Emma Sainovic

DiDa 130

Final Project

Introduction : My data frame displays information about how and where the government spends or receives money in the United States between the years 1992 and 2019 for each US state. I created a regression model to determine which variables had the greatest impact on the total expenditure of New York state, if any, and it turned out to be public education. I created many visualizations to show how significant of a variable education was with respect to total expenditure.

Background information : Provided by taxpayers, the government spends money on different programs, goods and services for the people of the United States. According to an article from FiscalData, the main two categories are called mandatory spending and discretionary budget.[1] Mandatory spending are things the government must always pay for by law and includes categories such as medicare and social security, while discretionary budget is for programs that need to be approved by case and in a timely manner before money is provided such as for education or affordable housing.[2] There is also another main category that occurs less frequently called supplemental spending which says the government may provide funding in emergency situations without a formal approval, such as back in 2020 when the government needed to spend more money because of the Covid-19 outbreak. [3]

From an article from TaxFoundation, they found that in general, less taxing is more beneficial. For instance, something that they found was that when taxes are less progressive, meaning taxing the rich less, leads to an increase in wages for every economic class.[4] They also found that if the government were to charge less taxes then unemployment may decrease with it.[5] The article lists several points on why

[1] "Fiscal Data Explains Federal Spending"

[2] Ibid

[3] Ibid

[4] Vermeer

[5] Ibid

charging taxpayers less taxes would do more good than harm.

Description of data : The data frame I am working with is titled “finance” and was created by Austin Cory Bart in 2021. It contains 31 columns which contain information such as state names, years, revenue, expenditure,and tax information. The data frame has information for every state from the year 1992 to the year 2019.

These are the columns I will be using from the data frame (Note: the numbers provided in each of these columns are per state per year) :

- ‘Details.Health.Health Total Expenditure’ which provides the total amount spent on public health services such as school health or water pollution control.
- ‘Details.Education.Education Total’ which provides the total amount spent on public educational services such as public colleges including special classes such as blind or deaf educational institutions
- ‘Details.Natural Resources.Parks.Parks Total Expenditure’ which provides the amount spent on recreational activities such as public beaches, playground, or play fields
- ‘Details.Correction.Correction Total’ which provides the amount spent on correctional purposes, where only prison is mentioned in the details
- ‘Totals.Expenditure’ which is the states total money paid by the government
- ‘Totals.Tax’ which is the number of taxes paid to the state

Research Question **:** In New York State, how do ‘Details.Health.Health Total Expenditure’, ‘Details.Education.Education Total’, ‘Details.Natural Resources.Parks.Parks Total Expenditure’ and ‘Details.Correction.Correction Total’ impact the ‘Totals.Expenditure’ between 1992 and 2019?

To find the results of my question, I will create a regression model. This is the **equation** :

formula = Totals.Expenditure ~ Details.Health.Health.Total.Expenditure

+ Details.Education.Education.Total

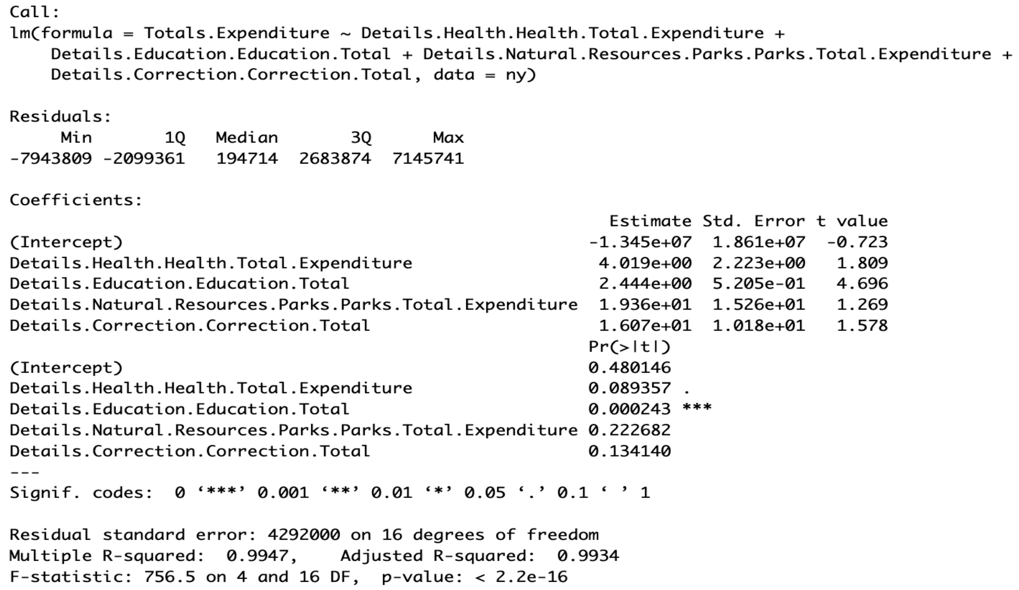
+ Details.Natural.Resources.Parks.Parks.Total.Expenditure

