DS100 - Final Project

Univeristy of California Santa Barbara

U.S. Educational Finances

Author:
Jason Yu

Instructor:
Professor Alexander Franks

Instructor:
Professor Yekaterina Kharitonova

June 12, 2020

Chapter 1

Introduction

1.1 Abstract

Education is an important foundation for many Americans, and funding for schools is a big aspect of this. It is not only important that this is crucial for students, but also for the future. Therefore, the revenue and distribution of profit from schools in all 50 states is imperative to address the inequalities of education funding in certain states. In this project, the use of translations, visualization, and linear regression allowed me to conclude that school spending on support services and other beneficial services will not only result in happier students, but also increased revenue. This project shows how important education funding is for the United States.

1.2 Introduction

The allocation of money circulating in the United States elementary and middle schools affects almost all Americans. Funding education takes part in a huge role in the outcomes of success, for not only the students but also the future. Many people believe that students tend to thrive in better funded schools than less funded schools. Education should be at a level playing field for all students by distributing an equal amount of funding between the 50 states (Biddle 2002). Thus, the amount of given revenue and generated profit is important to determine the changes in specific states that will need to partake, in order to achieve maximum success in United States schools. The data set used in this analysis contains the financial information surveyed from all public schools across the country. Data includes different categories of revenue, expenditures, number of students enrolled, the state that was surveyed, and the year that the survey was conducted.

1.3 Questions of Interest

1.3.1 Question 1:

What is a multilinear regression model that can predict total revenue of a school, and how accurate is it? Which predictors should be increased to increase total revenue?

1.3.2 Question 2

Can we use principal component analysis to reduce the number of dimensions while still having an accurate model?

1.3.3 Question 3

What is the general trend of school revenue over time?

1.3.4 Question4

Can we predict whether a school is profitable or not based on its spending?

Chapter 2

Data and Methods

2.1 Data

The data come from the United States Census Bureau but the data set used in the analysis was cleaned by Roy Garrard on kaggle.com. The Census Bureau allows others to use their data.
When selecting variables, we were under the objective of trying to predict the effect different spending had on the total revenue of the school. For this reason, I dropped the State and total_expenditure columns but kept all other numerical variables. Four of my predictors represent four different forms of spending: infrastructure, support services, capital, and other. I also included enrollment to prevent my model from under representing smaller schools or overemphasizing bigger schools.
We performed multiple operations on the data. The data set was missing enrollment for all states during 1992 so I dropped all the observations belonging to that year. I converted all the column names into lowercase for better readability I created two new columns: profit and profitable profit was obtained by subtracting `total_expenditure` and `total_revenue`. `profit` was obtained by casting a Boolean mask of whether `profit` was greater than 0.

2.1.1 Relevant Variables

- `total_revenue` - Total Elementary-Secondary Revenue, Target Variable
- `year` - Year data was collected
- `enroll` - Number of Students enrolled
- `infrastructure_expenditure` - Total Spending For Instruction
- `support_services_expenditure` - Total Spending for Support Services
- `other_expenditure` - Total Spending for Other Programs
- `capital_outlay_expenditure` - Total Capital Outlay Expenditure
- `profit` - Difference between Total Revenue and Total Expenditure
- `profitable` - Whether or not Profits are greater than 0

2.1.2 Principles of measurement

- Relevance:
The data is relevant because it captures the behavior spending habits of schools that lead to higher revenue. One problem, however, may be that the data does not account for the wealth of the area the schools are in or the academic performance of the kids in the area.
- Precision:
Precision is not an issue with this data all the observations are submitted in reasonably small units. Dollar values are given to the nearest dollar. enrollment counted by each individual student, and the dates are shown by the year of the survey.
- Non-Distorting:
Measuring the revenue and expenditure should not distort the data because it does not affect the money distribution or behavior of the schools in any way.
- Cost: Cost is not an issue as most of the data is collected electronically. In addition, the manual work is paid for by the census bureau.

