Ascertaining the Relationship between Fiscal Health and Voter Turnout

Final Report

Statistical Consulting for Data Scientists

Karishma Raghuram, Christopher Thornton, Lucas Ophoff,
Kellen Whetstone, Xin Huang

March 22, 2023

1 Problem Statement and Background

The state of California is a key staple in the overall fiscal health of the United States, as the largest economy in the country and the fifth largest in the world. This economic backbone is built by the diverse industries represented in the many counties across the state, which each possess their own distinct socioeconomic and political characteristics.

Understanding the factors that influence each county's financial stability is vital to maintain and bolster the economic well-being of the entire state. Initially, we hypothesized that civic engagement is closely related to a county's fiscal health, particularly voter turnout rates. One such factor that is hypothesized to be associated with a county's fiscal health is civic engagement, particularly voter registration rates. This theory suggests that higher voter turnout rates signal a more informed and active citizenry, which could in turn lead to more advantageous fiscal decisions and policies.

In this report, we will investigate the relationship between past voter turnout rates and the fiscal health of California counties. Specifically, we want to answer the research questions: Is the fiscal health of California counties associated with civic engagement? To answer this question, we will analyze voter turnout data from the Census Bureau, along with data from the State Auditor of California regarding the state's financial risk by county. The findings of this report hope to shed light on the relationship between civic engagement and county fiscal health , helping policymakers and citizens further understand the importance and impact of an engaged populace.

Because our data includes a multitude of metrics relating to fiscal health, along with civic engagement, we had to narrow our focus to determine which variables we wanted to specifically investigate. First off, we decided to compare 2008 data on voter registration, with a variety of variables from 2018 regarding the fiscal health of California counties. We chose our response variable as voter registration, because it is a good indicator of civic engagement. As for the year, we settled on 2008, because it is important to recognize that the impact of voter registration on fiscal health may not be immediately apparent, and it may take several years for changes in voter registration patterns to affect fiscal health. Additionally, 2008 data precedes any impact that might be seen from the 2008 Great Recession. As for the choice of 2018 data for fiscal health, by comparing voter registration from a certain time period with fiscal health 10 years later, we can capture the long-term impact of voter registration on fiscal health. Furthermore, the economy goes through cycles of growth and recession (like the 2008 Great Recession), and these cycles can have a significant impact on fiscal health. 10 years allows us the most amount of time to control for these cycles, along with avoiding economic impacts from the COVID-19 pandemic. With these choices, we can focus on the long-term impact of voter registration on fiscal health.

2 The Data

Our data is the fiscal health data for California. We have also combined this data with voter registration data and census information. This data is in our repository on GitHub by the name of final_data_simplified.csv. The link to access our data and code are hyperlinked on the final page, and presented in the form of a QR code.

With our predictor variable of voter registration by county in 2008, we identified a few response variables to investigate during our analysis. These variables, taken on a by county basis, included:

- General fund reserves ratio (Represents the ratio of general government expenditure to government balance)
- Pension funding ratio (Reflects ability of the pension's assets to cover future obligations)
- Pension obligations ratio (Reflects the ability of the pension's assets to cover current obligations)
- Debt burden ratio (Compares debt obligations to revenue)
- Revenue trends ratio (Year to year percentage change in revenue)
- Liquidity ratio (Ration of current assets to current liabilities)
- Pension related debt (Debt incurred to finance future pension obligations)
- Unemployment rate
- Per Capita income

Our method of analysis to investigate the relationships between our predictor and response variables were correlation tests.

We joined the voter registration data for years 2002 to 2020 with the fiscal health data (for years 2017 - 2020) provided on Bruinlearn. Because the fiscal health data had observations in regards to California cities and the voter registration data had observations in terms of California counties, we introduced a third data set– fiscal health from the State Controller– to link the county measurements into cities. Lastly, we added in demographic data from the Census Bureau to help test our hypothesis. Given that some of our variables related to pension are not calculated per capita, we adjusted the values to be reflective of this as the context of the value would otherwise be ambiguous and unrepresentative of the population trend.

