Factors Influencing Exam Performance for California Schools

William Walker
May 9, 2018
SDS358

TEXAS
The University of Texas at Austin

Introduction

Objectives: Examine the additive effects of the variables of interest of student-to-teacher ratio and computers per 20 students over and above the nuisance variables of percent of English learners, expenditure per student, and total enrollment in the prediction of total exam score. Also determine the best single predictor of total exam score.

Hypotheses: The addition of the variables of interest will significantly improve the model. Student-to-teacher ratio will be the best single predictor.

Data Source: "The California Test Score Data Set":
https://vincentarelbundock.github.io/Rdatasets/datasets.html
The data is also part of the Ecdat R package (Caschool).

Methods

Sample: The data consisted of 420 school districts in the state of California. The variables included average exam scores for reading and writing (for elementary schools in the district), percentage of English learners, expenditure per student, total enrollment, student to teacher ratio, and the number of computers per student (all quantitative). The average math and reading scores were added together to calculate the total score. Computers per student, expenditure per student, and total enrollment were scaled to make the coefficients easier to interpret. Initial investigations showed 1 participant with a high Cook's Distance score. This subject was then removed, and analysis was performed.

Analysis Method: Sequential Regression.

Descriptives

Response Variable:

	Mean	SD
Total Score	1308.3	0.568

Explanatory Variables:

	Mean	SD
Expenditure per 100 Students	53.1	6.3
Percent English Learners	15.8	18.3
Total Enrollment (scaled)	26.3	39.1
Student-Teacher Ratio	19.6	1.9
Computes per 20 students	2.7	1.3

Results

Initial Model table:

Coefficient	Estimate	SE	t-value (df = 415)	P-value	2.5%	97.5%
Intercept	1275.90	12.03	106.0	< 0.001	1252.25	1299.55
Expenditure per 100	0.97	0.22	4.39	< 0.001	0.53	1.40
Percent English Learners	-1.41	0.08	-17.30	< 0.001	-1.57	-1.25
Total Enrollment (scaled)	0.13	0.04	3.14	0.002	0.05	0.21

Overall Model Fit: $F(3,415) = 112.8$, $p < 0.05$; $R^2 = 44.91\%$; Adj. $R^2 = 44.51\%$

Coefficient	St. Beta	Partial_Corr_Sq	Part_Corr_Sq
Expenditure per 100 Students	0.16	4.43%	2.55%
Percent English Learners	-0.68	41.91%	39.74%
Total Enrollment (scaled)	0.12	2.32%	1.31%

Full Model Table

Coefficient	Estimate	SE	t-value (df = 413)	P-value	2.5%	97.5%
Intercept	1309.49	30.54	42.88	< 0.001	1249.47	1369.51
Expenditure per 100	0.62	0.28	2.20	0.29	0.065	1.17
Percent English Learners	-1.36	0.08	16.54	< 0.001	-1.53	-1.20
Total Enrollment (scaled)	0.15	0.04	3.62	< 0.001	0.069	0.23
Student-Teacher Ratio	-1.20	0.98	-1.22	0.22	-3.13	0.73
Computers per 20 Students	2.66	1.15	2.30	0.02	0.39	

Overall Model Fit: $F(5,413) = 70.06$, $p < 0.05$; $R^2 = 45.89\%$; Adj. $R^2 = 45.24\%$

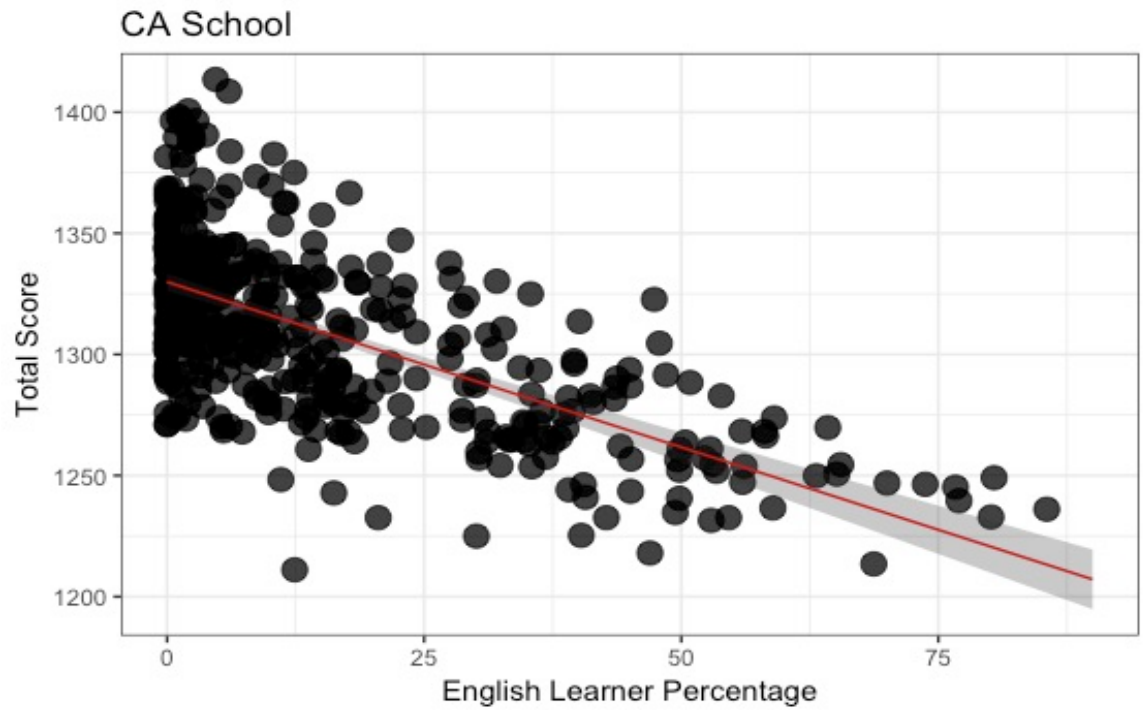
Coefficient	St. Beta	Partial_Corr_Sq	Part_Corr_Sq
Expenditure per 100 Students	0.10	1.15%	0.63%
Percent English Learners	-0.66	39.84%	35.84%
Total Enrollment (scaled)	0.15	3.08%	1.72%
Student-Teacher Ratio	-.060	0.36%	0.19%
Computers per 20 Students	0.091	1.27%	0.69%

```
Model 1: totalscore ~ elpct + expnstuscale + enrltotscale
Model 2: totalscore ~ expnstuscale + elpct + enrltotscale + str + computers20
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    415 333945
2    413 328010  2    5934.7 3.7362 0.02465 *
```

