The Relationship Between Library Access and High School Graduation

ECON 460: Group 3

Luke Alati, Darian Ahmadizadeh, Dung Pham, Natasha Denisyuk, Christian Becker

Notebook Link: https://github.com/beckerchTRJ/ECON460

I. Background:

It is universally understood that achieving a high school degree can improve an individual's long-term economic outcomes. Graduating high school can improve employment opportunities and income, increasing community well-being. The same can be said about a bachelor's degree, as well as a master's, PhD, and all other forms of higher education. However, high school graduation rates in the United States vary widely across demographics, regions, and socioeconomic groups, often due to unequal access to resources. If students have limited access to resources, graduating becomes much more challenging. Public libraries are one of many resources in a community.

Libraries offer free access to books, technology, and a quiet study environment. These services can bridge resource gaps for students, particularly in underserved communities. As an easily accessible hub for learning, libraries are a resource for many students who otherwise might not be able to access these resources. This leads us to our primary research question: How does public library access affect the high school graduation rate of a county in the United States?

If we find a strong, positive relationship between library access and high school graduation rate, it *could* suggest that investment in public libraries can improve graduation rate, and therefore economic outcomes. While our study does not claim to establish causality, further research could reveal a causal relationship between libraries and graduation rates. Public libraries play a crucial role in underserved communities by providing free access to books, the internet, study spaces, and tutoring programs, serving as hubs for learning and development. Our findings from this study can inform local policymakers on the importance of library funding in high-need areas. By recognizing libraries as critical community assets, policymakers could create legislation that prioritizes access to libraries.

II. Literature Review:

There is substantial academic literature on the positive impacts of libraries, including public and school libraries. The first source we looked at is *Linking School Libraries and Literacy (2010)* by Christina Clark. In this paper, Clark explores the relationship between school libraries and student academic success in the United Kingdom. Clark found that 68.7% of young people use the school library, a fairly high number, but declines with age. Additionally, Clark found a very strong relationship between library usage and reading level. Students who use the school library more often were reading at or above grade level. Interestingly, socioeconomic background was not statistically correlated with library usage. Students from all backgrounds generally used the school library.

In *The Academic Library Impact on Student Persistence (2011),* Mark Emmons and Frances C. Wilkinson examined the impact of university library funding on university student success in the United States. To do this, they looked at three measures of success: GPA, course pass rate, and general information retention. Emmons and Wilkinson found that increased funding and staffing of school libraries led to increased graduation rates. From regression analysis, they found that a 10% increase in library staff on average predicted a 1.55% graduation rate increase. Additionally, Students who attended library instruction sessions showed measurable improvements in academic performance, with higher GPAs and course pass rates.

The final academic paper we looked at is *Working Together (2019)* by Emma Power, et al. Power investigates the role of public libraries in addressing educational challenges faced by students in rural and remote areas. She highlights the unique contributions of libraries in creating educational opportunities for underserved communities. Additionally, Power discusses how libraries provide access to books, digital tools, and internet services which help students succeed.

III. Data:

Consistent with our main research question, we gathered data on a county level. Its granularity enables more observations and a closer look at local libraries' effects on the surrounding community and has the most comprehensive data for libraries. We also considered obtaining data on a ZIP code area-level; however, as ZIP code areas can be quite small, a library's reach may extend beyond the area and county-level data would be more appropriate.

