Analysis of the relationship between city size and municipal spending efficiency in California cities.*

Terrell Enoru

September 24, 2025

California's rising government debt over the past decade necessitates careful examination of municipal spending efficiency. To analyze spending efficiency I use data on expenditure per capita for all california cities over the past 20 years. I use regression analysis to model the relationship between population size(the predictor) and per capita spending(the response varible). Then hypothesis testing is used on the slope to determine whether large cities benefit from economies of scale. The results were that there was not enough statistical evidence to claim that expinditure per capita decreased as population increased which is significant because it implies that the continued expansion of large cities like Los Angeles and San Francsico that serve as the State's primary economic growth engine will not result in a less efficient expenditure.

1 Introduction

The motivation for this paper comes primarily from two sources. First seeing the rise in government expenditure and subsequent increase in debt prompted me to look into what lead to the increase in expenditure and how to make it more efficient. Subsequently when looking into government expenditure I found a paper which discovered that cities in Spanish experienced economy of scales effects up to a certain point (Benito, Bastida, and Guillamón 2010) which prompted me to think if that same efficiency exist in California. Some useful background on California state expenditure is that over the period from 2019 to 2024 state expenditure increased from around 300 billion to 494 billion. It currently is on pace to surpass to peak covid spending in 2021(499 billion) by the end of the year(2020) which is why examining how efficiently this money is being spent is so important.

*Project repository available at:https://github.com/tycebot/Modeling-population-and-spending.

The knowledge gap being addressed is that of understanding the relationship between population and expenditure on the city level in the United States. Particularly for California there is no analysis on population as a mechanisms driving city spending in California. Most of the analysis focuses on a subset of expenditure (like health care or education). What this paper looks to do is examine expenditure on the more granular city level with the hope that by finding efficiency on the city level they transfer to the state level. In order to address the knowledge gap the simple linear regression model was fit to measure the relationship between estimated population(the predictor) and expenditure per capita(response). Then a one sided hypothesis testing was done on the slope of the regression model to determine whether or not large cities benefit from economies of scale in their spending.

The structure of the the project is divided into four sections Section 2,Section 3,Section 4, and Section 5. Section 2 introduces the data used in this analysis including limitations,definitions, and why it's important. Section 3 explains the statistical methods used for the analysis. Section 4 discusses the results of the analysis as well as what can be gleaned from these results and what are the restrictions. Section 5 provides an overview on the project the results any shortcomings and then any future extensions.Finally Section 5 contains the citations for any external sources used in the paper.

2 Data

The California capital expenditure data set is provided by the California State Controller's Office(Matthew 2025). Each row in this data set represents the expenditure per capita of a particular city in California at the end of a fiscal year. The key variables are entity name which is the name of the city,fiscal year(July 1 of the current year to June 30 of the next year) which is the year for which each expenditure per capita is calculated ,total expenditure which is the gross expenditure of a particular city over the fiscal year, estimated population which is the population of the city based on the Housing Unit Method which extrapolates population by taking housing units adjusting for occupation and then using that to estimate household population then combining that with estimates for "group quarters population"(essentially everyone in nontraditional households like dorms and prisons) to get the overall population estimate(more details in reference),and expenditures per capita which is the total expenditures divided by estimated population.The data is grouped by city,year, as well as binned into three groups for large medium and small sized cities.The limitations for the data set are as follows the estimated population does a bad job of measuring homeless because of it's reliance on housing units,the years are limited to a twenty year period from 2003 to 2023,the estimated population can double count those that move until they register their new address, and the expenditures measures are both skewed over the covid period (2020-2023). There are 29 cities missing at least one year of these 8 are missing one year,8 are missing 2 years,4 are missing three years, and 7 are missing 5 or more.Since all the years missing come from the tail end of the collection range(eg if a city is missing 2 years they are always the oldest 2 years so 2003 and 2004) we will be filtering the cities to only those with data from 2005 to 2023 in order to make it so that

all cities have the same amount of data while keeping the vast majority (98%) of the data. We also converted any 0 in the expenditure,expenditure per capita, and total expenditure to null and then filtered out those rows of which there were 5.By analyzing how much money is spent by a city over an extended period of time for a cities with a variety of different populations we can reasonably ascertain the relationship between expenditure and the population. However the key assumptions are that the population measurement is accurate relative to other cities in the data set(meaning if it underestimates or overestimates the population then it does it consistently) and that the cities are correctly reporting their expenditures to the state.

