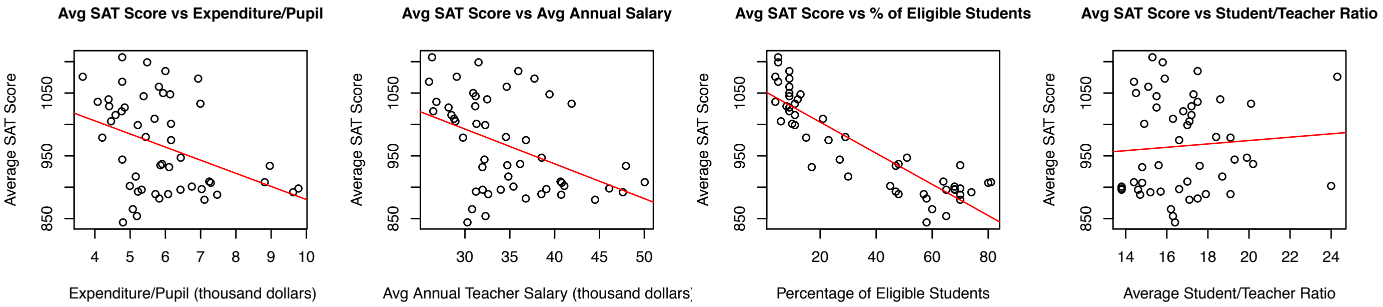
Linear Regression in Public Expenditure Analysis

Introduction

This report uses linear regression with the data about SAT scores assembled by Guber. It investigates the effect of several explantory variables on the SAT scores of students in public schools to ultimately consider the impact of an intervention in which limited resources are adjusted to improve student performance.

Analysis

Four scatterplots were created of Average Total SAT Score with Expenditure per Pupil, Average Annual Teacher Salary, Percentage of Eligible Students taking SAT and Average Student/Teacher Ratio respectively and a regression line was fit through each using a linear model obtained by the lm() function.



The table below shows the regression coefficient of explanatory variable, residual standard deviation and r^2 value for each of the 4 simple regression models

Table 1: Summary of 4 Simple Regression Models

Explanatory Variable	Explanatory Variable Coefficient	Residual SD	r^2 value
Expediture per Pupil	-20.89	69.9	0.1448
Average Annual Teacher Salary	-5.54	67.9	0.1935
Percentage of Eligible Students taking SAT	-2.48	34.9	0.7870
Average Student/Teacher Ratio	2.68	75.3	0.0066

The percentage of eligible students has the highest r^2 value with average SAT score, which means that it will be the best fitting simple linear model. The suitability of the linear model fit can also be seen graphically and from the fact that this model has a much lower (nearly half) residual standard error (RSS) than the other 3 models. The datapoints are roughly evenly distributed about the regression line and the residuals are relatively small. The co-efficient -2.48 indicates that a 1 percentage point increase in the percentage of eligible students taking the SAT results on average in a 2.48 unit decrease in average SAT score. This could indicate that as more students take the test, much more perform below the mean, which lowers the state's SAT Score.

Average student-teacher ratio has the lowest r^2 value with average SAT score, indicating the weakest linear relationship. Most of the datapoints are quite far from the regression line and this model has the highest RSS (75.3). The co-efficient 2.68 indicates that a 1 unit increase in average student/teacher ratio results on average in a 2.68 unit increase in average SAT score. However, this regression is not a good fit so the use of this model to summarize the association between these two variables does not make sense.The simple models with expenditure per pupil and average annual teacher salary as explanatory variables each have similar (low) r^2 values and high RSS values. Both these variables are negatively correlated with average SAT scores and while the datapoints in each are somewhat evenly distributed about the regression line, the residuals are quite large.

Now I will consider the effectiveness of two-variable models.

Two Variable Model: SAT Regressed on Expenditure Per Pupil and Fraction of Eligible Students

```
sat_expend_frac<-lm(data=s, formula=sat~expend+frac)
sat_expend_frac$coefficients
```

```
## (Intercept)      expend        frac
##  993.831659   12.286518   -2.850929
```

As shown above, the partial co-efficient of the fraction of eligible students taking the SAT (-2.851) is similar to its co-efficient (-2.480) in the simple linear model of SAT and eligible students. This partial co-efficent means that a 1 percentage point increase in the fraction of eligible students taking the SATs results on average in a 2.851 unit decrease in average SAT scores when expenditure per pupil is held constant. The partial co-efficient of expenditure (12.287) is positive in this model but the coefficient is negative (-20.89) in the simple linear model. The two-variable residual standard error is smaller and the R^2_{adj} value (0.812) is greater than that of each simple regression model.

