THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Evaluation of Education Input on Primary School Students

by
Meichen Qian 45331723

School of Information Technology and Electrical Engineering,
The University of Queensland.

Submitted for Master of Data Science Capstone
January 14, 2021

Abstract

The influential factors of education output have been studied for many years. However, not much research has been done on quantitatively assessing the relationship between school tuition fees and student's academic achievement, i.e. students' grades. Thanks to Biarri, we have the data of student's average NAPLAN grade of all schools in Australia, which enables us to quantitatively analyze this problem. In this project, My partner Nicholas and I were aimed to quantitatively model the relation between education input factors including tuition fees and student's NAPLAN results using multiple regression. The result of the project shows that tuition fee is a significant factor in the model of predicting student NAPLAN result, together with other factors like school indigenous enrollment, higher education degree ratio, etc. A higher tuition fee is correlated to a higher NAPLAN score. However, the coefficient of tuition fee is very small, floating at a scale of 0.001, which was too small to be important. Therefore we concluded that the tuition fee has a significant but not great enough effect on student's academic achievement.

Contents

1 Introduction

Whether paying for the extra tuition fee for going to private school really matters in students' academic outcomes has been debating for a long time. In this research, we were aimed to find out what a dollar of education is worth, namely the relationship between the education input and student's academic performance. To quantitatively measure this problem, we use school tuition fee to quantify the education input, and the student grade to model the academic performance, more specifically, the NAPLAN score of students in primary school. As the academic performance of students is influenced by multiple factors instead of one, we use multiple regression to model most of the influential factors, with school fees as one of them. By looking at the characteristics of the variable tuition fee, we speculate the relationship between tuition fees and students' NAPLAN scores and thus conclude with what a dollar of education is worth.

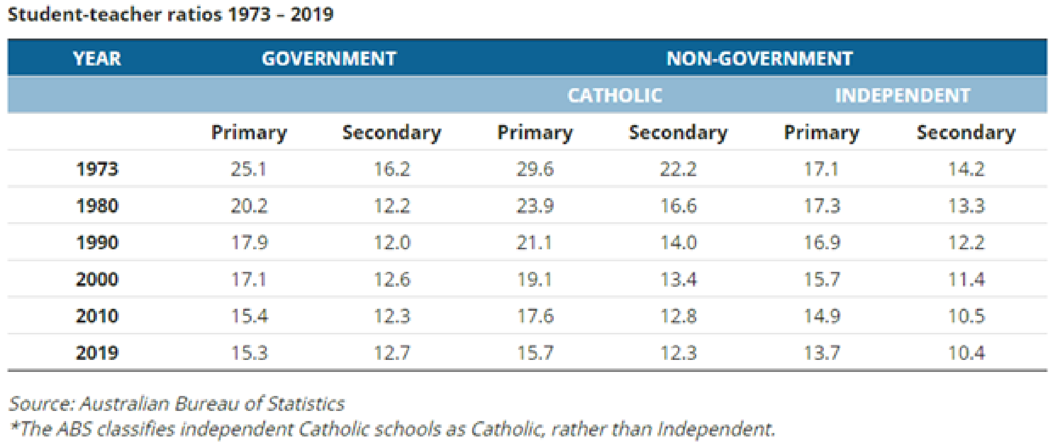
Ethical and privacy statement:

The datasets used in this project are NAPLAN score data, census data from the 2016 Australian Statistical Bureau, and tuition fee data of schools in Melbourne which was collected by My partner Nicholas and me on the official website of schools. The ethical problem was considered in three aspects. First, in the data collection step, the data collection of school fees was collected by hand on publically available data, and the NAPLAN data we are using are aggregated by the school, so no personal information is collected or used in this project. Second, data analysis considers appropriate assumptions and testing. The result was built with the supervision of mentors and tried to be reliable and not misleading. Third, the security of data was considered. The data was processed only on personal computers locally, and all shared raw data or intermediate results were deleted in time.

2 Project Review

2.1 Prior works on factors affecting education

The study on influential factors for education output has been on for years. The location of school would influence the education output where a school in rural areas has fewer study options("National Regional, Rural and Remote Education Strategy", 2018), and students usually need extra help in studying(Lamb, 2020). Comparing with locations, the socioeco-



Student-teacher ratios 1973 – 2019

YEAR	GOVERNMENT		NON-GOVERNMENT			
			CATHOLIC		INDEPENDENT	
	Primary	Secondary	Primary	Secondary	Primary	Secondary
1973	25.1	16.2	29.6	22.2	17.1	14.2
1980	20.2	12.2	23.9	16.6	17.3	13.3
1990	17.9	12.0	21.1	14.0	16.9	12.2
2000	17.1	12.6	19.1	13.4	15.7	11.4
2010	15.4	12.3	17.6	12.8	14.9	10.5
2019	15.3	12.7	15.7	12.3	13.7	10.4

Source: Australian Bureau of Statistics
**The ABS classifies independent Catholic schools as Catholic, rather than Independent.*

Figure 2.1: A comparison of student to teacher ratio in three different kinds of schools.

nomic status of the family can be a more important influential factor(Davis-Kean, 2005). Even if students come from different regions, their study outcome can be similar when keeping their social background to be consistent(Mc- Niece Jolliffe, 1998). Family with a higher educational background would have a positive effect on an individual's study(Chesters Watson, 2013). To better understand the problem of education in Australia, we then discuss the education system in Australia.

2.2 Education system background

2.2.1 School types

There are three types of schools in Australia, government school, catholic school, and independent school. Government schools do not charge students tuition fees. They operate on funding from the government and therefore provide free education to every kid in Australia. The government sector has the largest amount of schools in Australia. Independent schools, however, charge students a lot. For example, the tuition fees of an independent primary school in the Melbourne area is usually around 10,000 dollars per year, according to our project. Some schools also have strict entrance examinations to select students with higher entrance grades. Catholic schools charge less compared to independent schools. The higher tuition is matched by better learning conditions. The student to teacher ratio in independent schools are larger than the other two, meaning students in independent school may get more attention comparing to the other two kinds of schools. More info can be found in the table 2.1.