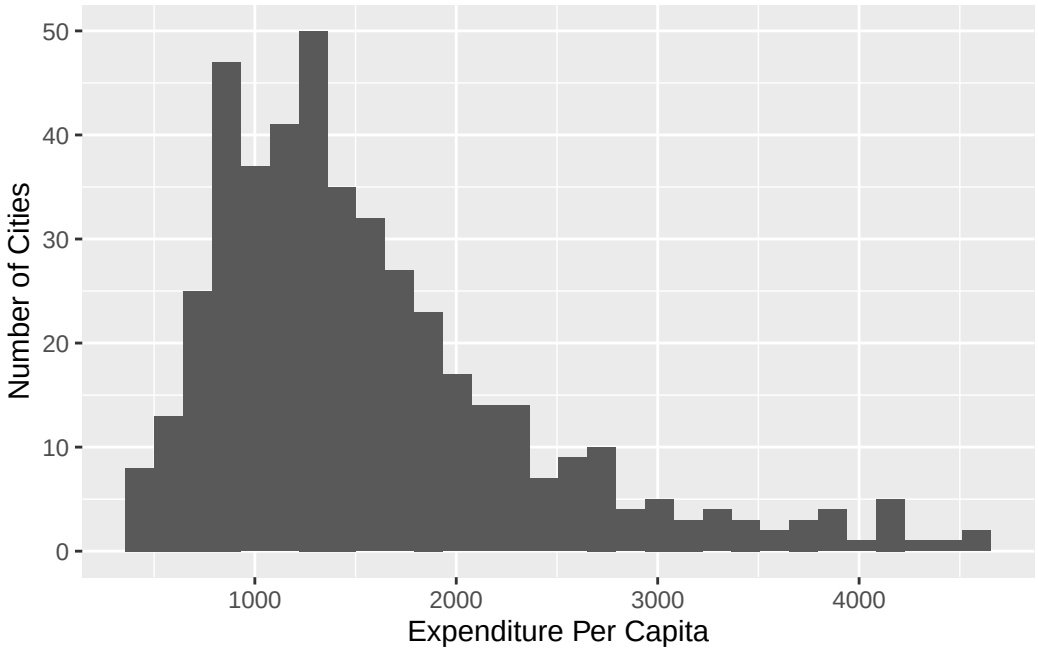

Figure 1: Histogram of average expenditure of each city.

3 Methods

We adopt a simple linear regression model with expenditure per capita as the response and population as the predictor. Let Y_i denote expenditure per capita and X_i denote population for the Y_i year in our dataset. The model can be written as $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ where β_0 represents the expected population when the expenditure is 0,β_1 represents the average increase in expenditure per capita for every individual added to the population, and ϵ_i represents the change in expenditure per capita not captured by the population The lm function is the a function used to create linear models.

In order to evaluate whether or not high population cities get an economy of scales bonus we

