A 1% Increase in The Proportion of Spending on Facilities Decreases Graduation Rates by 0.2%: An Analysis on Ontario School Districts and Related Predictors*

Andrew Goh

December 15, 2024

Using data from Ontario's school board achievement, financial, and demographic datasets, we analyze the relationship between various district-level factors and four-year graduation rates. Through polynomial regression analysis, we find that the percentage of households without post-secondary degrees and relative spending on facilities have significant negative correlations with graduation rates (-0.136, -0.197), while per-student expenses show a non-linear relationship. These findings suggest that socioeconomic factors and resource allocation decisions have statistically significant impacts on academic outcomes, with implications for education policy and resource distribution in Ontario school districts.

1 Introduction

The relationship between educational resources and academic outcomes has been a central focus of education policy research, particularly in publicly funded school systems. In Ontario, where school boards receive varying levels of funding and serve communities with diverse socioeconomic profiles, understanding how different factors influence student success is crucial for effective policy-making (Campbell 2020).

Recent studies have suggested that both financial resources and community characteristics play significant roles in determining student outcomes (Faulk 2010; Mehreen 2023). However, the specific mechanisms through which these factors interact, particularly in the Ontario context, remain unclear. This gap in understanding is particularly relevant given Ontario's significant

*Code and data are available at: https://github.com/mushroomcarbon/schools_and_money investment in education, with annual spending projected to be approximately $37.6 billion (OPSBA 2024).

Our paper examines how various district-level factors—including financial allocations, socioeconomic indicators, and operational characteristics—correlate with four-year graduation rates across Ontario school boards. We focus specifically on the relationships between graduation rates and five key variables: the percentage of households without post-secondary degrees, the percentage of low-income households, the percentage of the total budget spent on facilities, per-student expenses, and total enrolment.

Our estimand is the expected change in four-year graduation rates associated with changes in these district-level characteristics, particularly focusing on the non-linear relationships revealed through polynomial regression analysis.

The results suggest that socioeconomic factors, particularly the percentage of households without post-secondary degrees, show the strongest relationship with graduation rates, with each percentage point increase associated with a 0.136 percentage point decrease in graduation rates. We also find that the percentage of budget spent on facilities has a significant negative relationship with graduation rates, with each additional percentage point allocated to facilities associated with a 0.197 percentage point decrease in graduation rates. While per-student expenses show a U-shaped relationship with graduation rates, this relationship did not reach statistical significance at conventional levels. Our model explains approximately 36.5% of the variance in graduation rates, indicating that while these factors are important, there remain other unobserved variables influencing student outcomes.

The remainder of this paper is organized as follows: Section 2 describes our dataset and measurement approach, Section 3 outlines our polynomial regression methodology, Section 4 presents our findings, and Section 5 explores the implications and limitations of our analysis.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023) for the analysis and presentation of the project. The tidyverse (Wickham et al. 2019) ecosystem, particularly dplyr (Wickham et al. 2023), was used for data manipulation. For model development and evaluation, we utilized caret (Kuhn and Max 2008) for model training and glmnet (Friedman, Tibshirani, and Hastie 2010; Simon et al. 2011; Tay, Narasimhan, and Hastie 2023) for regularized regression methods. The arrow (Richardson et al. 2024) package was employed for efficient data reading and writing, while broom (Robinson, Hayes, and Couch 2023) was used for converting statistical objects into tidy data frames. Additional file management was facilitated by the here (Müller 2020) package. Styler (Müller and Walthert 2023) was used to style the code.

In order to build a model predicting the effect that various variables on the school-district-level have on their academic successes, we used three datasets from the Ontario government's Open Data portal: the School Board Achievements and Progress dataset (Government of Ontario 2024a), the School Board Financial Reports dataset (Government of Ontario 2024b), and the School Information and Student Demographics dataset (Government of Ontario 2024c), each recording different information on school district organised by district ID.

After cleaning the data by selecting the desired variables from the three different datasets, combining them into one file, and removing N/A variables, we obtain data on the Four Year Graduation Rate, Total Expenses, Expenditure on Facilities, and Total Enrolment values for 50 different Ontarian school districts, as well as the percentage of students that are identified as low income, the percentage of students without a parent that has a school degree/certificate, the percentage of total expenses spent on facilities, and the expenses per quota in each of the aforementioned districts.

As a measurement of academic success, the 4-year graduation rate was chosen due to its strong correlation with performance on the OSSLT, a standardized test used to assess Ontarian students' levels of academic literacy (Studies in Developmental Education (OASDI) 2024).

2.2 Measurement

Data from the School Board Progress Report is either self-reported by schools or obtained by the EQAO (Education Quality and Accountability Office), as described by the Ministry of Education of Ontario (Education 2024a). Schools are expected to self-report internal information such as class size changes and financial status, whereas indicators of academic progress on a district-wide-level, such as standardized testing results and graduation rates, fall under the authority of the EQAO. EQAO's data collection and processing methodology follows the Statistics Canada Quality Guidelines (Canada 2019), and the aforementioned standardized assessments that EQAO records data on are administered by EQAO itself (Quality and (EQAO) 2021).

Data regarding school districts' financial reports is collected by the Government of Ontario via the Education Financial Information System (EFIS). Using EFIS, school districts submit their financial information to the Ontario Ministry of Education, where they are then compiled into a single dataset.

The School Information and Demographics dataset is based on a combination of school-submitted information and EQAO data. In light of privacy concerns such as data from this dataset potentially being traced back to individual students or small groups of students, random error and suppressing is utilised in this dataset (Education 2024b). Specifically, when the number of students referred to by each individual cell in the table is less than 50, then the cell is marked as SP, or suppressed. Values concerning more students, on the other hand, have their percentages rounded up/down at random to a certain granularity, making sure that there is always a potential error bound of at most 5 students for each dataset.

Several important limitations that must be acknowledged exist in the data collection and reporting processes for Ontario's education datasets. First, the self-reporting nature of some metrics by schools introduces potential inconsistencies in how different institutions interpret and report their data. Second, the privacy protection measures, while necessary, create inherent imprecision through data suppression and random rounding, particularly affecting analysis of smaller student populations or subgroups. The granularity of financial reporting through EFIS may also vary between school districts, potentially impacting the comparability of financial metrics. Additionally, the combination of multiple data sources (self-reported, EQAO, and demographic data) may lead to temporal misalignment, as different metrics might be collected at different times throughout the school year. Finally, the standardized testing results from EQAO, while following Statistics Canada guidelines, may not fully capture the diverse learning outcomes and educational experiences of students, particularly those from marginalized communities or with special educational needs.

2.3 Data Summary

Table 1 presents an overview of the summary statistics of our key variables across Ontario school boards, and Table 2 and Table 3 presents a snapshot of the data itself.

Table 1: Summary Statistics of Key Variables

Variable	Mean[1]	Std. Dev.	Minimum	Maximum	Median
Four Year Graduation Rate	0.82	0.08	0.59	0.96	0.83
Total Expenses	0.45	0.58	0.03	3.89	0.27
Total Enrolment	28.45	37.88	0.71	231.48	15.40
percentage_spent_on_facilities	9.14	0.99	6.06	12.55	9.20
expenses_per_quota	18,730.25	6,036.02	13,523.54	44,148.26	16,523.47
percent_no_degree	5.71	2.14	1.85	10.66	5.40
percent_low_income	15.77	3.17	10.35	25.92	15.62

[1] Total Expenses shown in billions ($), Enrolment in thousands, percentages as raw values, expenses per student in dollars.

