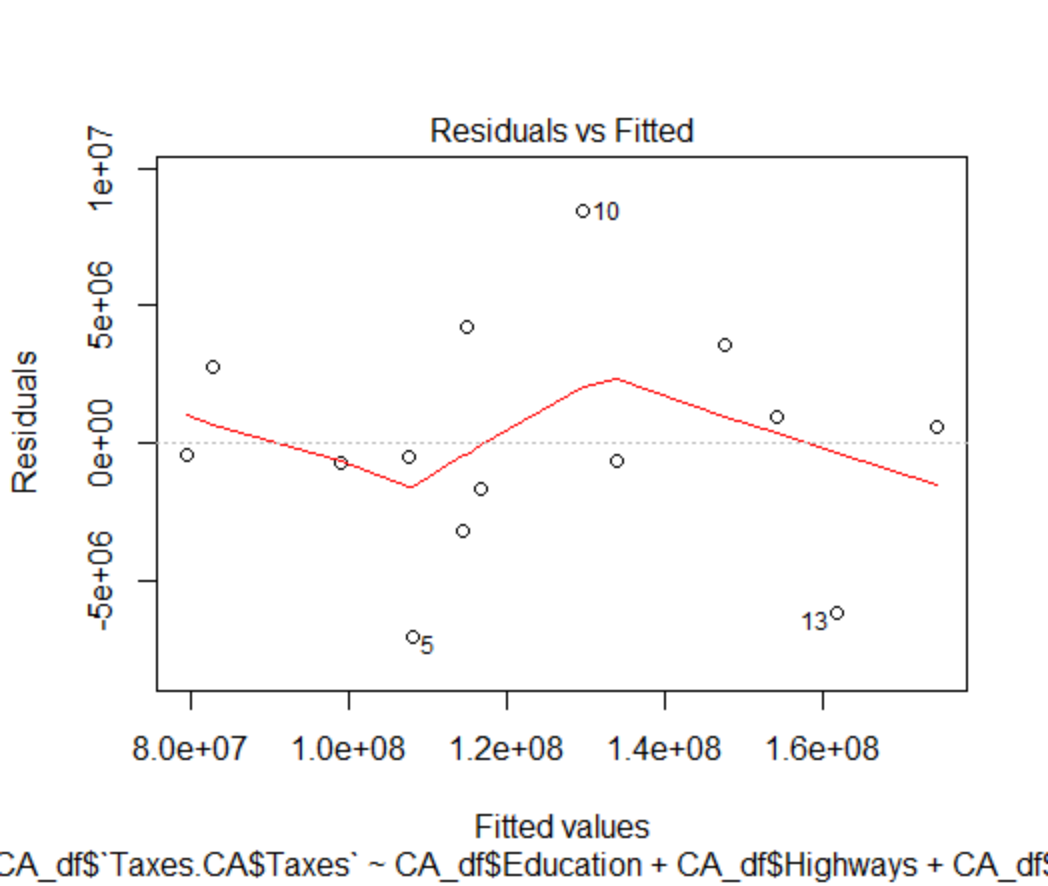
check if you can include any higher order terms or interactions, (I am not sure what exactly this means)