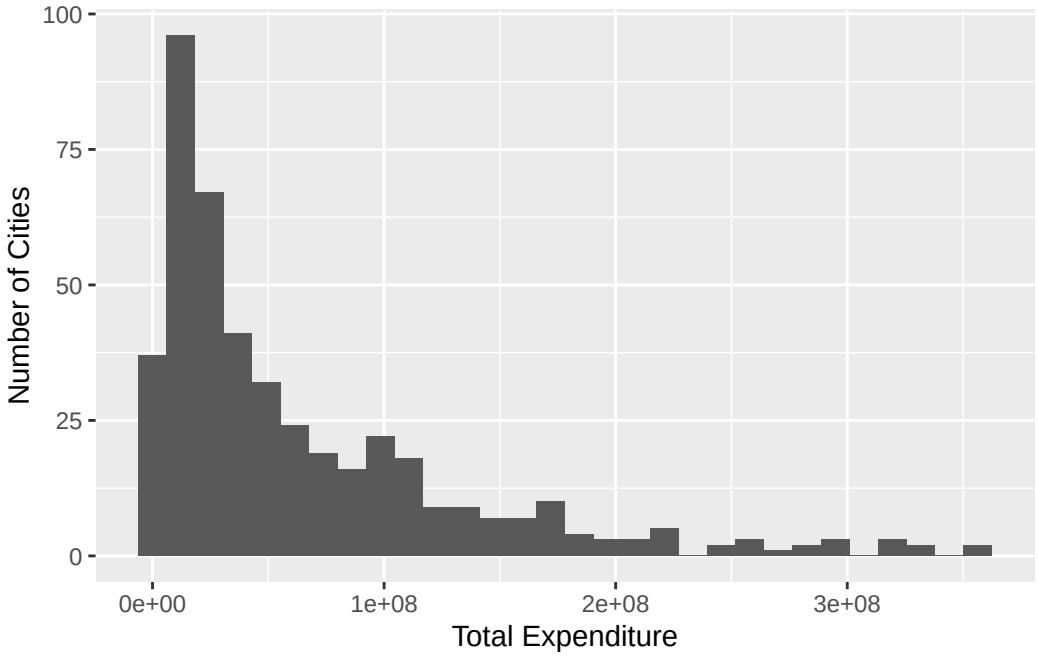

Figure 2: Histogram of total expenditure of each city.

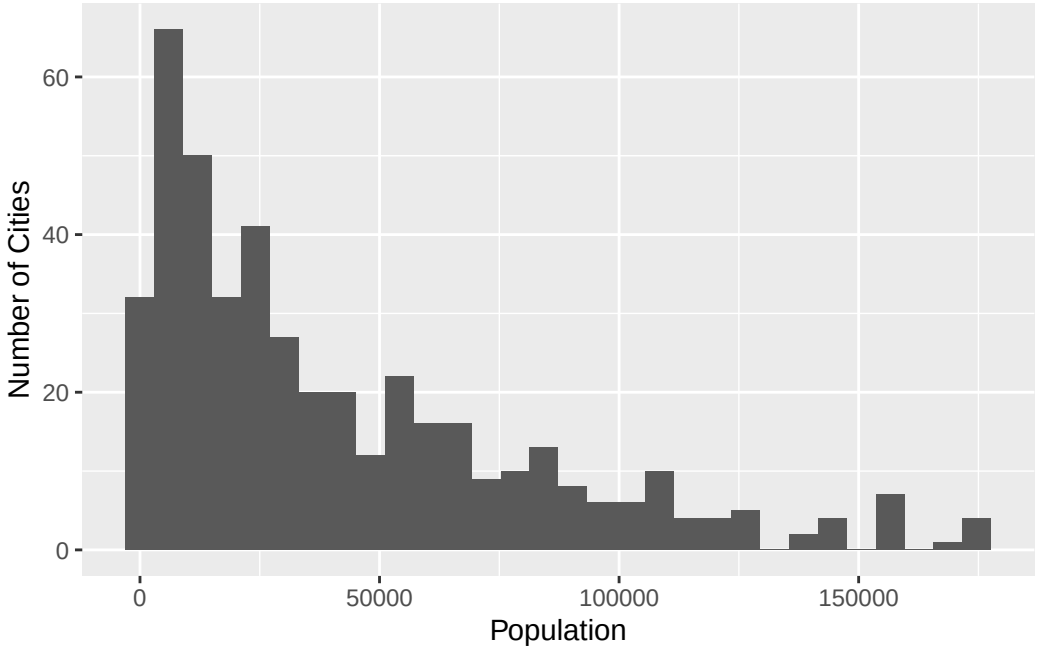

Figure 3: Histogram of population of each city.

use a one sided hypothesis test to determine whether there is a negative association between expenditure per capita(Y) and population(X). The type 1 error rate will be set at 0.05 and the null and alternative hypothesis are as follows

$H_0 : B_1 >= 0$

$H_1 : B_1 < 0$

I implemented this analysis using the R programming language(R Core Team 2025) and the tidyverse package(Wickham et al. 2019) for data cleaning and plot creation.

4 Results

The estimated slope parameter is b_1 = -0.007. In other words, for each individual added to the population the expected change in the expenditure per capita is -0.007. The estimated intercept is $b_0 = 7462.951$. In other words the expected population of a city with no expenditure per capita is 7462.951.

Since the t value is greater than the p value we fail to reject the null hypothesis that the slope is greater than or equal to 0.

The model and hypothesis test rely on 5 key assumptions which are that the regression function is linear, the error terms have equal variance, the error terms are independent, the error terms have a mean of zero, and the error terms are normally distributed. Based on the scatter plot of population against average expenditure the regression function does not seem to be linear.The relatively even spread of points above and below the zero line suggest the error terms have a mean of zero(there do seem to be a few more below the line but the ones above have a higher value on average).Equal variance does seem to hold for most of the fitted values but the range of residuals for fitted values below 1400 definitely seems smaller which violates the assumption of equal variance. Based on the curve pattern in the qqplot it can be inferred that the error terms are not normally distributed. Based on the up and down pattern in the residuals vs index plot it can be inferred that the errors are not independent. Since almost all the assumptions necessary for hypothesis testing and most for linear regression were broken(except for the mean of errors being 0) we cannot trust the outcomes from the inference on the slope.

