BUSINESS REPORT

Quantitative analysis for California Department of Education

TABLE OF CONTENTS

PART 1

1.1 EXPLORATORY ANALYSIS FOR THE AVERAGE 5TH GRADE READING AND MATH TEST SCORE

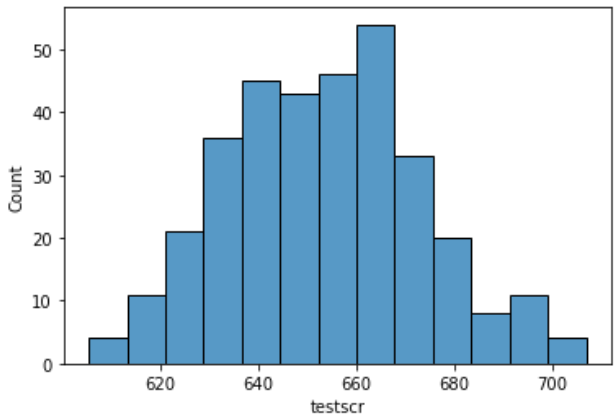

Fig 1.1.1

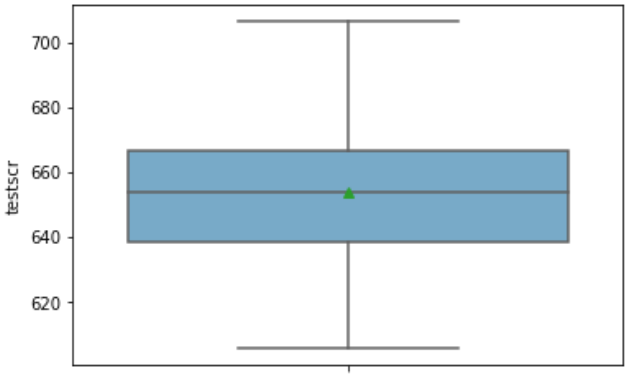

Fig 1.1.2

Statistical summary of sample test score	
Count	336
Min	605.550
25%	638.3375
50%	653.7500
75%	666.6625
Max	706.7500
Mean	653.6382
Variance	389.2172
Standard deviation	19.7286
Skewness	0.1717
Kurtosis	−0.3365

Table 1.1

From the above numerical summary, there are 336 observations with a considerable difference between the maximum test score and the minimum one at 101.2 (i.e 706.75-605.55). Additionally, the mean test score for the sample is observed at 653.6382 marks and the standard deviation at 19.7286 marks. Moreover, the 25, 50, and 75 percentiles are reported to be 638.3375, 653.75, and 666.6625 marks, respectively. Thus, the resulting interquartile range is 28.325 marks.

According to the histogram and the boxplot, the test score distribution is quite symmetric. This is further supported by the reported summary with the skewness of 0.1717 and kurtosis of -0.3365. Other than that, the boxplot also indicates no extreme outliers in the sample, which is favourable for later data analysis.

1.2 EXPLORATORY ANALYSIS FOR CLASS SIZE

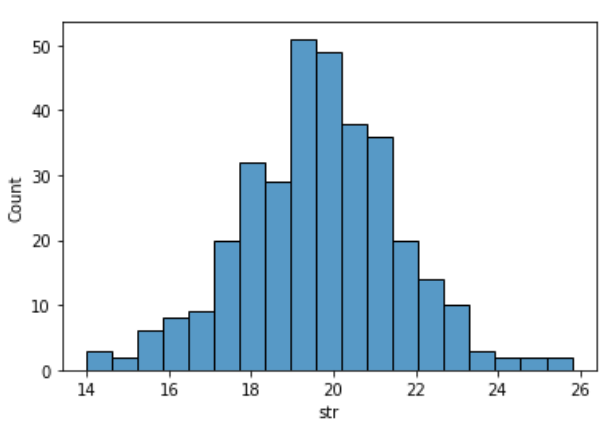

Fig 1.2.1

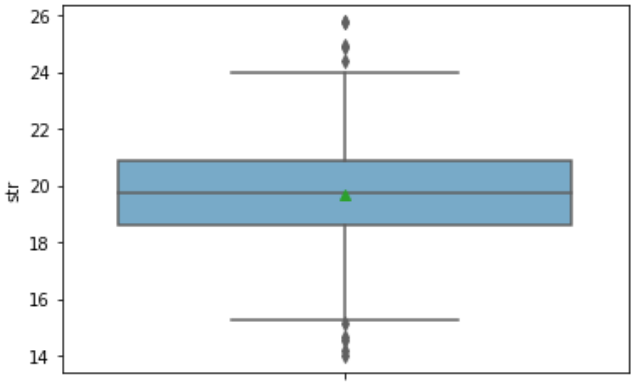

Fig 1.2.2

Statistical summary of sample class size	
Count	336
Min	14
25%	18.6305
50%	19.7601
75%	20.9024
Max	25.8000
Mean	19.6880
Variance	3.6850
Standard deviation	1.9196
Skewness	−0.0178
Kurtosis	0.5932

Table 1.2

From the above numerical summary, there are 336 observations with a considerable difference between the maximum student teacher ratio and the minimum one at 11.8 (i.e 25.8-14). Additionally, the mean ratio for the sample is observed at 19.688 and the standard deviation at 1.92. Moreover, the 25, 50, and 75 percentiles are reported to be 18.6305, 19.7601, and 20.9024, respectively. Thus, the resulting interquartile range is 2.2719.

According to the histogram and the boxplot, the test score distribution is quite symmetric. This is further supported by the reported summary with the skewness of -0.0178 and kurtosis of 0.5932. Other than that, the boxplot also indicates no extreme outliers in the sample, which is favourable for later data analysis.

1.3 EXPLORATORY ANALYSIS FOR THE RELATIONSHIP BETWEEN AVERAGE 5TH GRADE READING AND MATH TEST SCORE AND STUDENT TEACHER RATIO

	testscr	str
testscr	1	-0.2494
str	-0.2494	1

Table 1.3

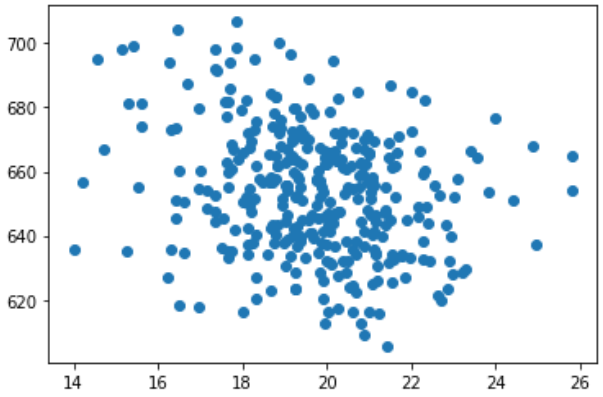

Fig 1.3

From the scatterplot, it may not be very clear about the type of relationship between these two variables. However, from the linear regression perspective, it might be said there is a negative linear relationship since when the student teacher ratio increases (i.e., a teacher teaches more students), the students' performances reflected through the test scores seem to be at a lower level. This is further supported by the reported negative correlation between these two variables of -0.2494. However, whether this correlation is significant is unclear without further statistical testing, which would be conducted below.

1.4 EXPLORATORY ANALYSIS FOR OTHER VARIABLES

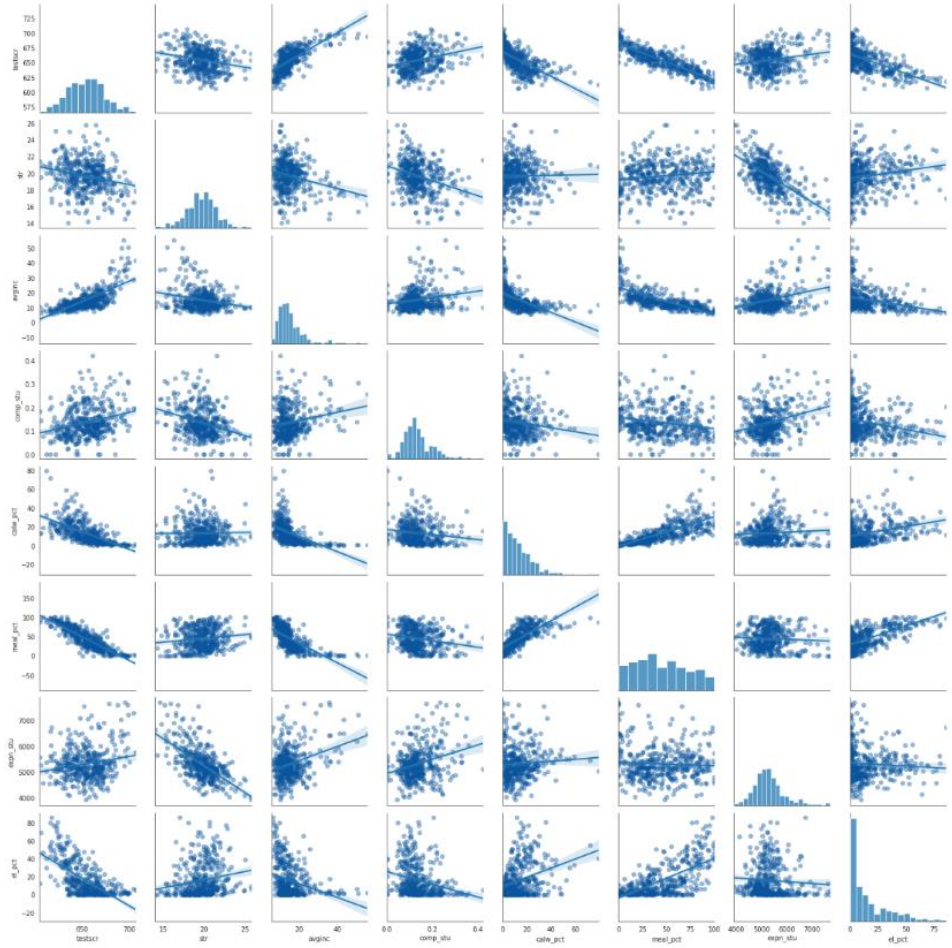

Fig 1.4

Statistical summary of other variables						
Variable	**enrl_tot**	**teachers**	**calw_pct**	**meal_pct**	**computer**	**comp_stu**
Count	336	336	336	336	336	336
Min	81	4.85	0	0	0	0
25%	373	19.6625	4.362	23.1831	45.75	0.0918
50%	955	50.375	10.52	44.3252	116.5	0.126
75%	3031	146.72	18.6574	67.8367	386.5	0.1726
Max	27176	1429	78.9942	100	3324	0.4208
Mean	2627.7054	128.6603	13.2708	45.5221	302.5744	0.1355
Variance	15416547.1	35280.5634	139.8558	772.9863	192609.7	0.0041
Standard deviation	3926.3911	187.8312	11.8261	27.8026	438.8732	0.0643
Skewness	2.9703	3.0608	1.7845	0.1473	2.9396	0.8103
Kurtosis	11.1498	12.5483	4.9639	−1.0576	11.7657	1.1994
Statistical summary of other variables						
Variable	**expn_stu**	**avginc**	**el_pct**	**read_scr**	**math_scr**	
Count	336	336	336	336	336	
Min	3926.0696	5.335	0	605.5	605.4	
25%	4888.856	10.591	1.9408	638.575	638.175	
50%	5203.1016	13.63	9.4205	655.4	651.7	
75%	5590.6052	17.508	27.8918	668.725	666.125	
Max	7711.5068	55.328	85.5397	704	709.5	
Mean	5305.4194	15.3177	16.432	654.3815	652.8949	
Variance	426575.7	57.1348	353.8599	435.0447	373.716	
Standard deviation	653.1277	7.5588	18.8112	20.8577	19.3317	
Skewness	1.0379	2.211	1.357	0.0122	0.3354	
Kurtosis	1.6488	6.1647	1.1905	−0.4803	−0.1803	

Table 1.4

Here enrl_tot and teachers would not be considered or graphed further since they are not the primary goals and have been reflected in the student teacher ratio. Similar argument can be applied for read_scr and math_scr, which have been reflected in average test score, and computer, which has been reflected in comp_stu as well. However, these 5 variables' statistical summary are shown for full information.

Most variables seem to be related to testcr, and most of them are linear. However, for avginc, calw_pct, and el_pct, the relationship with testscr seems to be curved and nonlinear. Logically speaking, avginc, expn_stu, and comp_stu are all positively correlated with testcr (these variables roughly indicate the level of extra support given to students), while el_pct, calw_pct, and meal_pct are negatively correlated with testcr (these are socioeconomic status variables, the larger, the lower its socio-economic status). The explanatory variables are also mostly related to each other, though usually in nonlinear and apparently complicated ways.

PART 2

2.1 ANALYSIS ON THE RELATIONSHIP BETWEEN CLASS SIZE AND AVERAGE 5TH GRADE READING AND MATH TEST SCORE

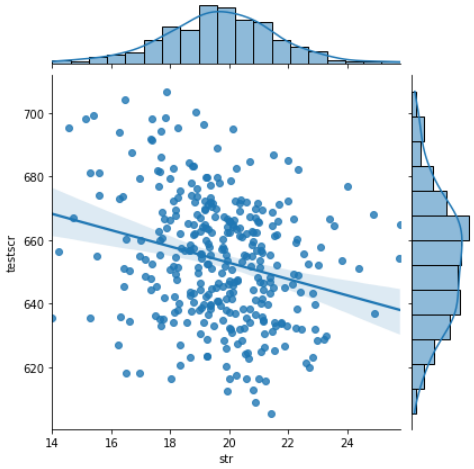

Fig 2.1.1

From the scatter plot illustrated above, there is a weak negative relationship recognized between the average 5th grade reading and math test score and class size since as the bigger the class size gets, the higher the average 5th grade reading and math test score is. Further evidence supports for this claim can be seen from the correlation coefficient of these two variable, which is registered at -0.249 approximately Besides, according to the result of testing hypotheses $H_0 : \rho_{str,testscr} = 0$ versus $H_1 : \rho_{str,testscr} \neq 0$, where $\rho_{str,testscr} \neq 0$ is the correlation coefficient between class size and average 5th grade reading and math test score, the p-value of 3.7048×10^{-06} claims that this correlation coefficient is significantly different than zero (Test 2.1.1).