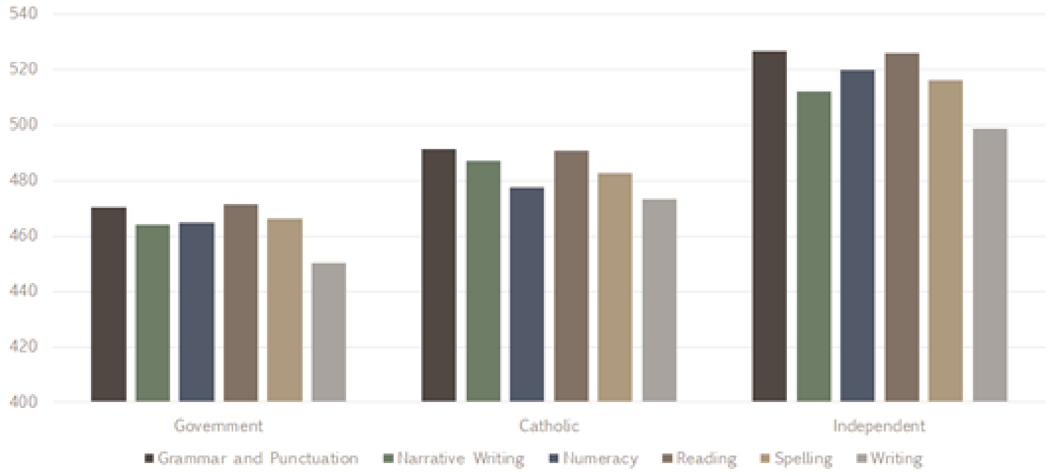


Figure 2.2: The average NAPLAN grade of students in three different school types. Students from independent schools have the highest average NAPLAN grade, which is about 60 points higher than the average grade for Government schools.

2.2.2 NAPLAN and ICSEA Score

NAPLAN is an annual assessment on Literacy and Numeracy for students in Years 3, 5, 7, and 9. It is a nationwide test, and is consistent between different schools, therefore becoming a good evaluation for children's development in Australia.

ICSEA, also know as the Index of Community Socio-Educational Advantage reflects on the level of the school's educational advantage. It is not dependent on the school itself. It evaluates the students' Socio-Educational background like parents' occupation, as well as the school's geographical location, etc. Based on that, it gives a score for each school on its educational advantage. (ICSEA, 2014)

Simple exploratory data analysis on ICSEA and NAPLAN is shown below. From figure 2.2, we can see that the NAPLAN score for independent schools is the highest, at the same time, its ICSEA score is also the most advanced, as shown in 2.3. This correlation between the ICSEA score and student average naplan grade motivates us to research on the relationship between student grades and ICSEA factors like school location, family economical background, etc.

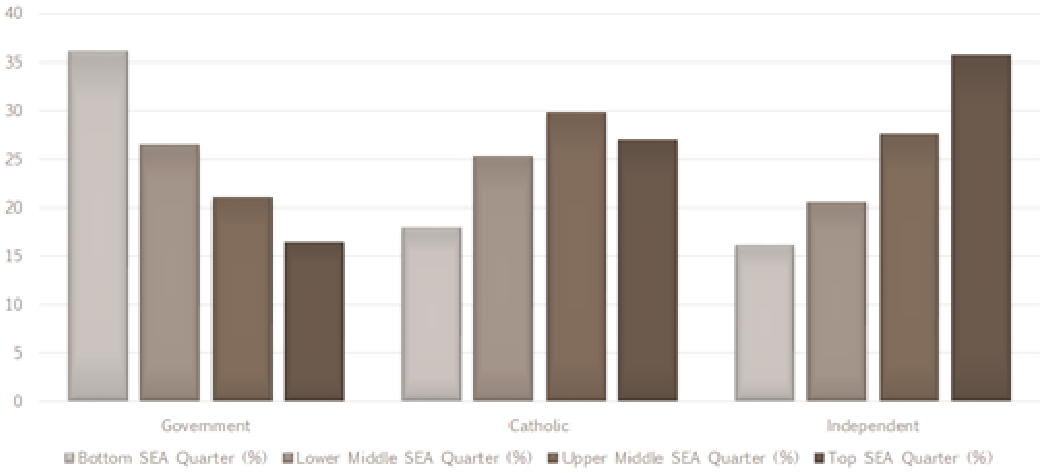


Figure 2.3: The socio-economical background of students in different kinds of schools. The data set divides students' socio-educational background into 4 quarters, while the top quarter has the best socio-educational background, the bottom one has the worst background. From the graph we can see that students in independent school are mostly coming from the top SEA Quarter, while government school students are mostly coming from the bottom Quarter.

2.3 Literature review: regression

2.3.1 Multiple linear regression

Linear regression is a commonly used method to model the linear relationship between explanatory variables and response variables. When involving multiple explanatory variables, the process is also called multiple linear regression. If we denote the response variable by Y and the explanatory variable by X_1, X_2,...,X_k, then an important class of linear model is

$$\phi(x_1, x_2, ..., x_k) = \beta_0 + \beta_1 x_1 + ... + \beta_K x_k \tag{1}$$

, which is linear in the parameters β_j (Seber & Lee, 2003). The multiple linear regression is based on several assumptions.

- There is a linear relationship between the independent variables and the dependent variable.
- The independent variables are not highly correlated with each other.
- The residuals of the model should be roughly normally distributed with a mean of 0 and the variance of it should be constant with the change of independent variables.

2.3.2 Stepwise feature selection

The stepwise feature selection is a method for feature selection that uses a sequence of steps to allow features to leave or enter a model one at a time. Usually, the feature selection is based on the p-value threshold. In each step, if the p-value of a feature is larger than the threshold, this feature would leave the model. If the p-value of a feature is smaller than the threshold, it would enter the model (Kuhn, and Johnson). This process would finally converge to a set of variables.

2.3.3 Regularized Regression

There are two types of regularized regression used in this project. One is Ridge regression and the other is lasso regression. These two regressions help to reduce the model complexity and prevent over-fitting of the model. (Hoerl Kennard, 1970) Consider the standard model of ordinary least squares (OLS) for multiple linear regression

$$Y = X\beta + \varepsilon \tag{2}$$

where $y \in R^n$, $\beta \in R^p$, and $X \in R^{nxp}$. We can expand this to $y_i = \sum_{j=1}^{p} \beta_i X_{ij} + \varepsilon_i, \forall i = 0, 1, ..., n$. Here β_j are non-random unknown parameters, X_{ij} are non-random and observable, and ε_i are random so y_i are random. This is a standard model of multiple linear regression. However the model has a common issue that it has the tendency to overfit the data when there is too much noise caused by correlated variables. In order to reduce the effect, Regularized methods are introduced, which introduces a penalty term on model complexity.