Table 2: Sample of First Five Rows from Ontario School Board Dataset (Part 1)

Board				Grad	Total Expenses
ID	Name	Region	City	Rate	($)
B28010	Algoma DSB	North Region	Sault Ste Marie	0.719	207869377
B67202	Algonquin and Lakeshore CDSB	East Region	Napanee	0.895	188706486

Board ID	Name	Region	City	Grad Rate	Total Expenses ($)
B66010	Avon Maitland DSB	West Region	Seaforth	0.802	243574611
B66001	Bluewater DSB	West Region	Chesley	0.715	277108177
B67164	Brant Haldimand Norfolk CDSB	West Region	Brantford	0.818	176090176

Table 3: Sample of First Five Rows from Ontario School Board Dataset (Part 2)

Facility Expenses ($)	Enrolment	Low Income (%)	No Degree (%)	Facilities (%)	$/Student
18775889	10265	17.572	5.268	9.033	20250.31
16466344	11670	16.166	4.160	8.726	16170.22
26131858	14865	13.491	9.979	10.728	16385.78
28049633	17605	16.302	9.447	10.122	15740.31
16557722	10910	13.956	5.721	9.403	16140.25

The summary statistics reveal considerable variation across Ontario school boards. Four-year graduation rates average 82%, ranging from 59% to 96%. The diverse scales of the school boards, whether it be with respect to finances or population, are also evident: total expenses vary dramatically, from $30 million to $3.89 billion (with a mean of $450 million), and so does total enrolment, which ranges from 710 students to 231,480 students, with a mean of 28,450 students. The percentage spent on facilities ranges from 6.06% to 12.55%, with a mean of 9.14%, while per-student expenses average $18,730, ranging from $13,524 to $44,148.

2.4 Response Variable

2.4.1 Four-Year Graduation Rate

The outcome variable is the four-year graduation rate for Ontario school boards, recorded as a percentage of students who graduate within four years of starting high school.

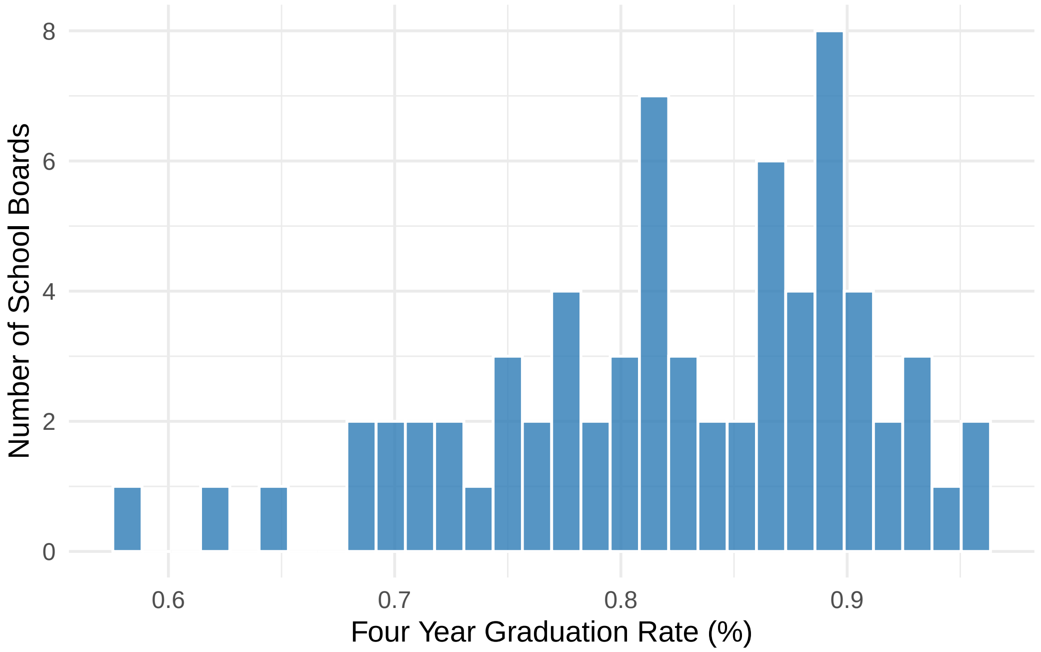

Figure 1: Distribution of Four-Year Graduation Rates Across Ontario School Boards

As displayed by Figure 1, graduation rates across Ontario school boards follow an approximately normal distribution, with most boards achieving rates between 80% and 90%. The mean graduation rate is 82%, with rates ranging from 59% to 96%.

2.5 Predictor Variables

The predictor variables in our analysis include the percentage of school board budget spent on facilities, per-student expenses, percentage of students without a parent holding a post-secondary degree, and percentage of students from low-income households. The distributions of these variables are as follows:

2.5.1 Facilities Spending

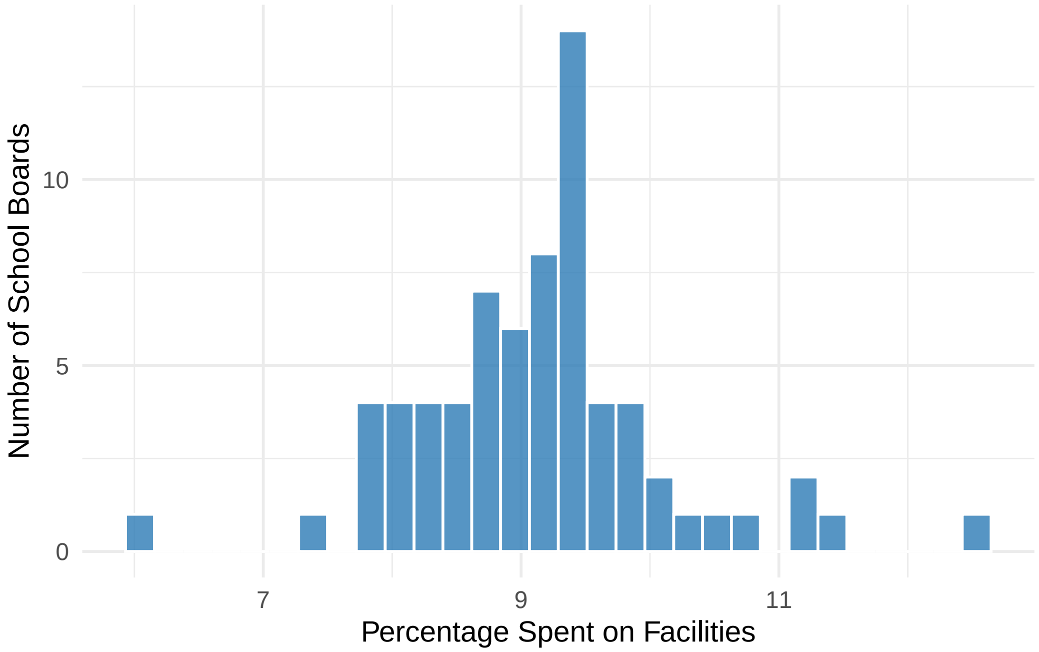

Figure 2: Distribution of facility spendings as a percentage of total spendings

The percentage of budget allocated to facilities maintenance and operations ranges from 6.06% to 12.55%, with a mean of 9.14%. This metric captures the varying infrastructure needs and priorities across different school boards.