As a result of combining many data sets together, we need to clean the data and understand which variables are more important to our model construction. In order to deal with missing values, we imputing the NA's with the average value of the other data points for the same variable in that given city. We then began the lengthy process of variable selection.

Initially, we tried to use VIF, but were unable to successfully run the function because some of the predictor variables in our full data set had aliased coefficients. Aliased coefficients occur as a result of perfect multicollinearity, a statistical term used to described correlation between predictors. Simply put, some of the predictors were exactly correlated. As such, we adjusted our process to begin with a correlation plot for the quantitative variables.

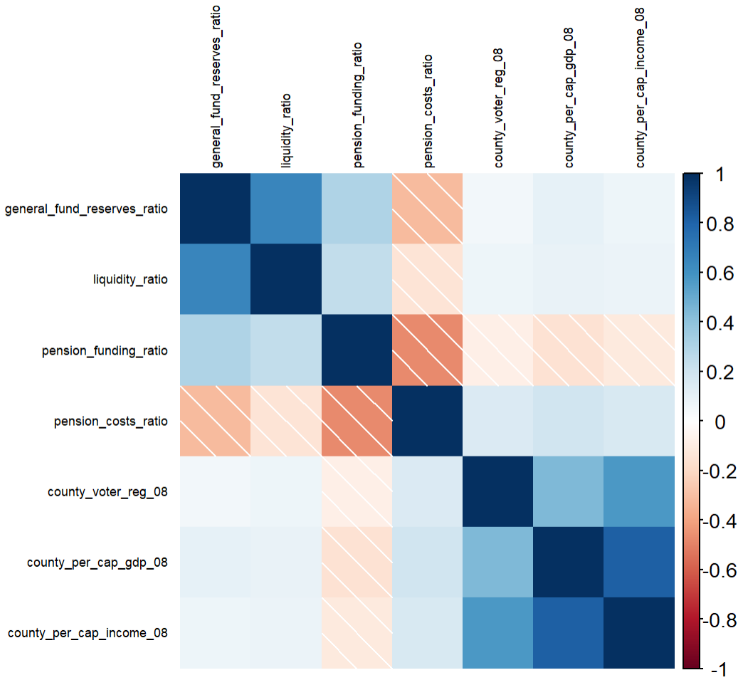


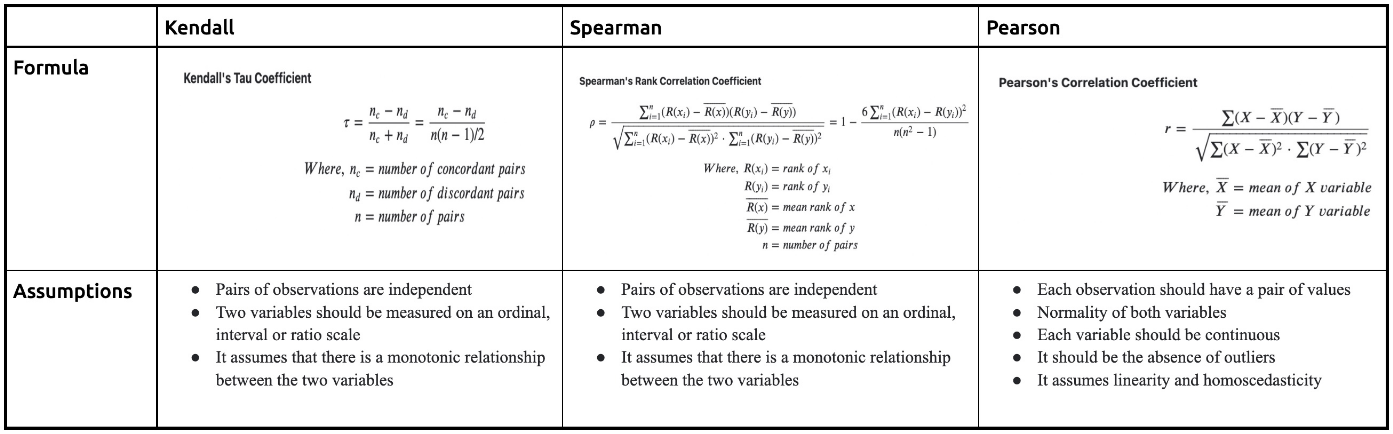
Figure 1: Correlation Plot Visualization