Ridge regression alters the cost function by adding a penalty equivalent to the square of the magnitude of the coefficients, which is shown as below:

$$\hat{\beta}^{\mathbf{ridge}} = \mathbf{arg}\min_{\beta} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - (\beta_{\mathbf{0}} + \beta^{\mathbf{T}} \mathbf{x_i}))^{\mathbf{2}} + \lambda \|\beta\|_{\mathbf{2}}^{\mathbf{2}} \tag{3}$$

Lasso regression is similar to ridge regression, but the penalty becomes the absolute value of the magnitude of the coefficients. The expression of lasso regression is as follows:

$$\hat{\beta}^{\mathbf{ridge}} = \mathbf{arg}\min_{\beta} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{y_i} - (\beta_{\mathbf{0}} + \beta^{\mathbf{T}} \mathbf{x_i}))^{\mathbf{2}} + \lambda \|\beta\|_{\mathbf{1}} \tag{4}$$

2.3.4 Grid Search

Grid search is a method for selecting the best combination of parameters. With a predefined parameter list, grid search iterates through all combinations of possible parameter values, calculates the error of each model, and chooses the one with the lowest error.

3 Data preparation and processing

3.1 NAPLAN data

The NAPLAN data set is the core data set of this industrial project. Provided by our project supervised company Biarri, the NAPLAN data set has two parts, one is the NAPLAN results of schools in Australia and the other is a school information set.

The NAPLAN result data set contains the NAPLAN results from 2008-2018 years of all schools. It is a large table data set, with 1032397 records (rows) and 38 attributes (columns). The attributes contain the year of the record, the school ID, school name, school address, domain, Student grade level, average NAPLAN score, and band percentage, etc. There are four student grade levels, which are grade 3, 5, 7, and 9. In this project, only grade levels 3 and 5 are used as a research sample to simplify the problem.

The School information set contains the information of schools like the year of the record, the school ID, school name, school address, school sector, school ICSEA score, school enrollment, number of teachers and staff, etc.

3.2 Census data

The census data comes from Australia's statistical collection undertaken by the Australian Bureau of Statistics (ABS) in 2016. The census data is a series of data information related to areas. Organized by ABS, the data information has 5 different profiles, including general community profile, indigenous profile, Place of enumeration, Time series profile, and working population profile. Each profile contains tables of data in it, for example, the general community profile has 59 tables in it. Each table contains different attributes, for example, the table "Indigenous Status by Age by Sex" contains 180 attributes, describing the indigenous population of different age groups and gender types.

The Australian Statistical Geography Standard (ASGS) provides a framework of statistical areas used by the Australian Bureau of Statistics (ABS), as shown in the graph 3.1. The largest scale of statistical area is Australia as a whole and then dividing Australia into smaller and smaller regions until reaching Mesh Blocks. In this project, we are using the data on the Statistical area 2 level, which fits with the address information of schools in the NAPLAN data set.

The target region in this project is Melbourne, so only the areas in Melbourne are considered. At the statistical area 4 level, the Melbourne areas are:

No.	SA4 – area
1	Melbourne - Inner
2	Melbourne - Inner East
3	Melbourne - North West
4	Melbourne - North East
5	Melbourne - Outer East
6	Melbourne - South East
7	Melbourne - West
8	Mornington Penninsula
9	Melbourne - Inner South

Using the regions decided on SA 4 level, we can get a list of region names on SA 2 level, and then select the target schools(which are schools in Melbourne) from the NAPLAN school list using region name as a filter.

3.3 Tuition fee data

The tuition fee data is a large table dataset, which contains the school name, location, and tuition fee of the school in 2020 for year 3 and 5 students. The data we aimed for was the tuition fee in 2020 of year 3 and 5 students in the Melbourne region. However, we cannot get the list of schools in Melbourne directly from the NAPLAN school data set. That is because, the NAPLAN school dataset only has information on schools on S/T (state/territory) level and SA2 level, while we cannot distinguish whether a school is in Melbourne or not using this information. The census data comes in at this point. We use the Melbourne area obtained at SA4 level, and find all SA2 levels in them. We then join the Melbourne SA2 regions with NAPLAN school location information and get the list of schools in Melbourne.

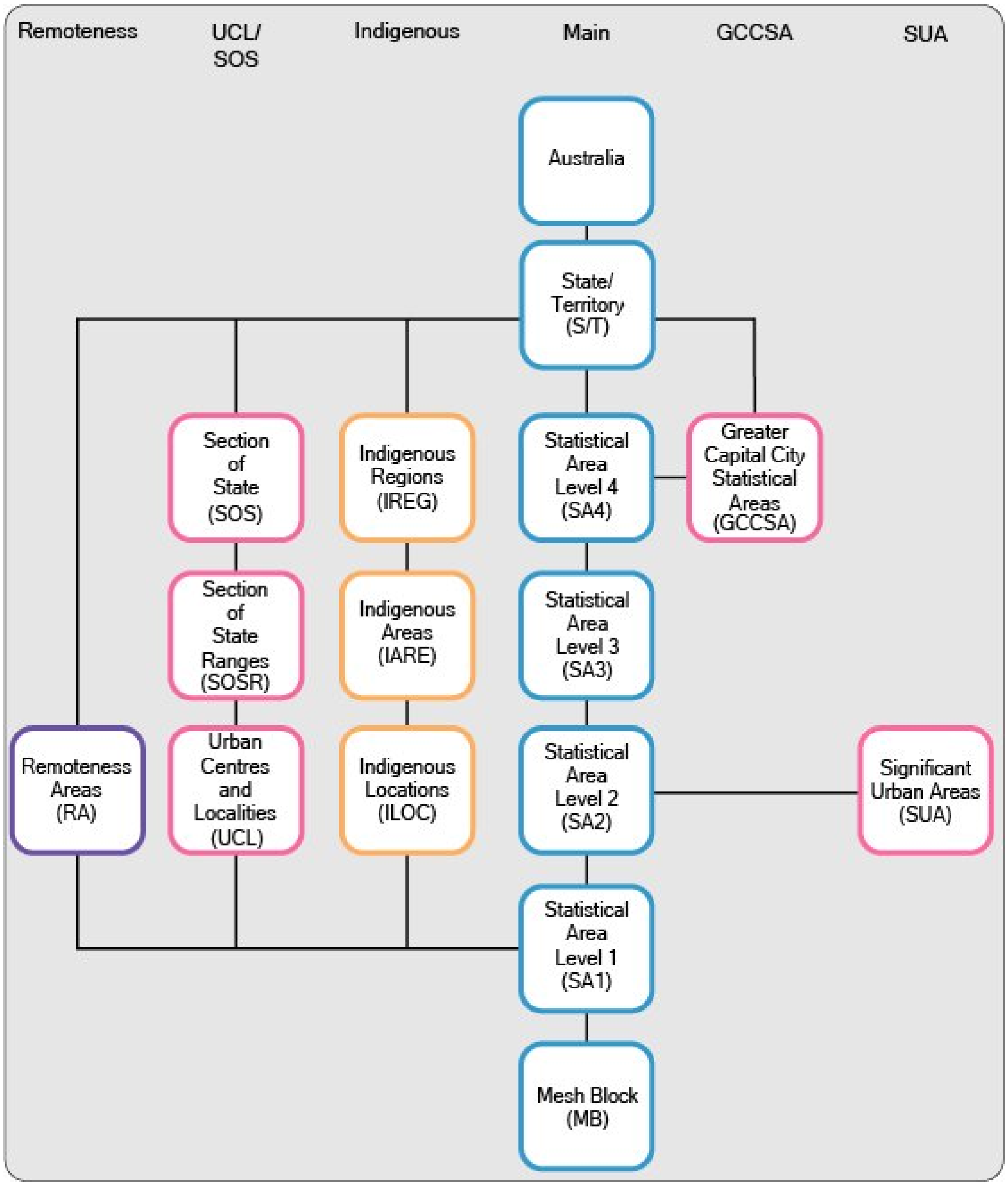
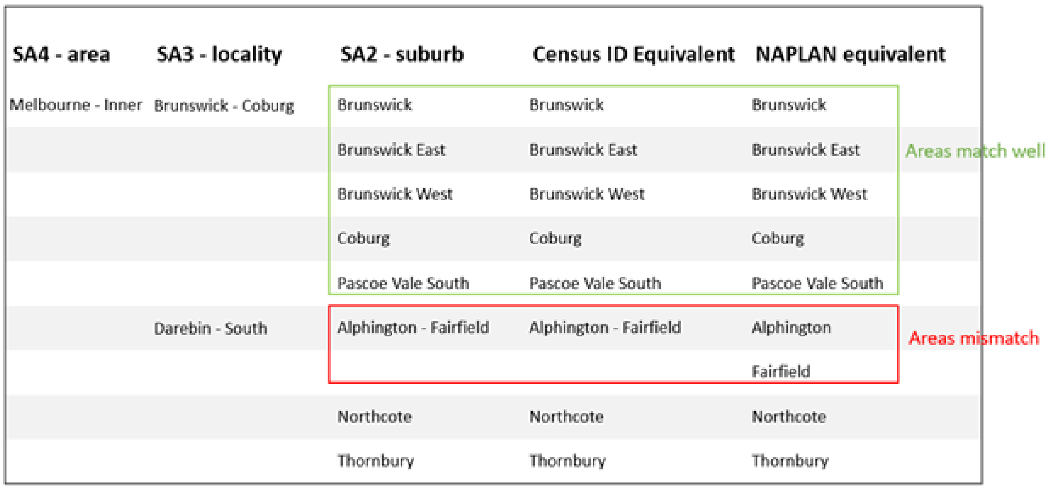