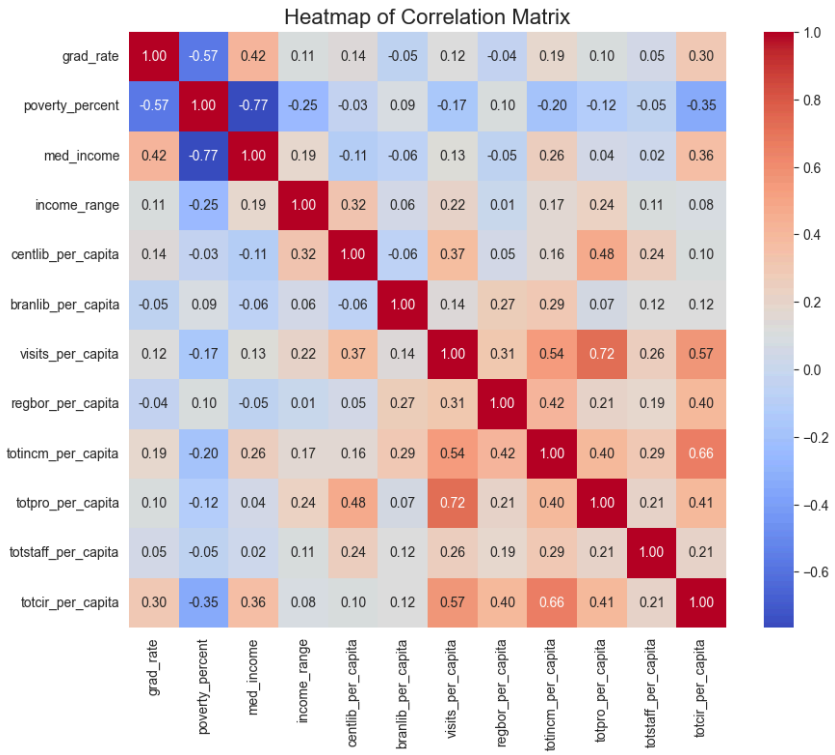
Estimated high school graduation rates by county for the 5-year period ending in 2022, sourced from the American Community Survey (ACS), serve as the dependent variable. Data for this is available for 3,218 counties in all 50 U.S. states, Puerto Rico, and Washington D.C., which ensures consistent estimation procedures. While 5-year estimates are less current than their 1-year or 3-year estimates, they are the most reliable and available for all counties, including those with populations fewer than 65,000 that otherwise do not have 1-year estimates (U.S. Census Bureau).

For independent variables relevant to library usage, we utilized data from the 2022 Public Libraries Survey (PLS), which includes information for around 9,000 library systems in the United States (e.g., circulation of resources, educational programs, and staffing). To create the relevant variables for our analysis, we aggregated a certain subset of available variables in the PLS dataset by county, and divided by the population of the county to create per-capita measurements to allow meaningful comparisons across countries. Appendix A details these variables and their definitions. In the process, we noticed slight variations between county classifications of the ACS and PLS data and performed alterations to match the two. Libraries for which this reconciliation is impossible, and/or those with missing, null, or negative values for its variables (often markers for missing or unavailable data) are dropped, leaving 8,401 observations

before aggregation across counties (counties may have more than one library system, so data needed to be aggregated). The dropped observations represent approximately 10% of the 9,248 libraries in the PLS data set, which left us with plenty of observations. Otherwise, the PLS data set has been cleaned and is well-documented, and thus did not require further cleaning work before deriving per-capita variables for the analysis to follow.

Anticipating the potential problem of omitted variables in regression, we include measures of socioeconomic condition. This clarifies the link between library usage and a community's economic health. We used the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) data from 2022, specifically information on median income and poverty rate among 5–17-year-old children for each county. The data was collected in the same period, with consistent county definitions to that of the ACS dataset, and therefore merged without needing further processing.

Figure A.1: Correlation Matrix of Included Variables



IV. Methodology:

Lasso and Ridge are linear regression models that use regularization to prevent overfitting by penalizing large coefficients. Ridge minimizes the sum of squared coefficients, while Lasso minimizes the sum of absolute coefficients. We selected these models because they excel at identifying the most important variables and shrinking the coefficients of less relevant ones. Our library dataset includes numerous variables, highlighting a need for regularization. Since it's unclear which variables to prioritize for our regression analysis, Lasso and Ridge offer an effective approach to variable selection.

With these models come some assumptions. The first assumption is that the relationships between the explanatory and the dependent variables are linear. The second assumption is that the residuals should be independent of each other. The third assumption is homoscedasticity, that the variance of the residuals remains constant across all levels of the explanatory variables. The fourth assumption is the predictors are scaled appropriately before running the models. The fifth assumption is that the sample size is sufficient, easily satisfied by the large number of counties.

