Econometrics A

Stata Assignment

Name: **Max Dunne**

Student No: 21365739

Date: 14th of November 2023

1 Part 1

In this project I will use the statistics software Stata to analyse the different effects of variables on the test scores of 500 schools from the state of California.
If used correctly, this analysis could serve to increase the average overall test scores and also to show departments of education where increased funding is necessary.

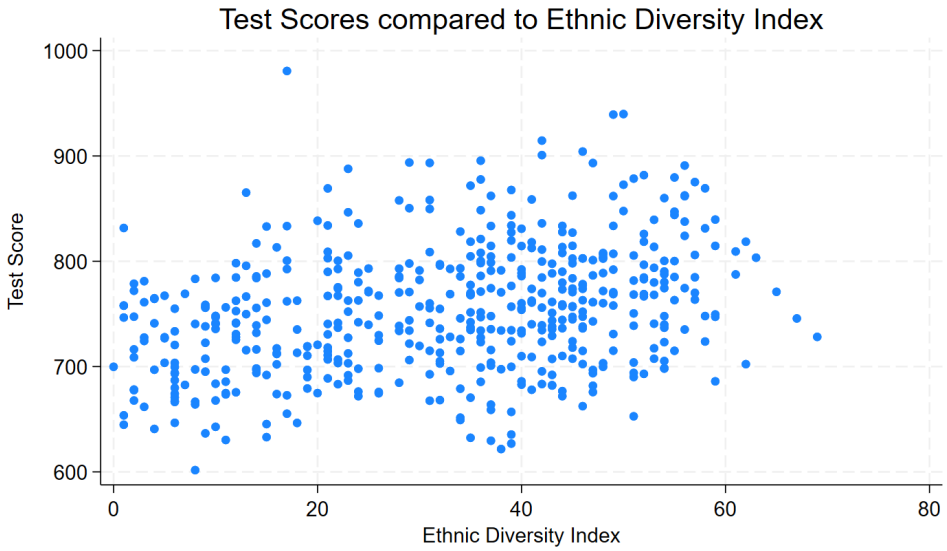
1.1 Table of Summary Statistics

Variable	Obs	Mean	Std. dev.	Min	Max
countyname	0				
districtname	0				
schoolname	0				
zipcode	500	93289.07	1543.599	90250	96117
testscore	500	753.8542	60.3035	601.7	980.7
str_s	500	24.02282	3.559646	14.90358	32.81482
charter_s	500	.056	.230152	0	1
frpm_frac_s	500	.6018114	.2758199	.018	1
enrollment_s	500	587.434	193.8105	128	1290
ell_frac_s	500	.317998	.2105888	.004	.916
edi_s	500	32.466	16.49221	0	69
te_fte_s	500	24.643	8.038632	6	60
te_avgyr_s	500	12.656	3.49544	2	26
ada_enroll~d	500	.9285269	.0753718	.4246284	1.005714
te_salar~w_d	500	41256.85	3493.816	32398	51302
te_salar~g_d	500	69182.28	5928.886	45882	88533
te_days_d	500	178.76	1.999599	175	181
te_serdays_d	500	182.662	2.733991	175	188
age_frac_5~z	500	.1970646	.0346791	.0478864	.2947368
pop_1_olde~z	500	43957.26	20148.37	182	104363
ed_frac_hs_z	500	.2272743	.0532198	.0531634	.364233
ed_frac_sc_z	500	.3178418	.067052	.0585009	.5786695
ed_frac_ba_z	500	.157373	.0800098	0	.3805932
ed_frac_gr~z	500	.0815993	.0638368	0	.4154756
med_income_z	500	28061.45	9402.315	13188	72149

Using Stata I was able to import data and summarize all of the variables.
We are given the sample mean, standard deviation and the maximum and minimum observation of each variable.
For example, the Ethnic Diversity Ratio (edi_s) has a standard deviation of 16.49221 which is just more than half of its mean (32.466), with ranges of 0 to 69, showing that some schools are much more ethnically diverse that others. This is a good example of showing where schools differ from

each other.
On the other hand, schools are extremely similar among other variables, such as Teaching Days (te_days_d), with a standard deviation (1.999599) of less than 2 percent of the mean (178.76), with minimums and maximums of 175 and 181. From this table we are able to note that the number of teaching days in each school only varies slightly.

1.2 Comparing Test Scores and the Ethnically Diverse Index



The relationship between test scores and ethnic diversity appears to be a postive linear relationship, hence an increase in ethnic diversity should cause an increase in test scores. This could be due to students coming from schools with a high EDI. Schools with a higher EDI have students of different cultures which can offer important social-emotional benefits, which could be linked to the increase in test scores.