+ Details.Correction.Correction.Total

Null Hypothesis : There is no significant relationship between the dependent variable and the independent variables.

Alternative Hypothesis : There is a significant relationship between the dependent variable and the independent variables

Results of the Regression Model are shown below :



```
Call:
lm(formula = Totals.Expenditure ~ Details.Health.Health.Total.Expenditure +
    Details.Education.Education.Total + Details.Natural.Resources.Parks.Parks.Total.Expenditure +
    Details.Correction.Correction.Total, data = ny)

Residuals:
     Min       1Q   Median       3Q      Max
-7943809 -2099361   194714  2683874  7145741

Coefficients:
                                                         Estimate Std. Error t value
(Intercept)                                            -1.345e+07  1.861e+07  -0.723
Details.Health.Health.Total.Expenditure                 4.019e+00  2.223e+00   1.809
Details.Education.Education.Total                       2.444e+00  5.205e-01   4.696
Details.Natural.Resources.Parks.Parks.Total.Expenditure  1.936e+01  1.526e+01   1.269
Details.Correction.Correction.Total                     1.607e+01  1.018e+01   1.578
                                                       Pr(>|t|)
(Intercept)                                            0.480146
Details.Health.Health.Total.Expenditure                0.089357 .
Details.Education.Education.Total                      0.000243 ***
Details.Natural.Resources.Parks.Parks.Total.Expenditure 0.222682
Details.Correction.Correction.Total                    0.134140
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4292000 on 16 degrees of freedom
Multiple R-squared:  0.9947,    Adjusted R-squared:  0.9934
F-statistic: 756.5 on 4 and 16 DF,  p-value: < 2.2e-16
```

Results : Y = -1.345e + 4.019e(x1) + 2.444e(x2) + 1.936e(x3) + 1.607e(x4)

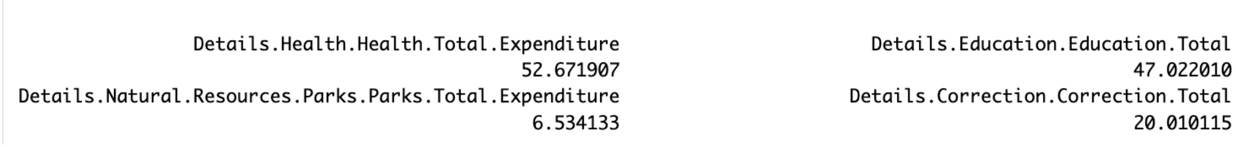
The p-value of the F-statistic is <2.2e^-16, which is much smaller than 0.05, which means that the result is highly significant and the null hypothesis is rejected. That means that at least one of the independent variables has a significant effect on the dependent variable. The column that holds the amount of tax money spent on education has an extremely small p-value, as well as three stars next to it, so we can conclude that this is the independent column with the largest effect on the dependent variable, total expenditure. This means that most of the money spent by the New York state government between 1992 and 2019 is most likely the education department.

Checking assumptions of the Regression Model

Normality : A W-value of 0.98986 indicates that the residuals are nearly perfectly distributed.