Our process involved multiple steps. First, we split the data into training and test sets. Then, we standardized each of the variables. Next, we used 20-fold cross-validation to select the optimal penalty term for each of the two models. Once this was selected, the models were trained and their performance evaluated.

Our methodology has several inherent benefits: Lasso and Ridge are relatively simple to run, are fairly explainable, handle multicollinearity well, and the regularization penalties reduce overfitting. However, as with any model design, there are drawbacks too: we do not account for temporal dependencies or non-linear relationships. Omitted variable bias likely plays a role in our model performance, too, with the exclusion of factors like school funding at the county level.

V. Results:

1. Regression Results: All Counties

We performed Ridge and Lasso regressions across all counties in the U.S. As shown in Table B.1 below, poverty percent is the most significant predictor, with negative coefficients of -2.98 (Ridge) and -2.92 (Lasso), underscoring the challenges posed by economic disparities.

Library-specific variables, such as central libraries per capita (coefficients of 0.88 for Ridge and 0.70 for Lasso) and total circulation per capita (0.84 for Ridge and 0.70 for Lasso) positively correlate with graduation rates, suggesting that well-resourced and actively utilized libraries contribute meaningfully to educational success. However, variables like visits per capita (-0.59 for Ridge, -0.43 for Lasso) and registered borrowers per capita (-0.22 for Ridge, -0.08 for Lasso) show negative coefficients, potentially reflecting resource strain or overcrowding in underserved areas.

Table 1: Key Predictors of Graduation Rates Across All U.S Counties

(For a comprehensive view of all results, please see Appendix B)

Feature	**Ridge Coefficient**	**Lasso Coefficient**
Poverty Percent	-2.982074	-2.925565
Central Libraries per Capita	0.875742	0.694940
Total Circulation per Capita	0.840902	0.695533
Visits per Capita	-0.585155	-0.427737
Registered Borrowers per Capita	-0.216382	-0.078291

2. Regression Results: Top 15% Counties

In the top-performing counties, the relationships between library variables and graduation rates become more positive but less significant. As shown in Table B.2 below, the income range becomes the most significant predictor, with coefficients of 0.14 (Ridge) and 0.20 (Lasso). For these top-performing counties, there are higher median values for income and library resources.

Library-specific variables such as central libraries per capita (0.09 for Ridge, 0.12 for Lasso) and visits per capita (0.05 for Ridge, 0.07 for Lasso) continue to show positive relationships that are smaller in magnitudes. This suggests that libraries still play a role in these communities, but their relative importance is diminished. This may be due to the availability of alternative educational resources.

Table B.2: Key Predictors of Graduation Rates in Top 15% Counties by Graduation Rates

Feature	**Ridge Coefficient**	**Lasso Coefficient**
Income Range	0.132925	0.197750
Central Libraries per Capita	0.089224	0.121547
Visits per Capita	0.052768	0.065464
Median Income	0.075009	0.097454
Total Circulation per Capita	0.033759	0.034611

3. Regression Results: Bottom 15% Counties

In the bottom-performing counties, where socioeconomic challenges are most acute, the influence of libraries appears to be less consistent and may be overshadowed by the negative impact of economic disparities. As shown in Table B.3 below, poverty percent is a strong negative predictor (-0.34 for Ridge, -0.74 for Lasso).

Interestingly, both median income and library circulation have negative relationships with the graduation rate for this subset of counties. There could be a degree of economic segregation in these counties, where more affluent students attend private schools, leaving the public school system worse off. However, this claim lies outside the scope of our study.

Another important point is that the model trained on the bottom counties performed far worse than the model trained on the top counties. Such results point to the substantial variation among many of these bottom counties. For instance, the county of Lagrange, Indiana has a high median income and low poverty, despite having the second-lowest graduation rate in our data. Upon closer inspection, this is due to a large Amish population that exits school after the 8th grade. Likewise, many border towns are present, too, which may have large amounts of immigrants who are not integrated into the school systems. Without including more variables, underperforming counties are challenging to model.