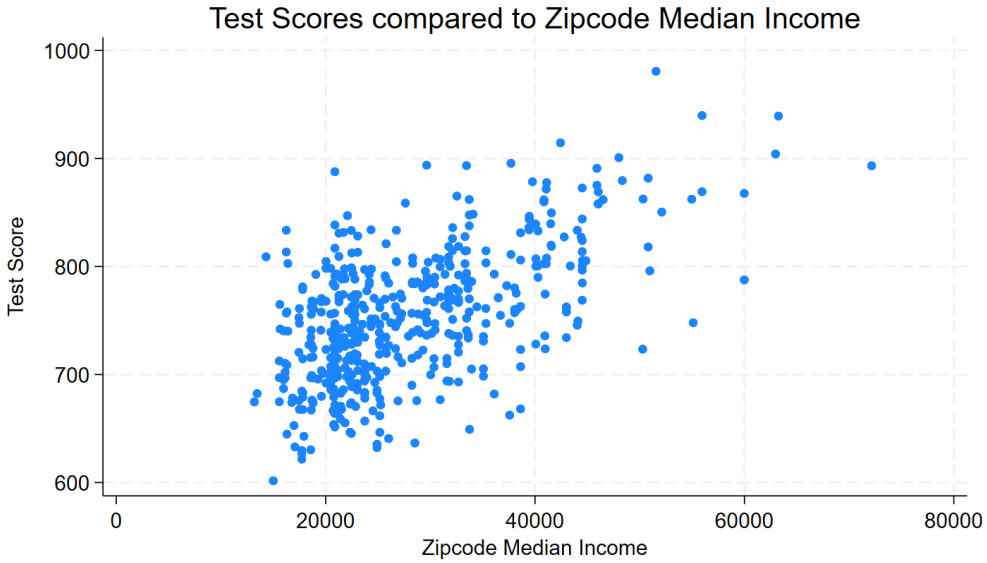
Variable	Obs	Mean	Std. dev.	Min	Max
testscore	277	767.1032	59.86052	621.7	939.8
edi_s	277	45.19856	7.653284	33	69

Variable	Obs	Mean	Std. dev.	Min	Max
testscore	223	737.3969	56.81533	601.7	980.7
edi_s	223	16.65022	9.21972	0	32

Above is the two tables, the first compares test score to EDI scores above the mean EDI and the second compares test score to EDI scores below the mean EDI.
The average test score of a student who attends a relatively school ethnically diverse is 30 points higher than the average test score of a student who attends a school that is relatively not ethnically diverse.

As this relationship is strong, I will use it later in Part 3.

1.3 Comparing Test Scores and Income



We can clearly see a positive relationship between test scores and the median income of a zip code, hence as the median income of a zipcode increases the test scores for schools in that zip code should increase. The reltionship appears to be logarithmic.

Household income has an obvious effect on education;

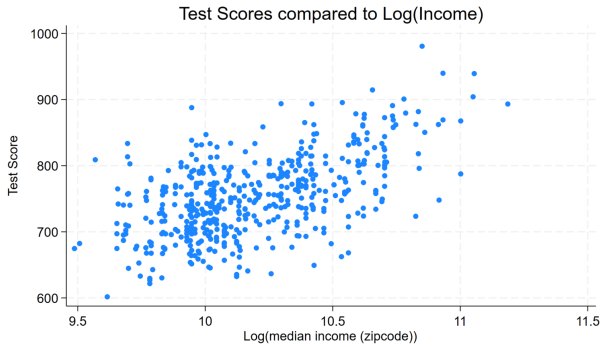
- It allows children to go to fee paying school. Fee paying schools are known to have the best teachers and educational facilities.

- It allows families to supply their children with extra help if necessary, i.e. extra tutoring or subscriptions to websites such as Chegg.

- High household income can increase a child's well-being which can only serve to improve their learning experience.

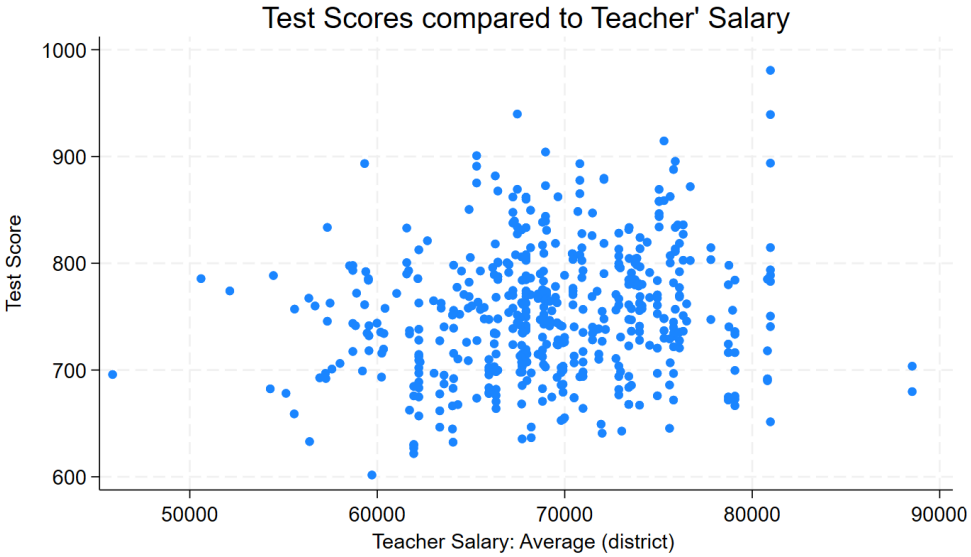
1.3.1 Is there a logarithmic relationship here?