perform necessary transformations on the response/predictors/both,

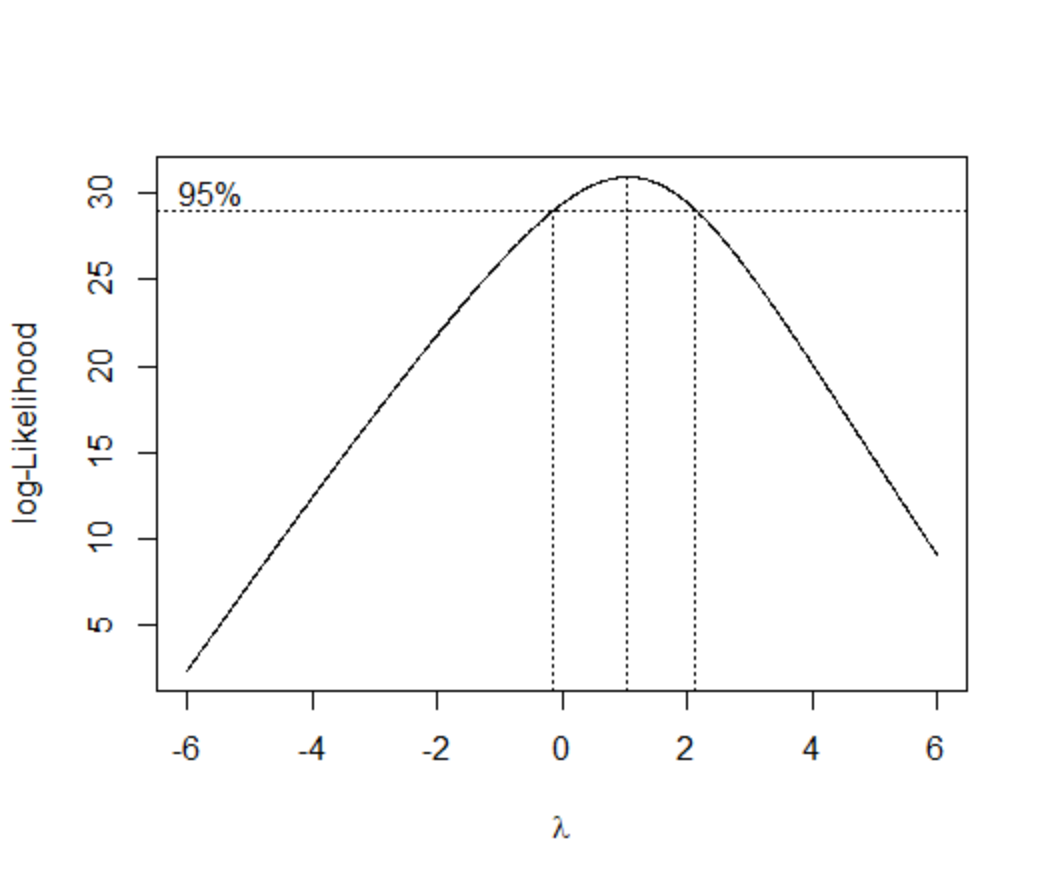
With our model decided we must verify that each of the aforementioned LINE conditions have been satisfied. The first observation is with a residual versus (vs) fitted values scatterplot. This plot shows the error for each prediction value (residuals) against the fitted values expected.



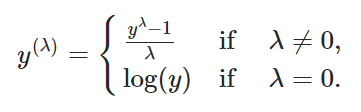
We expect the residual error terms to be normal and have equal variance, which should manifest as: bouncing randomly around 0 (no error) forming a rough horizontal band with no outliers. i