Table B.3: Key Predictors of Graduation Rates in Bottom 15% Counties by Graduation Rates

Feature	**Ridge Coefficient**	**Lasso Coefficient**
Poverty Percent	-0.344838	-0.743188
Income Range	-0.398101	-0.617222
Median Income	-0.189547	-0.484054
Total Circulation per Capita	-0.136290	-0.134625
Central Libraries per Capita	0.106410	0.178806

VI. Critical Reflection and Future Directions:

Ridge and Lasso regressions assume a linear relationship. However, the relationship between library engagement and graduation rates may not be perfectly linear. For example, an increase in library resources might result in a substantial improvement in graduation rates, but the impact may diminish beyond a specific threshold due to diminishing marginal returns. Our regressions, focusing on the bottom 15% of counties by graduation rate, had weaker coefficients for library-specific variables, indicating a potential non-linear relationship between these variables. For future research, we could explore non-linear models, such as decision trees or random forests, to study this relationship in more detail.

This project also raises two other questions for further consideration. First, for young library users, their educational outcomes are not observable until their (possible) high school graduation many years later. Therefore, we might extend this study by including data for past years—which is readily available for all three of our data sources—or analyzing our graduation rate data with time series methods to see the extent of a lag in the effect of public libraries, if any.

Another direction for expansion is considering non-public libraries, such as libraries belonging to a school or college, to see if there is a significant difference in the effectiveness of these libraries compared to public libraries. We hypothesize that our model might suffer from omitted variable bias if private library use significantly affects educational outcomes, and research in private libraries will help refine our model further. However, we realize possible challenges in collecting the data necessary for this goal, as well as designing an adequate model to directly compare public and private libraries, given the difference in audiences and operations between the two.

VII. Conclusion:

Our findings generally aligned with our expectations, with a positive relationship between public libraries and graduation rate. Specifically, we found that active library engagement supports better educational outcomes. Interestingly, we found that mere presence at the library may not yield the same benefits as borrowing materials. Students needed to actively engage with the library, by checking out books and other materials, to see a noticeable impact on graduation rate. Additionally, we found that central libraries had a more significant impact on graduation rate, as opposed to branch libraries. This may be because central libraries have better resource allocation, larger collections, and foster more engagement. As expected, the largest impacts on graduation rate were not library-related, but socioeconomic; the effect of libraries is trumped by the socioeconomic background and home life of a student.

In the top-performing counties, the relationships between library variables and graduation rates become more positive but less significant. In the bottom-performing counties, where socioeconomic challenges are most acute, the influence of libraries appears to be less consistent and may be overshadowed by the negative impact of economic disparities.

Our research provides exciting results, and further research could tease out a causal relationship between public libraries and graduation rates. If so, local governments around the United States could theoretically point to this relationship to influence policy decisions. Policymakers could prioritize funding public libraries, and specifically enhancing library engagement, to improve graduation rates. That being said, resource allocation targeting economic disparity may be a better use of funds, as these variables were the most significant in our regressions.