2.5.2 Per-Student Expenses

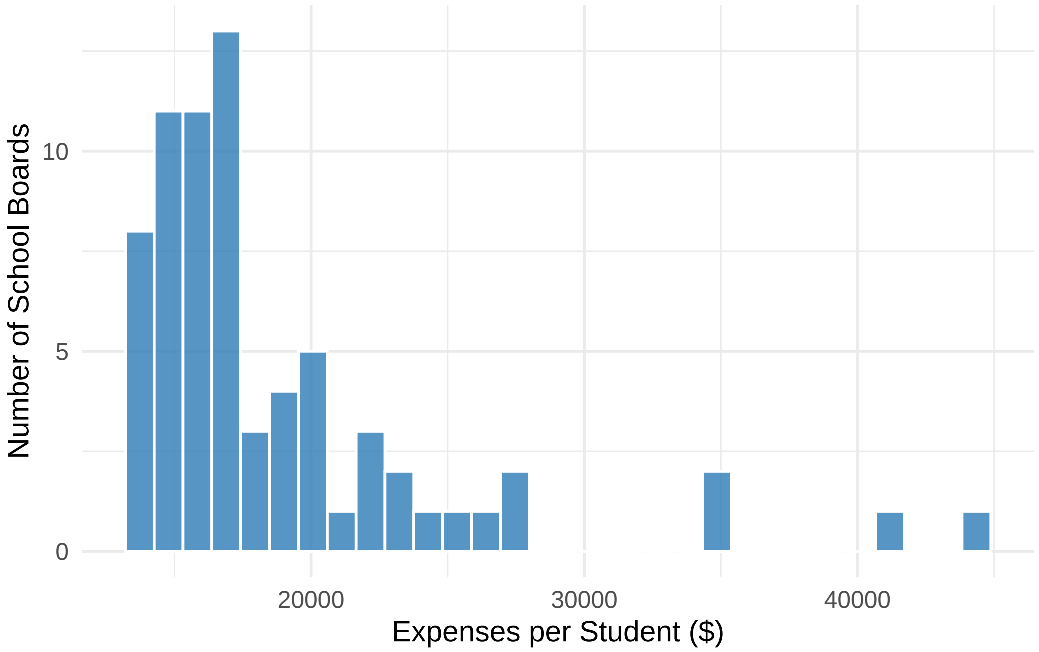

Figure 3: Distribution of expenses per student

As per Figure 3, per-student expenses show considerable variation across boards, averaging $18,730 with a range from $13,524 to $44,148. This variation reflects differences in operational costs, programs offered, and resource allocation across different regions.

2.5.3 Parental Education

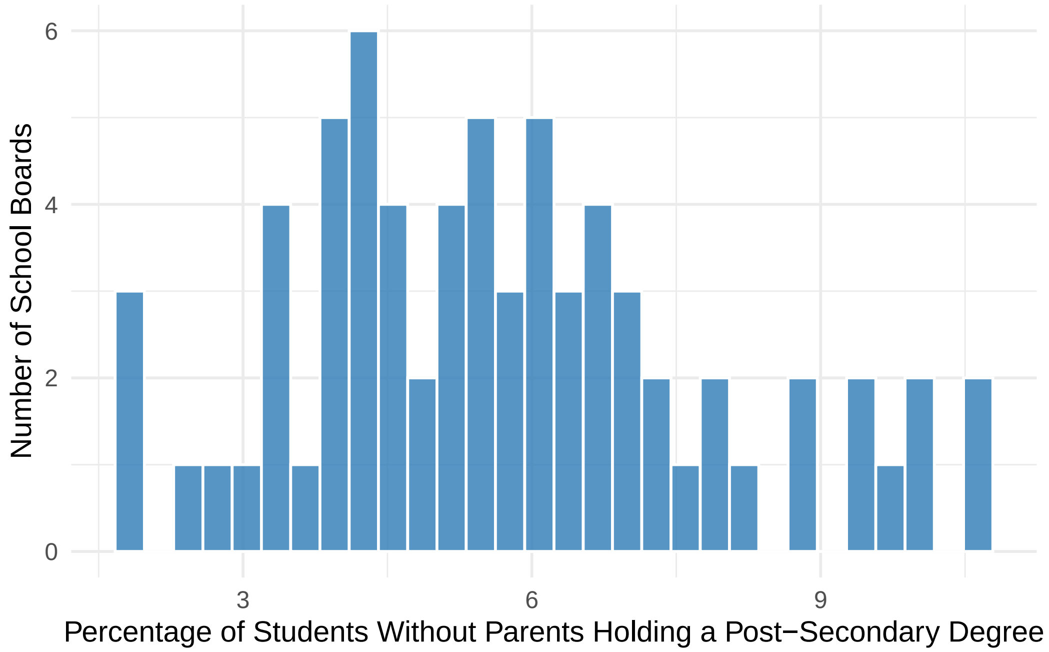

Figure 4: Percentage of students without parents holding a post-secondary degree

As can be seen in Figure 4, the percentage of students whose parents do not hold a post-secondary degree or certificate varies from 1.85% to 10.66%, with a mean of 5.71%. This metric serves as an indicator of the educational background of the school board's community.

2.5.4 Low-Income Status

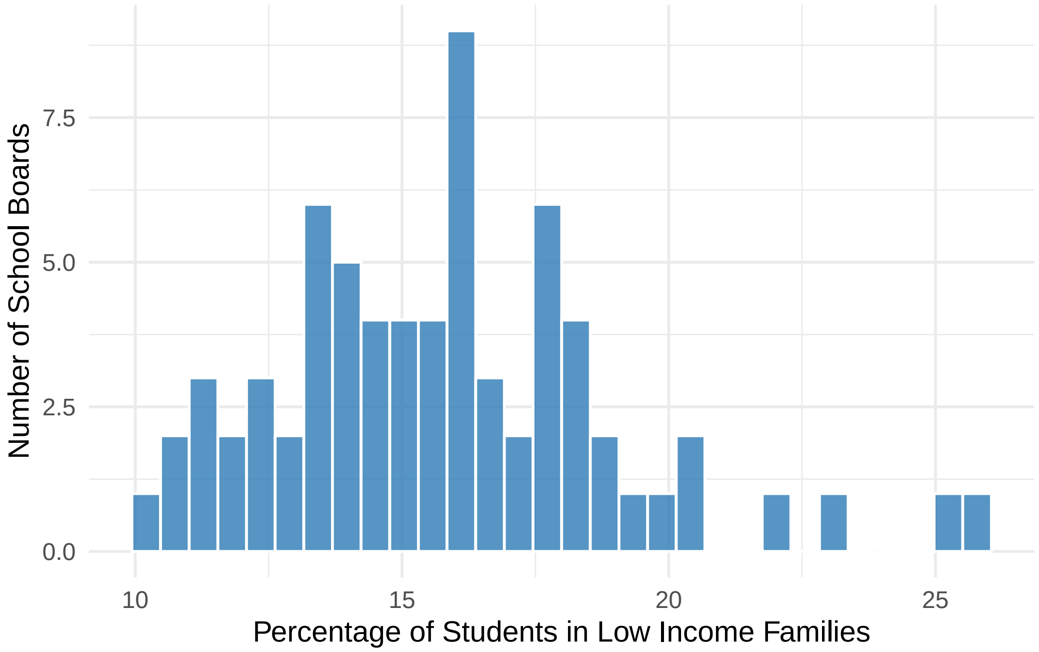

Figure 5: Distribution of students in low-income families

The percentage of students from low-income households ranges from 10.35% to 25.92%, with a mean of 15.77%. This variable describes the socioeconomic composition of each school board's student population, and we can see from Figure 5 the diversity of socioeconomic compositions *between* school districts.