Remoteness
UCL/ SOS
Indigenous
Main
GCCSA
SUA
Australia
State/ Territory (S/T)
Section of State (SOS)
Indigenous Regions (IREG)
Statistical Area Level 4 (SA4)
Greater Capital City Statistical Areas (GCCSA)
Section of State Ranges (SOSR)
Indigenous Areas (IARE)
Statistical Area Level 3 (SA3)
Remoteness Areas (RA)
Urban Centres and Localities (UCL)
Indigenous Locations (ILOC)
Statistical Area Level 2 (SA2)
Significant Urban Areas (SUA)
Statistical Area Level 1 (SA1)
Mesh Block (MB)

Figure 3.1: The graph shows the ASGS framework of statistical areas. From the graph we can see that the main scale divides Australia into S/T, and then at SA4 level divide into Capital city areas, from which we can tell the area coverage of Melbourne. In this project we are focusing on the blue Main list of statistical areas, especially the statistical area level 2, because at this level the location name corresponds to the location of schools in NAPLAN dataset.(ABS, 2020)



SA4 - area	SA3 - locality	SA2 - suburb	Census ID Equivalent	NAPLAN equivalent
Melbourne - Inner	Brunswick - Coburg	Brunswick	Brunswick	Brunswick
		Brunswick East	Brunswick East	Brunswick East
		Brunswick West	Brunswick West	Brunswick West
		Coburg	Coburg	Coburg
		Pascoe Vale South	Pascoe Vale South	Pascoe Vale South
	Darebin - South	Alphington - Fairfield	Alphington - Fairfield	Alphington
				Fairfield
		Northcote	Northcote	Northcote
		Thornbury	Thornbury	Thornbury

Figure 3.2: The mismatch problem happens when we try to match the location in NAPLAN data and census data. Because of that, the match was then done by hand, in order to merge the data from NAPLAN and Census dataset for further analysis.

In most of the cases, the SA2 suburb name from census data matches directly with the NAPLAN school location. However, there were also mismatches in this process. For example in the table 3.2, in census data, Alphington-Fairfield is an SA2 area. However, in the NAPLAN school list, the location Alphington and Fairfield were separated. Similar mismatches happen a lot. Therefore, it is hard to do the join with a simple programming method. Instead, it was done by hand in excel and found in a total of 370 equivalent NAPLAN regions in Melbourne on SA2 level. Using this result region names, we were then able to list all schools located in Melbourne.

The school list contains 964 schools, including government (601 schools), catholic (234 schools), and independent schools (129 schools). The school tuition fee of government schools was considered as 0 dollars. The tuition fee of the other two kinds of schools was then collected by my project partner Nicholas and me from the official website of each school. The school fee of a catholic or independent school usually contains two parts, the tuition fee, and levy fee. In this project, we collected these two kinds of fees separately at the same time.

Not all schools are posting their tuition fee information on websites. Most independent schools do post, and most of the Catholic schools do not. We got the school fee information from the official websites of 105/129 independent schools, and 109/234 catholic schools.