Appendix A: Variable Sources and Data

Figure A.1: Correlation Matrix of Included Variables

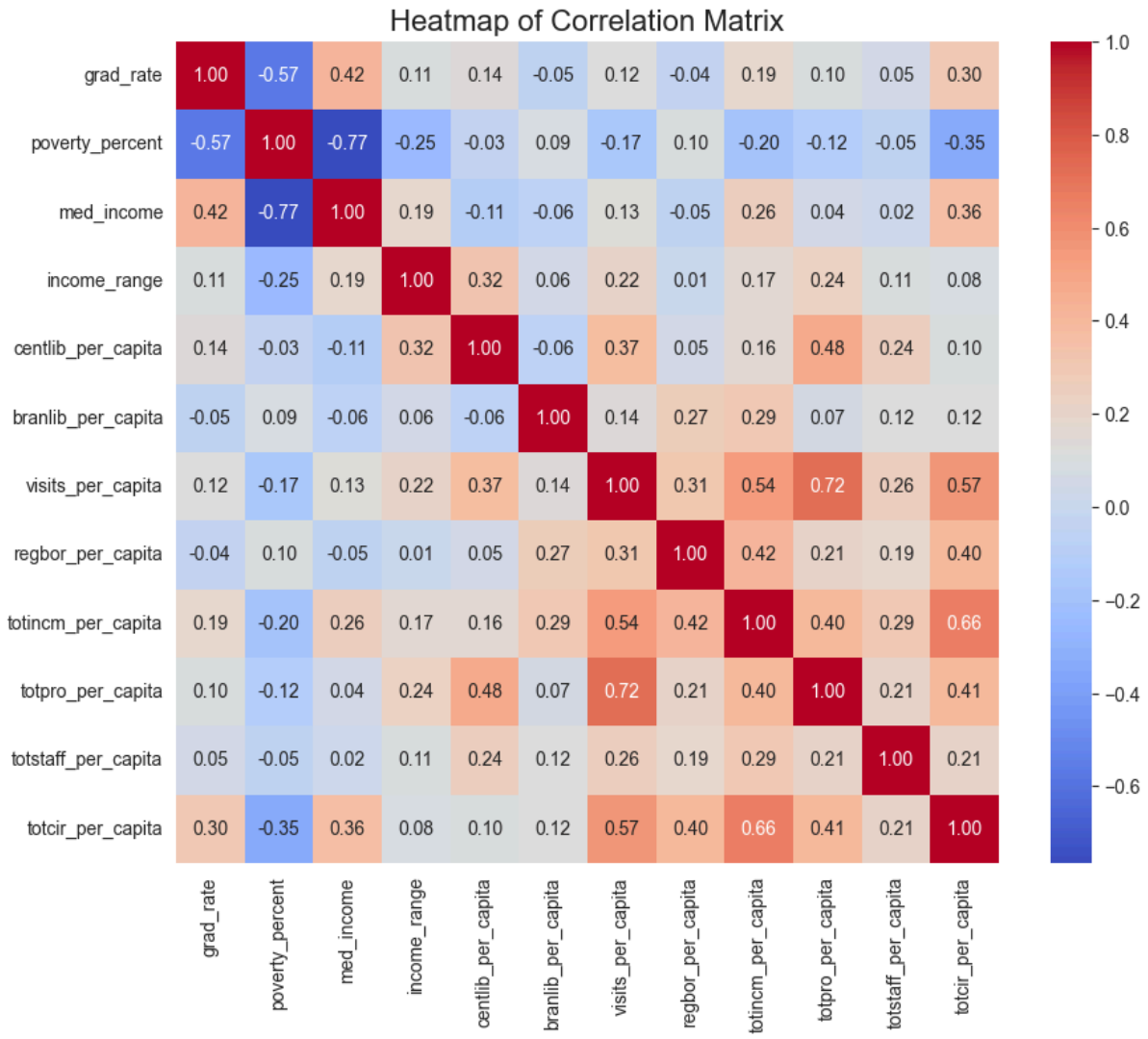


Table A.2: Variable Definitions and Sources

Variable	Definition	Source
grad_rate	County 5-year estimates of high-school graduation rate (%)	2022 ACS
visits_per_capita	Number of visits to all library systems in a county	2022 PLS
regbor_per_capita	Per-capita count of users registered to public library systems in a given county	2022 PLS
locgvt_per_capita	Per-capita funds from local governments directed towards public library systems in a given county	2022 PLS
totpro_per_capita	Per-capita number of "synchronous program sessions" (as defined by the PLS) held across public library systems in a given county	2022 PLS
totstaff_per_capita	Per-capita number of full-time staff across public library systems in a given county	2022 PLS
totcir_per_capita	Per-capita circulation volume (including electronic and print materials) across public library systems in a given county	2022 PLS
centlib_per_capita	Per-capita number of central libraries in a given county	2022 PLS
branlib_per_capita	Per-capita number of branch libraries in a given county	2022 PLS
poverty_percent	Percentage of children ages 5–17 in families and related to the householder living in poverty in a given county	2022 SAIPE
med_income	Median household income among households in the county	2022 SAIPE
income_range	The difference between the upper and lower 90% C.I. for median income	2022 SAIPE