In stata, I generated a new variable, this variable is the log of median income (zipcode). The graph comparing this new variable with test scores is shown below.



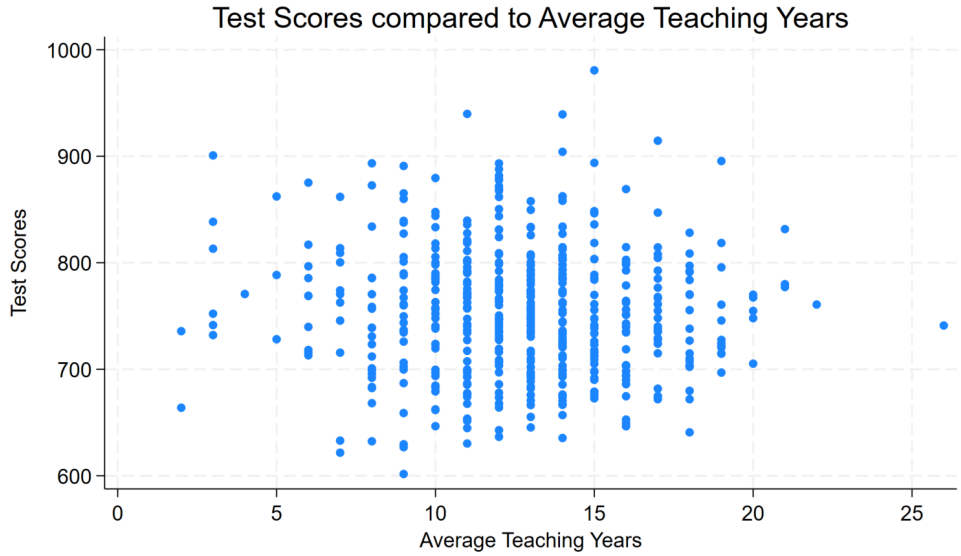
This graph clearly shows the linear relationship between test scores and the log of income, which in turn shows the logarithmic relationship between test scores and income. I will use this new variable in Part 3.

1.4 Comparing Test Scores and Teacher Salary



As we can see, there is clearly a positive correlation between test scores and teacher's salaries. As in all professions, the better you are in your field, the more income you get. This is the same in education, better teachers cost more money, this explains why test scores are correlated with a teacher's salary. However, this is not causal as if you were to increase the salary of a teacher, this would not directly increase the test scores of the students.

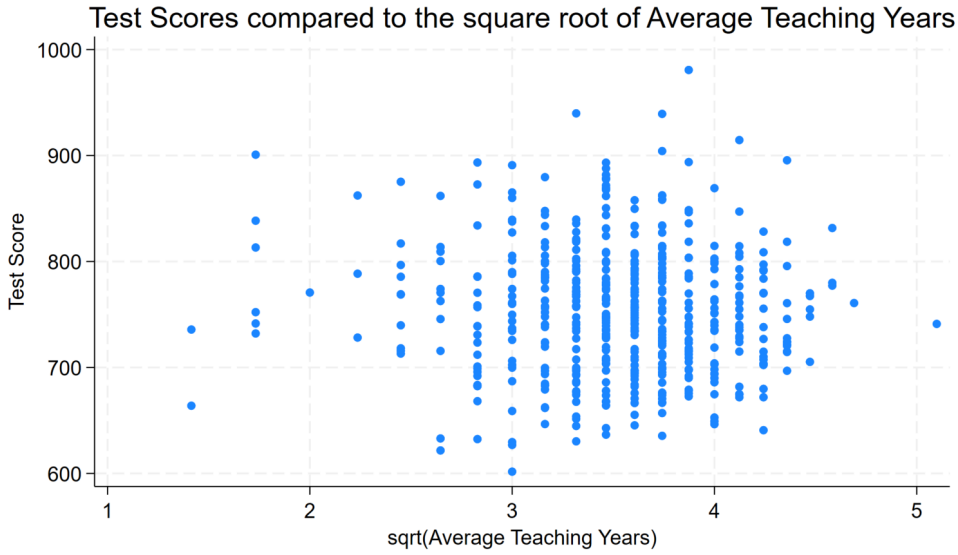
1.5 Comparing Test Scores and Average Teaching Years



The relationship is n-shaped which would hint at the relationship being quadratic. It appears that teachers are most effective in the middle of their careers. Perhaps, in a teacher's early career they haven't fully developed some of the skills necessary, such as controlling the classroom and later in a teacher's career, they may grow tired or lose some of their skills.