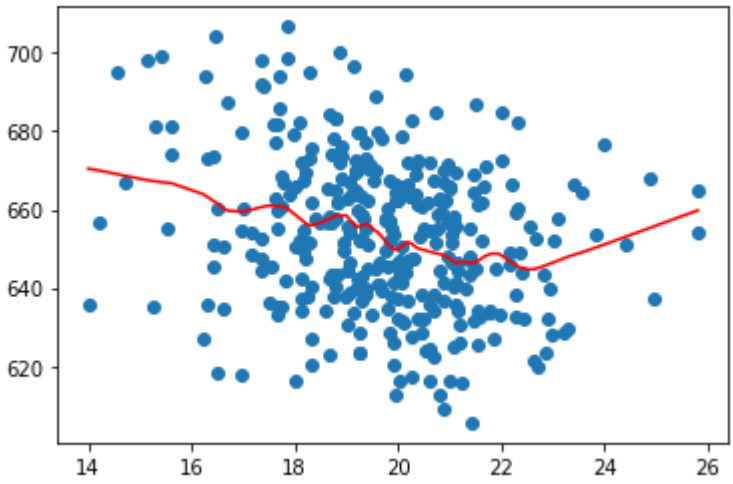

Fig 2.1.2

Observing the scatterplot illustrated above, there is a weak negative relationship shown by the relationship of the average 5th grade reading and math test score with the class size, the pattern show as the class size grow bigger the average 5th grade reading and math score will be lower. Besides, as the

class size crosses the threshold of 23 students per student, the score starts increases as suggested by the LOWESS line.

2.2 SIMPLE LINEAR REGRESSION MODEL

Adjusted R-squared	0.059
R-squared	0.062
SER	19.134

Predictor	Coefficient	Standard error	t	p-value	95 CI lower bound	95 CI upper bound
intercept	704.0954	10.772	65.361	0.000	682.905	725.286
str	-2.5628	0.545	-4.706	0.000	-3.634	-1.942

Table 2.2

According to the numerical summary of the model fitting, the estimated relationship between average 5th grade reading and math test score with class size can be illustrated as follows.

$$\widehat{testscr} = 704.0954 - 2.5628str$$

Thus, for any one student per teacher increase in the class size, the corresponding 2.5628 decrease will expect to be recognized in the mean of the average 5th grade reading and math test score. This logically makes sense since the larger the class size of a class, the more crowded the classroom will be, this could lead the class to be noisier and teacher will pay less attention to every individual student in the class and this may lead to lower average 5th grade reading and math test score.

2.3 TEST FOR RELATIONSHIP BETWEEN CLASS SIZE AND AVERAGE 5TH GRADE READING AND MATH TEST SCORE

The hypotheses are:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Test 2.3.1

By choosing $\alpha - 0.05$, it can obtain test statistics $t = \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})} = \frac{-2.5628}{0.545} = -4.706$, which has a degree of freedom of 334 if the null of coefficient is zero and the assumption is true. The chance that the resulting Student-t value would be more extreme than −4.706, in the direction of the alternative hypothesis which is less than or greater than 0. Specifically, the p-value is $2 \times P(t_{334} < -4.706) = 0.000$.

Hence, such p-value is less than 0.05, it is statistically sufficient to reject the null hypothesis and conclude that the correlation between average 5th grade reading and math test score and class size is significant.

2.4 STRENGTH OF FIT ASSESSMENT

The $R^2 = 0.062$ and $R^2_{adj} = 0.059$, which is relatively low and thus, can only demonstrate a weak strength fit to the data. Besides, $SER = 19.134$ marks, showing that the standard deviation of the residuals is approximately 19.134, which is significantly large for predicting and other practical purposes. Besides, it is necessary to have an education consultant to consult whether such an error is a relatively small or larger standard deviation when predicting the average score.

2.5 TEST ASSUMPTIONS ASSESSMENT

2.5.1 Linearity and exogeneity:

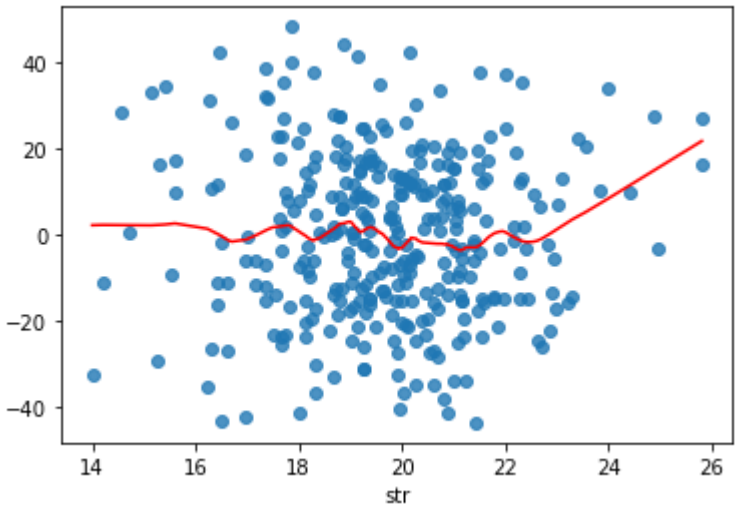

Fig 2.5.1

By observing the LOWESS line of residuals against fitted values, there is a tendency that the model may overpredict as class size cross the level of 23 students per teacher. Nevertheless, the line does not show any disastrous deviation from zero across most values of class size, which suggests the prediction is relatively accurate as the class size is smaller 23 students per teacher. Thus, the linearity relationship is acceptable in broad term

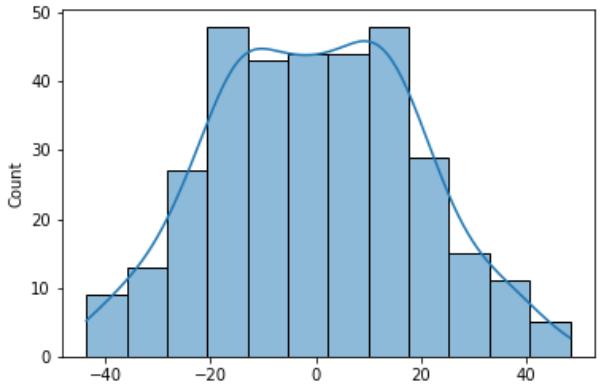

Fig 2.5.2

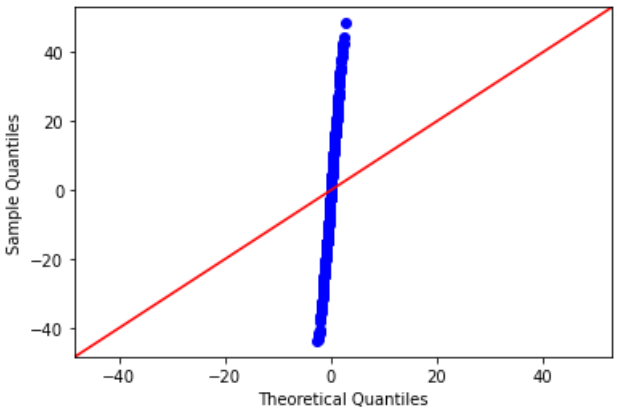

Fig 2.5.3

Statistical summary of sample residuals	
Min	-43.6846
Max	48.4336
Mean	-6.0430×10^{-13}
Variance	365.0136
Skewness	0.0330
Kurtosis	-0.5186

Table 2.5.1

Other than that, despite the fact that the QQ plot illustrates that the distribution of the residuals does not perfectly follow normal distribution the histogram of the residuals shows a symmetric distribution around zero, which is indicated more obviously by the mean of residuals of approximately -6.043×10^{-13} marks and the skewness of -0.033, which is around zero. Thus, this suggests exogeneity assumption may hold

Hence, the linearity and exogeneity assumptions may be satisfied in general.

2.5.2 Independence and identical distribution of data:

Without any further description of data collecting procedure, the validity of this assumption in this data set cannot be assured. Nevertheless, this assumption may assume to be true for average test score unless students cheat off each other and this should also be true for class size as well. Hence, the assumption of independence and identical distribution of data may be satisfied for this data set.

2.5.3 Existence of fourth moments of response variable and predictors:

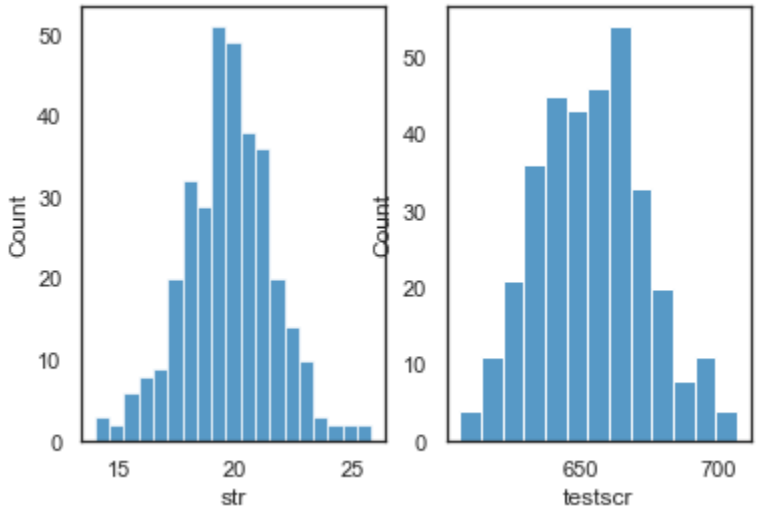

Fig 2.5.4

Considering average 5th grade reading and math test score, this can reasonable be a bounded variable. Specifically, it is the case since the math score and reading are limited from 0 to a certain upper bound logically which makes the average value of these two variables a bounded random variable. Furthermore, the histogram of average 5th grade reading and math test score shows a symmetric pattern with nearly no outliers, which can be examined more thoroughly via the skewness of 0.171 and the kurtosis of -0.336. Hence, the fourth moment of the average 5th grade reading and math test score can be regarded as a finite one.

Turning to class size, as the total population of a district should be bounded between 0 and a certain upper bound, the number of enrolments in a district should be a bounded random variable. The same reasoning can be also applied to the total number of teachers in a district, which makes the number of students per teacher a bounded random variable. Thus, the fourth moment of class size should be finite.

Hence, the existence of fourth moments of average 5th grade reading and math test score and class size are met in this data set.

2.5.4 Constant error variance:

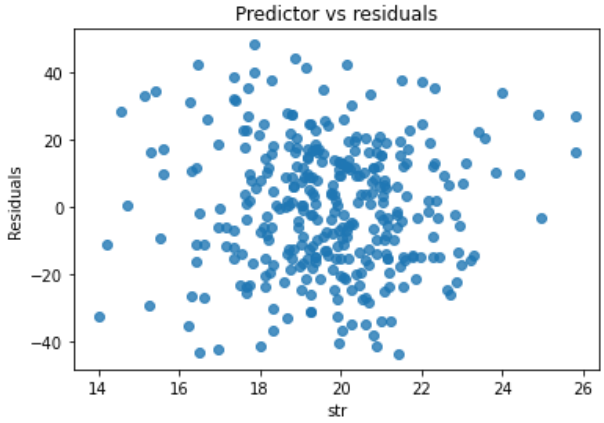

Fig 2.5.5

By observing the scatter plot of residuals against class size, the residuals are distributed nicely and evenly around zero with no clear non-linear pattern. Hence, the assumption of constant error variance may be satisfied.

PART 3:

3.1 POTENTIAL VARIABLE CAUSING OMITTED VARIABLE BIAS

	testscr	calw_pct	meal_pct	comp_stu	expn_stu	str	avginc	el_pct
testscr	1	-0.6289	-0.8694	0.2601	0.1986	-0.2494	0.7318	-0.6460
calw_pct	-0.6289	1	0.7292	-0.1237	0.0738	0.0341	-0.5013	0.3397
meal_pct	-0.8694	0.7292	1	-0.1732	-0.0581	0.1628	-0.6896	0.6630
comp_stu	0.2601	-0.1237	-0.1732	1	0.2966	-0.3285	0.1924	-0.2297
expn_stu	0.1986	0.0738	-0.0581	0.2966	1	-0.6286	0.3276	-0.0490
str	-0.2494	0.0341	0.1628	-0.3285	-0.6286	1	-0.2671	0.1854
avginc	0.7318	-0.5013	-0.6896	0.1924	0.3276	-0.2671	1	-0.3262
el_pct	-0.6460	0.3397	0.6630	-0.2297	-0.0490	-0.3262	-0.3262	1

Table 3.1.1

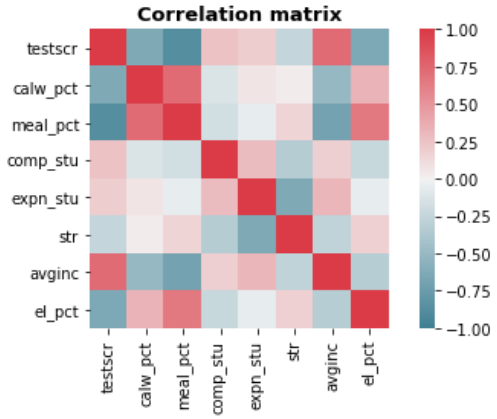

Fig 3.1

By choosing $\alpha = 0.05$, the report comes up with the tests of correlation between potential OVB variables and average test score and between potential OVB variables and class size with their corresponding results

H_0	H_1	r	p-value	Reject
$\rho_{avginc,testscr} = 0$	$\rho_{avginc,testscr} \neq 0$	0.7318	1.4504×10^{-57}	Yes
$\rho_{avginc,str} = 0$	$\rho_{avginc,str} \neq 0$	-0.2671	6.7546×10^{-07}	Yes
$\rho_{comp\_stu,testscr} = 0$	$\rho_{comp\_stu,testscr} \neq 0$	0.2601	1.3445×10^{-06}	Yes
$\rho_{comp\_stu,str} = 0$	$\rho_{comp\_stu,str} \neq 0$	-0.3285	6.7645×10^{-10}	Yes
$\rho_{expn\_stu,testscr} = 0$	$\rho_{expn\_stu,testscr} \neq 0$	0.1986	0.0002	Yes
$\rho_{expn\_stu,str} = 0$	$\rho_{expn\_stu,str} \neq 0$	-0.6286	2.3873×10^{-38}	Yes
$\rho_{meal\_pct,testscr} = 0$	$\rho_{meal\_pct,testscr} \neq 0$	-0.8694	2.8053×10^{-104}	Yes
$\rho_{meal\_pct,str} = 0$	$\rho_{meal\_pct,str} \neq 0$	0.1628	0.0028	Yes
$\rho_{calw\_pct,testscr} = 0$	$\rho_{calw\_pct,testscr} \neq 0$	-0.6289	2.1279×10^{-38}	Yes
$\rho_{calw\_pct,str} = 0$	$\rho_{calw\_pct,str} \neq 0$	0.0341	0.5333	No
$\rho_{el\_pct,testscr} = 0$	$\rho_{el\_pct,testscr} \neq 0$	-0.646	4.4718×10^{-41}	Yes
$\rho_{el\_pct,str} = 0$	$\rho_{el\_pct,str} \neq 0$	0.1854	0.0006	Yes

Table 3.1.2

According to the above summary table of correlation coefficient tests between class size and other relevant variables and between average 5th grade reading and math test score and other relevant variables, with the exception of percentage qualifying for CALWORKS whose significance relationship with class size is rejected, district average income, computer per student, expenditure per student, percentage qualifying for reduced-price lunch, and percentage of English learners all have their relationship with class size and average 5th grade reading and math test score proved to be significant at the critical level of 5%. Subsequently, it is necessary to examine the determinant relationship between average 5th grade reading and math test score and the mentioned five variables.