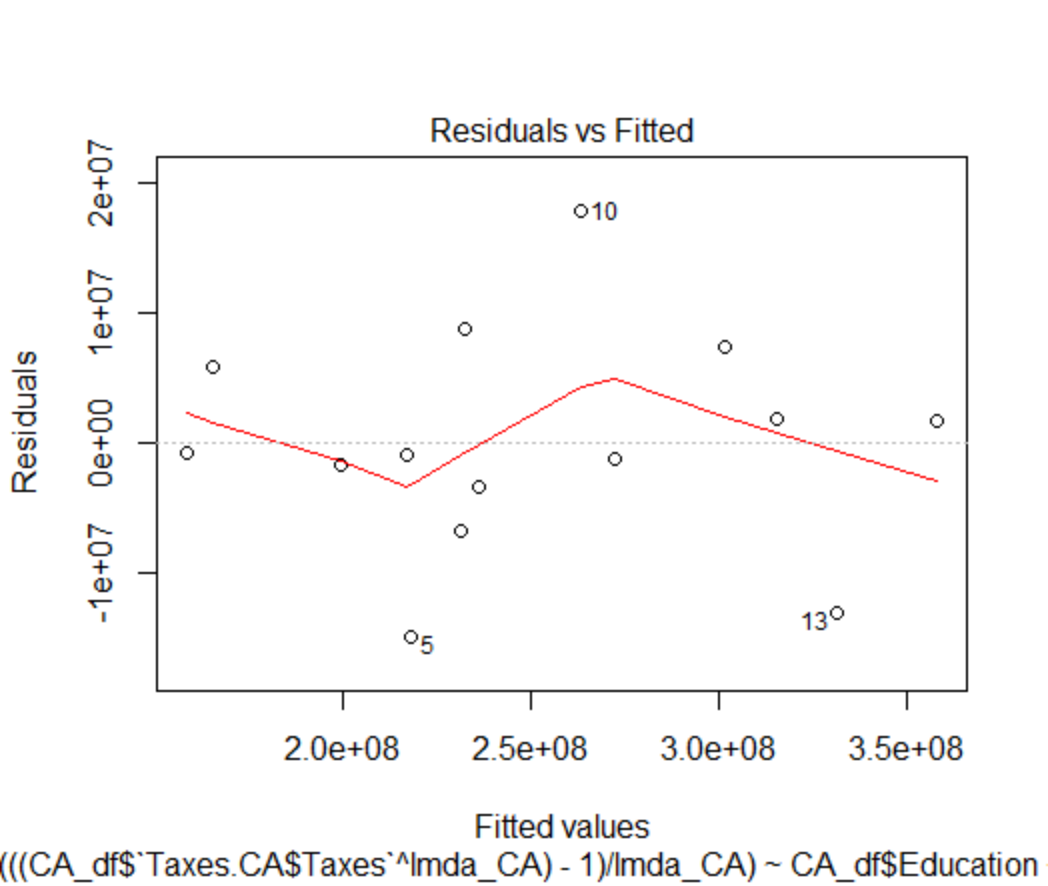
A conclusion cannot be drawn by observing the residual vs fit plot alone, as it depicts no major characteristic, such as fanning or non-linear behavior. Given this circumstance we chose to use the Box-Cox transformation which uses the estimation method of maximum likelihood iterating over exponent values on our response variable (taxes) to minimize error in the model.