2.5.5 Enrolment

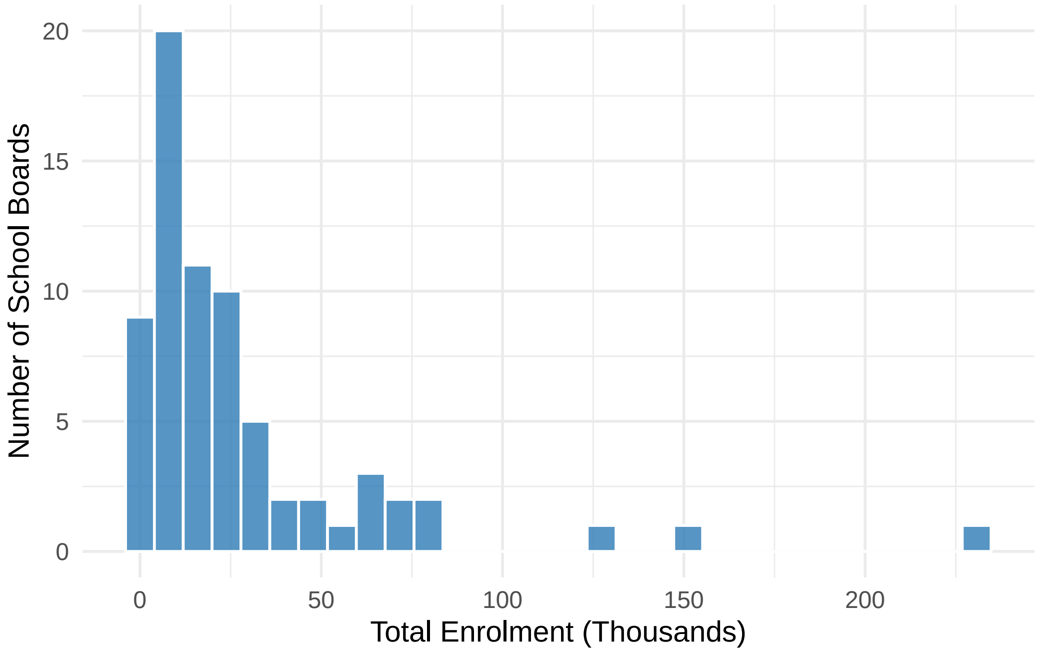

Figure 6: Distribution of total enrolment across school boards

Figure 6 shows that total enrolment across boards ranges from 710 to 231,480 students, with a mean of 28,450 students. This wide range reflects Ontario's diverse school board sizes, from small rural boards to large urban districts.

2.6 Feature Engineering, Selection, and Extraction

The selection of these specific predictor variables was guided by both theoretical relevance and statistical considerations regarding multicollinearity. Initial exploratory analysis revealed strong correlations between several potential predictors. For instance, total expenses and total enrolment showed an extremely high correlation ($r > 0.95$), as larger boards naturally have higher total expenses. Similarly, various socioeconomic indicators such as median household income, unemployment rates, and low-income status demonstrated substantial overlap in their variation. To address these multicollinearity concerns, we used feature engineering, extraction, and selection to isolate representative variables that capture distinct aspects of school board characteristics while minimizing redundancy:

Per-student expenses were calculated from total expenses and total enrolment and was chosen as the analysis metric in the financial dimension over total expenses to control for board size and provide a standardized measure of financial resource usage that can be compared in an isolated environment. Percentage spent on facilities was retained as it represents a unique aspect of resource allocation independent of overall spending levels, as well as being reminiscent of the author's initial driving question of whether increased spending on libraries would correlate with better academic performance. Among the (highly intercorrelated) various socioeconomic indicators, the percentage of low-income students and parental education levels were selected as they captured different dimensions of socioeconomic status while maintaining relatively low correlation with each other (r < 0.4). Total enrolment was included as a control variable to account for potential scale effects that might influence graduation rates independently of other factors. This, combined with expenses per capita, still include all the information that was present in the original total expenses feature, but in the form of two less-correlated variables that can be individually considered and analysed with an additive model.

Feature selection helps ensure our analysis avoids the statistical issues associated with multicollinearity while still capturing the key factors that may influence academic performance.

2.7 Note on Standardization

While the data is presented in its raw form above, all variables are standardized (mean = 0, standard deviation = 1) for our subsequent modeling analysis to ensure comparability of coefficients and improve numerical stability, as well as to improve results obtained from the ridge and lasso regression models (Tibshirani 1996). This standardization helps interpret the relative importance of different predictors while maintaining their underlying relationships.

3 Model

3.1 Model Set-up

We evaluate the performance of several regression models to understand the relationship between school board characteristics and graduation rates:

3.1.1 Linear Regression

We begin with our baseline model, a simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \epsilon_i \tag{1}$$

Where y_i is the Four Year Graduation Rate for school board i, ϵ_i is the error term, and x_{1i} to x_{5i} are the predictor variables: percentage of students without a parent holding a post-secondary degree, percentage of low-income students, percentage spent on facilities, expenses per student, and total enrolment.

3.1.2 Shrinkage Regression

To address potential multicollinearity and model overcomplexity between our predictors, we also implemented Ridge and Lasso regression models, respectively:

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

$$\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

Where λ is the regularization parameter chosen through cross-validation and the x-values are the same as above.

3.1.3 Polynomial Regression

Finally, a polynomial regression model of degree 2 was fitted to the data to account for non-linear relationships between the data. The model is structured as follows:

$$y_i = \beta_0 + \sum_{j=1}^{5} (\beta_{j1} x_{ji} + \beta_{j2} x_{ji}^2) + \epsilon_i \tag{4}$$

Where β_{j1} represents the linear term coefficient and β_{j2} represents the quadratic term coefficient for each predictor.

3.2 Model Selection

After comparing the results of the models and assessing how well they each fit to the data respectively, the polynomial regression model was chosen due to multiple reasons:

1. **Non-linear Relationships**: Initial exploratory data analysis revealed non-linear relationships between several predictors and graduation rates, particularly for expenses per student and total enrolment. This suggests that a polynomial regression model potentially can capture more trends in the data, more accurately, compared to a linear model, though the latter is simpler. The tradeoff in this case is sensible due to the added model complexity actively capturing statistically significant trends in the data.

2. **Model Comparison**: Comparing Root Mean Squared Error (RMSE) across models:

```
# A tibble: 4 x 2
  Model                 Metric
  <chr>                 <chr>
1 Linear Regression     RMSE:  0.838
2 Polynomial Regression RMSE:  0.791
3 Ridge                 RMSE:  0.856
4 Lasso                 RMSE:  0.85
```

The polynomial regression model showed the lowest MSE, which indicates that it has a better fit to the data and is hence a "better" model to use in this case (Willmott 1981).

3. **Interpretability**: While Ridge and Lasso regressions help with multicollinearity and variable selection respectively, they don't capture the non-linear relationships we observe in the data. The polynomial model allows us to interpret both linear and quadratic effects of our predictors on graduation rates.
4. **Statistical Significance**: As shown in our model summary, several quadratic terms (particularly for expenses per quota) are statistically significant, confirming the value of including these higher-order terms.