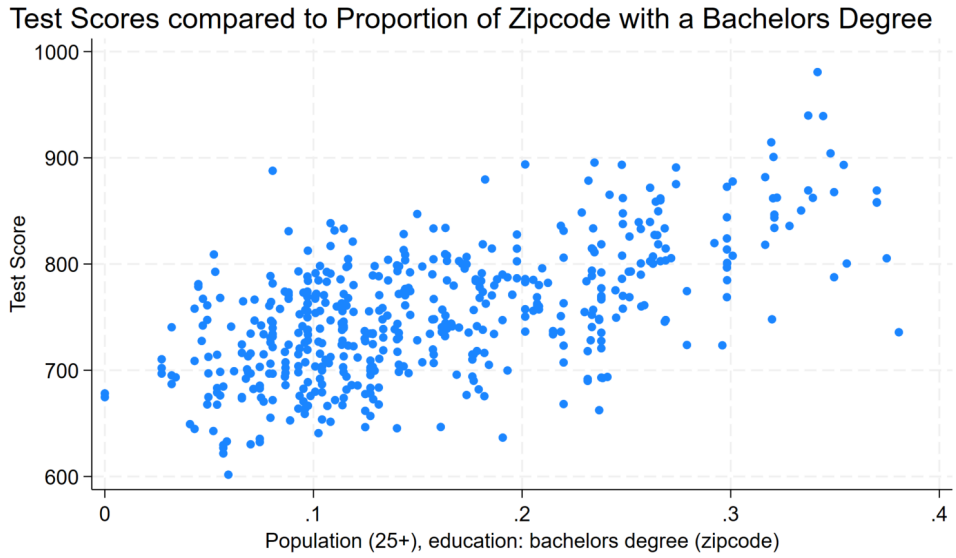
1.5.1 Making this Relationship Linear

As this relationship appears to be quadratic, I will create a new variable that is the square root of average teaching years. The relationship between test scores and this new variable is below.



This shows a linear relationship and I will use it in Part 3.

1.6 Comparing Test Scores and Proportion of Zip Code with a Bachelor's Degree



We can clearly see a positive linear relationship between test scores and the proportion of zip code with a bachelor's degree. If you are a student from a zip code where there is a high proportion of the population (25+), your parents and older siblings are more likely to have a bachelor's degree. A child's intelligence is highly correlated with the intelligence of their family, hence a student from this area is more likely to be intelligent for many reasons such as genetics, parents encouraging the child to become as educated as they are and being able to ask their family members for help with their homework etc. I will use this variable in Part 3.

2 Part 2

2.1 Regression of Test scores on the student teacher ratio, the share of students receiving free or reduced-price school meals, the share of English language learners, and area (zip code) median income.

Source	SS	df	MS	Number of obs	=	500
				F(4, 495)	=	131.31
Model	934200.447	4	233550.112	Prob > F	=	0.0000
Residual	880419.358	495	1778.62497	R-squared	=	0.5148
				Adj R-squared	=	0.5109
Total	1814619.81	499	3636.51264	Root MSE	=	42.174

testscore	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
str_s	-.6792563	.5341249	-1.27	0.204	-1.728688	.3701753
frpm_frac_s	-136.8878	12.00709	-11.40	0.000	-160.479	-113.2967
ell_frac_s	11.90755	11.64503	1.02	0.307	-10.97223	34.78733
med_income_z	.000937	.0002987	3.14	0.002	.0003501	.001524
_cons	822.4715	17.99854	45.70	0.000	787.1085	857.8344

2.2 Economical and Statistical Significance

We can interpret the model as:

- On average, when the student-teacher rises by 1, we expect test scores to fall by .6792563 points, holding all else constant.
- On average, if x percent of a school have free or reduced priced meals, the expected fall of test scores is $136.8878(\frac{x}{100})$ points, holding all else constant.
- On average, if x percent of a school are English language learners, the expected increase of test scores is $11.90755(\frac{x}{100})$ points, holding all else constant.
- On average, a 1 dollar increase in zip code median income will cause an increase of test scores by .000937 points, holding all else constant.
- Theoretically, if the student teacher ratio was 0 (∞ teachers?), no students received free or reduced price meals, no students are English language learners and zip code median income is 0 dollars, the expected test score would be 822.4715.

At the 95% significance level, we can say that an increase in zip code median income will cause an increase in test scores and an increase the proportion of students receiving free or reduced price meals will cause a decrease in test scores. We cannot say with 95% certainty that an increase in the student teacher ratio will cause a decrease in test scores as both positive and negative numbers exists in the 95% confidence interval for the coefficient; 95% Confidence Interval for $\beta_1 = (-1.728688, .3701753)$. Similarly, we cannot say, at 95% significance level, that an increase in the proportion of English language learners will cause an increase in test scores as both positive and negative numbers exists in the 95% confidence interval for the coefficient; 95% Confidence Interval for $\beta_3 = (-10.97223, 34.78733)$.

According to this model, in order to increase test scores, we should;

- decrease the student teacher ratio, hence, increase the number of teachers in schools,
- decrease the proportion of students receiving free or reduced price meals, hence reduced funding for subsidised meals.
- increase the proportion of English language learning students, we could potentially allow more foreign immigrants into the country so their children enroll in our schools.
- increase the zip code median income, we could increase the minimum wage or implement an expansionary fiscal policy.

Although the model would say otherwise, this is clearly not the entire case, for example, free or reduced price meals don't cause test scores to fall, they are correlated nonetheless. For this reason I believe there is omitted variable bias in this model.