2.1.3 Ethical Considerations

The United States Census Bureau conducts annual surveys to assess the finances of elementary and high schools. The analysis of the dataset should not harm any of the schools or states represented in any significant way. This dataset represents every public school in the U.S. Older schools may be overrepresented because newer schools have had less data collected over the years. Private schools are also not represented in this dataset. Since the Census Bureau collects data from every school system, there should not be any sampling bias. There could be a possibility that some schools do not report their finances accurately which would add a bias to our results.

2.2 Methods

2.2.1 Question 1 - What is a multi linear regression model that can predict total revenue of a school, and how accurate is it? Which predictors should be increased to increase total revenue?

1. To obtain a regression model for this question, we will use year, enroll, `infrastructure_expenditure`, `support_services_expenditure`, `capital_overlay_expenditure`, and `other_expenditure` as predictor variables and `total_revenue` as the target variable.
2. We fit the regression model using the scikit-learn `LinearRegression()` function and obtain the multi linear regression model for this question.
3. After, we observe the coefficients to determine which predictors have the largest positive values for their slope coefficients. This shows us which predictors are indicators of higher total revenue.
4. To determine how accurate our model is, we will inspect the residual plot to see how closely the model's predictions compare to the actual predictions. Prior to running the regression, I split the data into a training and testing data set to ensure proper comparisons.

2.2.2 Question 2- Can we use principal component analysis to reduce the number of dimensions while still having an accurate model?

1. To perform PCA, we will use scikit-learn's `PCA` package. First, we drop the state column as PCA does not like qualitative variables. With the remaining data, we use the StandardScaler() function to normalize my data so that the mean and variance are 0 and 1 respectively.
2. We then use the `PCA()` function to reduce my predictors to k number of components.
3. After fitting my model with the reduced data, we calculate the RMSE value of the original model and the new model to compare the loss values. We will then adjust the number of components to obtain a model that has at least 90% of the cost of the original mode. 4. After obtaining an appropriately reduced model, we will analyze the residual plot of reduced model and determine whether PCA can be used.

2.2.3 Question 3: What is the general trend of school revenue over time?

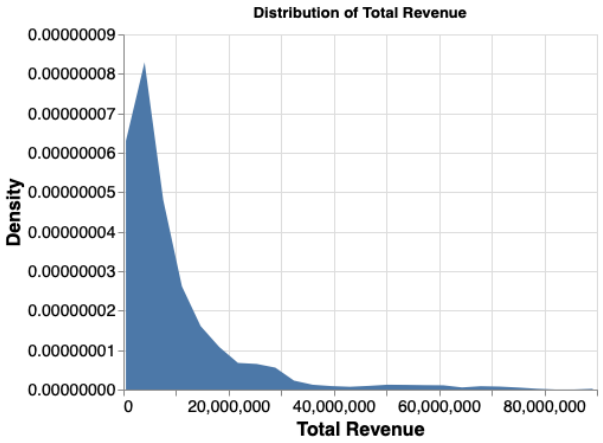
To observe the trend of total revenue for each state over time, we will construct a line plot of total revenue (`total_revenue`) and time (`year`).

2.2.4 Question 4: Is there an uneven distribution of profit amongst the schools in the United States?

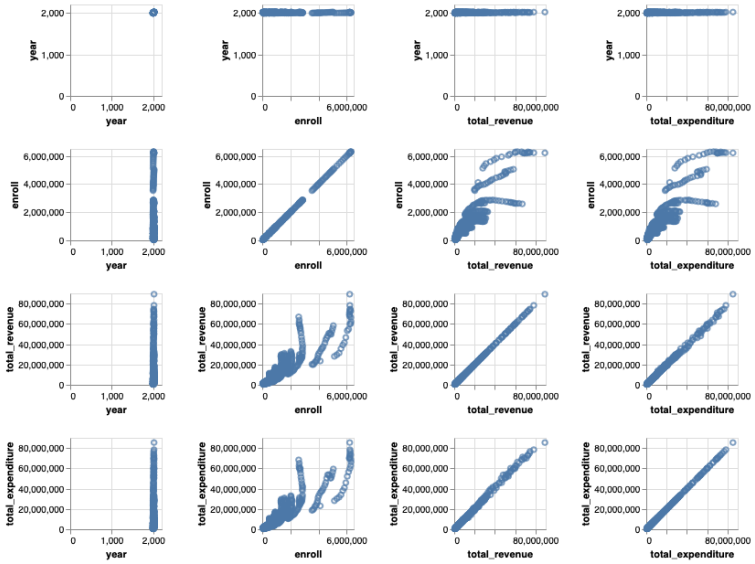
To visualize the distribution of profit amongst all schools, I make a histogram of the profit in my data set.