Full Model R^2 – Initial Model R^2 = 45.89% - 44.91% = 0.98%

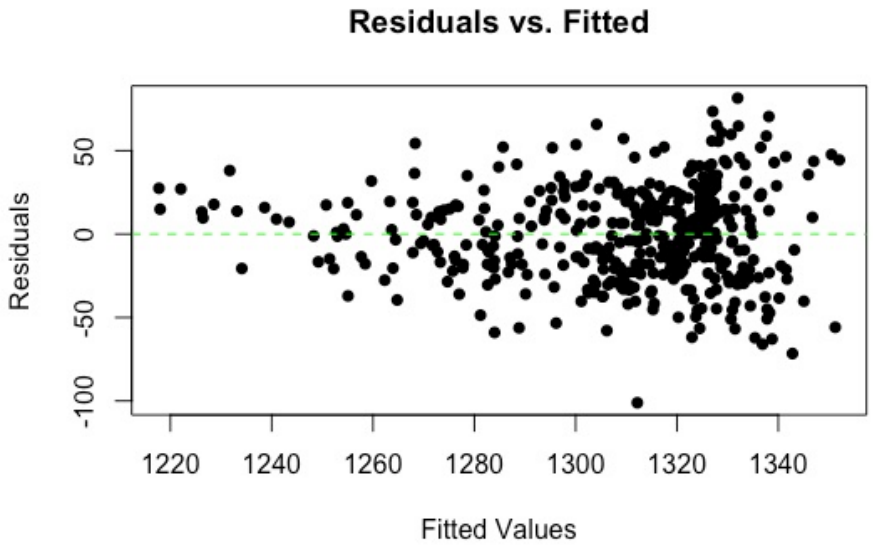
Correlation matrix

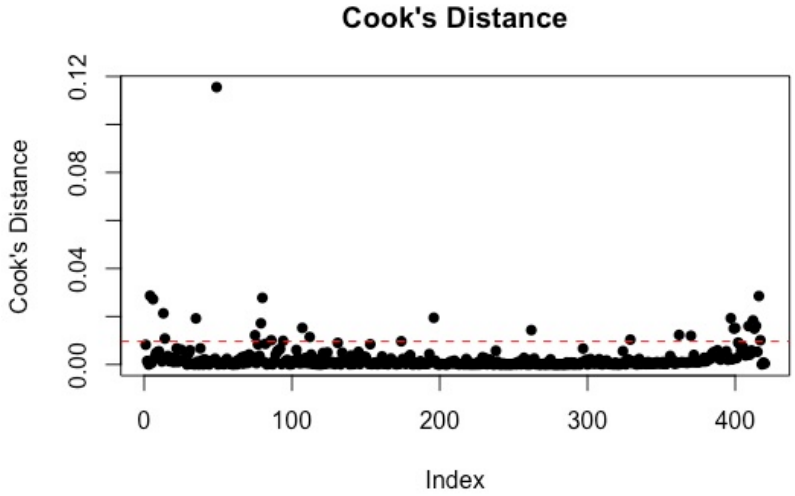
	totalscore	elpct	expnstuscale	enrltotscale	str	computers20
totalscore	1.00	-0.64	0.19	-0.15	-0.23	0.27
elpct	-0.64	1.00	-0.07	0.35	0.19	-0.25
expnstuscale	0.19	-0.07	1.00	-0.11	-0.62	0.29
enrltotscale	-0.15	0.35	-0.11	1.00	0.30	-0.21
str	-0.23	0.19	-0.62	0.30	1.00	-0.31
computers20	0.27	-0.25	0.29	-0.21	-0.31	1.00



Assumptions

Assumptions: Homoscedasticity was checked and confirmed by a Residual vs. Fitted plot. Cook's Distance was also examined for the removal of one outlier.





Discussion

Interpretation: The initial model was significant, $F(3,415) = 112.8$, $p < 0.05$, accounting for 44.91% of the variation in total exam score. The full model was also significant, $F(5,413) = 70.06$, $p < 0.05$, accounting for 45.89% of the variation in total exam score. The full model accounted for an additional 0.98% of the variance in total exam score. This change in R^2 was significant, $F(2, 413) = 3.74$, $p < 0.05$.

Limitations: This data set consists only of schools in California, so would be difficult to generalize the results across the United States. The data set is also from 1998-1999, so our conclusions may have changed over time.

Implications: Only an approximate additional 1% of the variance can be accounted for in the full model. Although this result was statistically significant, so this finding does not likely have any practical significance.