3.3 Model Validation and Diagnostics

3.3.1 Residual Analysis

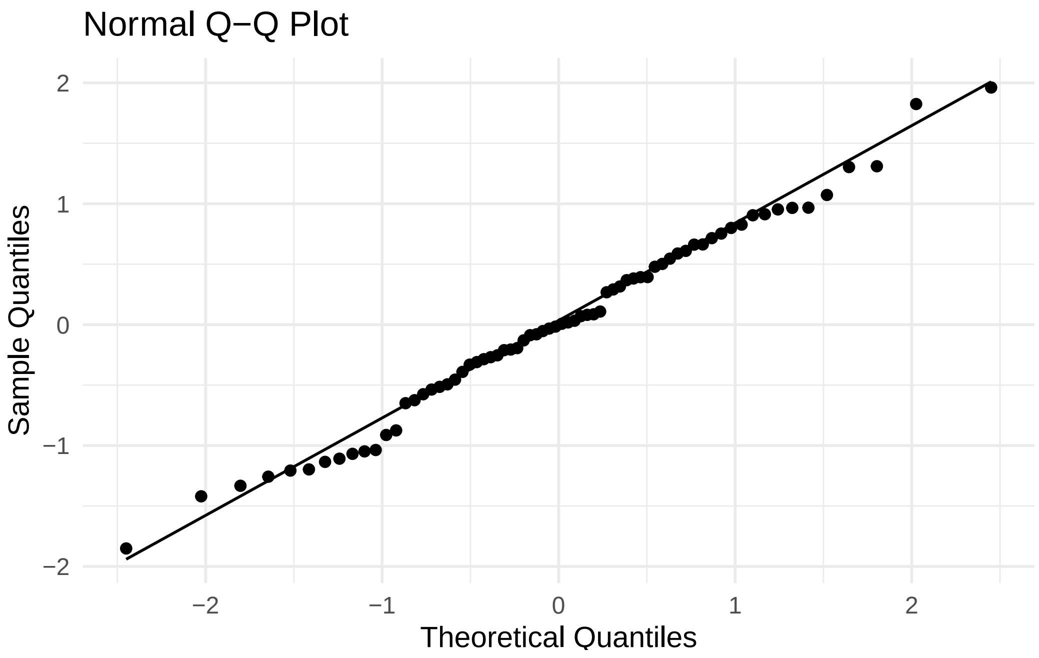

Figure 7: Normal QQ Plot for Polynomial Regression Model

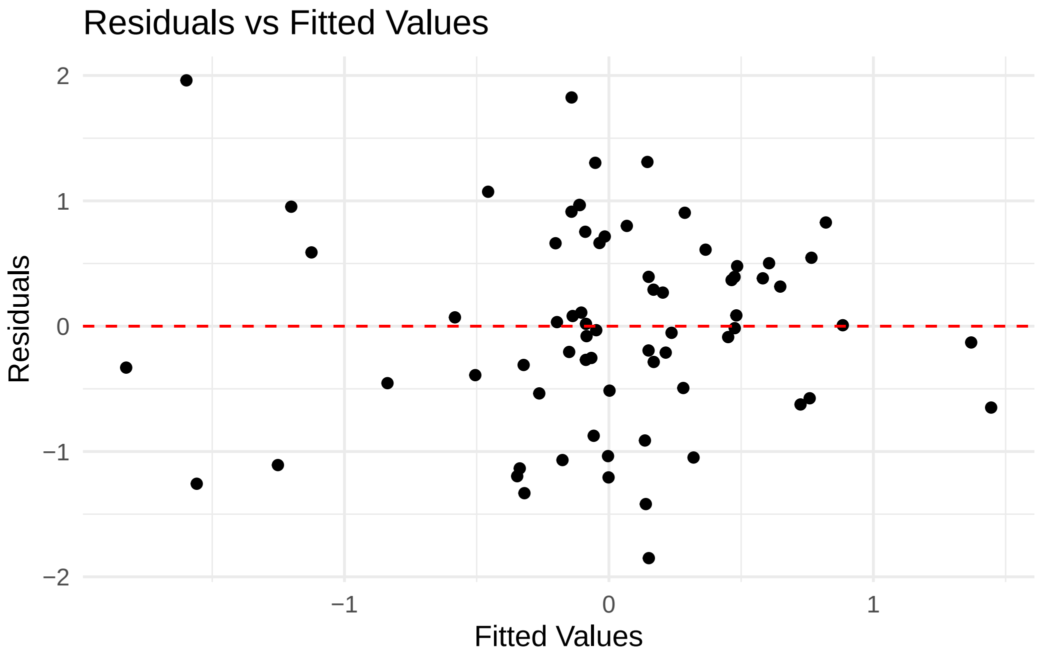

Figure 8: Residuals vs Fitted Values Plot for Polynomial Regression Model

Figure 7 and Figure 8 show our residual diagnostics. The Q-Q plot suggests that residuals are approximately normally distributed, while the Residuals vs Fitted plot shows no clear pattern, indicating homoscedasticity of residuals.

3.3.2 Model Performance Metrics

Table 4: Model Performance Metrics

Metric	Value
R-squared	0.365
RMSE	0.791

As shown in Table 4, our model explains approximately 36.5% of the variance in graduation rates. While this R-squared value might seem modest, it is reasonable given the complexity of factors that influence educational outcomes.

3.3.3 Multicollinearity Check

We examine Variance Inflation Factors (VIF) for the linear terms to assess multicollinearity:

Table 5: Variance Inflation Factors for Linear Terms

Variable	VIF
percent_no_degree	1.47
percent_low_income	1.54
percentage_spent_on_facilities	1.05
expenses_per_quota	1.25
‘Total Enrolment‘	1.35

VIF values above 5 would indicate problematic multicollinearity. Our values suggest that multicollinearity is not a major concern in our model.

3.3.4 Cross-Validation

To assess the model's predictive performance and guard against overfitting, we performed k-fold cross-validation:

Table 6: Cross-Validation Results

Metric	Value

CV RMSE	0.921
CV R-squared	0.323

The cross-validation results suggest that our model's performance is stable across different subsets of the data, with consistent RMSE values between training and testing sets.

3.3.5 Model Assumptions

Our polynomial regression model relies on several key assumptions:

1. **Linearity**: While we don't assume linear relationships between predictors and the response, we assume that the quadratic terms adequately capture the non-linear relationships.
2. **Independence**: We assume that graduation rates of different school boards are independent of each other, which is reasonable given the administrative separation between boards.
3. **Homoscedasticity**: As shown in our residual plot, the variance of residuals appears relatively constant across fitted values.
4. **Normality**: The Q-Q plot suggests that residuals are approximately normally distributed.

These diagnostics and validations support our choice of polynomial regression as the final model, though we acknowledge the limitations discussed in Section 5.

4 Results

We present the results from our polynomial regression analysis in order of statistical significance.

4.1 Primary Findings

Table 7: Polynomial Regression Coefficients

Variable	Estimate	Std. Error	P-value
(Intercept	0.000	0.103	1.000
percent_no_degree1	−3.627	1.151	0.003
percent_no_degree2	0.978	1.213	0.423
percent_low_income1	0.573	1.166	0.625

percent_low_income2	0.286	1.270	0.823
percentage_spent_on_facilities1	−2.435	1.040	0.023
percentage_spent_on_facilities2	−0.692	1.028	0.503
expenses_per_quota1	−0.561	1.333	0.675
expenses_per_quota2	1.932	1.102	0.085
‘Total Enrolment‘1	0.249	1.243	0.842
‘Total Enrolment‘2	−0.186	1.143	0.871

4.1.1 Parental Education

The percentage of households without post-secondary degrees shows the strongest relationship with graduation rates (β = -3.627, p = 0.003). After converting from standardized units back to real-world units, this relationship suggests that for every one percentage point increase in households without post-secondary degrees, graduation rates decrease by approximately 0.136 percentage points, holding other factors constant.