2.3 Testing for Omitted Variable Bias

I will now run a Ramsey test on Stata to test for omitted variable bias.

- H_0: Model has no omitted variables
- H_1: Model has omitted variables

```
F(3, 492) =   5.67
 Prob > F = 0.0008
```

Hence as the P-value of .0008 is less than .05, we can reject the null hypothesis and say with 95% certainty that there is omitted variable bias in this model.

2.4 How useful is this model?

When examining how useful a model is, we must check the adjusted R-squared and Root MSE for this model.

2.4.1 Adjusted R^2

The adjusted R-squared of a model measures how well the set of predictor variables explain the variation in the response variable. In this model, the adjusted R-squared is .5109, which means that 51.09% of the variance in test scores is explained by the student teacher ratio, the share of students receiving free or reduced-price school meals, the share of English language learners, and area (zip code) median income.

2.4.2 Root MSE

The Root MSE measures the difference between the values observed and values predicted by the model;

$$\sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

A model with a root MSE of value zero would be a perfect model where all the predicted values are equal to the observed value, the larger the value of root MSE, the less accurate our model is. In our model, the root MSE is 42.174 points. This is a relatively large root MSE which means our model isn't very accurate.

2.5 Is the estimate successful in uncovering an unbiased estimate for the effect of the student teacher ratio on test scores?

I will run an F-test on the student-teacher;

- H_O: The coefficient of the student teacher ratio (str_s) is equal to zero
- H_1: The coefficient of the student teacher ratio (str_s) is not equal to zero

```
F(  1,   495) =    1.62
     Prob > F =    0.2041
```

We attain a P-value of .2041, therefore we fail to reject the null hypothesis, hence the estimate is unsuccessful in uncovering an unbiased estimate for the effect of the student teacher ratio.

3 Part 3

For this section I will extend the model from part 2 to include more explanatory variables in attempt to increase the model's accuracy and lower the unexplained variance.

3.1 Regression 1

A regression of test scores was ran on the student teacher ratio, the share of students receiving free or reduced-price school meals, the share of English language learners, average teacher's salary (zipcode) and log(area (zip code) median income) and below is the data we attained.

Source	SS	df	MS	Number of obs	=	500
				F(5, 494)	=	110.84
Model	959417.671	5	191883.534	Prob > F	=	0.0000
Residual	855202.135	494	1731.17841	R-squared	=	0.5287
				Adj R-squared	=	0.5239
Total	1814619.81	499	3636.51264	Root MSE	=	41.607

testscore	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
str_s	-1.350358	.5512617	-2.45	0.015	-2.433465	-.2672514
frpm_frac_s	-138.0257	11.74353	-11.75	0.000	-161.0991	-114.9522
ell_frac_s	2.168379	11.76757	0.18	0.854	-20.95229	25.28905
te_salary_avg_d	.0014408	.0003377	4.27	0.000	.0007773	.0021043
log_med_income_z	19.40818	8.93922	2.17	0.030	1.844602	36.97176
_cons	571.1789	95.7894	5.96	0.000	382.974	759.3837

3.1.1 Test for Omitted Variable Bias

Firstly, I will run a Ramsey test to check for omitted varaible bias:

- H_0: Model has no omitted variables
- H_1: Model has omitted variables

```
F(3, 491) =   7.61
 Prob > F = 0.0001
```

We obtain a P-value of .0001, hence we must reject H_0, therefore omitted variable bias is at play in our model.

3.1.2 Adjusted R^2

This model has an adjusted R^2 of .5239, which means 52.39% of the variance of the test scores is explained by the predictor variables, this a slight increase from the model in Part 2.

3.1.3 Root MSE

The root MSE for this model was 41.607, an improvement on the previous model's 42.174, therefore our new model is slightly more accurate.

As our adjusted R^2 and root MSE have both improved, the addition of these variables have increased the model's explanatory power.

3.2 Regression 2

A regression of test scores was ran on the student teacher ratio, the share of students receiving free or reduced-price school meals, the share of English language learners, average teacher's salary (zipcode) and log(area (zip code) median income)and below is the data we attained.

Source	SS	df	MS	Number of obs	=	500
				F(6, 493)	=	93.66
Model	966618.358	6	161103.06	Prob > F	=	0.0000
Residual	848001.448	493	1720.08407	R-squared	=	0.5327
				Adj R-squared	=	0.5270
Total	1814619.81	499	3636.51264	Root MSE	=	41.474

testscore	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
str_s	-1.265612	.5510513	-2.30	0.022	-2.34831	-.1829128
frpm_frac_s	-138.8637	11.71301	-11.86	0.000	-161.8772	-115.8501
ell_frac_s	-7.507702	12.64728	-0.59	0.553	-32.35692	17.34151
te_salary_avg_d	.0014812	.0003372	4.39	0.000	.0008186	.0021437
log_med_income_z	23.973	9.185596	2.61	0.009	5.925255	42.02075
edi_s	-.2993379	.1463018	-2.05	0.041	-.5867898	-.011886
_cons	533.1253	97.27651	5.48	0.000	341.9976	724.253

3.2.1 Test for Omitted Variable Bias

Firstly, I will run a Ramsey test to check for omitted varaible bias:

- H_0: Model has no omitted variables
- H_1: Model has omitted variables

```
F(3, 490) =    5.39
 Prob > F = 0.0012
```

We obtain a P-value of .0012, hence we must reject H_0, therefore omitted variable bias is at play in our model.