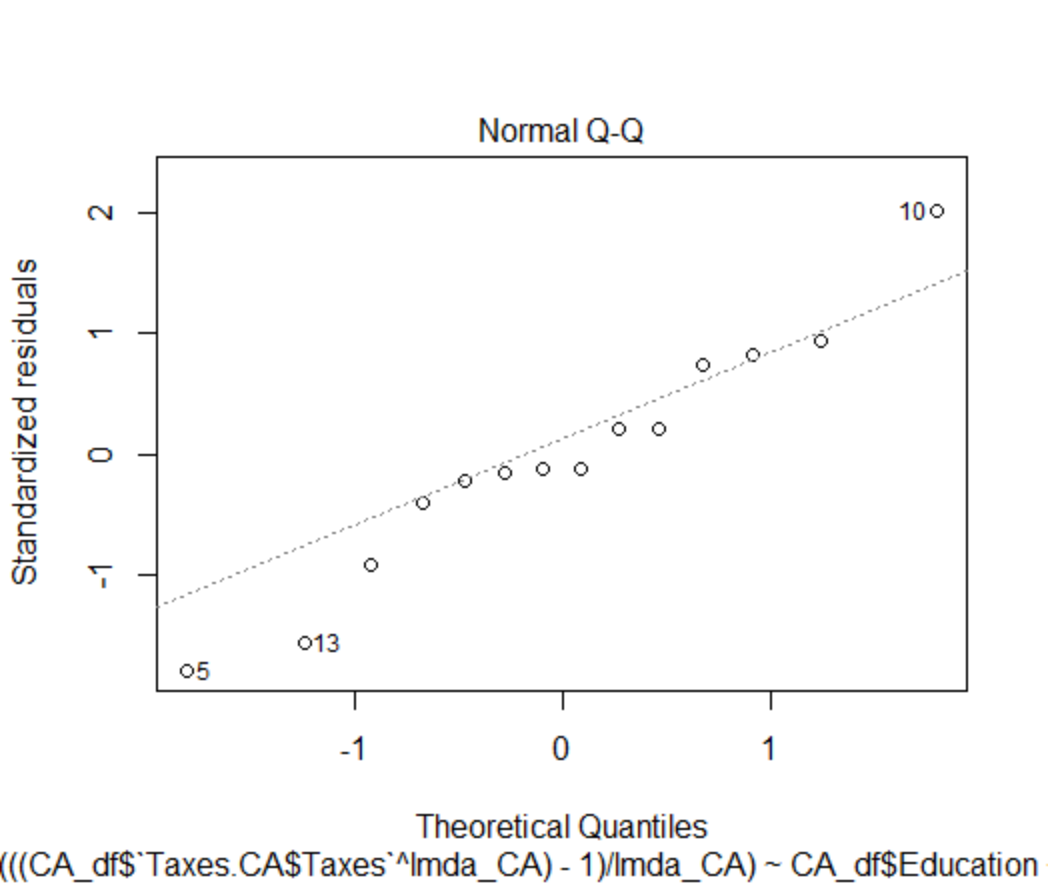
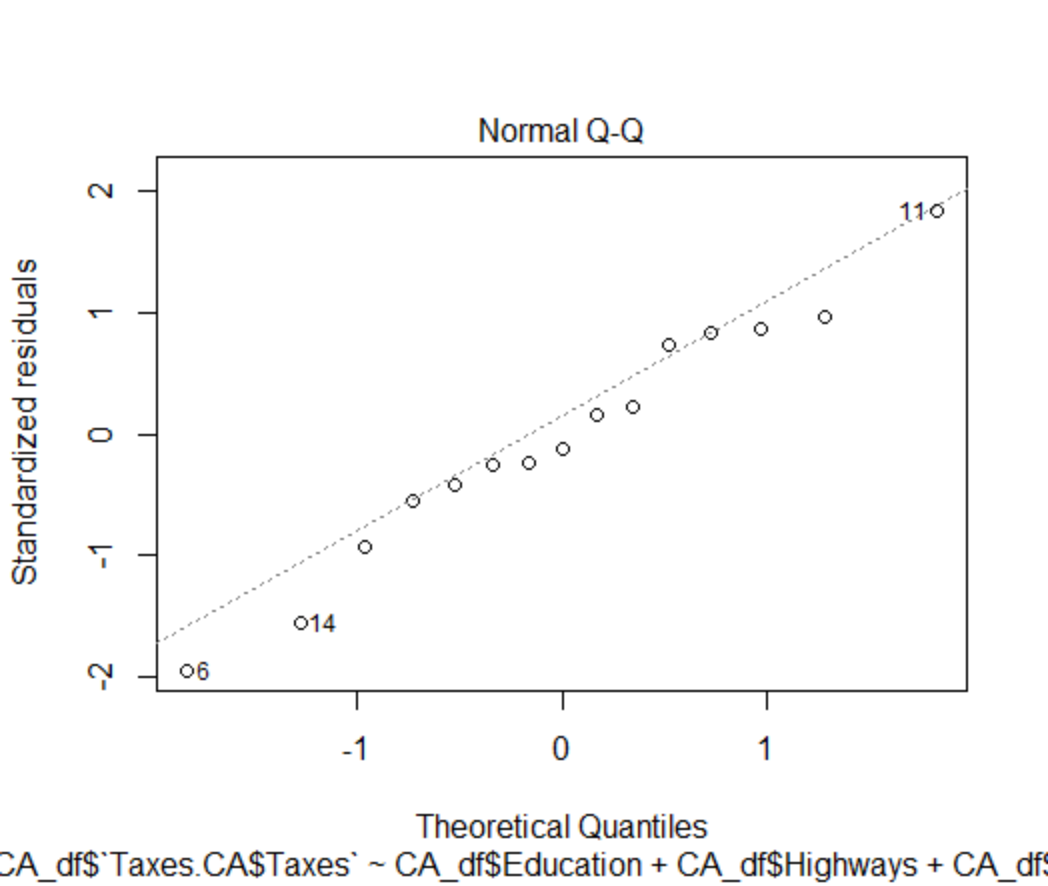
Utilizing this method gave us an exponent value (λ) of 1.04, which we use to transform the response. After applying the transformation the transformed residual vs fit plot can be seen below.