2.3 EDA

To begin, we create a density plot of total revenue to inspect how it is distributed. The distribution of revenue looks right skewed, which makes sense as most schools should be obtaining similar amounts of money with a fewer amount which have more unique circumstances.



Then, we construct a pairs plot to inspect if there are any relationships between year, enroll, total revenue, and total expenditure. Year has no real linear relationship with any other of the predictors. Enroll has a somewhat, but not strong linear relationship with total revenue and total expenditure. The most glaring observation is that total revenue and total expenditure have a very strong linear relationship.



2.3.1 Inferential or Predictive Methods/Models

In order to answer my first question of interest, I used linear regression in order to create a model that predicts a school's total revenue given its year, enrollment, and spending allocation.

Chapter 3

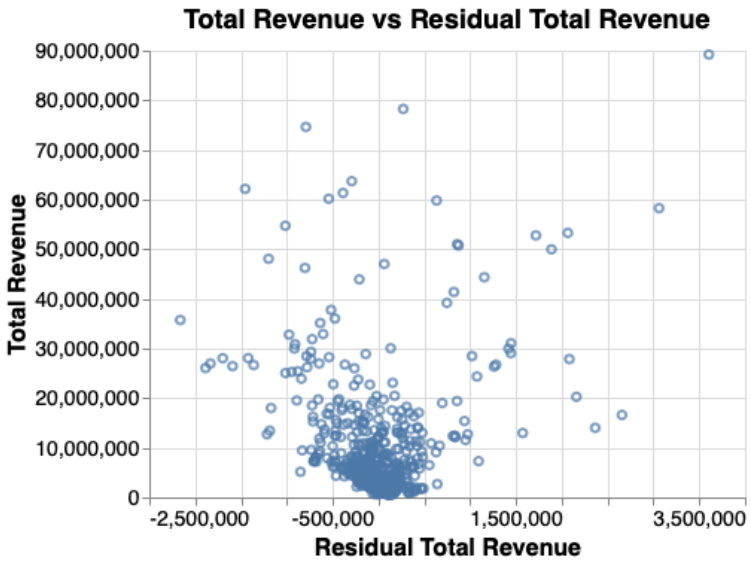
Analysis, Results, and Interpretation

3.1 Question 1: What is a multi linear regression model that can predict total revenue of a school, and how accurate is it? Which predictors should be increased to increase total revenue?

3.1.1 Interpretations

After obtaining the model, we observe the coefficients to notice any significant indicators of higher revenue.
Taking a look at the model's coefficients, `support_services_expenditure`, `other_expenditure`, and year seem to be indicators for the increase of `total_revenue`. Their coefficients are 1.5453, 1.9423, and -1058. This means that for about every $1 increase in support services expenditure, there is $1.5453 increase in total revenue. Similarly, every $1 increase in other expenditure will see about a $1.94 increase in total revenue. For every year increase, we expected a -1058 loss in revenue.
The `other_expenditure` predictor seemed interesting as it had the highest slope coefficient while being ambiguous in its name. Upon further research, we learned that the other expenditures represent spending outside of education purposes, such as food services, enterprise operations, and other benefits. This would perhaps be an interesting investigation to see which specific type of other expenditure is playing the biggest role in increasing the revenue.
At the beginning of the analysis, instructional expenditure was thought to be an important predictor. However, it was very insignificant in how it affected the response as its slope coefficient was only 0.7936 slope coefficient. This was surprising as it indicated that more spending for teachers resulted in less revenue.