Appendix B: Regression Analysis Results

Table B.1: Ridge and Lasso Regression Coefficients for Predicting Graduation Rates in All U.S. Counties

Feature	Ridge	Lasso
poverty_percent	-2.982074	-2.925565
centlib_per_capita	0.875742	0.694940
totcir_per_capita	0.840902	0.695533
visits_per_capita	-0.585155	-0.427737
income_range	-0.419217	-0.314131
totincm_per_capita	0.249516	0.053468
regbor_per_capita	-0.216382	-0.078291
totstaff_per_capita	-0.114141	-0.003888
med_income	-0.093407	0
totpro_per_capita	-0.098995	0
branlib_per_capita	-0.008305	0

Table B.2: Ridge and Lasso Regression Coefficients for Predicting Graduation Rates in Top 15% Counties

Feature	Ridge Coefficient	Lasso Coefficient
income_range	0.132925	0.197750
centlib_per_capita	0.089224	0.121547
visits_per_capita	0.052768	0.065464
med_income	0.075009	0.097454
totcir_per_capita	0.033759	0.034611
totpro_per_capita	0.039803	0.011654
totstaff_per_capita	0.053944	0.028492
branlib_per_capita	-0.020050	0
regbor_per_capita	0.021000	0
poverty_percent	-0.019450	0
totincm_per_capita	0.005189	0

Table B.3: Ridge and Lasso Regression Coefficients for Predicting Graduation Rates in Bottom 15% Counties

Feature	Ridge Coefficient	Lasso Coefficient
poverty_percent	-0.344838	-0.743188
income_range	-0.398101	-0.617222
med_income	-0.189547	-0.484054
totcir_per_capita	-0.136290	-0.134625
totincm_per_capita	-0.139685	-0.092002
centlib_per_capita	0.106410	0.178806
branlib_per_capita	0.041003	0
visits_per_capita	0.024150	0
totpro_per_capita	0.078223	0
totstaff_per_capita	-0.054362	0
regbor_per_capita	-0.020858	0

Table B.4: Ridge and Lasso MSE

	All Counties	Top 15%	Bottom 15%
Ridge Train MSE	18.960	1.010	16.545
Ridge Test MSE	19.096	1.069	22.324
Lasso Train MSE	19.068	0.998	16.366
Lasso Test MSE	19.226	1.067	21.996

Appendix C: Works Cited

Literature Review:

Clark, Christina *Linking School Libraries and Literacy*, https://files.eric.ed.gov/fulltext/ED513438.pdf. Accessed 9 Dec. 2024.

Emmons, Mark, and Frances C. Wilkinson. "The Academic Library Impact on Student Persistence." *College & Research Libraries*, vol. 72, no. 2, 1 Mar. 2011, pp. 128–149, https://doi.org/10.5860/crl-74r1.

Power, Emma, et al. "'Working Together': Public libraries supporting rural, regional, and remote low-socioeconomic student success in partnership with Universities." *Journal of the Australian Library and Information Association*, vol. 68, no. 2, 3 Apr. 2019, pp. 105–125, https://doi.org/10.1080/24750158.2019.1608497.

Data Sources:

Bureau, US Census. "Small Area Income and Poverty Estimates (SAIPE) Program." *Census.Gov*, 1 July 2022, www2.census.gov/programs-surveys/saipe/datasets/2022/2022-state-and-county/est22.xls

Bureau, US Census. "Using 1-Year or 5-Year American Community Survey Data." *Census.Gov*, 31 Oct. 2024, www.census.gov/programs-surveys/acs/guidance/estimates.html.

"Federal Reserve Economic Data." *FRED*, Federal Reserve Bank of St. Louis, fred.stlouisfed.org/release/tables?rid=330&eid=394818. Accessed 8 Dec. 2024.

"Public Libraries Survey." *Institute of Museum and Library Services*, www.imls.gov/research-evaluation/data-collection/public-libraries-survey. Accessed 8 Dec. 2024.