3.1.1 District average income:

Considering the determinant relationship, the more average income a district earns, the more enhanced life quality of that district gets, which boosts the education quality of students in that district. Logically, high education quality on average in a district can positively influence average 5th grade reading and math test score in that district. Thus, district average income may be a determinant of average 5th grade reading and math test score.

Other than that, the matrix of scatterplots also demonstrates a positive relationship between district average income and average test score and also a negative relationship between district average income and class size. In conclusion, district average income may cause omitted variable bias in the simple linear regression model in the previous part.

3.1.2 Computer per student:

Considering the determinant relationship, the more computers a typical student in a district has, the more effectively that student in that district can study and this may lead to higher average 5th grade reading and math test score. Thus, computer per student may be a determinant of average 5th grade reading and math test score.

Other than that, the matrix of scatterplots also demonstrates a positive relationship between computer per student and average test score and also a negative relationship between computer per student and class size. In conclusion, computer per student may cause omitted variable bias in the simple linear regression model in the previous part.

3.1.3 Expenditure per student:

Considering the determinant relationship, the more the student are invested in a district, the better education quality the students can get in a district, which leads to higher average 5th grade reading and math test score in that district. Thus, expenditure per student may be a determinant of average 5th grade reading and math test score.

Other than that, the matrix of scatterplots also demonstrates a positive relationship between expenditure per student and average test score and also a negative relationship between expenditure per student and class size. In conclusion, expenditure per student may cause omitted variable bias in the simple linear regression model in the previous part.

3.1.4 Percentage qualifying for reduced-price lunch:

Considering the determinant relationship, the less people are qualified for reduced-price lunch in a district, the more independent the citizen in that district can independently pay for their living, which means the quality of life of students in that district is better and final result is that students in that district can achieve better average 5th grade reading and math test score. Thus, percent qualifying for reduced-price lunch may be a determinant of average 5th grade reading and math test score.

Other than that, the matrix of scatterplots also demonstrates a negative relationship between percentage qualifying for reduced-price lunch and average test score and also a positive relationship between percentage qualifying for reduced-price lunch and class size. In conclusion, percent qualifying for reduced-price lunch may cause omitted variable bias in the simple linear regression model in the previous part.

3.1.5 Percentage of English learners:

Considering the determinant relationship, the more percentage of students in a district need to take English course, the more problem those students are experience with their linguistic ability, which leads to lower average 5th grade reading and math test score in that district. Thus, percentage of English learners may be a determinant of average 5th grade reading and math test score.

Other than that, the matrix of scatterplots also demonstrates a negative relationship between percentage of English learners and average test score and also a positive relationship between percentage of English learners and class size. In conclusion, percentage of English learners may cause omitted variable bias in the simple linear regression model in the previous part.

3.2 MULTIPLE REGRESSION MODEL INCLUDING POTENTIAL VARIABLES CAUSING OMITTED VARIABLE BIAS

Until now, with the above justification, all possible variables that can cause omitted variable bias are district average income, computer per student, expenditure per student, percentage qualifying for reduced-price lunch, percentage of English learners and these five random variables will be fitted along with class size to create a multiple regression model to predict average 5th grade reading and math test score.

Adjusted R-squared	0.813
R-squared	0.817
SER	8.525

Predictor	Coefficient	Standard error	t	p-value	95 CI lower bound	95 CI upper bound
intercept	653.1153	10.149	64.353	0.000	633.150	673.080
str	-0.0140	0.323	-0.043	0.965	-0.649	0.621
avginc	0.6601	0.097	6.826	0.000	0.470	0.850
comp_stu	16.9687	7.879	2.154	0.032	1.470	32.468
meal_pct	-0.4039	0.031	-12.874	0.000	-0.466	-0.342
expn_stu	0.0018	0.001	1.848	0.066	-0.000	0.004
el_pct	-0.1785	0.035	-5.109	0.000	-0.247	-0.110

Table 3.2

According to the numerical summary of the model fitting, the estimated relationship can be illustrated as following.

$\widehat{testscr} = 653.1153 - 0.0140\text{str} + 0.6601\text{avginc} + 16.9687\text{comp\_stu} - 0.4039\text{meal\_pct} + 0.0018\text{expn\_stu} - 0.1785\text{el\_pct}$

This estimated relationship provided the following information:

- As other variables remain constant, 1 student increase in class size will decrease the mean of average 5th grade reading and math test score by 0.0140 scores.
- As other variables remain constant, 1 dollar increase in district average income will increase the mean of average 5th grade reading and math test score by 0.6601 scores.
- As other variables remain constant, 1 computer increase in computer per student will increase the mean of average 5th grade reading and math test score by 16.9687 scores.
- As other variables remain constant, 1 percent increase in percentage qualifying for reduced-price lunch will decrease the mean of average 5th grade reading and math test score by 0.4039 scores.
- As other variables remain constant, 1 dollar increase in expenditure per student will increase the mean of average 5th grade reading and math test score by 0.0018 scores.
- As other variables remain constant, percent increase in percentage of English learners will decrease the mean of average 5th grade reading and math test score by 0.1785 scores.

3.3 TEST FOR A RELATIONSHIP BETWEEN CLASS SIZE AND TEST SCORE

The hypotheses are:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Test 3.3.1

Where β_1 reflects the relationship between class size and test score estimated in this 6-predictor model

By choosing the critical level $\alpha = 0.05$, we have the test statistic is $t = \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})} = \frac{-0.014}{0.323} = -0.043$, which has a Student-t distribution with 329 degrees of freedom, if the null and the assumptions are true

The chance that the resulting Student-t distribution value would be more extreme than -0.043, in the direction of the alternative hypothesis which is less than or greater than 0. Specifically, the p-value is $2 \times P(t_{329} < -0.043) = 0.965$, which is larger than 0.05. Hence, it is not statistically significant for us to reject the null hypothesis and conclude that there may be no significant impact of the class size on the average 5th grade reading and math test score.

Again, 95% confidence interval for the slope in the simple linear regression from part 2 is $(-3.634, -1.492)$, but the corresponding figure for this 6-predictor multiple linear regression is $(-0.649, 0.621)$, which contains 0. Therefore, the estimated effect of class size on test score in the 6-predictor model is insignificant, and it is reasonable that other 6 predictors have caused some omitted variable biased.

3.4 STRENGTH OF FIT ASSESSMENT

The $R^2 = 0.817$ and $R^2_{adjusted} = 0.813$, which is considerably high and thus can only demonstrate a decent strength fit to the data. Besides $SER = 8.525$ shows that the standard deviation of the residuals is approximately 8.825 marks, which may be significant for practical purposes. However, it is necessary for us to have a consultation with experts to determine whether such an error is a relatively small or large when predicting the average score.

3.5 GOODNESS OF FIT

3.5.1 Linearity and exogeneity:

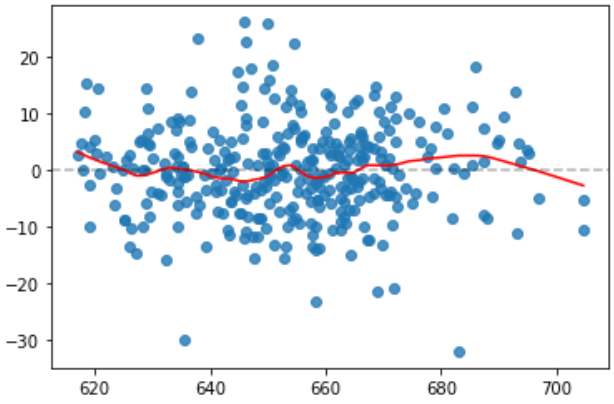

Fig 3.5.1

Observing the LOWESS line of the residual against the fitted values, the curve does not deviate much around zero and relatively flat and also from the scatter plot, the residuals seem to be nice and evenly distributed around the horizontal line 0. Thus, the assumption of linear relationship can be broadly satisfied.

Statistical summary of sample residuals	
Min	-32.1013
Max	26.2298
Mean	1.1386×10^{-12}
Variance	71.3681
Skewness	-0.0046
Kurtosis	0.9903

Table 3.5.1

From the statistical summary of the sample residuals, the distribution of sample residuals is relatively symmetric and not suffering from many outliers due to the skewness of -0.0046 and the kurtosis of 0.9903. Furthermore, as suggested by the sample mean of residual of 1.1386×10^{-12} which is approximately 0 and the fact that the LOWESS line does not deviate much from zero, it is reasonable to conclude that exogeneity assumption may be met.

3.5.2 Independence and identical distribution of data:

Without clear description of the data collection procedure, it cannot be confirmed with certainty whether the observations with their corresponding test score, class size, average income, computer per student, percent qualifying for reduced-price lunch, expenditure per student, and percent of English learners are independent and identically distributed or not. However, we can broadly assume that each observation is collected independently of one another. Therefore, the data inputs, which are six predictors, are also independent of one another. Hence, the assumption of independence and identical distribution of data may be satisfied.

3.5.3 Existence of fourth moments of response variable and predictors:

Statistical summary of 6 predictors						
Variable	**Min**	**Max**	**Mean**	**Variance**	**Skewness**	**Kurtosis**
testscr	605.55	706.75	653.6382	389.2172	0.1717	-0.3365
str	14	25.8	19.688	3.685	-0.0178	0.5932
avginc	5.335	55.328	15.3177	57.1348	2.211	6.1647
comp_stu	0	0.4208	0.1355	0.0041	0.8103	1.1994
meal_pct	0	100	45.52211	772.863	0.1473	-1.0576
expn_stu	3926.0696	7711.5068	5305.4194	426575.735	1.0379	1.6488
el_pct	0	85.5397	16.432	353.8599	1.357	1.1905

Table 3.5.3 – retrieved from Table 1.4

Considering the average 5th grade reading and math test score, this can reasonable be a bounded variable since the math and reading scores cannot be lower than 0 and should have a certain maximum point logically. Hence, the fourth moment of the average 5th grade reading and math test score can be regarded as a finite one.

Regarding the class size, as the population of a district should be bounded between 0 and a certain upper bound, the number of enrolments in a district should be a bounded random variable. The same reasoning can be also applied to the total number of teachers in a district, which makes the number of students per teacher a bounded random variable. Thus, the fourth moment of class size should be finite.

In terms of district average income, as income of a person should be bounded, it is reasonable to claim that district average income is a bounded random variable. Thus, the fourth moment of district average income should be finite. Still, the high kurtosis of 6.1647 of district average income should be monitored cautiously since it is significantly larger than 0, which is an indication of excessive outliers that may violate of the existence of forth moment.

Considering computer per student, as the number of the computers and the number of students in a district should be bounded below by 0 and some certain upper bounds, computer per student should be a bounded random variable. Thus, the fourth moment of computer per student is finite.

In terms of expenditure per student, as the number of students and total income of the schools in a district is limited, expenditure per student should be a bounded random variable. Thus, the fourth moment of expenditure per student is finite.

Regarding percentage qualifying for reduced-price lunch, this value is bounded between 0 and 100. Thus, the fourth moment of percentage qualifying for reduced-price lunch is finite.

Considering percentage of English learners, this value is bounded between 0 and 100. Thus, the fourth moment of percentage of English learners is finite.

Finally, all above patterns can be observed in figure 1.4.1 in which the histogram of average income indicates considerable positive outliers, whereas other histograms the distributions seem to be more focused with less abnormal high or low values.

Hence, the existence of fourth moments of response variable and predictors are met in this data set.

3.5.4 No perfect collinearity among predictors:

From the matrix and table of correlation coefficient, there are high positive correlation between percentage qualifying for reduced-price lunch and percentage of English learners of 0.663 as well as high negative correlation between percentage qualifying for reduced-price lunch and district average income of -0.6896 and between expenditure per student and class size of -0.6286. Nevertheless, these figures are not highly close to 1 or -1, so this assumption can be broadly satisfied. Still, it should be acknowledged that whether the true population collinearity is perfect or not is not completely obvious.

3.5.5 Constant error variance:

As indicated above by the scatter plot, the residuals are nice and evenly distributed around 0 and no significant fan-shaped patterns can be observed regarding the residuals. Thus, the assumption may be met.

3.6 LEVEL AND SOURCES OF MULTI-COLLINEARITY ASSESSMENT

Variable	str	avginc	meal_pct	comp_stu	expn_stu	el_pct
VIF	1.7677	2.4637	3.5071	1.1844	1.9276	1.9924

Mean = 2.1405

Table 3.6

While the VIFs for class size, district average income, computer per student, expenditure per student, percentage of English learners are lower than 3 as indicated above, that of percentage qualifying for reduced-price lunch is larger than 3. However, all figures observed are still lower than 5, so there may not be a fair degree of variance inflation and collinearity in the data. Furthermore, the average VIF is below 3, so, again, collinearity should not be a huge issue.

PART 4

4.1 MULTIPLE LINEAR REGRESSION – OMMITTED VARIABLE BIAS

This model has been fitted and presented in question 3

4.2 SIMPLE LINEAR REGRESSION

This model has been fitted and presented in question 2

4.3 LINEAR SPLINE MODEL:

4.3.1 Additional variables:

The spline variable for class size for one knot ('str_step'):

$$str\_step = (str\_step - 23)_+ = I(str\_step > 23) \times (str\_step - 23) = \begin{cases} str\_step - 23, str\_step > 23 \\ 0, str\_step < 23 \end{cases}$$

The motivation for creating this spline variable is that as observing the graph of LOWESS line of average test score against class size, the linearity relationship between these two variables is significantly different as class size ratio is below 23 percent as compared to the relationship as it becomes more linearly upward (i.e., steeper slope) when crossing the threshold of 23 percent. Thus, 23 is chosen to be the knot.