To determine how accurate our model is, we will inspect the residual plot to see how closely the model's predictions compare to the actual predictions.
From this residual plot we can say that the model fits our data fairly well. The majority of points are clustered around 0 for Residual Total Expenditure and we do not see too many points straying away. Our model does pretty well even for schools with very high revenue.



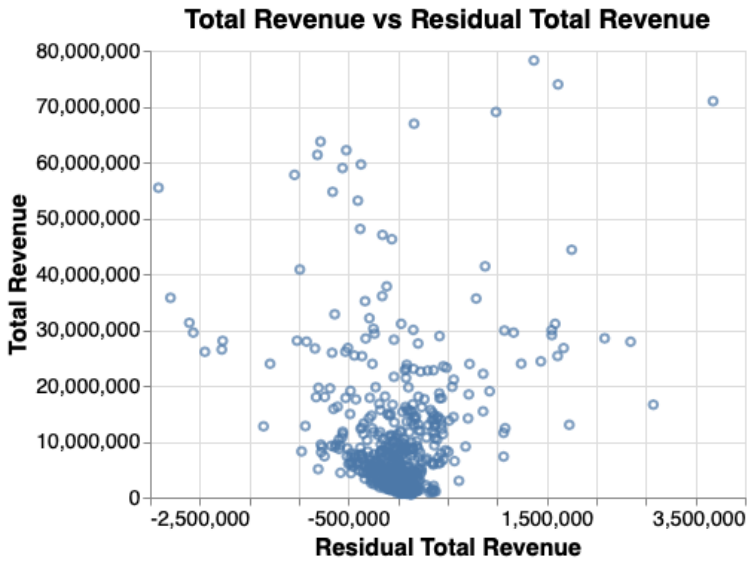
3.1.2 Ethical Issues

The results from this analysis could possibly harm teachers due to the observation that spending more on teachers results in less total profit.

3.2 Question 2: Can we use principal component analysis to reduce the number of dimensions while still having an accurate model?

3.2.1 Interpretations

Using RMSE as the cost function, 4 components were calculated to be the lowest dimension where 90% of the cost is preserved.
After regression on the reduced model, most points on the residual plot are still fairly closely clustered to the 0 residual threshold. The plot also significantly resembles the residual plot of the original model. Therefore, we can say that it is possible to reduce the dimensions of the model to 4 components while maintaining an accurate model.



In conclusion, we know that only 4 principal components are needed to predict the response, total revenue. With this in mind, the two variables that might be deemed less essential would be year and enroll.

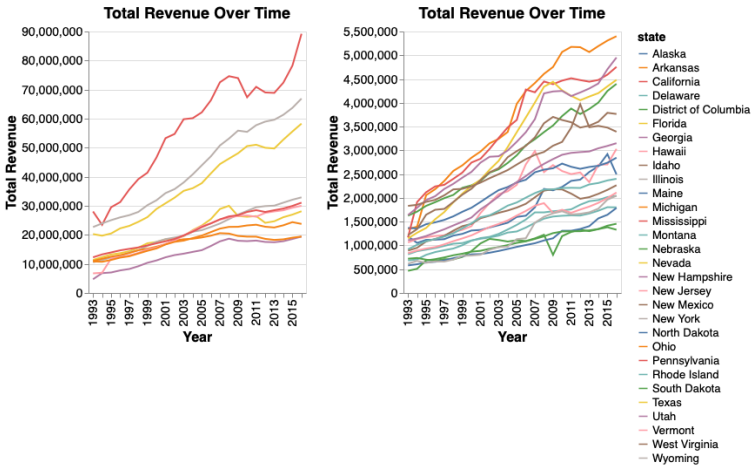
3.2.2 Ethical Issues

Since we concluded that enroll might not be a good predictor of revenue, there is a chance that schools may favor small sizes. However, given the nature of public school, the danger of schools keeping out students is not realistic and should not be concerning.