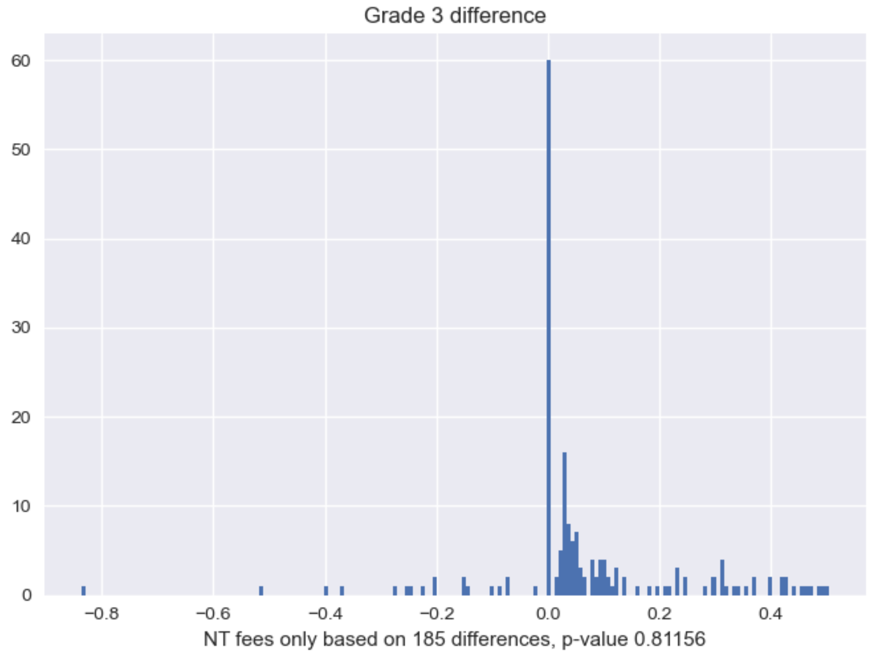


Figure 3.3: The result of comparing school fee data collected on the official website and from good school guide data set. We can see that about 60% of the data from the two different sources are almost the same, and p-value is too large to deny the assumption that these two are actually one identical data source. This graph comes from Nicholas' dataset.

During the process of data collection, we found out that the website "good school guide" also has school tuition fees information, including large amounts of schools that we do not have data for. To obtain more tuition fee data with accurate information, we first collected the tuition fees from the good school guide website and compared it with our known tuition fee data. The comparison result made by my partner Nicholas is shown in figure 3.3and 3.4.

From this graph, we can see that the school fee data from good school guides are very close to the accurate data that we obtained from school official websites. Therefore we decided to use data from it as supplementary information on tuition fees. With its supplement, we have 323/363 school's tuition fee information.

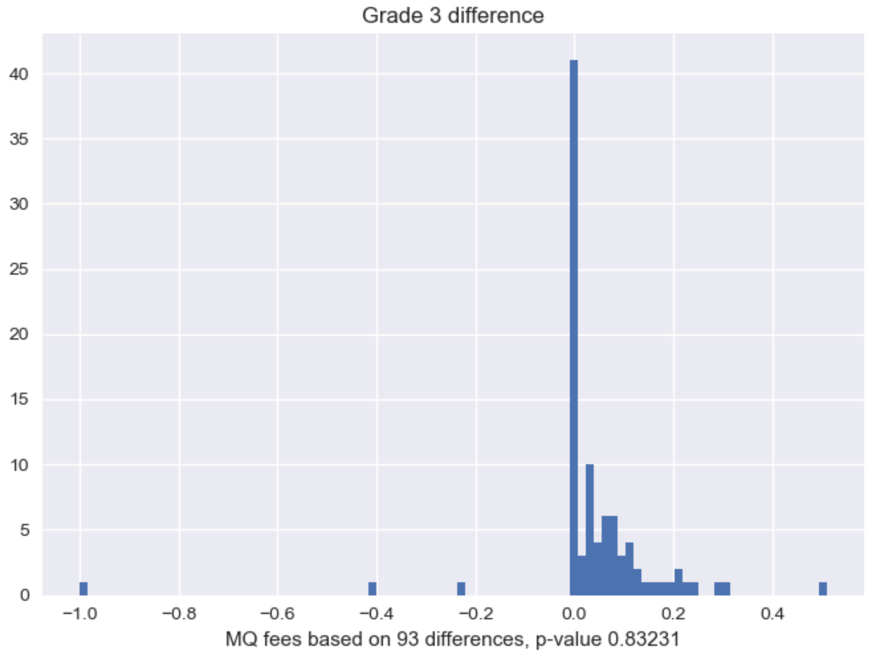


Figure 3.4: The result of comparing school fee data collected on the official website and from good school guide data set. We can see that about 60% of the data from the two different sources are almost the same, and p-value is too large to deny the assumption that these two are actually one identical data source. This graph comes from my dataset.

3.4 Creating a robust data pipeline

After preparing all the data sets, we then integrate the data into python as data frames for processing. The three datasets, census data, NAPLAN data, and school fee data were joined together for analysis.

There are a large number of attributes in census data, as described before in the census data section. Joining all of them together at one time consumes too much computing power. It cannot be done on our personal laptops and is also meaningless. We will never use all of the attributes together at one time, as it would cause overfitting. So what we would do is select particular attributes and analyze them. To simplify the process of selecting attributes so that we do not need to worry about where the attribute is and how to load it into the python data frame every time, my partner Nicholas created a python preprocessing script. The script preprogrammed the access method of all attributes at a time, including the process of fetching the desired data attribute from different folders, so that we can directly call the variables when modeling without worrying about data loading. It also integrates some data transformation methods in it.

The data transformation methods used in the preprocessing of this project can be roughly divided into two types. One is Data Normalization. Many of the attributes are not normally distributed, to fix that, the sklearn preprocessing package was used to transform data sets.

The other is transform on attributes. Some data attributes in census data are not suitable for use in models directly, for example, the overseas population of a region. This population in the census is the absolute population, which is influenced by the total number of population. A better representation of the overseas population is the ratio of the overseas population to the total number of people, which is the overseas population percentage. Similar ideas apply to other attributes like the indigenous population, etc. Some attributes are merged together to better represent a single characteristic of the region. For example, the number of people with working hours 0, 0-10, 10-20, etc, states the information of working hours in multiple attributes. These attributes are multicollinear with each other and the total population as the sum of them is roughly equal to the total population(with some random variance introduced by ABS before publishing the data). This variable was transformed by calculating the average working hours in this region.

Similar ideas apply to other attributes like average motor vehicles, which use the number of people with 0, 1, 2, and more motor vehicles to calculate the average number per dwelling.

4 Data Analysis

4.1 Multiple linear regression models

This project uses multiple regression to model the relationship between tuition fees, other factors, and student education output which is the NAPLAN result.

The data analysis of this project can be roughly divided into two stages. The first stage was more like trial and error, where we did exploratory data analysis on the variables and modeling the data at the same time. The second stage is more robust. With the help of stepwise feature selection and grid search, we were able to automate the feature selection and modeling procedure. These two stages are closely related, and the methods used during this process will be discussed in this section.

In the first stage, when constructing the linear model, assumptions of multiple linear regressions were also checked at the same time.

- Linear relationships between outcome variables and independent variables. If the relationship is not linear then the assumption is not met.

- Normally distributed residuals. The residual of the regression model should be normally distributed. This assumption can be tested using a normal quantile-quantile plot of residuals, namely the Q-Q plot.