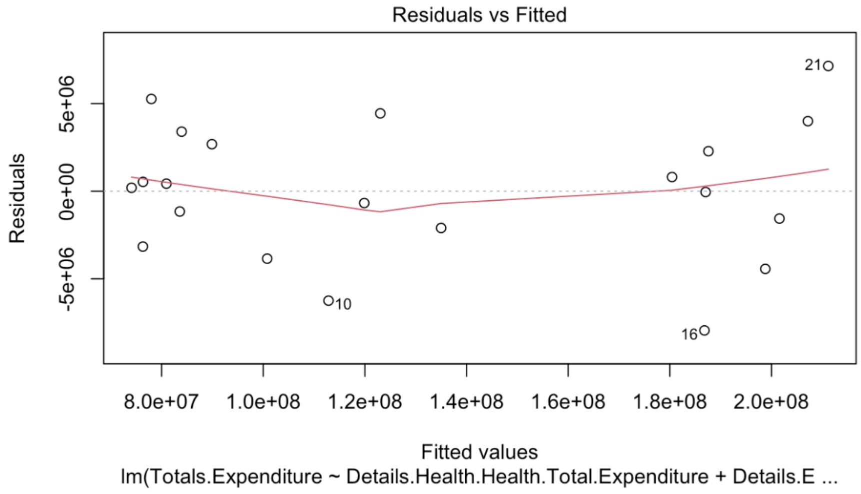
Homogeneity of Variance : The p-value of the variance score test is 0.21595, which is greater than 0.05, which means that the result is not significant and I fail to reject the null hypothesis. So, I know that the model is not violating the assumption of homogeneity of variance.

Multicollinearity : I used the vif() function to check for the linearity between the variables. The values are quite high which suggests that these variables may be too correlated with each other. However, this was expected to happen because each of the columns is describing the tax dollars spent on different categories by NY state.

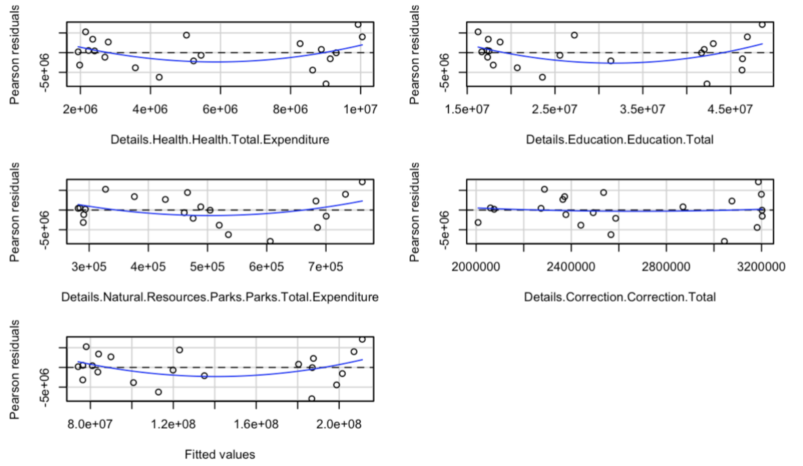


```
          Details.Health.Health.Total.Expenditure                 Details.Education.Education.Total
                                        52.671907                                         47.022010
Details.Natural.Resources.Parks.Parks.Total.Expenditure         Details.Correction.Correction.Total
                                         6.534133                                         20.010115
```

Linearity : I used the plot() function to check the linearity of residuals. The red line in the Residuals vs Fitted graph below indicates that there may be an issue because this line is not perfectly straight.

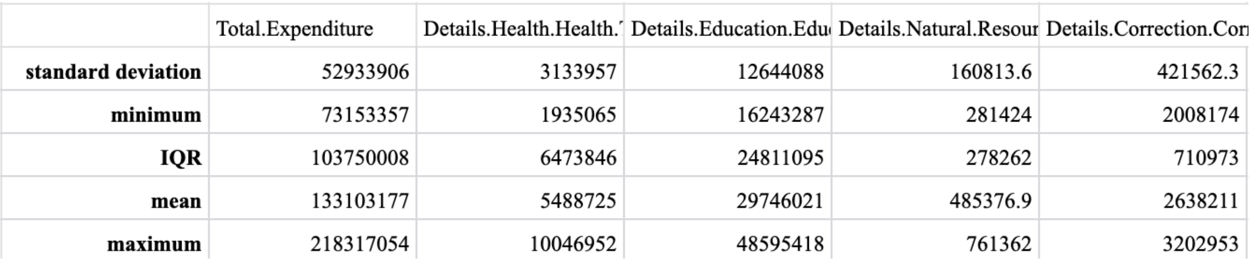


To investigate linearity further, I used the residualPlots() function to see the residuals relationship with each of the variables in the model.