4.3.2 Proposed population relationship:

$$testscr = \beta_0 + \beta_1 str + \beta_2 (str - 23)_+ + \varepsilon$$

4.3.3 Fitted model summary:

Adjusted R-squared	0.076
R-squared	0.082
SER	18.963

Predictor	Coefficient	Standard error	t	p-value	95 CI lower bound	95 CI upper bound
Intercept	716.3300	11.626	61.615	0.000	693.460	739.200
str	-3.2073	0.592	-5.421	0.000	-4.371	-2.043
str_step	10.7049	4.027	2.658	0.008	2.782	18.627

Table 4.3.3

According to the summary table, the estimated relationship is as following

$$\widehat{testscr} = 716.33 - 3.2073str + 10.7049str\_step$$

4.3.4 Strength of fit:

$R^2 = 0.082$ and $R^2_{\text{adjusted}} = 0.076$ may be very low and thus only demonstrating a slight strength fit to the data. Also, $\text{SER} = 18.963$ marks, which shows that the standard deviation of the residuals is approximately 18.963 marks, might be considerably high. It may be necessary for us to have a consultation with experts to determine whether such an error is a relatively small or large when predicting the average score. However, overall, this model is just marginally better than the SLR model and still poorly predicts the test scores.

4.3.5 Goodness of fit:

The LOWESS line of residuals against the fitted values suggests that there may be linearity, with the residuals seems to be nice and evenly distributed around the horizontal line $y = 0$. Also, the line is relatively flat and does not deviate significantly from zero, so LSA 1 and 2 seem satisfied. Testscr should be bounded so will have finite 4th moments. There is no clear evidence of heteroskedasticity, so LSA 6 seems OK. LSA 3 is about whether the data are independent and identically distributed which should be true for testscr (unless students cheat off each other) and should also be true for str too as long as the observations are independent.

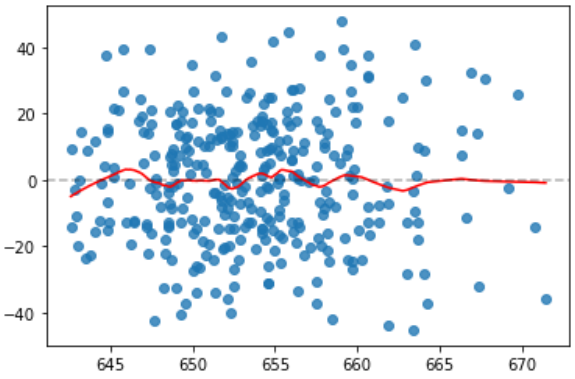

Fig 4.3.5

****Model 2 (SLR) and model 3 (Linear spline) are initially built to examine the pure effects of class size on test scores, which is the primary objective of the report, so they are intentionally simple and easy to interpret. Subsequently more transformations and interaction effects would be introduced to determine a decent model for predicting the response variable ('testscr'). However, as there is a trade-off between train sample predictability and over-fit issue, the report would try to maintain a reasonable number of predictors. In other words, all models would not be incorporated with every possible interaction or transformation but only some of them to maintain the understandability and over-fit.*

4.4 QUADRATIC POLYNOMIAL WITH INTERACTION EFFECT:

4.4.1 Additional variables:

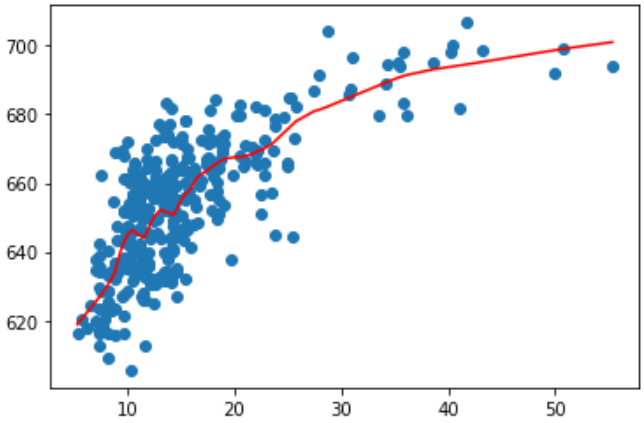

Fig 4.4.1

Squares of district average income ('avginc_power_2'): The LOWESS line test score against average income seems to suggest a quadratic relationship for $avginc < 25$. Specifically, in the range, the scatterplot can be said to form a shape of a concave down parabola, so it is reasonable there is a model examining the second-power polynomial relationship between district average income and the response variable.

District average income ('avginc'): This predictor is introduced to have a complete quadratic model. Also, this aligns with the principle of marginality, which states that a model including higher-order terms (such as an interaction) should also include all the lower-order relatives of that term.

District average income and English percentage interaction ('avginc_elpct'): People study English in California because they may have lacked language education before, which might indicate life hardships or poverty. Also, this idea can be applied to people with low average income as well. Therefore, it is necessary to examine whether those who study English and have low income may perform even poorer than the average overall, so the interaction effect is introduced here.

Percentage of English learners ('el_pct'): This predictor is introduced to align with the principle of marginality, which states that a model including higher-order terms (such as an interaction) should also include all the lower-order relatives of that term.

4.4.2 Proposed population relationship:

$$testscr = \beta_0 + \beta_1 str + \beta_2 el\_pct + \beta_3 avginc + \beta_4 avginc\_power\_2 + \beta_5 avginc\_elpct + \varepsilon$$

4.4.3 Fitted model summary:

Adjusted R-squared	0.743
R-squared	0.747
SER	10.003

Predictor	Coefficient	Standard error	t	p-value	95 CI lower bound	95 CI upper bound
Intercept	624.2537	6.869	90.876	0.000	610.741	637.767
str	-0.2656	0.300	-0.887	0.376	-0.855	0.324
el_pct	-0.1629	0.086	-1.891	0.059	-0.332	0.007
avginc	3.4532	0.345	10.007	0.000	2.774	4.132
avginc_power_2	-0.0366	0.006	-5.732	0.000	-0.049	-0.024
avginc_elpct	-0.0240	0.007	-3.441	0.001	-0.038	-0.010

Table 4.4.3

According to the summary table, the estimated relationship is as following

$$\widehat{testscr} = 624.2537 - 0.2656 str - 0.1629 el\_pct + 3.4532 avginc - 0.0366 avginc\_power\_2 - 0.0240 avginc\_elpct$$

4.4.4 Strength of fit:

$R^2 = 0.747$ and $R^2_{\text{adjusted}} = 0.743$ may be quite considerable and thus only demonstrating a moderate strength of fit to the data. Also, $\text{SER} = 10.003$ marks, which shows that the standard deviation of the residuals is approximately 10.003 marks, might be considerably moderate. It may be necessary for us to

have a consultation with experts to determine whether such an error is a relatively small or large when predicting the average score. However, overall, this model predicts the test scores quite well.

4.4.5 Goodness of fit:

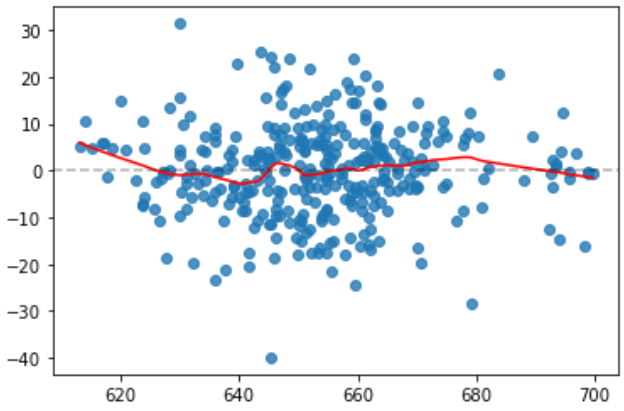

Fig 4.4.5

The LOWESS line of residuals against the fitted values suggests that there may be some nonlinearity, with the model under-predicting test scores for $660 < \widehat{testscr} < 690$ and over-predicting test scores for $630 < \widehat{testscr} < 650$ particularly. Nevertheless, the line is relatively flat and does not deviate significantly from zero, so LSA 1 and 2 are broadly satisfied. Testscr, el_pct, and avginc should be bounded so will have finite fourth moments. This also results that avginc_power_2 and and avginc_elpct should be finitely bounded as well. There is no clear evidence of heteroskedasticity, so LSA 6 seems OK. LSA 3 is about whether the data are independent and identically distributed or not which should be true for testscr (unless students cheat off each other) and should also be true for el_pct, avginc, and in turn avginc_power_2, avginc_elpct too as long as the observations are independent.

4.5 QUADRATIC POLYNOMIAL AND LOG TRANSFORMATION:

4.5.1 Additional variables:

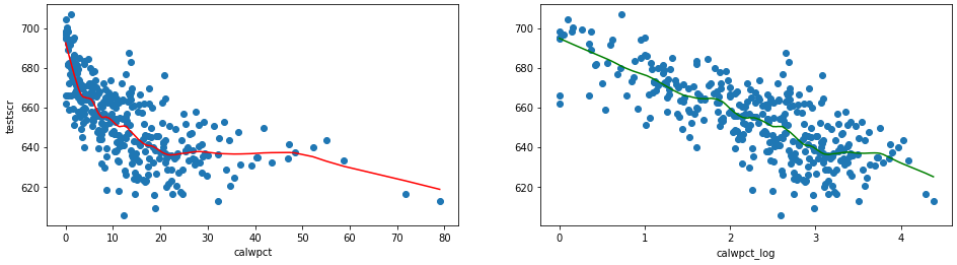

Fig 4.5.1

Log of percentage qualifying for CALWORKS ('calwpct_log): The LOWESS line between test score and percentage qualifying for CalWORKS seems to suggest a log relationship. Specifically, the scatterplot can be said to form a shape of a $y = -log(x)$ graph, so it is reasonable there is a model examining this relationship. However, as it is noted from the statistical summary of percentage qualifying for CALWORKS, the minimum of this data is 0, which is not feasible to conduct a log transformation. Therefore, following Fox's recommendation (2016), the data is added 1 before applying log (i.e., log(calw_pct+1)). The outcome

can be seen in the LOWESS line of average score against log of percentage qualifying for CALWORKS, in which the transformed relationship becomes more linear and favourable for modelling.

Avginc_power_2 and avginc: The argument and reason to include these two predictors are similar to those in model 3. Nevertheless, the interaction effect between percentage of English learners and district average income is left out to avoid overcomplication and assess the effectiveness of these two considerable quadratic and log transformations.

4.5.2 Proposed population relationship:

$$testscr = \beta_0 + \beta_1 str + \beta_2 avginc + \beta_3 avginc\_power\_2 + \beta_4 calwpct\_log + \varepsilon$$

4.5.3 Fitted model summary:

Adjusted R-squared	0.671
R-squared	0.675
SER	11.308

Predictor	Coefficient	Standard error	t	p-value	95 CI lower bound	95 CI upper bound
Intercept	666.0523	8.773	75.920	0.000	648.794	683.310
str	-1.0215	0.336	-3.039	0.003	-1.683	-0.360
avginc	2.4023	0.343	7.012	0.000	1.728	3.076
avginc_power_2	-0.0284	0.006	-4.521	0.000	-0.041	-0.016
calwpct_log	-9.0566	0.995	-9.105	0.000	-11.013	-7.100

Table 4.5.3

According to the summary table, the estimated relationship is as following

$$\widehat{testscr} = 666.0523 - 1.0215 str + 2.4023 avginc - 0.0284 avginc\_power\_2 - 9.0566 calwpct\_log$$

4.5.4 Strength of fit:

$R^2 = 0.675$ and $R^2_{adjusted} = 0.671$ may be quite considerable and thus only demonstrating a moderate strength of fit to the data. Also, $SER = 11.308$ marks, which shows that the standard deviation of the residuals is approximately 11.308 marks, might be considerably moderate. It may be necessary for us to have a consultation with experts to determine whether such an error is a relatively small or large when predicting the average score. However, overall, this model predicts the test scores quite well, but this is lower than that of quadratic polynomial with interaction effect.

4.5.5 Goodness of fit:

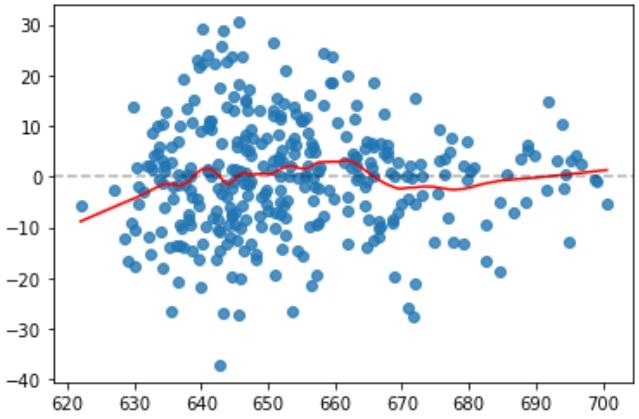

Fig 4.5.5

The LOWESS line of residuals against the fitted values suggests that there may be some nonlinearity, with the linear model under-predicting test scores for $\widehat{testscr} < 640$ and over-predicting test scores for $640 < \widehat{testscr} < 670$ particularly. Nevertheless, the line is relatively flat and does not deviate significantly from zero, so LSA 1 and 2 are broadly satisfied. Testscr, avginc and in turn avginc_power_2 should be bounded so will have finite 4th moments. Regarding calwpct_log, thanks to the transformation that forces lower limit of 1 and a certain upper limit, it should be finitely bounded as well. The variance of the residuals seems to be larger with low test scores, so LSA 6 may be violated. LSA 3 is about whether the data are independent and identically distributed or not which should be true for testscr (unless students cheat off each other) and should also be true for calw_pct, avginc and in turn avginc_power_2, calwpct_log too as long as the observations are independent.