- No Multicollinearity. Multiple linear regression requires the independent variables are not correlated with each other. As we are using census data, we must be really careful with this assumption. Many variables in the same census data table are correlated with each other, for example, the number of male plus female is the total population; the total population minus people with no vehicle is roughly equal to the number of people with vehicles (as the census data usually introduces small random variances to protect privacy). In this project, we were motivated by the correlation between students' grades and their socioeconomic advantages SEA. However, we found that it is not appropriate to add SEA in the multiple regression, as it correlates to most of the variables that we are studying, like tuition fees, percentage of indigenous people, etc. We removed the variable ICSEA score from our model after finding that in the first stage of our project.

- Homoscedasticity. The variance of residuals should be consistent across the values of independent variables, namely the variances should not change much with the change of independent variables. Sometimes the problem of not meeting this assumption can be solved by the variable transformation. For example, if the data is heteroscedastic, namely variance is increasing with the increase of variable, then a non-linear transformation of the variable can help to fix the problem.

In the second stage, we still use the multiple linear regression model to formulate the problem, but the way of constructing the model is becoming more robust. Feature selecting and parameter tuning methods are used in the second stage, as introduced as follows.

4.2 Criterion for model selection

There are two criteria for model selection: one is the model Metrics, and the other is the interpretability of the model. The metrics used in this project contains R squared, mean squared error, VIF, etc.

R squared represents the proportion of the explained variance of the dependent variable, ranging from 0 to 1. A larger r squared usually means better result.

Mean squared error estimates the average squared error of a model, and the smaller MSE means a better model.

VIF measures the multicollinearity of different independent variables in a model. In this project, a variable with VIF larger than 10 would be eliminated.

4.3 Stepwise feature selection

The stepwise feature selection is a feature selection technique for selecting features in linear regression. In this project, feature selection was based on the p-value and VIF value of a variable. The threshold of p-value is 0.1, which means a variable with p-value larger than 0.1 would be eliminated, and p-value larger than 0.1 would enter the model. For VIF, a variable with VIF larger than 10 would be removed from the model.

4.4 Regularisation

In order to reduce the effect of overfitting, ridge and lasso regression are used to regularize the model. As discussed in section 2.3.3, these two types of regression add penalty terms that is related to the magnitude of the coefficients, and therefore help to reduce the effect of overfitting. The parameter of ridge and lasso regression was determined using Grid search approach.

4.5 Grid search approach

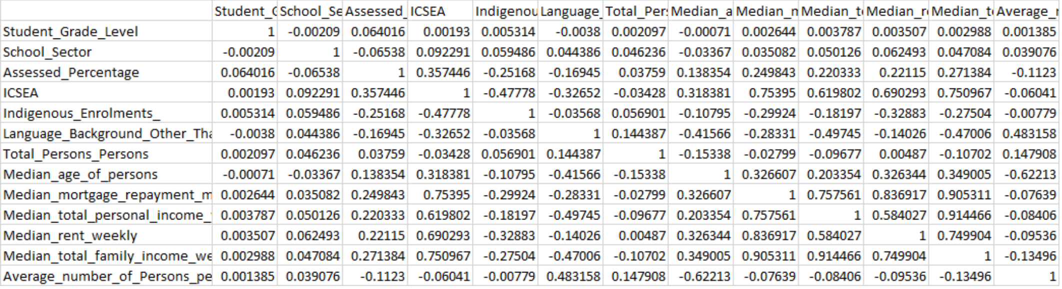
In this project, the grid search approach was used to select the best performing parameters in the model. In ridge/lasso regression, the term alpha is called a hyperparameter. The value of alpha has an influence on the penalty term and therefore influences the model performance. Grid search approach automatically iterates over possible values of hyperparameter using cross validation and finds the best value of hyperparameter. A process of grid search can be described as follows:

- Taking the model and possible values of hyperparameters.
- Take each possible hyperparameter value, plug into the model and calculate the unbiased error of that model.
- The unbiased error was calculated by splitting the data into training validation and test sets. We train the model on the training set, use MSE for each model on the validation set to find the best parameter value that minimizes the MSE, and use the test set to calculate the unbiased error of the model.
- Use cross validation to split the dataset and repeat the former step to find the model with lowest error, which indicates the best value of hyperparameter.

5 Results and discussion

5.1 Result from the first stage

In the first stage, the variables in multiple linear regression were selected manually, and transformed according to the variable's characteristics. Since the variables in multiple regression should be independent from each other, the correlation matrix of the variables were tested, and variables with high correlation were eliminated. 5.1 is a sample piece of correlation matrix. From the matrix we can see that, the ICSEA score is highly correlated with most of the variables. ICSEA is a combination of factors, as introduced in section 2.2.2. To better study the detailed influential factors of education output, the ICSEA variable was removed.



	Student_G	School_Se	Assessed_	ICSEA	Indigenou	Language_	Total_Per	Median_a	Median_n	Median_t	Median_r	Median_t	Average_n
Student_Grade_Level	1	-0.00209	0.064016	0.00193	0.005314	-0.0038	0.002097	-0.00071	0.002644	0.003787	0.003507	0.002988	0.001385
School_Sector	-0.00209	1	-0.06538	0.092291	0.059486	0.044386	0.046236	-0.03367	0.035082	0.050126	0.062493	0.047084	0.039076
Assessed_Percentage	0.064016	-0.06538	1	0.357446	-0.25168	-0.16945	0.03759	0.138354	0.249843	0.220333	0.22115	0.271384	-0.1123
ICSEA	0.00193	0.092291	0.357446	1	-0.47778	-0.32652	-0.03428	0.318381	0.75395	0.619802	0.690293	0.750967	-0.06041
Indigenous_Enrolments_	0.005314	0.059486	-0.25168	-0.47778	1	-0.03568	0.056901	-0.10795	-0.29924	-0.18197	-0.32883	-0.27504	-0.00779
Language_Background_Other_Tha	-0.0038	0.044386	-0.16945	-0.32652	-0.03568	1	0.144387	-0.41566	-0.28331	-0.49745	-0.14026	-0.47006	0.483158
Total_Persons_Persons	0.002097	0.046236	0.03759	-0.03428	0.056901	0.144387	1	-0.15338	-0.02799	-0.09677	0.00487	-0.10702	0.147908
Median_age_of_persons	-0.00071	-0.03367	0.138354	0.318381	-0.10795	-0.41566	-0.15338	1	0.326607	0.203354	0.326344	0.349005	-0.62213
Median_mortgage_repayment_m	0.002644	0.035082	0.249843	0.75395	-0.29924	-0.28331	-0.02799	0.326607	1	0.757561	0.836917	0.905311	-0.07639
Median_total_personal_income_	0.003787	0.050126	0.220333	0.619802	-0.18197	-0.49745	-0.09677	0.203354	0.757561	1	0.584027	0.914466	-0.08406
Median_rent_weekly	0.003507	0.062493	0.22115	0.690293	-0.32883	-0.14026	0.00487	0.326344	0.836917	0.584027	1	0.749904	-0.09536
Median_total_family_income_we	0.002988	0.047084	0.271384	0.750967	-0.27504	-0.47006	-0.10702	0.349005	0.905311	0.914466	0.749904	1	-0.13496
Average_number_of_Persons_pe	0.001385	0.039076	-0.1123	-0.06041	-0.00779	0.483158	0.147908	-0.62213	-0.07639	-0.08406	-0.09536	-0.13496	1

Figure 5.1: A sample piece of correlation matrix obtained from correlation calculation. This matrix is used to eliminate one of the variables that are highly correlated with each other, in order to comply with the assumptions for Multiple linear regression.