perform residual analyses,



The new plot is nearly identical in form to the previous, though the magnitudes have been increased, the trend is still near zero. This graph satisfies the independent error terms and equal variance of error or the ‘I’ and ‘E’ terms from LINE. Next we will look at a normal Q-Q plot to show normality or the ‘N’ from LINE. Serial correlation residual vs order plot could be used not sure we want to do it though.

The Normal Q-Q plots above (left: transformed, right: best subset model) which is a plot of theoretical quantiles of a normal distribution versus our sample quantile, shows that our plot is not perfectly normal but still within a satisfactory range to be used.

Can you think of any way to fix this or make it more linear? If you feel like trying you could remove the outlier points and reproduce the model summary. That’s the only way that I can think of that might fix it, but I think it would remove too many data points.

**Findings & Final Model**

propose a final model,

With our LINE conditions satisfied by our model, our predictor taxes, transformed by our lamda (λ) transformation, is predicted by spending on education, highways, and police protection. This model is our final model to be used in answering our research question.

> CA.transformed <- lm((((CA\_df$`Taxes.CA$Taxes`^lmda\_CA)-1)/lmda\_CA) ~ CA\_df$Education + CA\_df$Highways + CA\_df$`Police protection`)

Call:

lm(formula = (((CA\_df$`Taxes.CA$Taxes`^lmda\_CA) - 1)/lmda\_CA) ~

CA\_df$Education + CA\_df$Highways + CA\_df$`Police protection`)

Residuals:

Min 1Q Median 3Q Max

-14946650 -2983261 -884681 4843519 17803244

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -9.170e+07 2.162e+07 -4.242 0.00171 \*\*

CA\_df$Education 5.446e+00 4.957e-01 10.986 6.67e-07 \*\*\*

CA\_df$Highways 3.955e+00 1.432e+00 2.761 0.02010 \*

CA\_df$`Police protection` -7.821e+01 2.881e+01 -2.714 0.02177 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9727000 on 10 degrees of freedom

Multiple R-squared: 0.9803, Adjusted R-squared: 0.9743

F-statistic: 165.6 on 3 and 10 DF, p-value: 8.037e-09

answer research questions.

As can be seen above, all of our predictors are significant to a 0.05 alpha or 95% confidence, with a coefficient of determination suggesting 98.03% of all variation of CA taxes is explained by looking at just these three response variables. Intuitively this is likely to be an error as state economics are unlikely to be able to be this simplified with that level of accuracy. We believe the normality of error is the culprit as seen by the QQ plot.

Of the predictors, spending on education can be seen as the most significant by its relatively large t-value and vastly smaller p-value. California is known for its large education system which is the largest employer in the state, and includes our own UC system as well as many other state funded schools. This, as well as California’s economic support of digital technology, boasting prominent hubs such as Silicon Valley, which in-turn demands a skilled labor force, thus attracting educated residents and employing the qualified populous, and providing a boost to the economy of the state, thus effecting taxes.

The spending on highways is the next most significant predictor in the model, as seen similarly by its t-value and p-value. Roads are a vital part of economic progress. To provide some context, according to the American Trucking Association, trucking revenues in the US were $700 billion in 2017, with trucks alone moving just under 11 billion tons of freight. If roads are of low quality it may take longer to transport goods and services, which would damage economic output, and therefore performance, influencing taxes. At an anecdotal level this does make logical sense.