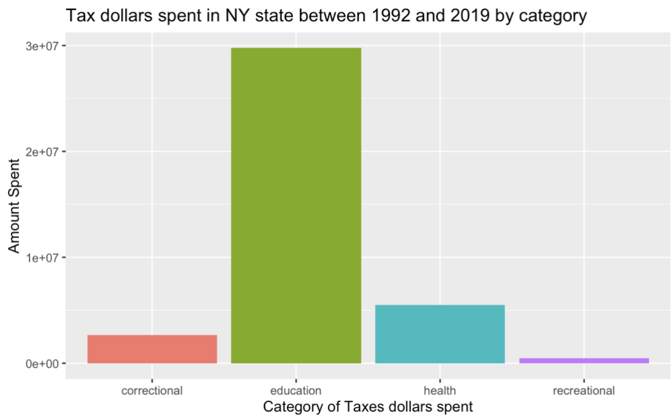


The plots indicate that some of the variabls may be causing a problem with the model. So, the linearity of the model could be better than it is.

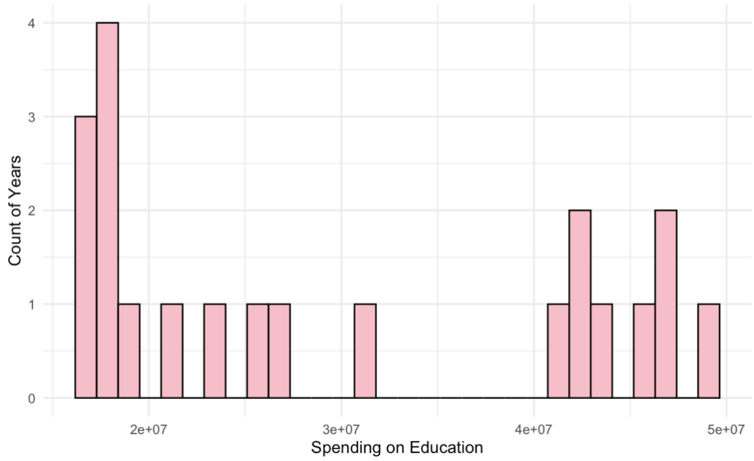
Statistics Summary Table : The table shows the standard deviation, minimum, IQR, mean, and maximum values of each of the variables used in the model.



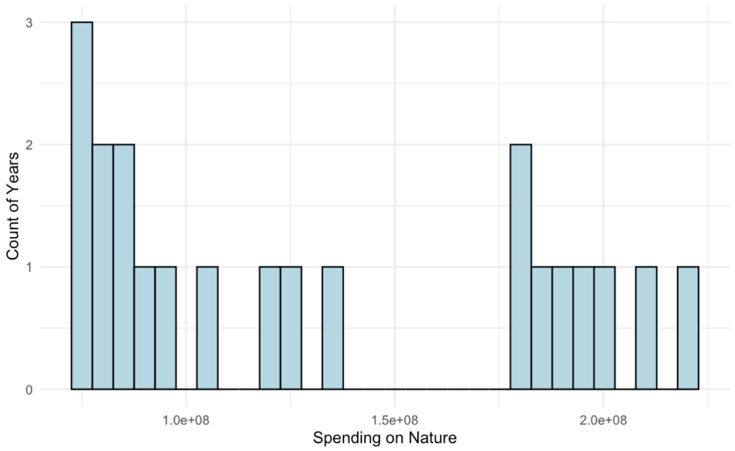
	Total.Expenditure	Details.Health.Health.	Details.Education.Edu	Details.Natural.Resour	Details.Correction.Cor
standard deviation	52933906	3133957	12644088	160813.6	421562.3
minimum	73153357	1935065	16243287	281424	2008174
IQR	103750008	6473846	24811095	278262	710973
mean	133103177	5488725	29746021	485376.9	2638211
maximum	218317054	10046952	48595418	761362	3202953



Bar graph : The bar graph below shows the mean amount of dollars spent in New York state on the y-axis between 1992 and 2019 per category shown on the x-axis. As I had predicted earlier from the model itself, it seems that New York spent most of taxpayer money on public education in those years compared to other places they spent money. The green bar, ‘education’, is significantly taller than the bars for correctional facilities, health programs, and nature resources.

