Variable Names and Definitions

My research focuses on student academic success and how socioeconomic factors, charter schools and development affect them. This regression uses data from the California Assessment of Student Performance and Progress. This organization provides large quantities of raw data on student performance, writing proficiency, math proficiency and other details surrounding student progress. The data is organized on a per school structure, where percentages and dummy variables are all referencing a certain school's performance, type, or student body.

Variable Names	**Definitions**
perc_std_met	Dependent Variable, what percentage of students meet or exceed standards averaged across math and reading competency per school.
perc_free_lunch	Percentage of students who qualify for free lunch at each school. Serves as a poverty estimate for each school as qualifying for free lunch implies students do not have the proper resources to afford lunch.
school_type	Dummy variable that shows if a school is elementary, middle, or high school.
charter	Dummy variable that shows if a school is a charter school or not.

The definitions for what defines standards met or exceeded are explained on the CAASPP website. Standards met or exceeded is defined by a specific score range that gradually increases alongside the grade level for both writing and math assessments. This research focuses on both math and reading scores, which both follow the same met/not met structure.

Summary Statistics

Variable	Obs	Mean	Std. dev.	Min	Max
per_std_met	7,961	.**4101875**	.1816982	.**0143**	.**9919375**
perc_free_lunch	7,961	.**558101**	.2530454	.**0046512**	**1**
school_type	7,961	1.461123	.7438371	1	3
charter	7,961	.1094084	.3121705	0	1
middle	7,961	.1563874	.363245	0	1
high	7,961	.1523678	.3593996	0	1

This is the summary statistics for all variables in my regression. Notice some key details in the table. For one, the average school has less than half of their students meeting or exceeding the standard for math and reading(41.01 percent). Also notice that the average school has about half of their students using free lunch(55.81 percent). Keep in mind that this regression is only looking at public school data so wealthier students that might be going to private schools are not included in the data. The maximum percentage of students meeting or exceeding the standard is around 99 percent while the lowest is almost zero, showing wide discrepancies between school test rates. This same discrepancy can be seen with the percentage of students on free lunch.

Regression Results and Tests

Adj R^2=0.5551	N=7,961

Variable	Coefficient	Std. err.	P>|t|
perc_free_lunch	-0.534496	0.00537	0.000
school_type	-0.013609	0.00183	0.000
charter	0.0011615	0.00437	0.791
_cons	0.7282479	0.00428	0.000

These results tell us quite a bit about how test scores are affected by other variables. Firstly, it seems as though a 1 percent increase in percent of students eligible for free lunch tends to decrease test scores by about half a percentage point which represents the massive impact poverty has on student performance. School type also paints a worrying picture. Given that the base for school_type is elementary school, it seems that as a child moves from elementary to middle to high school each increment leads to worse test score outcomes. Charter schools did not have an impact on student performance as the p stat is not significant at the 5 percent level. The adjusted R^2 is also quite high considering how many factors affect student achievement.

Misspecification Error

First test is the **Ramsey test,** which tells me if my regression is misspecified or not.

F(3, 7954) = 155.61	Prob > F = 0.0000

I have a very low Prof>F, which means I have no omitted variables.

The second test is the **linktest**, which also tells me if my regression is misspecified or not.

Variable	Coefficient	Std. err.	P>|t|
_hat	-0.4326696	0.0677373	0.000
_hatsq	1.65257	0.0773195	0.000
_cons	0.2793229	0.0137315	0.000

Unfortunately, our hatsq value is statistically significant at the 5 percent level, meaning that our regression does suffer from a misspecification problem. This is not surprising, as I am sure there are many hundreds of variables that affect a student's academic performance. I maintain the belief that our model analyzing test scores and poverty is a reasonable one.

Heteroskedasticity Errors

In order to test for heteroskedasticity in our model, we need to run both the Breusch-Pagan test as well as the White's test. Our Breusch-Pagan test gives us these values:

chi2(1)=41.53	Prob>chi2=0.000

Unfortunately our Prob>chi2 value is significant at the 5 percent level, which means our model suffers from heteroskedasticity. Our White's model shows similar results:

chi2(1)=140.77	Prob>chi2=0.000

Addressing Heteroskedasticity Errors(robust regression)

In order to solve these problems, I will use a robust regression model which will be less efficient but still unbiased and consistent. This is because the robust regression does not operate under the assumption of homoscedasticity.

R^2=0.5552	N=7,961

Variable	Coefficient	Std. err.	P>|t|
perc_free_lunch	-0.534496	0.005583	0.000
school_type	-0.013609	0.0019128	0.000
charter	0.0011615	0.0048915	0.812
_cons	0.7282479	.0046001	0.000

This regression paints the same picture as our original one. Poverty greatly affects school performance as does school type. Charter schools are still not significant to a child's academic performance.

Sources of Data:

"English Language Arts/Literacy and Mathematics Smarter Balanced Summative Assessments." 2023–24 Research Files for Smarter Balanced ELA and Mathematics - CAASPP Reporting (CA Dept of Education). Accessed November 21, 2024. https://caaspp-elpac.ets.org/caaspp/ResearchFileListSB?ps=true&lstTestYear=2024&lstTestType=B&lstCounty=00&lstDistrict=00000.