The only spending category that is significant to the model and has a negative slope estimate is police protection. While it is the least significant of the predictors in the model, the negative impact on taxes is still of interest in light of current political discussions and protests.

These findings could suggest that it actually damages the economy when states fund the police force. Possible reasons for this negative economic impact could range from police abuses targeting certain communities and not allowing them to economically flourish, wrongful imprisonments, and recidivism within the criminal corrections system which can turn an otherwise economically productive citizen into a captive of the state leading to economic drain.

NOTE: I used this same process on 4 other states, only 2 showed a negative sloped variable as significant, corrections (prison costs) and public welfare both in the state of Texas.

**Conclusion**

In conclusion, these findings do provide interesting insights into the effects of fiscal policy, such as education being a significant positive growth factor in California, and that police protection expenditure impacts tax growth negatively in California, which corroborates current political discussions. To draw a formal conclusion however, we have determined that, despite the interesting results, there is simply not enough annual data for the model to be as accurate as is necessary for making any strong claims.

**Appendix**

Variable definitions - <https://www.census.gov/govs/definitions/index.html>

Data Source - <https://www.census.gov/programs-surveys/state/data/tables.All.html>

Code:

require("readxl")

require("dplyr")

library(tidyverse)

getwd()

setwd('S:/Sams things/homework/UCSB 2018-/PSTAT 126- regression analysis/Group project/states2/states/data')

filenames <- list.files()

filenames # should see the name of each year of data

#### function to get variables ###

get\_variable <- function(var,area){

x <- data.frame()

count = 0

for(year in 2003:2011)

{

count = count + 1

data <- read\_excel(filenames[count])

index <- which(data$Item == var)

if(length(index)!=1)

index = index[1]

val <- data[index,toupper(area)]

new <- data.frame(year = year, `var` = as.numeric(val))

x <- bind\_rows(x,new)

}

for(year in 2012:2018)

{

count = count + 1

data <- read\_excel(filenames[count])

index <- which(data$`(Thousands of Dollars)` == var)

if(length(index)!=1)

index = index[1]

val <- data[index,area]

new <- data.frame(year = year, var = as.numeric(val))

x <- bind\_rows(x,new)

}

names(x)[2] <- var

return(x)

}

data <- read\_excel(filenames[1])

data['Item']

get\_state <- function(state){

print(filenames)

}

######################################################################

######## Define Statistics to be used ################################

######################################################################

## Definitions: https://www.census.gov/govs/definitions/index.html ###

######### Potential Predictors CA: Forms of income/growth ###########

Total\_revenue.CA <-get\_variable('Total revenue','California')

Taxes.CA <- get\_variable('Taxes','California')

Ind\_inc\_tax.CA <- get\_variable('Individual income tax', 'California')

Corp\_inc\_tax.CA <- get\_variable('Corporate income tax', 'California')

Sales\_tax\_gen.CA <- get\_variable('General sales', 'California')

Sales\_tax\_sel.CA <- get\_variable('Selective sales', 'California')

Sales\_tax\_net.CA <- cbind(Sales\_tax\_gen.CA[1], (Sales\_tax\_gen.CA[2] + Sales\_tax\_sel.CA[2])) # Net sales tax

intergov\_Rev.CA <- get\_variable('Intergovernmental revenue', 'California')

## Potenial Response Variables CA: Forms of spending/inv. ##########

Total\_expenditure.CA <- get\_variable('Total expenditure','California')

intergov\_exp.CA <- get\_variable('Intergovernmental expenditure', 'California')

Dir\_exp.CA <- get\_variable('Direct expenditure', 'California') # All expenditure - intergov

Education.CA <- get\_variable('Education', 'California')

Welfare.CA <- get\_variable('Public welfare', 'California')

Hospitals.CA <- get\_variable('Hospitals', 'California')

Health.CA <- get\_variable('Health', 'California')

Highways.CA <- get\_variable('Highways', 'California')

Police.CA <- get\_variable('Police protection', 'California')

Prisons.CA <- get\_variable('Correction', 'California')