3.2.2 Adjusted R^2

This model has an adjusted R^2 of .5270, which means 52.7% of the variance of the test scores is explained by the predictor variables, this a slight increase from the previous model.

3.2.3 Root MSE

The root MSE for this model was 41.474, this a slight decrease from the previous model, hence our model is now more accurate.

Similar to before, the addition of a new variable has increased our model's explanatory power.

3.3 Regression 3

A regression of test scores was ran on the student teacher ratio, the share of students receiving free or reduced-price school meals, the share of English language learners, average teacher's salary (zipcode), sqrt(average teaching years) and log(area (zip code) median income)and below is the data we attained.

Source	SS	df	MS	Number of obs	=	500
				F(7, 492)	=	80.22
Model	967192.364	7	138170.338	Prob > F	=	0.0000
Residual	847427.442	492	1722.4135	R-squared	=	0.5330
				Adj R-squared	=	0.5264
Total	1814619.81	499	3636.51264	Root MSE	=	41.502

testscore	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
str_s	-1.222937	.5563572	-2.20	0.028	-2.316066	-.1298078
frpm_frac_s	-138.1086	11.79369	-11.71	0.000	-161.2808	-114.9364
ell_frac_s	-8.886426	12.87922	-0.69	0.491	-34.19148	16.41863
te_salary_avg_d	.0015566	.0003618	4.30	0.000	.0008457	.0022675
log_med_income_z	23.65432	9.208376	2.57	0.010	5.561726	41.74691
edi_s	-.3015538	.1464511	-2.06	0.040	-.5893005	-.013807
te_avgyr_s_quad	.4482957	.7765605	0.58	0.564	-1.077488	1.97408
_cons	180.1333	619.1703	0.29	0.771	-1036.411	1396.677

3.3.1 Test for Omitted Variable Bias

Firstly, I will run a Ramsey test to check for omitted varaible bias:

- H_0: Model has no omitted variables
- H_1: Model has omitted variables

```
F(3, 489) =   4.94
 Prob > F = 0.0022
```

We obtain a P-value of .0022, hence we must reject H_0, therefore omitted variable bias is at play in our model.

3.3.2 Adjusted R^2

This model has an adjusted R^2 of .5264, which means 52.64% of the variance of the test scores is explained by the predictor variables, this a slight decrease from the previous model,.

3.3.3 Root MSE

The root MSE for this model was 41.502, this a slight increase from the previous model, hence our model is now less accurate.

The addition of sqrt(average years of teaching) has caused our model to lose some of its explanatory power, therefore I will not include it in the next model.

3.4 Regression 4

A regression of test scores was ran on the student teacher ratio, the share of students receiving free or reduced-price school meals, the share of English language learners, average teacher's salary (zipcode), sqrt(average teaching years),proportion of zip code (25+) with a bachelors degree and log(area (zip code) median income)and below is the data we attained.

Source	SS	df	MS	Number of obs	=	500
				F(7, 492)	=	82.40
Model	979282.258	7	139897.465	Prob > F	=	0.0000
Residual	835337.548	492	1697.84054	R-squared	=	0.5397
				Adj R-squared	=	0.5331
Total	1814619.81	499	3636.51264	Root MSE	=	41.205

testscore	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
str_s	-1.052281	.553021	-1.90	0.058	-2.138855	.0342933
frpm_frac_s	-126.8111	12.44572	-10.19	0.000	-151.2644	-102.3578
ell_frac_s	-15.23398	12.87977	-1.18	0.237	-40.54012	10.07216
te_salary_avg_d	.0012854	.0003426	3.75	0.000	.0006123	.0019585
log_med_income_z	-1.687365	13.09819	-0.13	0.898	-27.42266	24.04793
edi_s	-.3404523	.1461302	-2.33	0.020	-.6275686	-.053336
ed_frac_ba_z	145.043	53.10817	2.73	0.007	40.69616	249.3897
_cons	776.7909	131.5311	5.91	0.000	518.3589	1035.223

3.4.1 Test for Omitted Variable Bias

Firstly, I will run a Ramsey test to check for omitted variable bias:

- H_0: Model has no omitted variables
- H_1: Model has omitted variables

```
F(3, 489) =   3.05
 Prob > F = 0.0282
```

We obtain a P-value of .0282, hence we must reject H_0, therefore omitted variable bias is at play in our model.

3.4.2 Adjusted R^2

This model has an adjusted R^2 of .5331, which means 53.31% of the variance of the test scores is explained by the predictor variables, this a slight increase from the previous model,.

3.4.3 Root MSE

The root MSE for this model was 41.205, this a slight decrease from the previous model, hence our model is now more accurate.