After gleaning which variables were strongly correlated, we began to manually remove them. Once this process was complete, we fit a simple full model relating "General Funds Reserves Ratio" as the response variable to the other quantitative variables as predictors. We then iteratively found the maximum VIF score of the simple model to continue variable selection.

Given the high degree of correlated predictors, we ultimately reduced our data set to include 19 number of independent variables. To address our main thesis for this project, we opted for a correlation test. However, rather than simply using a singular type of correlation test, we decided to compare and contrast the outcome of three different types: Pearson, Kendall, and Spearman.

Before we examine the results of our correlation tests, let us review the distinct attributes of the Pearson, Kendall, and Spearman tests.

Correlation measures the strength and direction of the relationship between two variables and the direction of said relationship. It is measured on a scale of -1 to 1. Perfectly correlated variables translate to a numerical value of 1 or -1, while uncorrelated variables result in a score of 0. A negative number insinuates that as one variable increases in value, the other decreases. The converse is true of positive correlation values: as one variable increases in value, the other increases simultaneously.



	Kendall	Spearman	Pearson
Formula	Kendall's Tau Coefficient $\tau = \frac{n_c - n_d}{n_c + n_d} = \frac{n_c - n_d}{n(n-1)/2}$ $Where,\ n_c = number\ of\ concordant\ pairs$ $n_d = number\ of\ discordant\ pairs$ $n = number\ of\ pairs$	Spearman's Rank Correlation Coefficient $\rho = \frac{\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})(R(y_i) - \overline{R(y)})}{\sqrt{\sum_{i=1}^{n}(R(x_i) - \overline{R(x)})^2 \cdot \sum_{i=1}^{n}(R(y_i) - \overline{R(y)})^2}} = 1 - \frac{6\sum_{i=1}^{n}(R(x_i) - R(y_i))^2}{n(n^2-1)}$ $Where,\ R(x_i) = rank\ of\ x_i$ $R(y_i) = rank\ of\ y_i$ $\overline{R(x)} = mean\ rank\ of\ x$ $\overline{R(y)} = mean\ rank\ of\ y$ $n = number\ of\ pairs$	Pearson's Correlation Coefficient $r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \cdot \sum(Y - \overline{Y})^2}}$ $Where,\ \overline{X} = mean\ of\ X\ variable$ $\overline{Y} = mean\ of\ Y\ variable$
Assumptions	● Pairs of observations are independent ● Two variables should be measured on an ordinal, interval or ratio scale ● It assumes that there is a monotonic relationship between the two variables	● Pairs of observations are independent ● Two variables should be measured on an ordinal, interval or ratio scale ● It assumes that there is a monotonic relationship between the two variables	● Each observation should have a pair of values ● Normality of both variables ● Each variable should be continuous ● It should be the absence of outliers ● It assumes linearity and homoscedasticity

Figure 2: Kendall, Spearman, and Pearson Correlations

Now that the three different types of correlation tests have been discussed, it is important to compare them to each other and understand how the results inform us.

Holistically, parametric testing is more informative and more powerful in comparison to non-parametric methods. As such, Pearson's correlation test is a stronger method of analyzing the data in comparison to Spearman's Rank Correlation Coefficient and Kendall's Tau Coefficient. That being said, the assumptions of Pearson's Correlation test stipulate that the variables must be normally distributed; furthermore, the outcome produced is indicative of a linear relationship between variables.