4.1.2 Facilities Spending

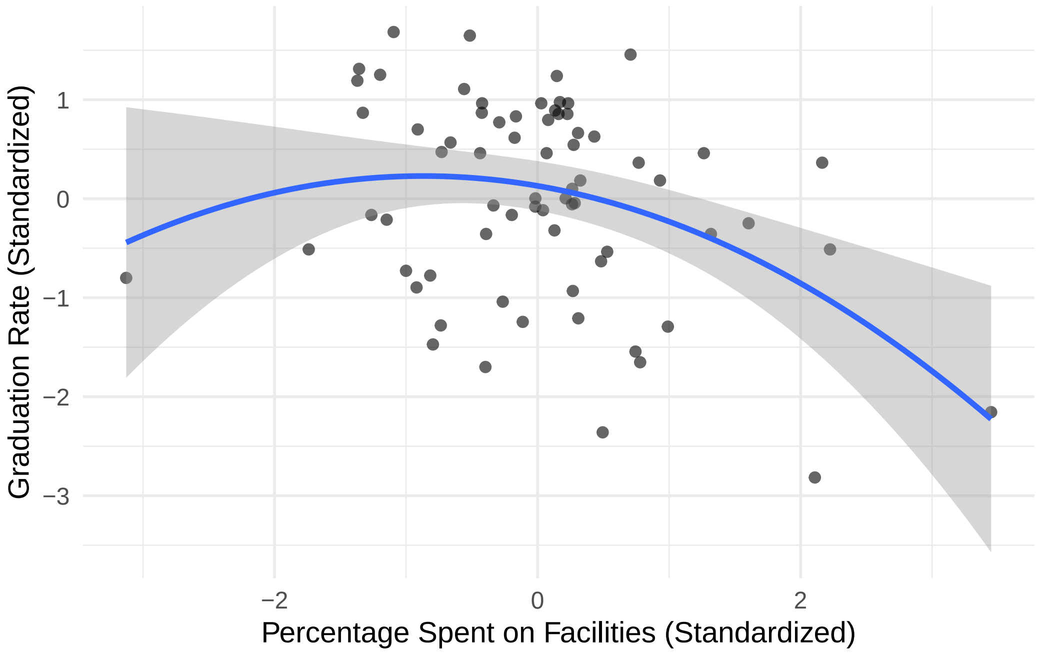

Figure 9: Relationship Between Facilities Spending and Graduation Rates

The percentage spent on facilities shows a significant negative linear relationship (β = -2.435, p = 0.023), as illustrated in Figure 9. After converting from standardized units back to real-world units, this suggests that for each additional percentage point of budget allocated to facilities, graduation rates decrease by about 0.197 percentage points, holding other factors constant. In other words, boards allocating a larger proportion of their budget to facilities tend to have lower graduation rates, possibly indicating trade-offs between infrastructure spending and other educational resources.

4.1.3 Per-Student Expenses

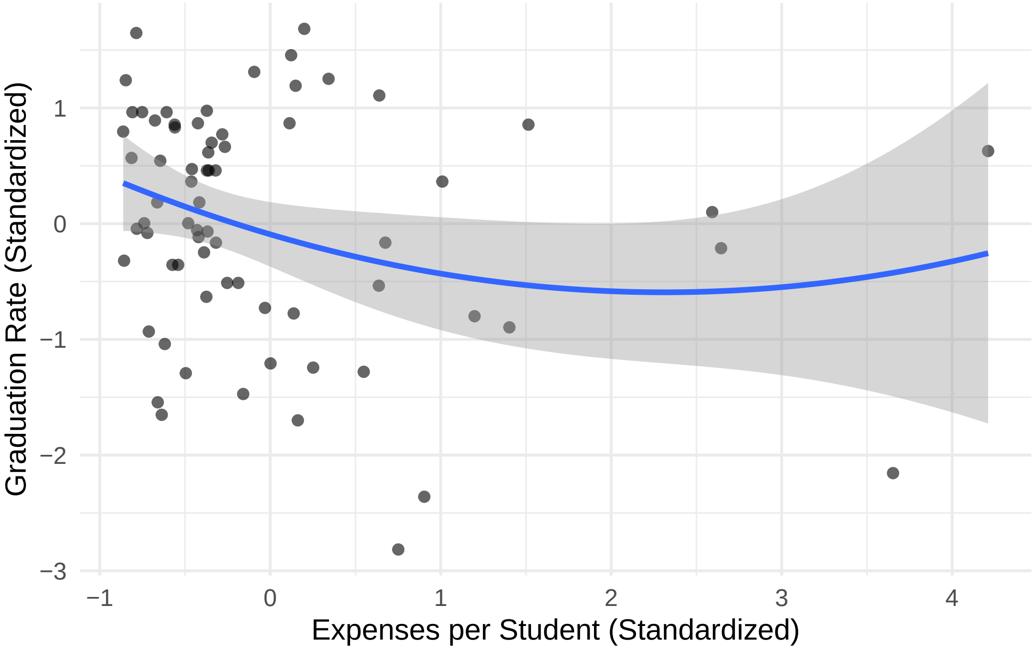

Figure 10: Non-linear Relationship Between Per-Student Expenses and Graduation Rates

Per-student expenses show a marginally significant quadratic relationship (β = 1.932, p = 0.085), as shown in Figure 10. After converting from standardized units back to real-world units, while not significant at the conventional $p < 0.05$ level, there appears to be a U-shaped trend in the relationship between per-student spending and graduation rates. This suggests that graduation rates tend to decrease with initial increases in per-student spending up to a certain threshold, after which additional spending is associated with increasing graduation rates. However, given the lack of statistical significance, these trends should be interpreted with caution.

4.2 Secondary Findings

4.2.1 Total Expenses and Enrolment

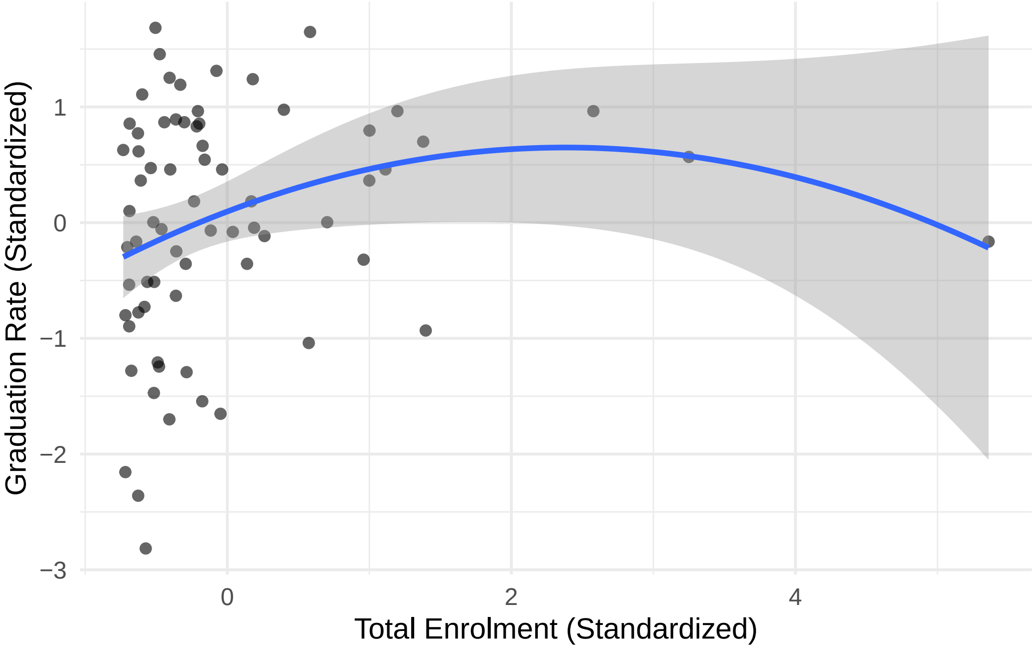

Figure 11: Relationships Between School Board Size Metrics and Graduation Rates

As shown in Figure 11, Total Enrolment showed no statistically significant relationship with graduation rates (linear term: $\beta = 0.249$, p = 0.842; quadratic term: $\beta =$ -0.186, p = 0.871). This suggests that the size of the school board, measured by enrollment, has minimal impact on graduation outcomes, with each additional thousand students associated with a non-significant change of 0.00053 percentage points in graduation rates once converted back into real-world units.