The removal of the variable sqrt(average teaching years) and the addition of the variable "Population (25+), education: bachelors degree (zip code)" has increased our adjusted R^2 and decreased our Root MSE and has therefore increased our model's explanatory power. This is the most accurate model I obtained, but, it still has omitted variable bias, hence the zero conditional mean is assumption is violated

3.5 Does the relationship between class size and test scores differs between schools with above and below median shares of students receiving free or reduced-price school meals?

First of all, I found the median of the proportion of a students that receive free or reduced priced meals to be .6511. I then ran two regressions of our model from Part 3, one with the proportion of a students that receive free or reduced priced meals above the median and one below, they are shown below respectively.

Source	SS	df	MS	Number of obs	=	250
				F(7, 242)	=	5.37
Model	75359.8986	7	10765.6998	Prob > F	=	0.0000
Residual	485119.832	242	2004.62741	R-squared	=	0.1345
				Adj R-squared	=	0.1094
Total	560479.731	249	2250.92261	Root MSE	=	44.773

testscore	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
str_s	-1.501965	.817775	-1.84	0.067	-3.112831	.1089006
frpm_frac_s	-153.7948	39.95832	-3.85	0.000	-232.5053	-75.08424
ell_frac_s	-24.62646	20.86375	-1.18	0.239	-65.72419	16.47128
te_salary_avg_d	.0010133	.0005499	1.84	0.067	-.0000698	.0020964
log_med_income_z	-13.34007	24.82319	-0.54	0.591	-62.23716	35.55702
edi_s	-.4932203	.2558695	-1.93	0.055	-.9972359	.0107954
ed_frac_ba_z	138.6643	102.9342	1.35	0.179	-64.09702	341.4256
_cons	956.3244	258.3213	3.70	0.000	447.4791	1465.17

Source	SS	df	MS	Number of obs	=	250
				F(7, 242)	=	37.35
Model	358040.918	7	51148.7026	Prob > F	=	0.0000
Residual	331383.928	242	1369.35507	R-squared	=	0.5193
				Adj R-squared	=	0.5054
Total	689424.846	249	2768.77448	Root MSE	=	37.005

testscore	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
str_s	-.7052592	.7432729	-0.95	0.344	-2.169369	.7588509
frpm_frac_s	-153.7241	18.60147	-8.26	0.000	-190.3656	-117.0827
ell_frac_s	-20.18885	17.67383	-1.14	0.254	-55.00303	14.62534
te_salary_avg_d	.0014091	.0004358	3.23	0.001	.0005506	.0022675
log_med_income_z	-1.97388	14.84708	-0.13	0.894	-31.21989	27.27213
edi_s	-.0570963	.1987225	-0.29	0.774	-.4485429	.3343502
ed_frac_ba_z	148.3435	59.20467	2.51	0.013	31.72122	264.9657
_cons	759.2029	149.0941	5.09	0.000	465.5151	1052.891

We can clearly see the relationship between test score and the student teacher ratio with the proportion of a students that receive free or reduced priced meals above the median is greater than that of below the median but can we say with 95% certainty that one relationship differs from each other?

I created a dummy variable for when the proportion is above the median. I then created an interaction variable between the student teacher ratio and the dummy variable and ran a regression.

Linear regression

Number of obs	=	500	
F(9, 490)	=	60.70	
Prob > F	=	0.0000	
R-squared	=	0.5461	
Root MSE	=	40.998	

testscore	Coefficient	Robust std. err.	t	P>|t|	[95% conf.	interval]
str_s	-.4392074	.7479375	-0.59	0.557	-1.908768	1.030353
frpm_frac_s	-156.5205	16.86784	-9.28	0.000	-189.6627	-123.3783
frpm_frac_s_binary	46.76051	27.62589	1.69	0.091	-7.519312	101.0403
interaction	-1.186271	1.075303	-1.10	0.270	-3.299045	.9265031
ell_frac_s	-15.86623	12.81557	-1.24	0.216	-41.04649	9.314032
te_salary_avg_d	.0012258	.0003538	3.46	0.001	.0005307	.0019209
log_med_income_z	-4.567455	13.40541	-0.34	0.733	-30.90663	21.77172
edi_s	-.2713505	.1457973	-1.86	0.063	-.5578156	.0151146
ed_frac_ba_z	140.6434	60.24019	2.33	0.020	22.28246	259.0044
_cons	803.0144	135.6653	5.92	0.000	536.4568	1069.572

The coefficient of the interaction term measures the difference in the relationship between student teacher ratios and test scores above and below the median for the proportion of students receiving free or reduced priced meals. I then ran a test on this interaction:

- H_0: There difference between the relationships $= 0$
- H_1: There difference between the relationships $\neq 0$

```
F(  1,   490) =    1.22
     Prob > F =    0.2705
```

We obtain a P-Value of .2705, thus we fail to reject H_1, hence we cannot say with 95% certainty that the relationship between class size and test scores differs between schools with above and below median shares of students receiving free or reduced-price school meals.