Given that many of our predictors in the subsetted data set fail to meet normality based on the Normal Q-Q plots (see Figure 3.1) and Kolmogorov-Smirnov Test, Pearson's methodology is less useful and applicable. The Normal Q-Q plot assesses normality through identifying skewness, outliers, kurtosis, etc. The Kolmogorov-Smirnov test ascertains normality through hypothesis testing: in this case, the null hypothesis is that the data is normally distributed. As such, larger p-values indicate a lack of normality in the data. Converse to Pearson's correlation, Kendall's and Spearman's coefficients focus on monotonic relationships between variables; monotonic refers to variables that move in the same direction but not at the same speed.

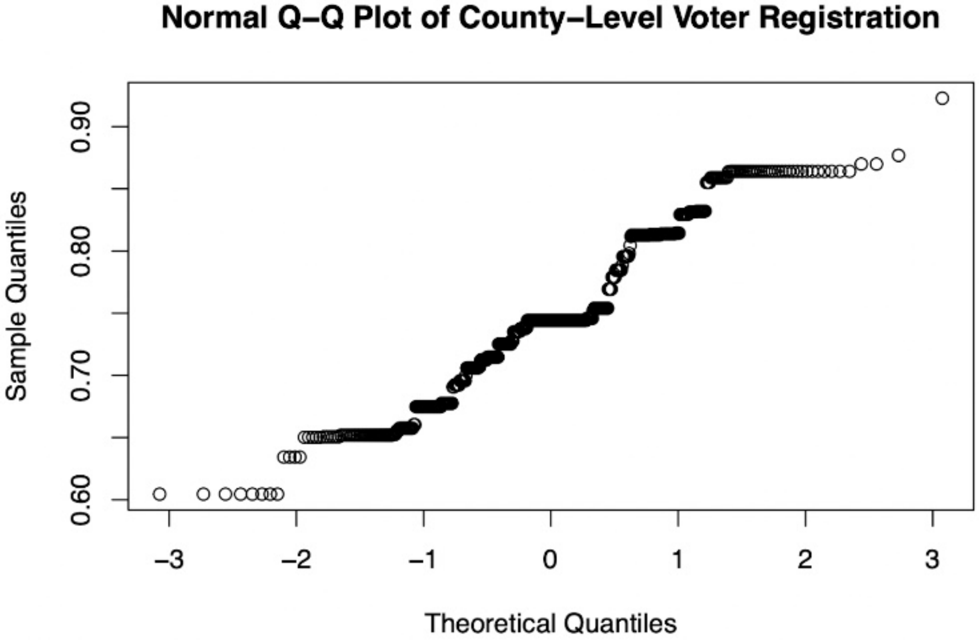
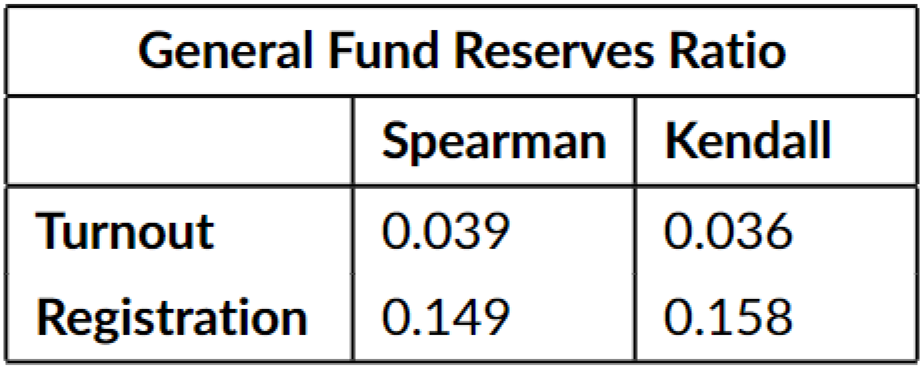