4.6 FORWARD SELECTION MODEL:

4.6.1 Additional variables:

$$elpct\_step = (el\_pct - 20)_+ = I(el\_pct > 20) \times (el\_pct - 20) = \begin{cases} el\_pct - 20\,, el\_pct > 20 \\ 0, el\_pct < 20 \end{cases}$$

The motivation for creating this spline variable is that as observing the graph of LOWESS line of average test score against percentage of English learners, the linearity relationship between these two variables is significantly different as percentage of English learners is still below 20 percent as compared to the relationship as it crosses the threshold of 20 percent. Thus, 20 is chosen to be the knot.

$$mealpct\_elpct = meal\_pct \times el\_pct$$

The motivation for creating this interaction variable is that it is necessary to examine whether the impact of poor linguistic skills of students in a district may be different across different level of poor financial status of that district. Thus, the interaction variable

4.6.2 Proposed population relationship:

As suggested by the forward selection algorithm based on the criteria of maximizing R^2_{adj}, the population relationship may be as following equations.

$$testscr = \beta_0 + \beta_1 str + \beta_2 str\_step + \beta_3 meal\_pct + \beta_4 avginc + \beta_5 avginc\_power\_2 + \beta_6 el\_pct + \beta_7 elpct\_step + \beta_8 calwpct\_log + \beta_9 comp\_stu + \beta_{10} expn\_stu + \varepsilon$$

4.6.3 Fitted model summary:

Adjusted R-squared	0.822
R-squared	0.827
SER	8.328

Predictor	Coefficient	Standard error	t	p-value	95 CI lower bound	95 CI upper bound
intercept	660.3993	10.526	62.742	0.000	639.692	681.106
str	-0.2980	0.340	-0.876	0.382	-0.967	0.372
str_step	4.6257	1.821	2.540	0.012	1.043	8.208
meal_pct	-0.3071	0.045	-6.842	0.000	-0.395	-0.219
avginc	0.9038	0.309	2.928	0.004	0.297	1.511
avginc_power_2	-0.0054	0.005	-1.016	0.310	-0.016	0.005
el_pct	-0.3722	0.086	-4.321	0.000	-0.542	-0.203
elpct_step	0.2545	0.114	2.241	0.026	0.031	0.478
calwpct_log	-2.6851	0.991	-2.709	0.007	-4.635	-0.736
comp_stu	15.6889	7.820	2.006	0.046	0.305	31.072
expn_stu	0.0017	0.001	1.750	0.081	-0.000	0.004

Table 4.6.3

According to the summary table, the estimated relationship is as following

$$\widehat{testscr} = 660.3993 - 0.298str + 4.6257str\_step - 0.3071meal\_pct + 0.9038avginc - 0.0054avginc\_power\_2 - 0.3722el\_pct + 0.2545elpct\_step - 2.6851calwpct\_log + 15.6889comp\_stu + 0.0017expn\_stu$$

4.6.4 Strength of fit:

As $R^2 = 0.827$ and $R^2_{adj} = 0.822$, this model expresses a relatively strong fit to the training data. Other than that, the model also provides $SER = 8.328$. Hence, the model produced by forward selection algorithm has a strong degree of fit for the training data.

4.6.5 Goodness of fit:

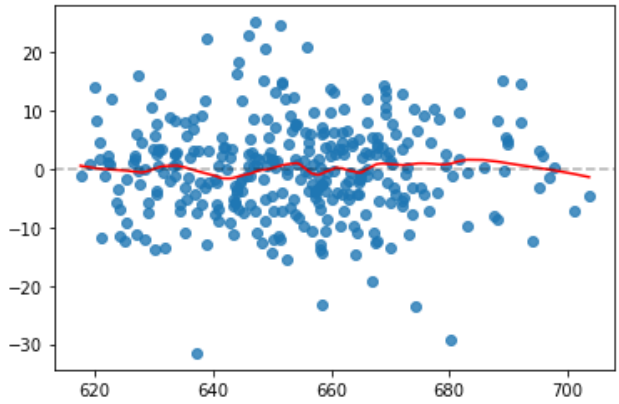

Fig 4.6.5

By observing the LOWESS line of residuals against the fitted values, it can be seen that the line is relatively flat and does not deviate significantly from zero. Other than that, the residuals are observed to be scatter nice and even around zero and do not illustrate any considerable different variances across different fitted values, which means there is no sign of heteroskedasticity. Hence, LSA1, LSA2, and LSA6 are satisfied.

On the other hand, LSA 3 is about whether the data are independent and identically distributed, which should be true for average test score unless students cheat off each other and this should also be true for the predictors included in the model. Besides, since all the original variables are proved to be bounded variables, their transformed variables should also be bounded variables. Thus, fourth moments of the response variable and predictors should be finite, which makes LSA4 satisfied.

4.7 CONDENSED FORWARD SELECTION MODEL:

4.7.1 Proposed population relationship:

Although the model produced by forward selection algorithm provide impressive figures for the measurement of strength of fit, it should be notice that there are 10 predictors in such variable and this may lead to the issue of over reacting to the training data due to the effort of maximizing the fitted. Hence, it is necessary to examine other choices from the forward selection algorithm. Specifically, it is desirable to come up with a model with the least number of predictors that can produce an R^2_{adj} of at least 80%. As a result, by examining the filtering process of the algorithm from the top to the bottom, the following relationship is suggested.

$$testscr = \beta_0 + \beta_1 str + \beta_2 meal\_pct + \beta_3 avginc + \beta_4 el\_pct + \varepsilon$$

Table 4.7.1

Adjusted R-squared	0.809					
R-squared	0.812					
SER	8.616					
Predictor	**Coefficient**	**Standard error**	**t**	**p-value**	**95 CI lower bound**	**95 CI upper bound**
intercept	673.4392	5.995	112.342	0.000	661.647	685.231
str	-0.5264	0.258	-2.039	0.042	-1.034	-0.018
meal_pct	-0.3864	0.031	-12.591	0.000	-0.447	-0.326
avginc	0.7383	0.092	8.064	0.000	0.558	0.918
el_pct	-0.1922	0.035	-5.502	0.000	-0.261	-0.123

4.7.2 Fitted model summary:

According to the summary table, the estimated relationship is as following

$$\widehat{testscr} = 673.4392 - 0.5264 str - 0.3864 meal\_pct + 0.7383 avginc - 0.1922 el\_pct$$

4.7.3 Strength of fit:

As $R^2 = 0.812$ and $R^2_{adj} = 0.809$, this model expresses a relatively strong fit to the training data. Other than that, the model also provides $SER = 8.616$ marks. Hence, the model produced by forward selection algorithm has a strong degree of fit for the training data.

4.7.4 Goodness of fit:

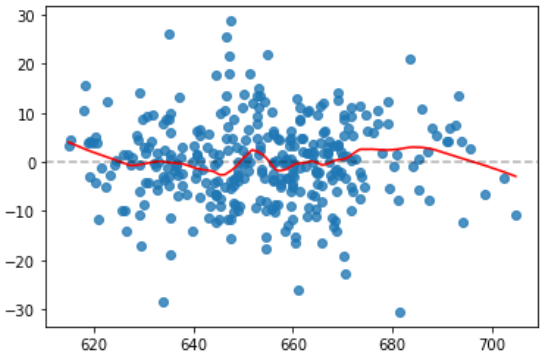

Fig 4.7.4

By observing the LOWESS line of residuals against the fitted values, it can be seen that the line is relatively flat and does not deviate significantly from zero in spite of the tendency of positive residuals as the fitted values range from 670 to 690 marks. Other than that, the residuals are observed to be scatter nice and even around zero and do not illustrate any considerable different variances across different fitted values, which means there is no sign of heteroskedasticity. Hence, LSA1, LSA2, and LSA6 are satisfied.

On the other hand, LSA 3 is about whether the data are independent and identically distributed, which should be true for average test score unless students cheat off each other and this should also be true for the predictors included in the model. Besides, since all the original variables are proved to be bounded variables, their transformed variables should also be bounded variables. Thus, forth moments of the response variable and predictors should be finite, which makes LSA4 satisfied.

4.8 CONDENSED FORWARD SELECTION MODEL WITH INTERACTION EFFECTS:

4.8.1 Proposed population relationship:

As two interaction variables are formed, it is necessary to take into account the different impacts may occur when the same change in percentage of English learners is made across different percentage qualifying for reduced-price lunch and district average income. Hence, the following relationship is suggested.

$$testscr = \beta_0 + \beta_1 str + \beta_2 meal\_pct + \beta_3 avginc + \beta_4 el\_pct + \beta_5 avginc\_elpct + \beta_6 mealpct\_elpct + \varepsilon$$

Adjusted R-squared	0.809					
R-squared	0.813					
SER	8.614					
Predictor	**Coefficient**	**Standard error**	**t**	**p-value**	**95 CI lower bound**	**95 CI upper bound**
intercept	673.6908	6.196	108.726	0.000	661.502	685.880
str	-0.4975	0.259	-1.922	0.055	-1.007	0.012
meal_pct	-0.4026	0.036	-11.240	0.000	-0.473	-0.332
avginc	0.7462	0.109	6.871	0.000	0.533	0.960
el_pct	-0.2844	0.214	-1.330	0.184	-0.705	0.136
mealpct_elpct	0.0015	0.002	0.798	0.425	-0.002	0.005
avginc_elpct	-0.0015	0.008	-0.184	0.854	-0.017	0.014

Table 4.8.1

4.8.2 Fitted model summary:

According to the summary table, the estimated relationship is as following.

$\widehat{testscr} = 673.6908 - 0.4975str - 0.4026meal\_pct + 0.7462avginc - 0.2844el\_pct + 0.0015mealpct\_elpct - 0.0015avginc\_elpct$

4.8.3 Strength of fit:

As $R^2 = 0.813$ and $R^2_{adj} = 0.809$, this model expresses a relatively strong fit to the training data. Other than that, the model also provides $SER = 8.614$ marks. Hence, the model produced by forward selection algorithm has a strong degree of fit for the training data.

4.8.4 Goodness of fit:

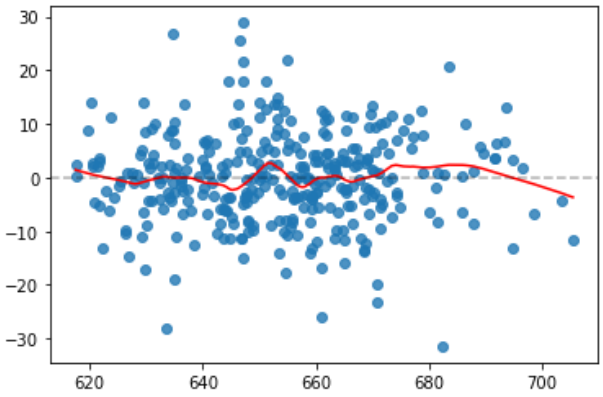

Fig 4.8.4

By observing the LOWESS line of residuals against the fitted values, it can be seen that the line is relatively flat and does not deviate significantly from zero in spite of the tendency of positive residuals as the fitted values range from 670 to 690 marks. Other than that, the residuals are observed to be scatter nice and even around zero and do not illustrate any considerable different variances across different fitted values, which means there is no sign of heteroskedasticity. Hence, LSA1, LSA2, and LSA6 are satisfied.

On the other hand, LSA 3 is about whether the data are independent and identically distributed, which should be true for average test score unless students cheat off each other and this should also be true for the predictors included in the model. Besides, since all the original variables are proved to be bounded variables, their transformed variables should also be bounded variables. Thus, forth moments of the response variable and predictors should be finite, which makes LSA4 satisfied.

4.9 COMPARISON OF FIT AMONG PROPOSED MODELS

Summary of interaction/nonlinear/transformation effects	
Variable name	**Variable meaning**
$avginc\_elpct$	$avginc \times el\_pct$
$mealpct\_elpct$	$meal\_pct \times el\_pct$
$avginc\_power\_2$	$avginc^2$
$calwpct\_log$	$log(calw\_pct + 1)$
str_step	$(str - 23)_+$
$elpct\_step$	$(el\_pct - 20)_+$

Table 4.9.1

Model	Number of predictors	Predictors	SER	Adjusted R-squared	Goodness of fit	Logical predictions	Effect of class size
Multiple linear regression omitted variable bias	6	str, avginc, comp_stu, meal_pct, expn_stu, el_pct	8.525	0.813	Good	Yes	-0.014
Simple linear regression	1	str	19.133	0.059	Good	Yes	-2.5628
Linear spline	2	str, str_step	18.963	0.076	Excellent	No	-3.2073, 10.7049
Quadratic polynomial with interaction	5	str, el_pct, avginc, avginc_power_2, avginc_elpct	10.003	0.743	Moderate	Yes	-0.2656
Quadratic polynomial and log transformation	4	str, avginc, avginc_power_2, calwpct_log	11.308	0.671	Moderate	Yes	-1.0215
Forward selection	10	str, str_step, meal_pct, avginc, avginc_power_2, el_pct, elpct_step, calwpct_log, comp_stu, expn_stu	8.328	0.822	Excellent	No	-0.298, 4.3277
Condensed forward selection	4	str, meal_pct, avginc, el_pct	8.616	0.809	Good	Yes	-0.5264
Condensed forward selection with interactions	6	str, meal_pct, avginc, el_pct, mealpct_elpct, avginc_elpct	8.614	0.809	Good	Yes	-0.4975

Table 4.9.2

Logical predictions: Theoretically, all models should be able to capture part of the variations in the test score at some levels, but it is important that they have to be interpreted in a meaningful way. Among 8 models above, all seem to be quite logical except the linear spline and forward selection models which contain the str_step and elpct_step. Specifically, this transformation effect assumes that the student teacher ratio and percent of English learners should be continuous random variables, which is not logical in real situation. Since number of students, teachers, and people are in integer values, the ratio or percentage should only be presented by some finite values. For example, no irrational numbers should be found on all ratio and percentage sets that contain all possible values. Therefore, in terms of logic, it is reasonable that these two models are not satisfied, and that is the reason there are two exceptions in the 'logical predictions' column.

From the comparison table, it is suggested that the condensed forward selection model should be the optimal one for predicting and understanding test score due to several reasons. Firstly, although the model is quite simple when the total number of predictors included is only 4, the variation it captures

from the data reaches the impressive level of 80.9%. Furthermore, the standard prediction error of the model also stands out as compared to most of the remaining models with larger number of predictors as it can achieve a SER of 8.616 marks. Therefore, the model predicts the dependent variable test score quite well while its interpretation and understandability, without many complicated transformations, is not overly difficult for most users. Besides, as discussed above, the goodness of fit of this model is justified to be relatively good. Finally, the nominated model addresses the parsimony concept as it does not rely on huge number of predictors and transformed variables, which may cause over-fit problem, to capture a significant percent of variation of the data.

4.10 DIAGNOSTICS AND COLLINEARITY ASSESSMENT ON THE OPTIMAL MODEL

4.10.1 Linearity and exogeneity:

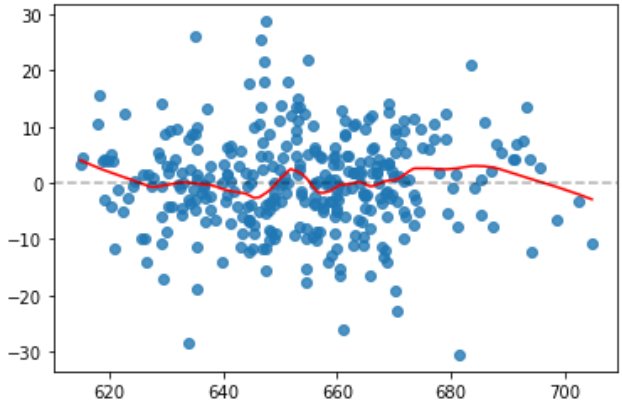

Fig 4.10.1.1

By observing the LOWESS line of residuals against fitted values, there is a tendency that the model may overpredict as the fitted values are either smaller 620 marks or larger 680. Nevertheless, the line does not show any disastrous deviation from zero across most possible value of the mark, which suggests the prediction is relatively accurate.