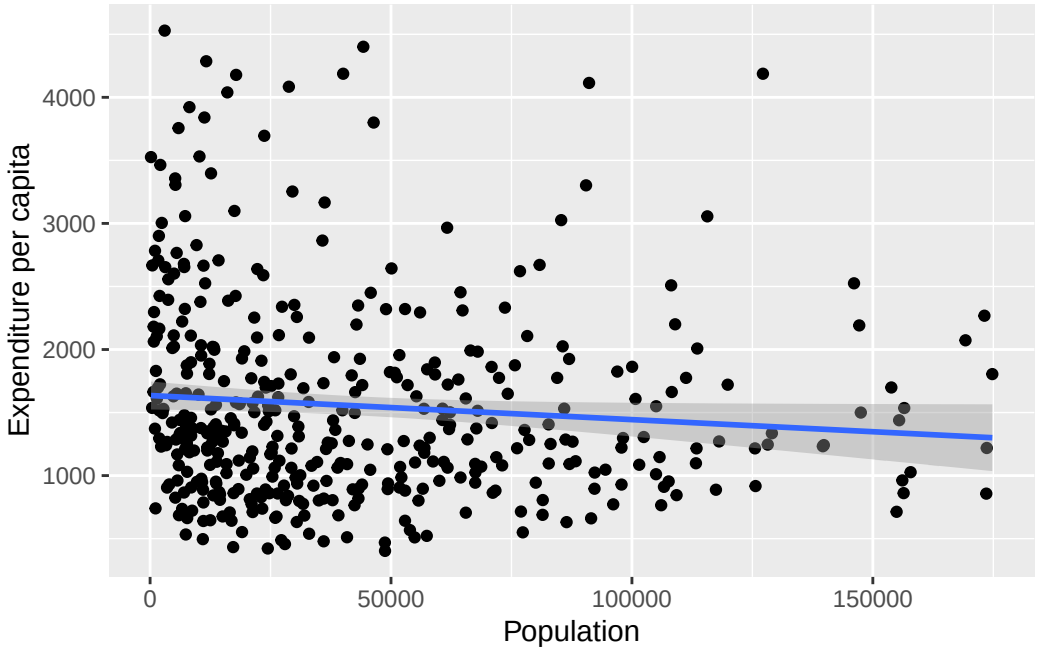

Figure 4: Scatter plot of population (x-axis) and expenditure per capita (y-axis) for a simple linear regression with expenditure per capita as the response and population as the predictor.

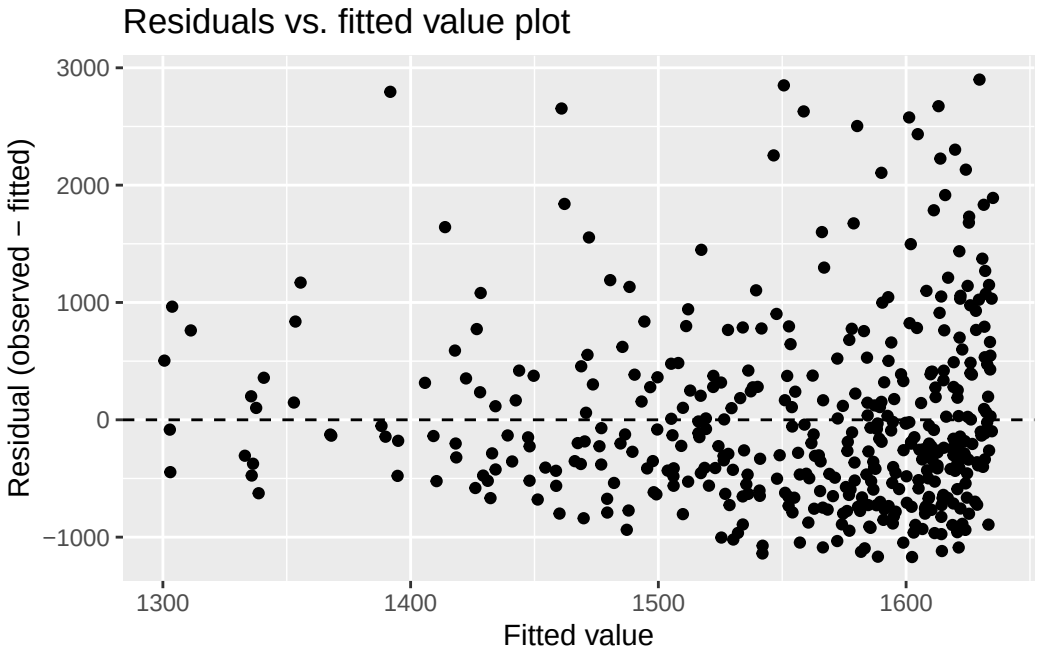

Figure 5: Scatter plot of fitted values (x-axis) and residuals (y-axis) for a simple linear regression with expenditure per capita as the response and population as the predictor.