Nat\_resources.CA <- get\_variable('Natural resources', 'California')

Parks.CA <- get\_variable('Parks and recreation', 'California')

Debt.CA <- get\_variable('Debt at end of fiscal year', 'California')

Cash.CA <- get\_variable('Cash and security holdings', 'California')

Net\_cash.CA <- cbind(Cash.CA[1], (Cash.CA[2] - Debt.CA[2])) # Net Cash

Ann\_Def.CA <- cbind(Total\_expenditure.CA[1], (Total\_revenue.CA[2] - Total\_expenditure.CA[2]))

#####################################################################

### Statistics that have been created ###########################

# annual deficit ==> Total Rev - Total exp

# Net cash position ==> Cash and security holdings - Debt at end of fiscal year

# Income tax net ==> Individual income tax + corporate income tax

#####################################################################

#####################################################################

#####################################################################

#### Research questions - Proposed Response #########################

#####################################################################

# Growth of states income - Total rev or Taxes (doesn't include money flow between states and Fed)

# Growth of Cash position - Net\_cash

################# Taxes final decision ##############################

######################################################################

######################################################################

######################################################################

## scatter plot matrix

pairs(~CA\_df$`Taxes.CA$Taxes` + CA\_df$`Cash and security holdings` + CA\_df$`Total expenditure`

+ CA\_df$`Direct expenditure`)

pairs(~CA\_df$`Taxes.CA$Taxes` + CA\_df$`Public welfare` + CA\_df$`Police protection`

+ CA\_df$Hospitals )

pairs(~CA\_df$`Taxes.CA$Taxes` + CA\_df$Education + CA\_df$Correction + CA\_df$Highways)

######################################################################

######################################################################

######################################################################

### Model for Growth using taxes: CA #################################

CA\_df <- cbind(Total\_expenditure.CA, Dir\_exp.CA[2], Net\_cash.CA[2], Ann\_Def.CA[2], Education.CA[2],

Prisons.CA[2], Welfare.CA[2], Police.CA[2], Prisons.CA[2], Hospitals.CA[2], Hospitals.CA[2],

Highways.CA[2], Taxes.CA$Taxes )

CA\_df

# remove 2007 from model

CA\_df <- CA\_df[-c(5),]

CA\_df

################### Step-wise regression #############################

Model\_growth.CA <- lm(CA\_df$`Taxes.CA$Taxes` ~ 1) # base model is fully reduced

### use add1() to iterate through predictors

add1(Model\_growth.CA, ~. + CA\_df$`Cash and security holdings` + CA\_df$`Total expenditure`

+ CA\_df$`Direct expenditure` + CA\_df$Education + CA\_df$Correction

+ CA\_df$`Public welfare` + CA\_df$`Police protection`

+ CA\_df$Hospitals + CA\_df$Highways, test = 'F' ) # Education most sig

#### shows which variable to add next: notes 12 #######################

#### update model ####

Model\_growth.CA1 <- update(Model\_growth.CA, ~.+ CA\_df$Education) # add best predictor from initial pass

################################# test which predictor to add next ##

add1(Model\_growth.CA1, ~. + CA\_df$`Cash and security holdings` + CA\_df$`Total expenditure`

+ CA\_df$`Direct expenditure` + CA\_df$Correction

+ CA\_df$`Public welfare` + CA\_df$`Police protection`

+ CA\_df$Hospitals + CA\_df$Highways, test = 'F' ) # flags Highways as the only sig stat.