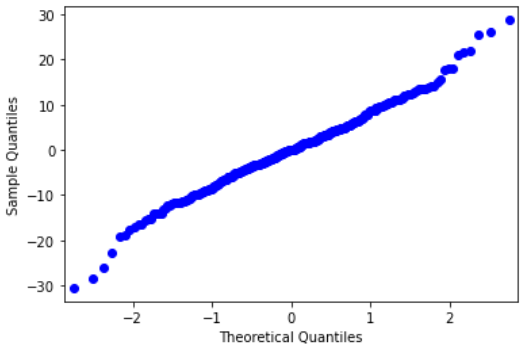

Fig 4.10.1.2

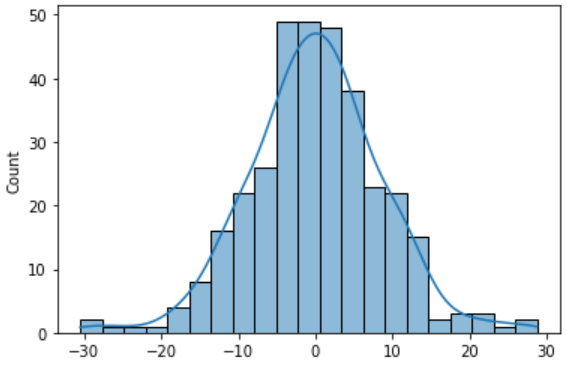

Fig 4.10.1.3

Statistical summary of sample residuals	
Min	-30.5243
Max	28.7165
Mean	-2.2907×10^{-13}
Variance	73.3574
Skewness	-0.0265

Kurtosis	0.9689

Table 4.10.1

Other than that, despite the fact that the QQ plot illustrates that the distribution of the residuals does not perfectly follow normal distribution the histogram of the residuals shows a symmetric distribution around zero, which is indicated more obviously by the mean of residuals of approximately -2.291×10^{-13} marks and the skewness of -0.027, which is around zero.

Hence, the linearity and exogeneity assumptions are well satisfied in general.

4.10.2 Independence and identical distribution of data:

Without any further description of data collecting procedure, the validity of this assumption in this data set cannot be assured. Nevertheless, this assumption may assume to be true for average test score unless students cheat off each other and this should also be true for the predictors included in the model, which are class size, percentage qualifying for reduced-price meal, district average income as well as percentage of English learners. Hence, the assumption of independence and identical distribution of data may be satisfied for this data set.

4.10.3 Existence of fourth moments of response variable and predictors:

Considering average 5th grade reading and math test score, this can reasonable be a bounded variable. Specifically, it is the case since the math score and reading are limited from 0 to a certain upper bound logically which makes the average value of these two variables a bounded random variable. Furthermore, the histogram of average 5th grade reading and math test score shows a symmetric pattern with nearly no outliers, which can be examined more thoroughly via the skewness of 0.171 and the kurtosis of -0.336. Hence, the fourth moment of the average 5th grade reading and math test score can be regarded as a finite one.

Turning to class size, as the total population of a district should be bounded between 0 and a certain upper bound, the number of enrolments in a district should be a bounded random variable. The same reasoning can be also applied to the total number of teachers in a district, which makes the number of students per teacher a bounded random variable. Thus, the fourth moment of class size should be finite.

Regarding district average income, the high kurtosis of 5.7880 of district average income should be monitored cautiously since it is significantly larger than 0, which is an indication of excessive outliers and the possibility of violation of the existence of forth moment. Nevertheless, as income of a person should be a finite quantity, it is reasonable to claim that district average income is a bounded random variable. Thus, the fourth moment of district average income should be finite.

Considering percentage qualifying for reduced-price lunch and percentage of English learners, these random variables bounded between 0 and 100. Thus, the fourth moment of percentage qualifying for reduced-price lunch and percentage of English learners are finite.

Hence, the existence of fourth moments of response variable and predictors are met in this data set.

4.10.4 Constant error variance:

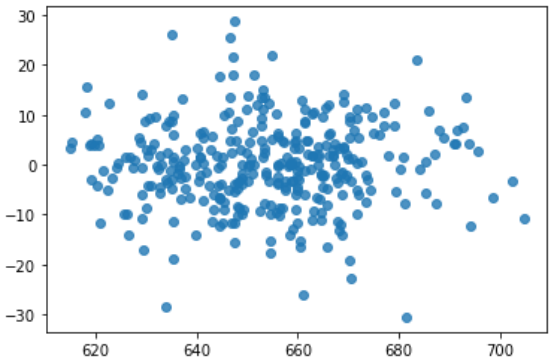

Fig 4.10.4

By observing the scatter plot of residuals against fitted values, the residuals are distributed nicely and evenly around zero with no clear non-linear pattern. Hence, the assumption of constant error variance may be satisfied.

4.10.5 No perfect collinearity:

Variable	str	meal_pct	avginc	el_pct
VIF	1.108	3.284	2.161	1.949

Mean=2.126

Table 4.10.5

While the VIFs for class size, district average income, percentage of English learners are lower than 3 as indicated above, that of percentage qualifying for reduced-price lunch is larger than 3. However, all figures observed are still lower than 5, so there may not be a fair degree of variance inflation and collinearity in the data. Furthermore, the average VIF is below 3, so, again, collinearity should not be a huge issue. Hence, the assumption of no perfect collinearity may be satisfied.

4.11 NONLINEAR EFFECTS IN THE OPTIMAL MODEL

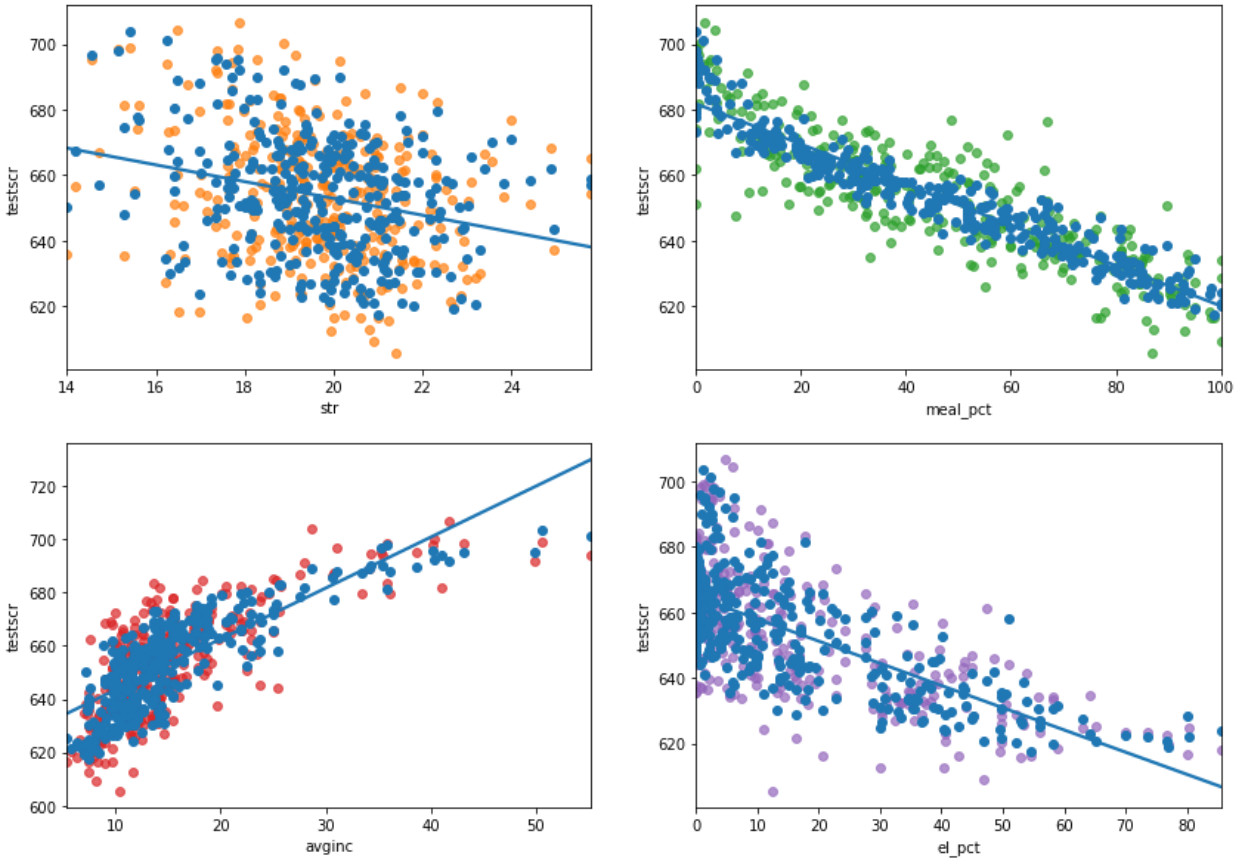

Fig 4.11

By observing the scatterplots of average test score against each predictor, the prediction plots suggest strong effects from percentage qualifying for reduced-price lunch where increasing that by 1% results in a decrease in average test score by 0.3864 marks on average, holding other predictors in model constant and also district average income, which has a positive effect on average test score, where increasing average income by 1 unit leads to an increase in average test score by 0.7383 marks on average, holding other predictors in the model constant. Besides, the impact of percentage of English learners is also significant but not that strong since by increasing this by 1% leads to a decrease of 0.1922 marks in average test score on average.

On the other hand, it can be seen that the effect of percentage qualifying for reduced-price meal and percentage of English learners are relatively linear. Nevertheless, it is possible to observe a diminishing increase in the average test score as district average income increases at different current level despite the fact that this variable is included in the model in linear form. Specifically, this makes sense in logical perspective as one individual cannot be more excellent in learning in the same manner as they become richer at any welfare status. Finally, the effect of class size on the average test score in this model turns out to be significant due to its 95% confident interval does not contain 0.

PART 5

5.1 RESULT SUMMARY

For the first goal, regarding the train data set, by including class size, percentage qualifying for reduced-price lunch, district average income, and percentage of English learners, the model may give desirably highly accurate, despite not perfect, statistics measure without too many variables included that may cause over-fitting. Also, the model does not introduce many transformations and its parsimony is well satisfied, which makes this combination of factors should be the optimal one. Regarding the second goal, as suggested by the comparison table, in case all relevant factors are taken into account to avoid omitted variable bias, the negative relationship between these two considered factors is not significant. Nevertheless, it should be noted that the biased effect of the coefficients does not necessarily impact the precision of predicting capability for test score, considering only 4 predictors in the optimal model might be adequate for somewhat accurate prediction. In that case, the relationship between class size and average test score is evidently a weak negative one. Finally, until now, as the department is awarded some extra state or federal funding, they should focus on either improving the district average income due to its largest absolute value of regression estimated coefficient or reducing percentage qualifying for reduced-price lunch due to its high negative correlation with average test score. However, investment in increasing district average income should be cautiously monitored as this factor has diminishing return effect on the average test score, which can be observed in the scatterplot, so the department might need invest more later if they want to create the same improving progression in average test score.

5.2 PREDICTION OF IMPACT OF CLASS SIZE ON AVERAGE TEST SCORE

5.2.1 Prediction using the condensed forward selection model:

As the condensed forward selection model provides the following estimated relationship.

$$\widehat{testscr} = 673.4392 - 0.5264str - 0.3864meal\_pct + 0.7383avginc - 0.1922el\_pct$$

The estimated impact of reducing the student-teacher ratio by 1 unit is expected to be increasing 0.5264 marks on average. Considering 95% confident interval, as suggested by the regression summary table for this model, the impact is confidently reported to be increasing the average test score by within 0.018 and 1.034 marks. Nevertheless, as discussing above, as this model does not take into account potential variables that can cause omitted bias variable, this measure may be not reliable.

5.2.2 Prediction using the multiple linear regression omitted variable bias:

As the multiple linear regression omitted variable bias model provides the following estimated relationship.

$$\widehat{testscr} = 653.1153 - 0.0140\text{str} + 0.6601\text{avginc} + 16.9687\text{comp\_stu} - 0.4039\text{meal\_pct} + 0.0018\text{expn\_stu} - 0.1785\text{el\_pct}$$

The estimated impact of reducing the student-teacher ratio by 1 unit is expected to be increasing 0.014 marks on average. Considering 95% confident interval, as suggested by the regression summary table for this model, the impact is confidently reported to be ranging from decreasing the average test score by 0.621 marks to increasing the average test score by 0.649 marks. Besides, as this model takes all potential variables that can cause omitted bias variable, the resulting prediction should be more accurate than the previous one.

5.2.3 Prediction using the simple linear regression model:

As the simple linear regression model provides the following estimated relationship.

$$\widehat{testscr} = 704.0954 - 2.5628str$$

Thus, for any one student per teacher decrease in the class size, the corresponding 2.5628 increase will expect to be recognized in the mean of the average 5th grade reading and math test score. This logically makes sense since the smaller the class size of a class, the less crowded the classroom will be, this could lead the class to be quieter and teacher will pay more attention to every individual student in the class and this may lead to higher average 5th grade reading and math test score.

Considering 95% confident interval, as suggested by the regression summary table for this model, the impact is confidently reported to be increasing the average test score by within 1.942 and 3.634 marks. Nevertheless, as discussing above, as this model does not take into account potential variables that can cause omitted bias variable, this measure may be not reliable.

PART 6

6.1 FIVE BEST MODEL SPECIFICATIONS

As indicated above, the condensed forward selection model (model 7) should be the first one to be included due to its achievement of strong fit as well as good fit to the training data in spite of containing a small number of predictors which is four. Similarly, the condensed forward selection model with interactions (model 8) will be considered in this part due to its parsimony and goodness of fit. As the previous options face a risk of the bias estimated coefficient corresponding to class size, the omitted variable bias model (model 2) will be also assessed in this section. Despite having good fit to the training data, the simple linear regression model and the linear spline model have poor strength of fit, which is reflected by their small values of adjusted R squared, the quadratic polynomial with interaction effect model (model 4) and the quadratic polynomial with log transformation model (model 5) will be evaluated instead. The following table shows 4 observation with their 5 predictions from 5 chosen models.