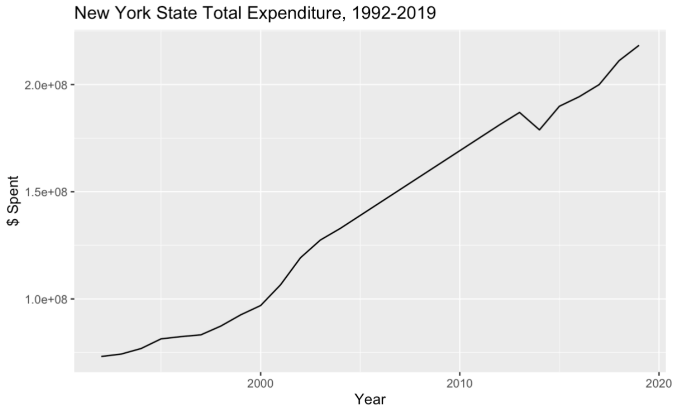


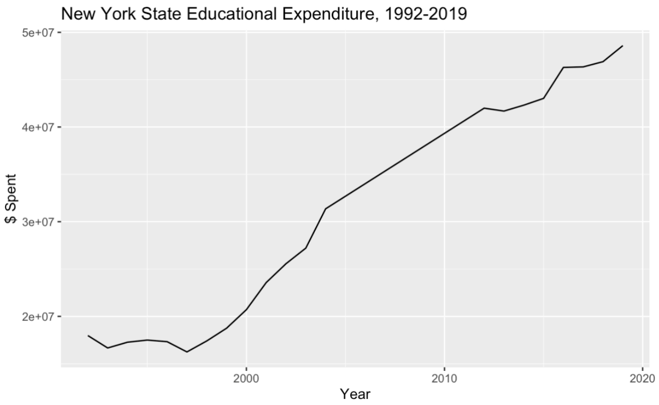
Histogram 1 (pink bars) : The histogram below shows the ranges of taxpayer money spent on education in New York, the x-axis and the number of times, or amount of years, that this much money was spent. There is a noticeably large amount of years, the second bar, where New York spent a certain amount of money that seems to be on the lower end. Then there is a large gap in the middle indicating they increased spending by a large amount, but not for many years because the bar is shorter. This may tell us that New York was possibly seeing low test scores, high dropout rates, low graduation rates, or something else that indicates that not enough funding was put into education, so they decided after many years to put more funding into education.



Histogram 2 (blue bars) : The histogram below shows the range of the total amount of taxpayer tax money spent on the x-axis and the number of years where they spent an amount within that range on the y-axis. The first and tallest bar shows that the lowest total amount they ever spent was similar during three different years. This bar is followed by several shorter bars which indicates that the amount spent between 1992 and 2019 fluctuates quite a bit.

Line graphs :


New York State Total Expenditure, 1992-2019


New York State Educational Expenditure, 1992-2019

The line graph on the left represents the total expenditure in New York State from 1992 to 2019, and the line graph on the right represents the educational expenditure in New York state from 19921 to 2019. The general direction of both graphs is a positive upward trend, with a small difference in the earlier years where the right graph has a bit of a dip. This similarity further shows that a large portion of the total expenditure is due to educational expenditure because their graphs are similar within the same time frame. I couldn't come to this conclusion from the histograms alone because I didn't know which years were being represented by the bars, only the amount.

Conclusion : I have rejected the Null Hypothesis based on the p-value of the regression model as well as the graph visuals. I have concluded that the total amount of tax money New York state spends every year is mainly due to the amount being spent for educational purposes. The other columns representing tax money spent on recreational activities and areas, correctional programs, and health programs, had somewhat of an effect on the independent variable, total expenditure, however, education expenditure exceeded these other variables by a significant amount. According to the articles mentioned in the background information, it looks like the government spending in this time span was mainly in the discretionary budget.

Work Cited

Source 1 : “*Fiscal Data Explains Federal Spending*.” Fiscaldata.treasury.gov, fiscaldata.treasury.gov/americas-finance-guide/federal-spending/#the-difference-between-mandatory-discretionary-and-supplemental-spending.

Source 2 : Vermeer, Timothy. “The Impact of Individual Income Tax Changes on Economic Growth.” *Tax Foundation*, 14 June 2022, taxfoundation.org/research/all/state/income-taxes-affect-economy/.