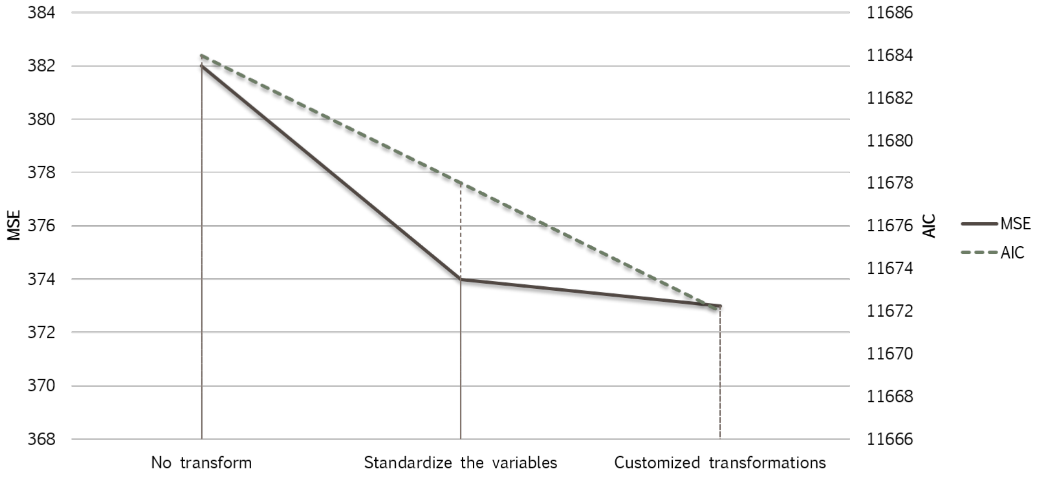
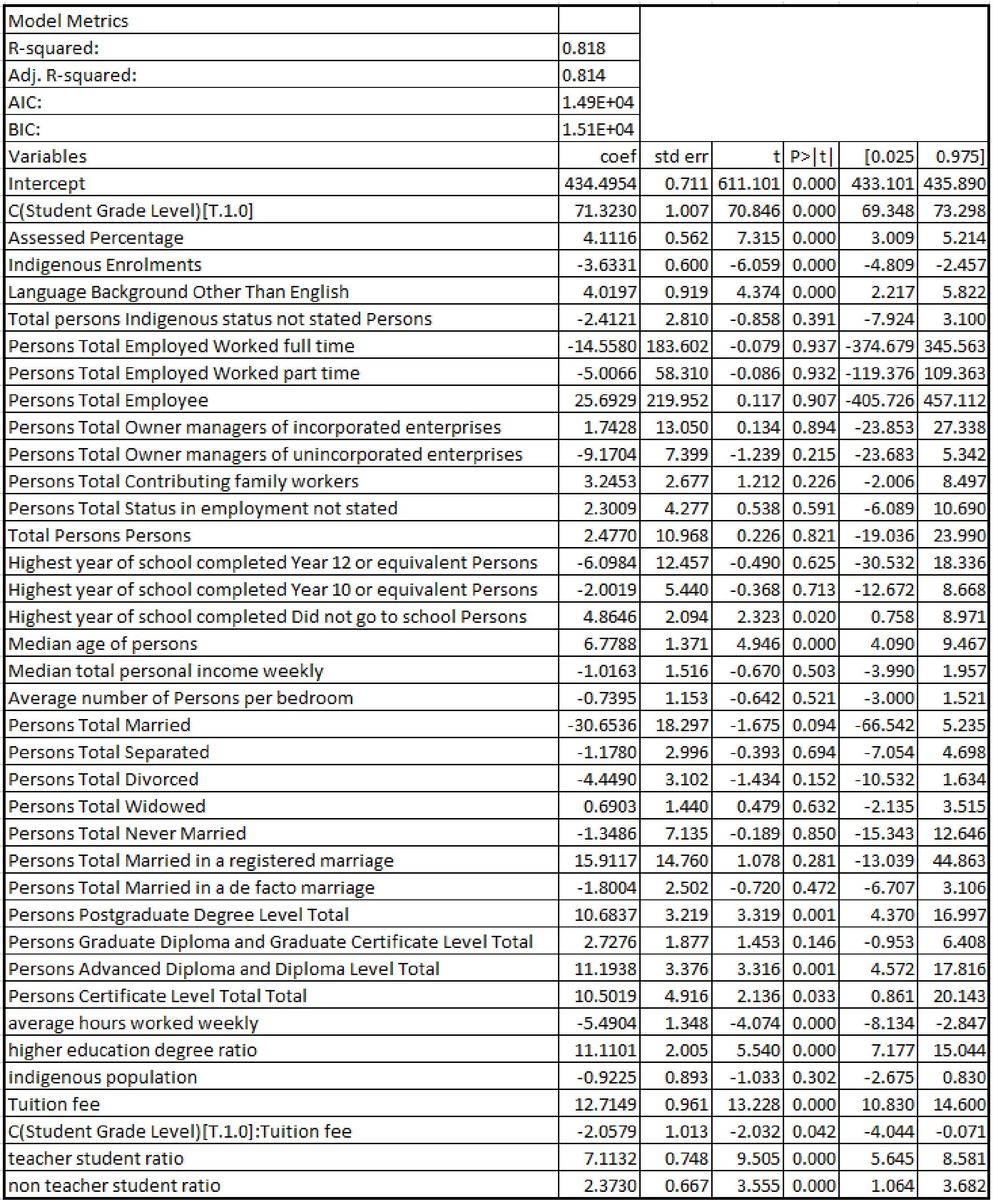


Figure 5.2: The decrease of AIC and MSE when applying appropriate transformations on individual variables.