Observation	Testscr	Model 2	Model 4	Model 5	Model 7	Model 8
1	690.799988	684.531377	679.026296	683.921528	679.985515	680.898731
2	640.849976	636.452667	642.352493	634.484960	637.349974	637.083439
3	606.750000	618.747222	618.141702	636.866378	618.263439	620.207860
4	621.750000	631.556576	627.284934	635.326533	632.644410	632.438300

Table 6.1

6.2 SUMMARY OF PREDICTION ACCURACY

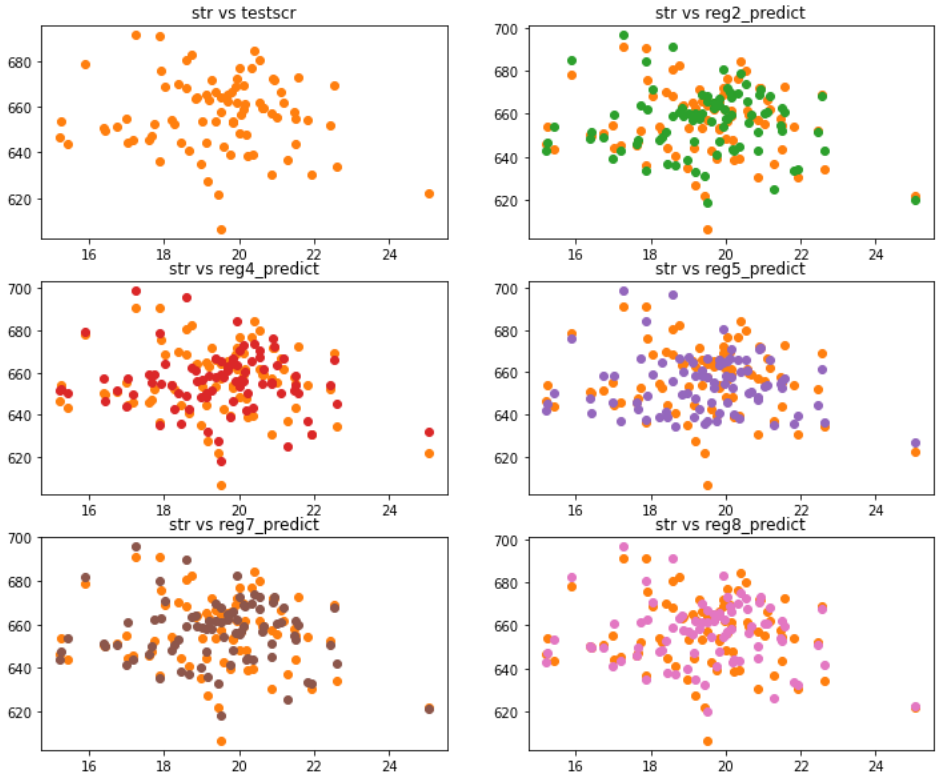

Fig 6.2

By observing the plots of prediction, as being the model covering the omitted variable bias issue, it generates similar predictions. Although two models that capture the nonlinear effect of district average income and percentage qualifying for CALWORKS do admirable and decent work, they cannot get close to the prediction of the omitted variable bias model. The final two models, which are the modified version of forward selection models, have similar prediction capability. Nevertheless, the model with four predictors gives a closer prediction than the other one, which reflects the improvement forecast capability of the condensed forward selection model.

Model	Number of predictors	Predictors	RMSFE	MAFE	Forecast R squared
Omitted variable bias	6	str, avginc, comp_stu, meal_pct, expn_stu, el_pct	7.967	6.335	0.754
Quadratic polynomial with interaction	5	str, el_pct, avginc, avginc_power_2, avginc_elpct	9.831	7.769	0.627
Quadratic polynomial with log transformation	4	str, avginc, avginc_power_2, calwpct_log	11.662	9.147	0.506
Condensed forward selection	4	str, meal_pct, avginc, el_pct	7.784	6.083	0.762
Condensed forward selection with interactions	6	str, meal_pct, avginc, el_pct, mealpct_elpct, avginc_elpct	7.827	6.161	0.760

Table 6.2

According to the above table, as condensed forward selection model is expected for its parsimony, this model turns out to be the best performer among five selected model with the highest forecast R^2 of 76.2% at the same time it obtains the smallest RMSFE and MAFE of 7.784 marks and 6.083 marks respectively. On the other hand, the condensed forward selection with interactions model also has high forecast R^2 of 76.0% and small RMSFE and MAFE of 7.827 marks and 6.161 marks respectively despite not being the optimal one. Although the omitted variable bias model may potentially address the biased estimation of coefficient corresponding to class size, this model turns out to be not as good performer as the previous two models mentioned since biased coefficient does not mean biased prediction.

Finally, two models with quadratic polynomials term have poor figures for RMSFE in spite of relatively moderate figures for forecast R^2. Hence, these two models are the poorest models among five selected models in predicting the test data set.

6.3 RESULTS AND CONCLUSIONS REDISCUSSION

As suggested by the comparison table and the above results analysis, the optimal model in this test data set is still the condensed forward selection model, which provides the smallest RMSFE and MAFE as well as the largest forecast among five models considered. Besides, this model also has the smallest number of predictors among considered models, so not only high accuracy may be ensured but parsimony might be guaranteed. Furthermore, according to the second goal of this study, the relationship between class size and average test score is evidently a weak negative one, and this is further supported by the

significance of class size in the multivariable linear regression model. In the case, beside manipulating the class size to enhance the average test score, the department can consider any change to percentage qualifying for reduced-price meal, district average income, and percentage of English learners to achieve the goal. Nevertheless, investment in increasing district average income should be cautiously monitored as this factor has diminishing return effect on the average test score, which can be observed in the scatterplot, so the department might need invest more later if they want to create the same improving progression in average test score.

PART 7

7. FINAL REPORT

There are three primary goals to achieve in this report, and they have been progressively resolved through different parts. Firstly, regarding the optimal model determination, the report has tried to fit the train data through different models and conclusively chooses the condensed forward selection to be a reasonable choice. Particularly, this model is highly accurate (above 80%) but with only few variables to predict and explain the average test scores. Obviously, the proposed combination of factors may not be as easily interpretable as the simple linear regression or as very accurate as the forward selection model, it is believed that this choice provides a reasonable balanced trade-off between understandability and accuracy. In short, by including class size, percentage qualifying for reduced-price lunch, district average income, and percentage of English learners, the model may give the most desirable statistics measuring the accuracy prediction. Secondly, in terms of relationship between average test score and class size, initial exploratory analysis in part 1 has suggested negative relationship between these two variables, and this has been further supported in subsequent parts with formal statistical testing. Nonetheless, in terms of strength, this negative relationship may be quite weak since even not significant when all factors considered. Hence, overall, it can be concluded that the impact of changes in class size on average test score is not large.

From addressing two first primary goals above, good findings are found to have some recommendations for the Department of Education in California. Specifically, as argued before, without other factors considered except class size and average test score, if money is spent hiring more teacher to reduce class size, the average test score for a district is suggested to increase. However, when all necessary factors are considered, such increment in the average test score is not significant. Hence, such decision to hire additional teachers is not recommended in this report. Instead, following from the optimal model, instead of hiring additional teachers, the Department is encouraged to either attempt to decrease the percentage qualifying for reduced-price lunch or increase the average income level by taking any social welfare enhancement measurements. One probable measurement may be developing and implementing rapid and sustained economic growth policies and programs, in areas such as health, education, nutrition and sanitation, allowing the poor to participate and contribute to the growth. Another one for consideration is investing in and implement agricultural programs which China has done to help 800 million people out of poverty since 1978. It should be noted that this does not mean other variables are not important or neglected. The report just provides reasonable recommendations based on the proposed findings and optimal model. The Department should consider other variables as well where appropriate upon different economic and social contexts in each region.

Finally, for future studies, this report has one limitation: it has not considered the impact of psychological well-being of the student on learning outcome, which has been suggested by psychologists and educationists as a critical determinant of learning ability. Therefore, a good further consideration can be examining the relationship between district average IQ level or crime rate with the average test score since those two variables may be effective predictors for average test score.

APPENDIX:

```
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
import statsmodels.api as sm
lowess = sm.nonparametric.lowess
import statsmodels.formula.api as smf
from statsmodels.stats.anova import anova_lm
from statsmodels.stats.multicomp import pairwise_tukeyhsd
from statsmodels.stats.multicomp import MultiComparison

data=pd.read_excel("C:/Users/Admin/OneDrive - The University of Sydney
(Students)/Desktop/QBUS2810/Assignment 3/caschool.xlsx")

state=500449260+500354333+500080960+490238266

train = data.sample(frac=0.8, random_state=state)
test = data[data.index.isin(train.index)==False].copy()

train=train.reset_index(drop=True)
test=test.reset_index(drop=True)
```

Fig 1.1.1

```
sns.histplot(train['testscr'], kde=False)
```

Fig 1.1.2:

```
ax = sns.boxplot(y='testscr', data=train, palette='Blues', showmeans=True)
```

Table 1.1:

```
stats.describe(train['testscr'])
```

```
train['testscr'].describe()
```

Fig 1.2.1:

```
sns.histplot(train['str'], kde=False)
```

Fig 1.2.2:

```
ax = sns.boxplot(y='str', data=train, palette='Blues', showmeans=True)
```

Table 1.2:

```
stats.describe(train['str'])

train['str'].describe()
```

Fig 1.3:

```
plt.scatter(train['str'],train['testscr'])
```

Table 1.3:

```
train[['testscr','str']].corr()
```

Fig 1.4:

```
variables=['testscr','str','avginc','comp_stu','calw_pct','meal_pct','expn_stu','el_pct']
with sns.axes_style('white'):
    g=sns.pairplot(data[variables], kind='reg',
                   plot_kws={'scatter_kws' :{'color': sns.color_palette('Blues')[-1], 'alpha': 0.4}})
plt.tight_layout()
```

Table 1.4:

```
for i in ['enrl_tot','teachers', 'calw_pct',
'meal_pct','computer','comp_stu','expn_stu','avginc', 'el_pct', 'read_scr','math_scr']:
    print('statistical summary of {}'.format(i))
    print(train[i].describe())

for i in ['enrl_tot','teachers', 'calw_pct',
'meal_pct','computer','comp_stu','expn_stu','avginc', 'el_pct', 'read_scr','math_scr']:
    print('statistical summary of {}'.format(i))
    print(stats.describe(train[i]))
```

Fig 2.1.1:

```
sns.jointplot(x=train['str'], y=train['testscr'], kind="reg")
```

Test 2.1.1:

```
p, p_value = stats.pearsonr(train['str'],train['testscr'])
p, p_value
```

Fig 2.1.2:

```
plt.scatter(train['str'],train['testscr'])
z1 = lowess(train['testscr'],train['str'], frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
```

Table 2.2:

```
reg1 = smf.ols(formula = 'testscr ~ str', data =train).fit()
reg1.summary()
reg1.mse_resid**0.5
```

Test 2.3.1:

```
hypothesis = 'str = 0'
t_test = reg1.t_test(hypothesis)
print(t_test)
```

Fig 2.5.1:

```
sns.regplot(x = 'str', y = reg1.resid, data=train, fit_reg = False)
ax.set_xlabel('str')
ax.set_ylabel('Residuals')
ax.set_title('Predictor vs residuals')
z1 = lowess(reg1.resid,train['str'], frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
```

Fig 2.5.2:

```
sns.histplot(reg1.resid, kde=True)
```

Fig 2.5.3:

```
sm.qqplot(reg1.resid, line='45')
```

Table 2.5.1:

```
stats.describe(reg1.resid)
```

Fig 2.5.4:

```
f, axes = plt.subplots(1, 2)
sns.histplot(train['str'], kde=False, ax=axes[0])
sns.histplot(train['testscr'], kde=False, ax=axes[1])
```

Fig 2.5.5:

```
fig, ax = plt.subplots()
sns.regplot(x='str', y=reg1.resid, data=train, fit_reg=False)
ax.set_xlabel('str')
ax.set_ylabel('Residuals')
ax.set_title('Predictor vs residuals')
```

Table 3.1.1:

```
train[variables].corr()
```

Fig 3.1:

```
fig, ax = plt.subplots()
variables=['testscr','enrl_tot','teachers', 'calw_pct', 'meal_pct', 'computer',
'comp_stu','expn_stu','str','avginc','el_pct']
cmap = sns.diverging_palette(220, 10, as_cmap=True)
sns.heatmap(train[variables].corr(), vmax=1, vmin=-1, center=0, square=True, ax=ax, cmap=cmap)
ax.set_title('Correlation matrix', fontweight='bold', fontsize=13)
plt.tight_layout()
```

Table 3.1.2:

```
p, p_value = stats.pearsonr(train['str'],train['avginc'])
p, p_value
p, p_value = stats.pearsonr(train['testscr'],train['avginc'])
p, p_value
p, p_value = stats.pearsonr(train['str'],train['comp_stu'])
p, p_value
p, p_value = stats.pearsonr(train['testscr'],train['comp_stu'])
p, p_value
p, p_value = stats.pearsonr(train['str'],train['expn_stu'])
p, p_value
p, p_value = stats.pearsonr(train['testscr'],train['expn_stu'])
p, p_value
p, p_value = stats.pearsonr(train['str'],train['meal_pct'])
p, p_value
p, p_value = stats.pearsonr(train['testscr'],train['meal_pct'])
p, p_value
p, p_value = stats.pearsonr(train['str'],train['calw_pct'])
p, p_value
p, p_value = stats.pearsonr(train['testscr'],train['calw_pct'])
p, p_value
p, p_value = stats.pearsonr(train['str'],train['el_pct'])
p, p_value
p, p_value = stats.pearsonr(train['testscr'],train['el_pct'])
p, p_value
```

Table 3.2:

```
reg2 = smf.ols(formula = 'testscr ~ str+avginc+meal_pct+comp_stu+expn_stu+el_pct', data =train).fit()
reg2.summary()
reg2.mse_resid**0.5
```

Test 3.3.1:

```
hypothesis = 'str = 0'
t_test = reg2.t_test(hypothesis)
print(t_test)
```

Fig 3.5.1:

```
sns.regplot(x=reg2.fittedvalues, y=reg2.resid, data=train, fit_reg = False)
ax.set_xlabel('fitted values')
ax.set_ylabel('Residuals')
ax.set_title('fitted values vs residuals')
z1 = lowess(reg2.resid,reg2.fittedvalues, frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
plt.axhline(color='Black', alpha=0.3, linestyle='--')
```

Table 3.5.1:

```
stats.describe(reg2.resid)
```

Table 3.6:

```
features = train[['str','avginc','meal_pct','comp_stu','expn_stu','el_pct']]
features = sm.add_constant(features)
features.head()
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif = []
for i in range(6):
    vif.append(variance_inflation_factor(features.values, i+1))
print(vif)
np.mean(vif)
```

New variable generating:

```
train['avginc_elpct']=train['avginc']*train['el_pct']

train['mealpct_elpct']=train['el_pct']*train['meal_pct']

train['avginc_power_2']=train['avginc']**2

train['calwpct_log']=np.log(train['calw_pct']+1)

train['str_step']=np.where(train['str']>23,1,0)*(train['str']-23)

train['elpct_step']=np.where(train['el_pct']>20,1,0)*(train['el_pct']-20)
```