4 Part 4

- "All Models are wrong but some are useful," although my model has omitted variable bias and an adjusted R^2 of just over half, there is still value to it, it shows the relationship between test scores and income, diversity, class sizes among others.
- We cannot say with 95% certainty that on average an increase in class sizes causes an decrease in test scores, but, even if we could this is not sufficient proof to label this a causal relationship; we can say with 95% certainty that on average an increase in the proportion of students receiving free or reduced price meals will cause a decrease in test scores, is this a causal relationship? Obviously not.
- If I were to run this survey, I would split the school years into different sized classes, hence we would have 500 different schools where we can compare test scores between classes where all other independent variables would be equal to each other. This could potentially show a causal effect.

5 Appendix

Name: Max Dunne Student Number: 21365739

import excel "C:A Assignment.xlsx", sheet("Sheet1") firstrow

Part 1

label var countyname "county name"

label var districtname "district name"

label var schoolname "school name"

label var zipcode "zipcode of school"

label var testscore "test score (sum of math and english/language arts, 5th grade)"

label var str_s "student teacher(FTE) ratio (school)"

label var charter_s "charter school (0-1, school)"

label var frpm_frac_s "free or reduced price meals (fraction, school)" label var enrollment_s "enrollment (school)"

label var ell_frac_s "english language learners (fraction, school)"

label var edi_s "ethnic diversity index (school)"

label var te_fte_s "number (fte) teachers (school)"

label var te_avgyr_s "average years teaching (school)"

label var ada_enrollment_ratio_d "avg. daily attendance divided by enrollment (district)"

label var te_salary_low_d "Teacher Salary: lowest salary offered (district)"

label var te_salary_avg_d "Teacher Salary: average (district)"

label var te_days_d "Teaching days (district)"

label var te_serdays_d "Teaching service days (district)"

label var age_frac_5_17_z "Population(1+) fraction age 5-17 years (zipcode)"

label var pop_1_older_z "Population total: 1 year and older"

label var ed_frac_hs_z "Population (25+), education: high school (zipcode)"

label var ed_frac_sc_z "Population (25+), education: some college or AA (zipcode)"

label var ed_frac_ba_z "Population (25+), education: bachelors degree (zipcode)"

label var ed_frac_grd_z "Population (25+), education: graduate or professional degree (zipcode)"

label var med_income_z "Population (15+), median income (zipcode)"

sum countyname districtname schoolname zipcode testscore str_s charter_s frpm_frac_s enrollment_s ell_frac_s edi_s te_fte_s te_avgyr_s ada_enrollment_ratio_d te_salary_low_d te_salary_avg_d te_days_d te_serdays_d age_frac_5_17_z pop_1_older_z ed_frac_hs_z ed_frac_sc_z ed_frac_ba_z ed_frac_grd_z med_income_z

twoway (scatter testscore edi_s)

sum testscore edi_s if edi_s ¿ 32.466

sum testscore edi_s if edi_s ¡ 32.466

twoway (scatter testscore med_income_z)

gen log_med_income_z = log(med_income_z)

twoway (scatter testscore log_med_income_z)

twoway (scatter testscore frpm_frac_s)

twoway (scatter testscore te_salary_avg_d)

twoway (scatter testscore ada_enrollment_ratio_d)

sum testscore ada_enrollment_ratio_d if ada_enrollment_ratio_d ¿ 32.466

sum testscore ada_enrollment_ratio_d if ada_enrollment_ratio_d ¡ 32.466

twoway (scatter testscore te_avgyr_s)

gen te_avgyr_s_root = sqrt(te_avgyr_s)

twoway (scatter testscore te_avgyr_s_root)

twoway (scatter testscore ed_frac_ba_z)

Part 2

reg testscore str_s frpm_frac_s ell_frac_s med_income_z

ovtest

test str_s = 0

Part 3

reg testscore str_s frpm_frac_s ell_frac_s te_salary_avg_d log_med_income_z

ovtest

reg testscore str_s frpm_frac_s ell_frac_s te_salary_avg_d log_med_income_z edi_s

ovtest

reg testscore str_s frpm_frac_s ell_frac_s te_salary_avg_d log_med_income_z edi_s ed_frac_ba_z

ovtest

reg testscore str_s frpm_frac_s ell_frac_s te_salary_avg_d log_med_income_z edi_s ed_frac_ba_z ada_enrollment_ratio_

ovtest

egen median_frpm_frac_s = median(frpm_frac_s)

reg testscore str_s frpm_frac_s ell_frac_s te_salary_avg_d log_med_income_z edi_s ed_frac_ba_z if frpm_frac_s ¿ median_frpm_frac_s

reg testscore str_s frpm_frac_s ell_frac_s te_salary_avg_d log_med_income_z edi_s ed_frac_ba_z if frpm_frac_s ¡ median_frpm_frac_s

gen frpm_frac_s_binary = (frpm_frac_s ¿ median_frpm_frac_s)

gen interaction = str_s * frpm_frac_s_binary

reg testscore str_s frpm_frac_s frpm_frac_s_binary interaction ell_frac_s te_salary_avg_d log_med_income_z edi_s ed_frac_ba_z, robust