4.2.2 Model Fit

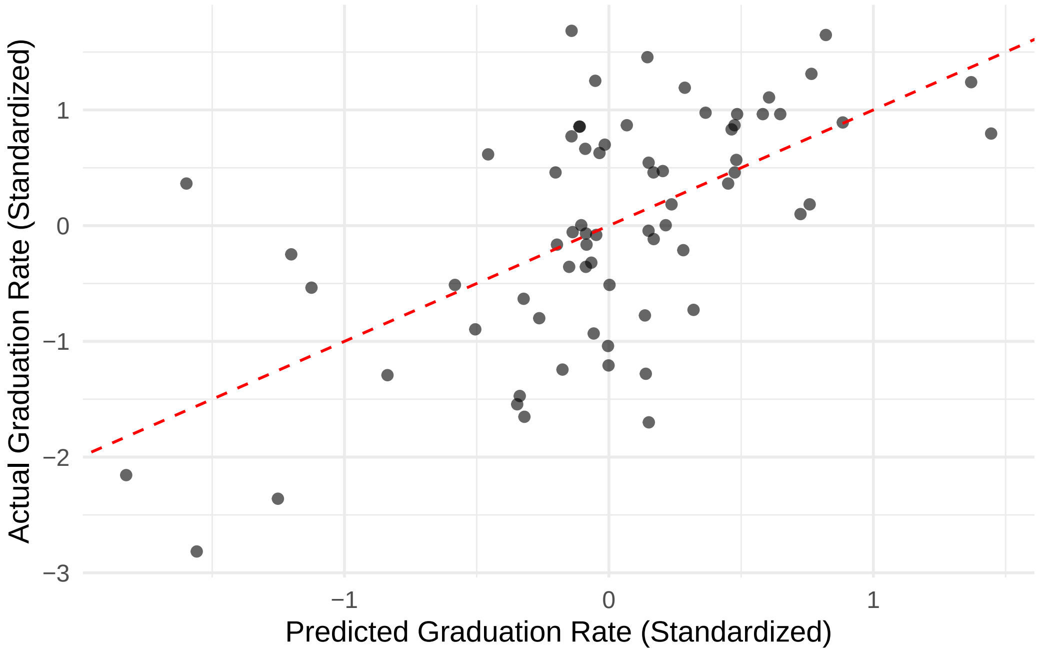

Figure 12: Predicted vs Actual Graduation Rates

Figure 12 shows the relationship between predicted and actual graduation rates. The model explains approximately 36.5% of the variance in graduation rates ($R^2 = 0.3648$), suggesting that while our identified factors are important, there remain other unobserved variables that influence graduation rates.

5 Discussion

5.1 Key Findings and Implications

The key takeaway from the model results is the strong negative relationship between the percentage of households without post-secondary degrees and graduation rates. This relationship suggests an intergenerational aspect to educational achievement, where students from communities with lower educational attainment face additional challenges in completing their secondary education.

The non-linear relationship between per-student expenses and graduation rates is particularly noteworthy. The U-shaped relationship we observed suggests that simply increasing funding may not always lead to better outcomes. This finding aligns with previous research suggesting that the effectiveness of educational spending depends heavily on how resources are allocated (Institute 2024). School boards might benefit from examining the spending patterns of high-performing boards that operate in the “efficient” region of this curve.

5.2 Resource Allocation Trade-offs

The negative relationship between facilities spending and graduation rates raises important questions about resource allocation. While maintaining adequate facilities is crucial for educational delivery, our findings suggest that boards allocating a larger proportion of their budget to facilities tend to have lower graduation rates. This could indicate that: some boards might be forced to prioritize urgent infrastructure needs over other educational resources, older facilities requiring more maintenance might be concentrated in areas facing other socioeconomic challenges, or that there might potentially be an optimal balance between infrastructure investment and direct educational spending, and further research in this direction is advised.

5.3 Limitations

Several limitations of our analysis should be noted:

5.3.1 Data Constraints

Our analysis is cross-sectional, looking at a single academic year (2023-2024). Therefore, trends existing in this model might not exist in previous years, and vice versa. Additionally, due to privacy concerns, lack of granularity in the data and similar reasons, more specific predictors such as the teacher-to-student ratio and student-to-faculty ratio as well as data specifically recording expenditures on libraries are lacking. This potentially led to the model not being able to utilise enough, or ideal, predictors in its training.

5.3.2 Methodological Considerations

Our polynomial regression model, while capturing non-linear relationships, may not fully represent more complex interactions between variables that are of higher degrees. The R-squared value of 0.404 suggests that substantial variation in graduation rates remains unexplained by our model, which again is likely due to the lack of data as mentioned above.

5.4 Future Research Directions

Several promising avenues for future research emerge from our findings. A longitudinal analysis examining how these relationships evolve over time could provide insights into the long-term effects of different spending patterns and policy changes. At a more granular level, investigating specific educational programs and interventions could help identify which approaches are most effective at improving graduation rates, particularly in communities with lower educational attainment. Future research could also focus on developing models to help school boards optimize their resource allocation decisions, particularly regarding the balance between facilities maintenance and other educational spending. Finally, studying how provincial funding formulas might be adjusted to better support boards facing particular challenges, such as aging infrastructure or high concentrations of socioeconomic disadvantage, would provide valuable policy insights.

5.5 Broader Context

Our findings contribute to the ongoing discussion about educational equity in Ontario. The strong relationship between community educational attainment and graduation rates suggests that breaking cycles of educational disadvantage may require interventions that extend beyond the school system itself. This might include community-based programs to support adult education, enhanced early childhood education initiatives, targeted support for communities with historically lower educational attainment, and/or integrated approaches that address both educational and broader socioeconomic factors

The complex relationships we've identified between financial resources and educational outcomes also suggest that policy makers should consider more nuanced approaches to school funding, moving beyond simple per-student funding formulas to consider the specific contexts and challenges faced by different school boards.