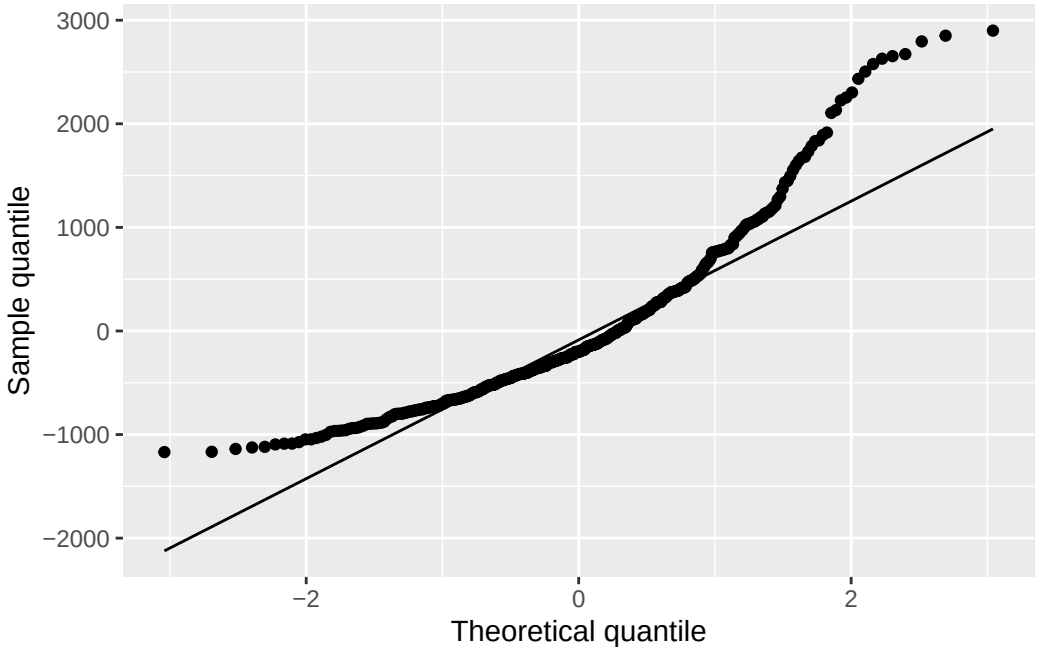

Figure 6: Quantile-quantile plot of residuals (y-axis) for a simple linear regression with expenditure per capita as the response and population as the predictor.