3.3 Question 3: What is the general trend of school revenue over time?

3.3.1 Interpretations

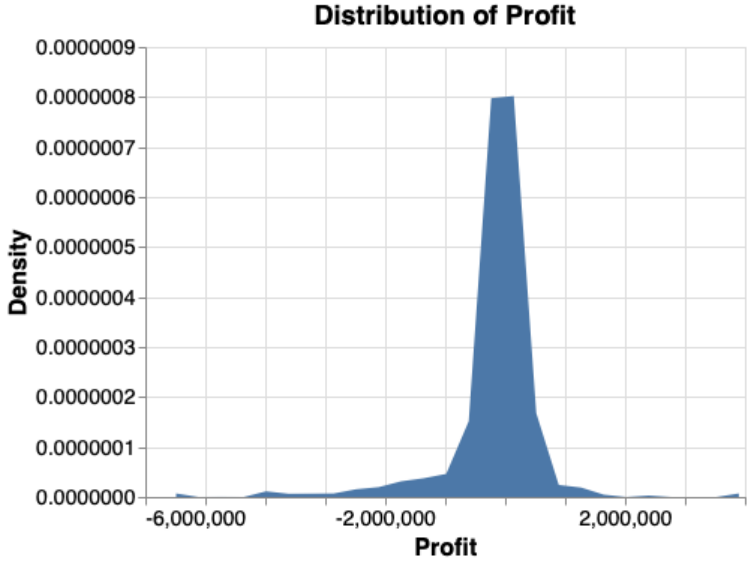
Based on the line plots, we observe that the general trend of school revenue is increasing over time. Note that there are two line plots as we decided to split the top 10 and bottom 20 states in revenue. This is to present a more clear picture as the total plot was very crowded and difficult to interpret.



3.4 Question 4: Is there an uneven distribution of profit amongst the schools in the United States?

3.4.1 Interpretations

From the density plot, the distribution of profit between schools seems well distributed. Most schools hover around 0 profit. Also, there are more schools that net negative profit than positive, which makes sense as schools are not meant to be profitable.



3.4.2 Ethical Issues

This analysis of profits should not have any ethical implications. The conclusion signifies that there are many schools that are in need of more profit.

3.5 Conclusion

In conclusion, our analysis suggests a school's total revenue is best estimated by its allocation of spending. We learned that the number of students that attend a school is not a very significant predictor of how much funding the school receives. From the linear regression, the other spending category provided substantial increases in total revenue. Schools that wish to learn more about ways to improve revenue may look into investigating which of the other expenditure categories are most effective. These could include spending more on food, business operations, or other forms. In addition, spending more on support services could also be a good way to improve their revenue. We also identified that profits are evenly distributed amongst schools but there are still more schools with negative profit that positive profit.

We originally attempted to predict whether a school was profitable or not but could not succeed with the data set given. Schools that were profitable had almost completely overlapping spending patterns as those that were not

We believe the results from this analysis are able to be trusted for multiple reasons. First, the data is collected with a solid process, leading to minimal biases. Also, the conclusions reached at each question can be supported with real world experiences and logical reasoning.

Chapter 4

Citations

1. Data:
https://www.kaggle.com/noriuk/us-educational-finances
https://www.census.gov/programs-surveys/school-finances/data/tables.html
2. Lab 8: Linear Regression fitting code
3. Lab 9: Used PCA code
4. Homework 3: Pairs plot code
5. Homework 4: Used code for scree plot
6. Website Used for Introduction:
Fiddle, Bruce J. "A Research Synthesis / Unequal School Funding in the United States." Unequal School Funding in the United States - Educational Leadership, 2002, www.ascd.org/publications/educational-leadership/may02/vol59/num08/Unequal-School-Funding-in-the-United-States.aspx.