Table 4.3.3:

```
reg3 = smf.ols(formula = 'testscr ~ str+str_step', data =train).fit()
reg3.summary()
reg3.mse_resid**0.5
```

Fig 4.3.5:

```
sns.regplot(x=reg3.fittedvalues, y=reg3.resid, data=train, fit_reg = False)
ax.set_xlabel('fitted values')
ax.set_ylabel('Residuals')
ax.set_title('fitted values vs residuals')
z1 = lowess(reg3.resid,reg3.fittedvalues, frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
plt.axhline(color='Black', alpha=0.3, linestyle='--')
```

Fig 4.4.1:

```
plt.scatter(train['avginc'],train['testscr'])
z1 = lowess(train['testscr'],train['avginc'], frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
```

Table 4.4.3:

```
reg4 = smf.ols(formula = 'testscr ~ str+el_pct+avginc+avginc_power_2+avginc_elpct', data =train).fit()
reg4.summary()
reg4.mse_resid**0.5
```

Fig 4.4.5:

```
sns.regplot(x=reg4.fittedvalues, y=reg4.resid, data=train, fit_reg = False)
ax.set_xlabel('fitted values')
ax.set_ylabel('Residuals')
ax.set_title('fitted values vs residuals')
z1 = lowess(reg4.resid,reg4.fittedvalues, frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
plt.axhline(color='Black', alpha=0.3, linestyle='--')
```

Fig 4.5.1:

```
fig = plt.figure(figsize=(16,4))
ax1 = fig.add_subplot(1,2,1)
ax2 = fig.add_subplot(1,2,2)

ax1.scatter(train['calw_pct'],train['testscr'])
z1 = lowess(train['testscr'],train['calw_pct'], frac=1./5)
ax1.plot(z1[:,0],z1[:,1],'red')
ax1.set_xlabel('calwpct')
ax1.set_ylabel('testscr')

train['calwpct_log']=np.log(train['calw_pct']+1)

ax2.scatter(train['calwpct_log'],train['testscr'])
z2 = lowess(train['testscr'],train['calwpct_log'], frac=1./5)
ax2.plot(z2[:,0],z2[:,1],'green')
ax2.set_xlabel('calwpct_log')

plt.show()
```

Table 4.5.3:

```
reg5 = smf.ols(formula = 'testscr ~ str+avginc+avginc_power_2+calwpct_log', data =train).fit()
reg5.summary()
reg5.mse_resid**0.5
```

Fig 4.5.5:

```
sns.regplot(x=reg5.fittedvalues, y=reg5.resid, data=train, fit_reg = False)
ax.set_xlabel('fitted values')
ax.set_ylabel('Residuals')
ax.set_title('fitted values vs residuals')
z1 = lowess(reg5.resid,reg5.fittedvalues, frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
plt.axhline(color='Black', alpha=0.3, linestyle='--')
```

Forward selection:

```
variable_set=['testscr','str','avginc','expn_stu','el_pct','comp_stu','calw_pct','meal_pct','str_step','
elpct_step','elpct_step','avginc_power_2','calwpct_log','avginc_elpct','mealpct_elpct']

def forward_selected(data, response, nominated = []):
    remaining = set(data.columns)
    selected = nominated
    remaining.remove(response)
    remaining ^= set(selected)
```

```
    current_score, best_new_score = 0.0, 0.0
    if nominated:
        formula = "{} ~ {} + 1".format(response, ' + '.join(nominated))
        current_score= smf.ols(formula, data).fit().rsquared_adj
        best_new_score = current_score
        print("you nominated variable(s) %s, the adj_r2 is: %f" %(nominated, current_score))
    while remaining and current_score == best_new_score:
        scores_with_candidates = []
        for candidate in remaining:
            formula = "{} ~ {} + 1".format(response,
                                           ' + '.join(selected + [candidate]))
            score = smf.ols(formula, data).fit().rsquared_adj
            scores_with_candidates.append((score, candidate))
        scores_with_candidates.sort()
        best_new_score, best_candidate = scores_with_candidates.pop()
        if current_score < best_new_score:
            print("adding %s increases adj_r2 from %f to %f" %(best_candidate, current_score,
best_new_score))
            remaining.remove(best_candidate)
            selected.append(best_candidate)
            current_score = best_new_score
    formula = "{} ~ {} + 1".format(response,
                                   ' + '.join(selected))
    model = smf.ols(formula, data).fit()
    print("final model is %s, with adj_r2 of %f" %(formula, model.rsquared_adj))
    return model

forward_selected(train[variable_set], 'testscr', ['str'])
```

Table 4.6.3:

```
reg6 = smf.ols(formula = 'testscr ~ str + meal_pct + avginc + el_pct + calwpct_log + elpct_step +
str_step + comp_stu + expn_stu + avginc_power_2', data =train).fit()
reg6.summary()
reg6.mse_resid**0.5
```

Fig 4.6.5:

```
sns.regplot(x=reg6.fittedvalues, y=reg6.resid, data=train, fit_reg = False)
ax.set_xlabel('fitted values')
ax.set_ylabel('Residuals')
ax.set_title('fitted values vs residuals')
z1 = lowess(reg6.resid,reg6.fittedvalues, frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
plt.axhline(color='Black', alpha=0.3, linestyle='--')
```

Table 4.7.1:

```
reg7 = smf.ols(formula = 'testscr ~ str+meal_pct+avginc+el_pct', data =train).fit()
reg7.summary()
reg7.mse_resid**0.5
```

Fig 4.7.4:

```
sns.regplot(x=reg7.fittedvalues, y=reg7.resid, data=train, fit_reg = False)
ax.set_xlabel('fitted values')
ax.set_ylabel('Residuals')
ax.set_title('fitted values vs residuals')
z1 = lowess(reg7.resid,reg7.fittedvalues, frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
plt.axhline(color='Black', alpha=0.3, linestyle='--')
```

Table 4.8.1:

```
reg8 = smf.ols(formula = 'testscr ~ str+meal_pct+avginc+el_pct+mealpct_elpct+avginc_elpct', data
=train).fit()
reg8.summary()
```

```
reg8.mse_resid**0.5
```

Fig 4.8.4:

```
sns.regplot(x=reg8.fittedvalues, y=reg8.resid, data=train, fit_reg = False)
ax.set_xlabel('fitted values')
ax.set_ylabel('Residuals')
ax.set_title('fitted values vs residuals')
z1 = lowess(reg8.resid,reg8.fittedvalues, frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
plt.axhline(color='Black', alpha=0.3, linestyle='--')
```

Fig 4.10.1.1:

```
sns.regplot(x=reg7.fittedvalues, y=reg7.resid, data=train, fit_reg = False)
ax.set_xlabel('fitted values')
ax.set_ylabel('Residuals')
ax.set_title('fitted values vs residuals')
z1 = lowess(reg7.resid,reg7.fittedvalues, frac=1./5)
plt.plot(z1[:,0],z1[:,1],'red')
plt.axhline(color='Black', alpha=0.3, linestyle='--')
```

Fig 4.10.1.2:

```
sm.qqplot(reg7.resid)
```

Fig 4.10.1.3:

```
sns.histplot(reg7.resid, kde=True)
```

Table 4.10.1:

```
stats.describe(reg7.resid)
```

Fig 4.10.4:

```
sns.regplot(x=reg7.fittedvalues, y=reg7.resid, data=train, fit_reg = False)
ax.set_xlabel('fitted values')
ax.set_ylabel('Residuals')
ax.set_title('fitted values vs residuals')
```

Table 4.10.5:

```
features = train[['str','meal_pct','avginc','el_pct']]
features = sm.add_constant(features)
features.head()
from statsmodels.stats.outliers_influence import variance_inflation_factor
vif = []
for i in range(4):
    vif.append(variance_inflation_factor(features.values, i+1))
print(vif)
np.mean(vif)
```

Fig 4.11:

```
tableau=['#1F77B4', '#FF7F0E', '#2CA02C', '#DB2728', '#9467BD', '#8c564b', '#e377c2', '#7f7f7f']
fig = plt.figure(figsize=(14,10))
ax1 = fig.add_subplot(2,2,1)
ax2 = fig.add_subplot(2,2,2)
ax3 = fig.add_subplot(2,2,3)
ax4 = fig.add_subplot(2,2,4)

sns.regplot(train['str'], train['testscr'], ci=None, scatter_kws={'s': 35, 'color': tableau[1], 'alpha':
0.7}, ax=ax1)
ax1.scatter(train['str'], reg6.fittedvalues)
```

```
sns.regplot(train['meal_pct'], train['testscr'], ci=None, scatter_kws={'s': 35, 'color': tableau[2],
'alpha': 0.7},ax=ax2)
ax2.scatter(train['meal_pct'], reg6.fittedvalues)

sns.regplot(train['avginc'], train['testscr'], ci=None, scatter_kws={'s': 35, 'color': tableau[3],
'alpha': 0.7},ax=ax3)
ax3.scatter(train['avginc'], reg6.fittedvalues)

sns.regplot(train['el_pct'], train['testscr'], ci=None, scatter_kws={'s': 35, 'color': tableau[4],
'alpha': 0.7},ax=ax4)
ax4.scatter(train['el_pct'], reg6.fittedvalues)
```

New variable generating:

```
test['avginc_elpct']=test['avginc']*test['el_pct']

test['mealpct_elpct']=test['el_pct']*test['meal_pct']

test['avginc_power_2']=test['avginc']**2

test['calwpct_log']=np.log(test['calw_pct']+1)

test['str_step']=np.where(test['str']>23,1,0)*(test['str']-23)

test['elpct_step']=np.where(test['el_pct']>20,1,0)*(test['el_pct']-20)
```

Predictions generating:

```
reg2_predict=reg2.predict({'str':test['str'],'avginc': test['avginc'],'meal_pct':
test['meal_pct'],'comp_stu': test['comp_stu'],'expn_stu': test['expn_stu'],'el_pct': test['el_pct']})

reg4_predict=reg4.predict({'str': test['str'],'el_pct': test['el_pct'],'avginc':
test['avginc'],'avginc_power_2': test['avginc_power_2'],'avginc_elpct':test['avginc_elpct']})

reg5_predict=reg5.predict({'str': test['str'],'avginc': test['avginc'],'avginc_power_2':
test['avginc_power_2'],'calwpct_log': test['calwpct_log']})

reg7_predict=reg7.predict({'str':test['str'],'meal_pct':test['meal_pct'],'avginc':
test['avginc'],'el_pct':test['el_pct']})

reg8_predict=reg8.predict({'str':test['str'],'meal_pct':test['meal_pct'],'avginc':
test['avginc'],'el_pct':test['el_pct'],'mealpct_elpct':test['mealpct_elpct'],'avginc_elpct':
test['avginc_elpct']})
```

Table 6.1

```
test['reg2_predict']=reg2_predict
test['reg4_predict']=reg4_predict
test['reg5_predict']=reg5_predict
test['reg7_predict']=reg7_predict
test['reg8_predict']=reg8_predict
test[['testscr','reg2_predict','reg4_predict','reg5_predict','reg7_predict','reg8_predict']].head(4)
```

Fig 6.2:

```
tableau=['#1F77B4', '#FF7F0E', '#2CA02C', '#DB2728', '#9467BD', '#8c564b', '#e377c2', '#7f7f7f']
fig = plt.figure(figsize=(12,10))
ax1 = fig.add_subplot(3,2,1)
ax2 = fig.add_subplot(3,2,2)
ax3 = fig.add_subplot(3,2,3)
ax4 = fig.add_subplot(3,2,4)
ax5 = fig.add_subplot(3,2,5)
ax6 = fig.add_subplot(3,2,6)

ax1.scatter(test['str'], test['testscr'], color=tableau[1])
ax1.set(title='str vs testscr')
```

```
ax2.scatter(test['str'], test['testscr'], color=tableau[1])
ax2.scatter(test['str'], reg2_predict , color=tableau[2])
ax2.set(title='str vs reg2_predict')

ax3.scatter(test['str'], test['testscr'], color=tableau[1])
ax3.scatter(test['str'], reg4_predict, color=tableau[3])
ax3.set(title='str vs reg4_predict')

ax4.scatter(test['str'], test['testscr'], color=tableau[1])
ax4.scatter(test['str'], reg5_predict, color=tableau[4])
ax4.set(title='str vs reg5_predict')

ax5.scatter(test['str'], test['testscr'], color=tableau[1])
ax5.scatter(test['str'], reg7_predict , color=tableau[5])
ax5.set(title='str vs reg7_predict')

ax6.scatter(test['str'], test['testscr'], color=tableau[1])
ax6.scatter(test['str'], reg8_predict, color=tableau[6])
ax6.set(title='str vs reg8_predict')
plt.show()
```

Table 6.2

```
RMSFE2=(sum((test['testscr']-reg2_predict)**2)/len(reg2_predict))**0.5
MAD2=sum(abs(test['testscr']-reg2_predict))/len(reg2_predict)
ols2 = smf.ols(formula='testscr~reg2_predict', data=test).fit()
ols2.summary()
Forecast_R_squared_2= 0.754
print(RMSFE2, MAD2, Forecast_R_squared_2)

RMSFE4=(sum((test['testscr']-reg4_predict)**2)/len(reg4_predict))**0.5
MAD4=sum(abs(test['testscr']-reg4_predict))/len(reg4_predict)
ols4 = smf.ols(formula='testscr~reg4_predict', data=test).fit()
ols4.summary()
Forecast_R_squared_4= 0.627
print(RMSFE4, MAD4, Forecast_R_squared_4)

RMSFE5=(sum((test['testscr']-reg5_predict)**2)/len(reg5_predict))**0.5
MAD5=sum(abs(test['testscr']-reg5_predict))/len(reg5_predict)
ols5 = smf.ols(formula='testscr~reg5_predict', data=test).fit()
ols5.summary()
Forecast_R_squared_5= 0.506
print(RMSFE5, MAD5, Forecast_R_squared_5)

RMSFE7=(sum((test['testscr']-reg7_predict)**2)/len(reg7_predict))**0.5
MAD7=sum(abs(test['testscr']-reg7_predict))/len(reg7_predict)
ols7 = smf.ols(formula='testscr~reg7_predict', data=test).fit()
ols7.summary()
Forecast_R_squared_7=0.762
print(RMSFE7, MAD7, Forecast_R_squared_7)

RMSFE8=(sum((test['testscr']-reg8_predict)**2)/len(reg8_predict))**0.5
MAD8=sum(abs(test['testscr']-reg8_predict))/len(reg8_predict)
ols8 = smf.ols(formula='testscr~reg8_predict', data=test).fit()
ols8.summary()
Forecast_R_squared_8=0.760
print(RMSFE8, MAD8, Forecast_R_squared_8)
```