I will also consider the two-variable models of SAT regressed on fraction of eligible students along with teacher salary and student teacher ratio respectively to check if a similar effect is found.

Two Variable Model: SAT Regressed on Teacher Salary and Fraction of Eligible Students

```
## (Intercept)      salary        frac
##  987.900464    2.180396   -2.778698
```

Two Variable Model: SAT Regressed on Student Teacher Ratio and Fraction of Eligible Students

```
## (Intercept)       ratio        frac
## 1118.508664   -3.726364   -2.547379
```

Once again, the partial co-efficients of the fraction of eligible students in both two-variable models are similar to its co-efficient (-2.480) in the simple linear model. There is a sign flip for both the other explanatory variables from the one to two variable model. The partial co-efficient of teacher salary (2.180) is positive in the two variable model while the co-efficient is negative (-5.54) in the simple linear model. The partial co-efficient of student teacher ratio (-3.726) is negative in the two variable model while the co-efficient is positive (2.68) in the simple linear model. The residual standard errors are smaller and the R^2_{adj} values are greater for both two-variable models than for any of the simple regression models.

Since all 3 variables flipped signs, I decided to examine the partial correlations between SAT score and the following 3 variables, controlling for the fraction of eligible students taking the exam, as indicated below.

Table 2: Partial Correlations between Sat Score and Explanatory Variables (Controlling for Fraction of Eligible Students)

Expenditure per Pupil	Teacher Salary	Student Teacher Ratio
0.391	0.295	-0.239

It can be seen that all 3 partial correlations are relatively close in magnitude. They reveal that there is a decent part of the model not explained by eligible fraction that is explained by the other variables. However, while the partial correlations of SAT score with Expenditure per pupil and Teacher salary (controlling for fraction of eligible students) are positive, the one with student teacher ratio is negative. This makes sense because as we increase student teacher ratio, we can expect the SAT scores to go down as a result of lesser attention being paid to the performance of each student.

I will also examine the three-variable model of SAT regressed on teacher salary, student teacher ratio and eligible student fraction

Three Variable Model: SAT Regressed on Teacher Salary, Student-Teacher Ratio & Eligible Fraction

```
sat_salary_ratio_frac<-lm(data=s, formula=sat~salary+ratio+frac)
coefficients(sat_salary_ratio_frac)
```

```
## (Intercept)      salary       ratio        frac
## 1057.898162    2.552470   -4.639428   -2.913350
```

The R^2_{adj} value has gone up in the three variable model. While the partial coefficient of the fraction of eligible students remains similar from the two to three variable models, the partial coefficient of teacher salary increases from 2.18 to 2.55 (absolute change of 0.37) and the partial coefficient of student teacher ratio decreases from -3.73 to -4.64 (absolute change of 0.91). The standard error of the co-efficient of teacher salary is 1.005 and that of student teacher ratio is 2.122, so the changes in the co-efficients aren't real because the absolute value of the changes are much smaller than the standard errors of the regression estimates. This indicates that the variables teacher salary and student teacher ratio are very weakly correlated.

To assess the tradeoffs between student teacher ratio and teacher salary I will use standardized co-efficients.

Table 3: Standardized Coefficients for the Three Variable Model

Fraction of Eligible Students	Teacher Salary	Student Teacher Ratio
-1.04	0.203	-0.141

These standardized coefficients show a "tradeoff" in the sense that a 1 standard unit increase in teacher salary results in a 0.203 unit increase in average SAT scores holding other variables constant while a 1 standard unit increase in student teacher ratio results in a 0.141 unit decrease in average SAT scores holding other variables constant.

Discussion

Among the simple regression models, the fraction of eligible students explains the greatest variability in average SAT score because these two variables have the strongest linear correlation. The flip in sign of expenditure per pupil, teacher salary and student teacher ratio from the one to two variable model indicates that each of these variables is correlated with the fraction of eligible students. This is why including each of these variables with the fraction of eligible students changes the effect of the variables on the total output. The two-variable residual standard error is smaller and the R^2_{adj} value is greater for each two-variable model as compared to each simple model, indicating that by using two variables we have actually increased the ratio of Reg_{SS} to T_{SS}. However, the increase in R^2_{adj} is not by much more than the simple model with fraction of eligible students.

After examining the 3 variable model, we see that teacher salary has a larger standardized coefficient (in terms of absolute value) than student teacher ratio or fraction of eligible students when we consider the combined effect of these variables on Average SAT Score

Conclusion

If we were to conduct an intervention with limited resources, there is a trade off between decreasing class size and increasing teacher salaries. The models examined in this report reveal that when the effect of multiple explanatory variables on SAT scores is considered, teacher salary has the largest effect on SAT scores, which is why it is important to increase teacher salaries to improve student performance.