Appendix

A Model details

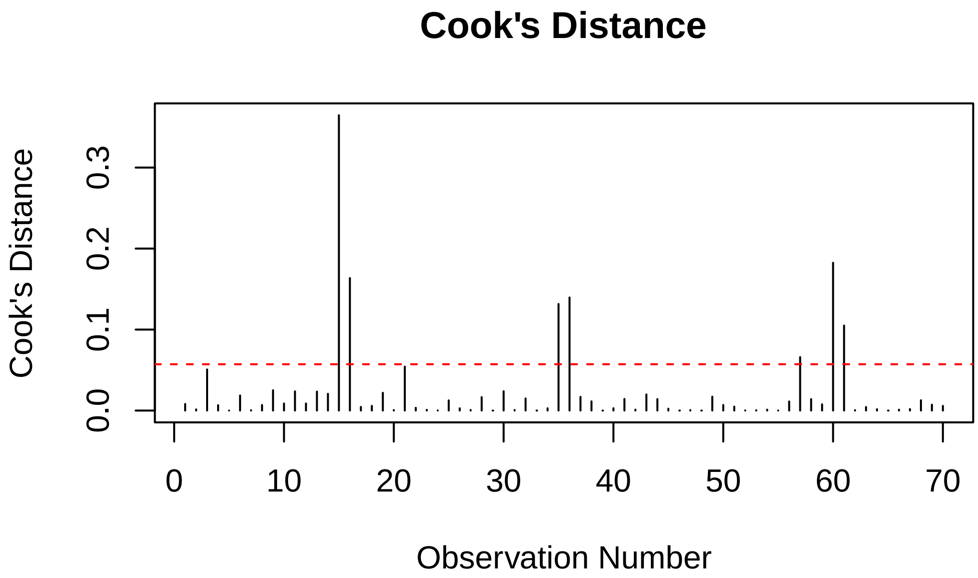

Figure 13: Cook's Distance Plot

B Model Diagnostics

Figure 13 shows Cook's distance for each observation, helping identify potentially influential data points. The dashed red line indicates a commonly used threshold of 4/n for identifying influential observations.

This diagnostic helps assess the validity of our model assumptions and identify potential areas of concern in our analysis.

C Data Collection and Sampling Considerations

C.1 Overview of Data Collection Process

The data used in this analysis comes from a complex, multi-source collection process that merges administrative data from Ontario school boards with demographic information and standardized test results. Understanding this process is crucial for interpreting our results and identifying potential sources of bias.

C.2 Administrative Data Collection

The Education Financial Information System (EFIS) serves as the primary collection mechanism for financial data. School boards submit standardized financial reports through EFIS, which then undergoes several quality control steps. These steps include automated validation checks for mathematical accuracy and internal consistency, cross-reference checks against previous years' submissions, manual review by Ministry staff for anomalous patterns, and opportunities for boards to explain unusual variations.

C.3 Sampling Frame and Coverage

While our dataset represents the complete population of public school boards in Ontario, it's important to note that this itself is a subset of all educational institutions in Ontario. Specifically, our sampling frame excludes private schools, First Nations schools, and adult education centers. This coverage limitation affects the generalizability of our findings.

C.4 Implications for Analysis

These data collection and measurement considerations have several implications for our analysis. Our focus on English-language public boards means our findings may not generalize to other educational contexts in Ontario, while financial data, being self-reported, may contain systematic measurement error. Furthermore, different variables are measured at different times throughout the school year, potentially introducing temporal misalignment in our cross-sectional analysis, and as stated by Data Ontario (Ontario 2024), data missing from the raw dataset is not missing completely at random (MCAR), which could introduce bias and/or artificially inflated variance in our estimates.

These limitations suggest several directions for future data collection efforts. Future work should consider expanding coverage to include private institutions, implementing standardized reporting periods across all variables, developing more robust validation procedures for self-reported financial data, and creating linked datasets that allow for longitudinal analysis.

References

Campbell, Carol. 2020. "Educational Equity in Canada: The Case of Ontario's Strategies and Actions to Advance Excellence and Equity for Students." *School Leadership & Management* 41 (4–5): 409–28. https://doi.org/10.1080/13632434.2019.1709165.

Canada, Statistics. 2019. "Statistical Quality Assurance Framework." https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-eng.htm.

Education, Ontario Ministry of. 2024a. "Frequently Asked Questions (FAQs) about the BPR." https://www.app.edu.gov.on.ca/eng/bpr/faq.html.

———. 2024b. "School Information Finder Glossary." https://www.app.edu.gov.on.ca/eng/sift/glossary.asp.

Faulk, Dagnet. 2010. "Sources of Financial Support and Academic Performance in Economics Principles Courses." *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1711172.

Friedman, Jerome, Robert Tibshirani, and Trevor Hastie. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33 (1): 1–22. https://doi.org/10.18637/jss.v033.i01.

Government of Ontario. 2024a. "School Board Achievements and Progress." https://data.ontario.ca/dataset/school-board-achievements-and-progress/resource/9d81dc27-32ef-4864-94b5-f09950d00c72.

———. 2024b. "School Board Financial Reports: Estimates, Revised Estimates and Financial Statements." https://data.ontario.ca/dataset/school-board-financial-reports-estimates-revised-estimates-and-financial-statements/resource/56598892-c180-4f0d-bde5-63f971d6ef9a.

———. 2024c. "School Information and Student Demographics." https://data.ontario.ca/dataset/school-information-and-student-demographics/resource/e0e90bd5-d662-401a-a6d2-60d69ac89d14.

Institute, Learning Policy. 2024. "How Money Matters for Schools." https://learningpolicyinstitute.org/sites/default/files/product-files/How_Money_Matters_BRIEF.pdf.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

Mehreen, Faiza. 2023. "The Impact of Socio-Economic Status on Academic Achievement." *Journal of Social Sciences Review* 3 (2): 695–705. https://doi.org/10.54183/jssr.v3i2.308.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

Müller, Kirill, and Lorenz Walthert. 2023. *Styler: Non-Invasive Pretty Printing of r Code*. https://github.com/r-lib/styler.

Ontario, Government of. 2024. "Cell Name Reference Book Guide." https://data.ontario.ca/dataset/f9393864-17ae-43f4-a0e1-062fddc6c99c/resource/1c1c20bc-b0cb-452f-b025-96e3dfcfd9a0/download/cell_name_reference_book_guide.pdf.

OPSBA. 2024. "2024-25 Ontario Budget: OPSBA Overview." https://www.opsba.org/wp-content/uploads/2024/03/2024-25-Ontario-Budget-OPSBA-Overview.pdf.

Quality, Education, and Accountability Office (EQAO). 2021. "EQAO Data Quality

Framework." https://www.eqao.com/wp-content/uploads/2021/01/eqao-data-quality-framework.pdf.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://github.com/apache/arrow/.

Robinson, David, Alex Hayes, and Simon Couch. 2023. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://broom.tidymodels.org/.

Simon, Noah, Jerome Friedman, Robert Tibshirani, and Trevor Hastie. 2011. "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent." *Journal of Statistical Software* 39 (5): 1–13. https://doi.org/10.18637/jss.v039.i05.

Studies in Developmental Education (OASDI), Ontario Association for. 2024. "Graduation Requirements for k-12 Education in Ontario." https://www.oasdi.ca/k-12-education-in-ontario/graduation-requirements/#:~:text=Diploma%20(OSSD).-,The%20Ontario%20Secondary%20School%20Diploma%20(OSSD),40%20hours%20of%20community%20service.

Tay, J. Kenneth, Balasubramanian Narasimhan, and Trevor Hastie. 2023. "Elastic Net Regularization Paths for All Generalized Linear Models." *Journal of Statistical Software* 106 (1): 1–31. https://doi.org/10.18637/jss.v106.i01.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–88. http://www.jstor.org/stable/2346178.

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.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Willmott, Cort J. 1981. "On the Validation of Models." *Physical Geography* 2 (2): 184–94. https://doi.org/10.1080/02723646.1981.10642213.