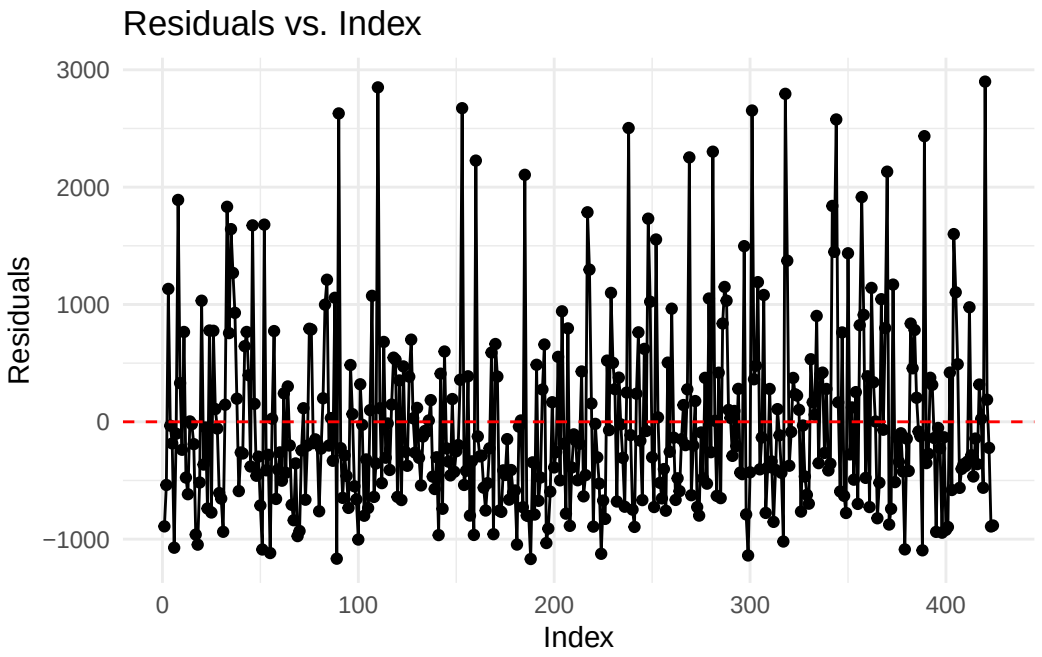

Figure 7: Scatter plot of index (x-axis) and residuals (y-axis) for a simple linear regression with expenditure per capita as the response and population as the predictor.

5 Discussion

In conclusion this paper was looking to examine whether there were economy of scale bonuses for big cities in terms of spending efficiency. It did this by looking at a data set of expenditure by capita and population for every city in California over a 20 year period from 2023. A regression model was then fit to this data with expenditure per capita as the response variable and population as the predictor which yielded a slope of -0.007. A one sided hypothesis test was then conducted on the slope with a type 1 error rate of 0.05 . From the results of our regression model and hypothesis test we can conclude that there is not enough statistical evidence to reject the null hypothesis that the slope is greater than or equal to zero.

The key finding from my research was that there is not a significant economy of scale bonus granted to larger cities when spending. Although based on the slope there also does not seem to be a positive association between population and expenditure per capita. This would lead us to suggest that there is not a strong linear association between population and expenditure per capita. The implications of this are twofold. First the mandatory new costs like added bureaucracy that is necessitated when administering a larger city does not seem to have a particularly negative impact on city expenditure. Second the economies of scale one would assume a larger city would gain(and which were found in a previous paper on the subject

in Spain(Benito, Bastida, and Guillamón 2010)) either do not exist or get eaten up by some other confounding factor like the bureaucracy theory I put forth earlier. Ultimately whatever the cause the results suggest that the continued movement of people to large cities should not have a negative impact on spending efficiency.

There are a few key weaknesses to this study. First because of the scale of California's big cities there are a few huge outliers when it comes to population particularly Los Angeles,San Diego, San Francisco, and San Jose have populations that dwarf the rest of the state making it so that they have an out sized impact on the relationship between population and per capita expenditure. Second because the population is calculated by residents it makes it so that cities like Vernon which mainly serve as hubs for business and industry rather than residence can have a relatively meager population(159) but very significant expenditure(277024003). This points to a broader problem with using expenditure per capita as a proxy for spending efficiency which is that it completely ignores the other side of fiscal efficiency which is revenue. Even if the expenditure per person decreased due to some economy of scale bonus(which was not found to be true) if the revenue growth was slower than the expenditure growth (even if the rate of expenditure growth was decreasing) then even though the spending efficiency would look great the actual fiscal efficiency would tell the opposite story.

Potential improvements that can be done to this study is combine the expenditure data set with the revenue data set. We could then infer the profit from the difference between the revenue and expenditures and use that to better measure the relationship between population and fiscal health. Potential extensions of this work would be to apply the profit analysis mentioned previously to states outside of California. The results could then be compared to see if the lack of an economies of scales bonus was due to factors specific to California or more widely common. Finally combining population with some kind of economic metric(possibly gdp per capita) would do a better job at answering the broader question of as a city grows(both economically and in population) how does that effect the health of it's finances.

Works Cited

2020. Urban Institute. https://www.urban.org/policy-centers/cross-center-initiatives/state-and-local-finance-initiative/projects/state-fiscal-briefs/california.

Benito, Bernardino, F. Bastida, and María-Dolores Guillamón. 2010. "Urban Sprawl and the Cost of Public Services: An Evaluation of Spanish Local Governments." *Lex Localis-Journal of Local Self-Government* 8: 245–64. https://doi.org/10.4335/8.3.245-264/%282010/%29.

Matthew. 2025. "City Expenditures Per Capita - Dataset - California Open Data." *California Open Data Portal*. https://data.ca.gov/dataset/city-expenditures-per-capita.

R Core Team. 2025. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.