################################ update model ########################

Model\_growth.CA2 <- update(Model\_growth.CA1, ~. + CA\_df$Highways)

############################# Check all predictors still sig #########

summary(Model\_growth.CA2) # all sig

############################ Check for next variable #################

add1(Model\_growth.CA2, ~. + CA\_df$`Cash and security holdings` + CA\_df$`Total expenditure`

+ CA\_df$`Direct expenditure` + CA\_df$Correction

+ CA\_df$`Public welfare` + CA\_df$`Police protection`

+ CA\_df$Hospitals , test = 'F' ) # police protection

############################# Update model ###########################

Model\_growth.CA3 <- update(Model\_growth.CA2, ~. + CA\_df$`Police protection`)

########################## Check significance of predictors in model #

summary(Model\_growth.CA3) #### all sig continue iterations

add1(Model\_growth.CA3, ~. + CA\_df$`Cash and security holdings` + CA\_df$`Total expenditure`

+ CA\_df$`Direct expenditure` + CA\_df$Correction

+ CA\_df$`Public welfare`

+ CA\_df$Hospitals , test = 'F' ) # None

############################# Final model ###########################

Model\_growth.CA3

######################################################################

######################################################################

######################################################################

######################## Akaike's Information Criterion ##############

######################################################################

###################### CA ############################################

CA0 <- lm(CA\_df$`Taxes.CA$Taxes` ~ 1) # base

CA.upper <- lm( CA\_df$`Taxes.CA$Taxes` ~ + CA\_df$`Cash and security holdings` + CA\_df$`Total expenditure`

+ CA\_df$`Direct expenditure` + CA\_df$Education + CA\_df$Correction

+ CA\_df$`Public welfare` + CA\_df$`Police protection`

+ CA\_df$Hospitals + CA\_df$Highways )

rstudent(CA.upper)

dffits(CA.upper)

########################### Identify subsets ##########################

step(CA0, scope = list(lower = CA0, upper = CA.upper))

############################ Final Model ##############################

CA1 <- lm(CA\_df$`Taxes.CA$Taxes` ~ CA\_df$Education + CA\_df$Highways +

CA\_df$`Police protection`) # same model as previous

######################################################################

######################################################################

######################################################################

############################ Best Subset regression #################

install.packages('leaps')

library(leaps)

######################################################################

################################## CA ################################

CA\_predictors <- subset(CA\_df, select = -c(`Taxes.CA$Taxes`, year))

CA\_predictors

CA\_df

################### regsubsets() ####################################

CA\_bsr <- regsubsets(CA\_predictors, CA\_df$`Taxes.CA$Taxes`)

################################ Summary and results #################

summary.CA\_bsr <- summary(CA\_bsr)

summary.CA\_bsr$which

summary.CA\_bsr$rsq # second jump is the last significant jump, indicating a 3 predictor model as best

summary.CA\_bsr$adjr2 # same as above ( in findings)

# no significant jumps anywhere, best subset with 3 subsets is the same as the previous

########################## Model #####################################

CA\_bsr\_fit <- lm(CA\_df$`Taxes.CA$Taxes` ~ CA\_df$Education + CA\_df$Highways + CA\_df$`Police protection`)

rstudent(CA\_bsr\_fit)

dffits()

######################################################################

######################################################################

########################### Best subset models #######################

############################## rsq + adjr2 ###########################

############ CA ###

CA\_bsr\_fit # education, highways, police protection

######################### Step wise regression models ################

######################### F - test ###################################

Model\_growth.CA5 # Education, police protection, Highways

############################# AIC - test ############################

CA1 # Education, police protection, Highways

########################### Plots ####################################

plot(CA1) # residuals seem alright, Normal Q-Q is questionable

#################### Box cox Transform ###############################

library(MASS)

library(car)

box\_CA1 <- boxcox(CA\_df$`Taxes.CA$Taxes` ~ Education + Highways + `Police protection`, lambda = seq(-6,6,0.01), data = CA\_df)

box\_CA1

cox\_CA <- data.frame(box\_CA1$x, box\_CA1$y)

cox\_CA

which.max(cox\_CA$box\_CA1.y) # 705 index

lmda\_CA <- cox\_CA[705,]

lmda\_CA <- lmda\_CA$box\_CA1.x

CA.transformed <- lm((((CA\_df$`Taxes.CA$Taxes`^lmda\_CA)-1)/lmda\_CA) ~ CA\_df$Education + CA\_df$Highways + CA\_df$`Police protection`)

plot(CA.transformed)