Figure 3

Note that the Kolmogorov-Smirnov Test for Voter Registration in '08 yielded a D statistic of 0.1360 and a p-value of 4.92 E-08 compared to an α level of 0.05. Thus, we can reject the null hypothesis and conclude that the variable is not normally distributed. Since the Kolmogorov-Smirnov Test implies that our data is not normally distributed, we cannot use the Pearson correlation test. As such, we focus on the Kendall and Spearman tests. While the two are nearly identical, Kendall's method looks at the difference between the percentage of concordant and discordant pairs amongst all possible pairwise events while Spearman's method is used in place of linear correlation. As such, either method is suitable for non-normally distributed data.

3 Results

From our Kendall and Spearman correlation testing results, we found that the majority of our variables were insignificant. Interestingly, three variables were only significantly correlated with either voter registration or voter turnout, but not both: 1) General Fund Reserves Ratio, 2) Debt Burden Ratio, and 3) OPEB Funding Ratio. The p-values for the correlation between the General Fund Reserves Ratio, and Voter Registration and Voter Turnout respectively through both the Kendall and Spearman methods are shown below in Figure 4.



General Fund Reserves Ratio		
	Spearman	Kendall
Turnout	0.039	0.036
Registration	0.149	0.158

Figure 4

Furthermore, within our correlation results, we observed significance in most of the pension related variables, excluding the Pension Funding Ratio. Given that Pension Obligations and Pension Cost Ratios were significant, we concluded that pension funds are able to pay out the promised amount. Further analysis of other time frames may be necessary to construct a more thorough understanding of this relationship.

4 Discussion

In understanding the nuances between these three methodologies, we see that Spearman's Rank Correlation is quite similar to Pearson's Correlation Coefficient. That being said, parametric methods of analysis are stronger than non-parametric forms if the conditions are met; specifically, parametric tests are more statistically significant on average compared with their nonparametric counterparts when conditions are met as they are more likely to correctly detect an effect or relationship. As such, we could potentially utilize a Box-Cox transformation to adjust our data to be normally distributed, and subsequently rerun a Pearson's Correlation test to see the difference in results. Using a Box-Cox transformation also allows us to mitigate kurtosis, skewness, and outliers.

In understanding what other methods of analysis we can use to examine our dataset, we may decide in the future to not simply see the relationship between variables, but rather see the causal relationship between them (i.e. Based on fiscal health, can we predict voter turnout rates? Perhaps we choose to pursue the converse as well: based on voter turnout rates, can we understand the fiscal health of that respective county?) Broadly speaking, the two major types of evaluation for labeled data are regression and classification. Given the nature of our data, we can elect to pursue a predictive model such as logistic regression, a form of classification which requires a binary outcome, absence of multicollinearity, linearity for continuous variables, and a lack of strong outliers. Through this, we can attempt to categorize California cities as below average, average, or above average regarding fiscal health based on voter turnout.

Note, however, that statistical analyses of economic and political trends suffer from one significant limitation: since social developments strongly depend on preceding economic or political circumstances, many social data points are indicative of past rather than contemporary trends.

This is in part due to the election cycle, and that many economics related data points reflected in a census are lagging indicators, such as unemployment rate. Lagging indicators do not predict trends, but rather reflect what changes have already happened within various macroeconomic factors. Lagging factors are countered with coincident indicators and leading indicators; the former category is indicative of the current macroeconomic state while the latter is predicative of the future macroeconomic state. As such, being more cognitive of these nuances when choosing data and running our analysis will allow for a more rich and nuanced understanding of the relationship between fiscal health and voter turnout rates.

5 Github

Listed here is the weblink to our team GitHub repository. Here, you can find the combined final data set, code, model outputs: : Team 5 Github Link

Figure 5: QR Code