Transformations were applied to variables, and after comparing the result before and after transformation, we found that the transformation of variables can improve the performance of the model when considering AIC or MSE as the metrics, as shown in 5.2.

The multiple regression models were then constructed in python. A sample of the ‘OLS Regression Results’ is in the following result form5.3. From the result we can see the coefficients of each variable. The p-values of the variables indicates the significance of the variables, where a lower p-value (less than 0.1) indicates higher significance. We can also get the AIC and BIC value of the model which indicates the error of the model. Given a set of candidate models, the preferred model is the one with the minimum AIC/BIC value.



Model Metrics						
R-squared:	0.818					
Adj. R-squared:	0.814					
AIC:	1.49E+04					
BIC:	1.51E+04					
Variables	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	434.4954	0.711	611.101	0.000	433.101	435.890
C(Student Grade Level)[T.1.0]	71.3230	1.007	70.846	0.000	69.348	73.298
Assessed Percentage	4.1116	0.562	7.315	0.000	3.009	5.214
Indigenous Enrolments	-3.6331	0.600	-6.059	0.000	-4.809	-2.457
Language Background Other Than English	4.0197	0.919	4.374	0.000	2.217	5.822
Total persons Indigenous status not stated Persons	-2.4121	2.810	-0.858	0.391	-7.924	3.100
Persons Total Employed Worked full time	-14.5580	183.602	-0.079	0.937	-374.679	345.563
Persons Total Employed Worked part time	-5.0066	58.310	-0.086	0.932	-119.376	109.363
Persons Total Employee	25.6929	219.952	0.117	0.907	-405.726	457.112
Persons Total Owner managers of incorporated enterprises	1.7428	13.050	0.134	0.894	-23.853	27.338
Persons Total Owner managers of unincorporated enterprises	-9.1704	7.399	-1.239	0.215	-23.683	5.342
Persons Total Contributing family workers	3.2453	2.677	1.212	0.226	-2.006	8.497
Persons Total Status in employment not stated	2.3009	4.277	0.538	0.591	-6.089	10.690
Total Persons Persons	2.4770	10.968	0.226	0.821	-19.036	23.990
Highest year of school completed Year 12 or equivalent Persons	-6.0984	12.457	-0.490	0.625	-30.532	18.336
Highest year of school completed Year 10 or equivalent Persons	-2.0019	5.440	-0.368	0.713	-12.672	8.668
Highest year of school completed Did not go to school Persons	4.8646	2.094	2.323	0.020	0.758	8.971
Median age of persons	6.7788	1.371	4.946	0.000	4.090	9.467
Median total personal income weekly	-1.0163	1.516	-0.670	0.503	-3.990	1.957
Average number of Persons per bedroom	-0.7395	1.153	-0.642	0.521	-3.000	1.521
Persons Total Married	-30.6536	18.297	-1.675	0.094	-66.542	5.235
Persons Total Separated	-1.1780	2.996	-0.393	0.694	-7.054	4.698
Persons Total Divorced	-4.4490	3.102	-1.434	0.152	-10.532	1.634
Persons Total Widowed	0.6903	1.440	0.479	0.632	-2.135	3.515
Persons Total Never Married	-1.3486	7.135	-0.189	0.850	-15.343	12.646
Persons Total Married in a registered marriage	15.9117	14.760	1.078	0.281	-13.039	44.863
Persons Total Married in a de facto marriage	-1.8004	2.502	-0.720	0.472	-6.707	3.106
Persons Postgraduate Degree Level Total	10.6837	3.219	3.319	0.001	4.370	16.997
Persons Graduate Diploma and Graduate Certificate Level Total	2.7276	1.877	1.453	0.146	-0.953	6.408
Persons Advanced Diploma and Diploma Level Total	11.1938	3.376	3.316	0.001	4.572	17.816
Persons Certificate Level Total Total	10.5019	4.916	2.136	0.033	0.861	20.143
average hours worked weekly	-5.4904	1.348	-4.074	0.000	-8.134	-2.847
higher education degree ratio	11.1101	2.005	5.540	0.000	7.177	15.044
indigenous population	-0.9225	0.893	-1.033	0.302	-2.675	0.830
Tuition fee	12.7149	0.961	13.228	0.000	10.830	14.600
C(Student Grade Level)[T.1.0]:Tuition fee	-2.0579	1.013	-2.032	0.042	-4.044	-0.071
teacher student ratio	7.1132	0.748	9.505	0.000	5.645	8.581
non teacher student ratio	2.3730	0.667	3.555	0.000	1.064	3.682

Figure 5.3: The result of a multiple regression with variables selected by hand. From the result we can see that the p-value of some of the variables are 0, and some are larger than 0.1. The ones with p-values larger than 0.1 are not considered as significant variables, like total number of persons.

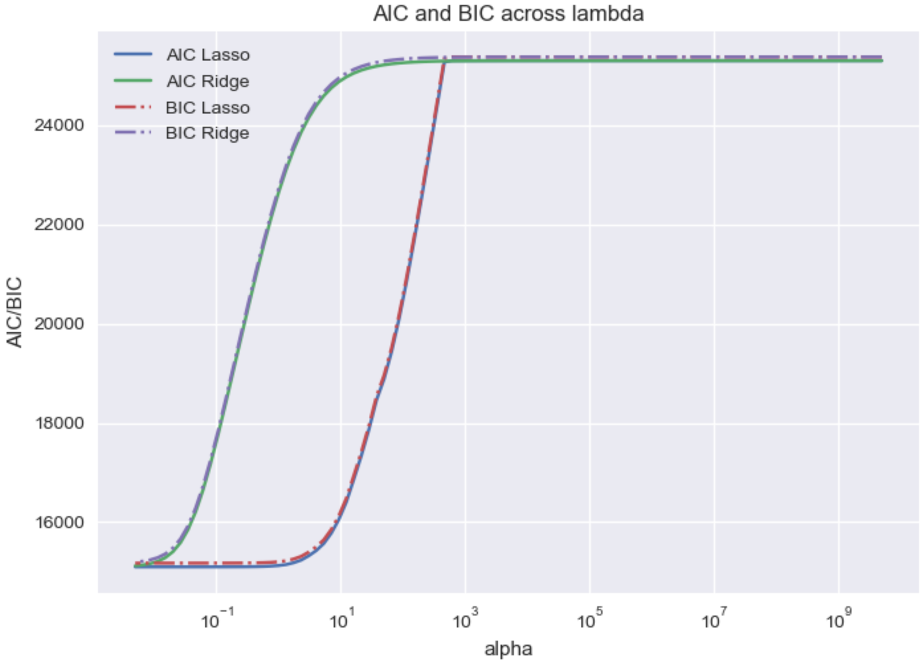


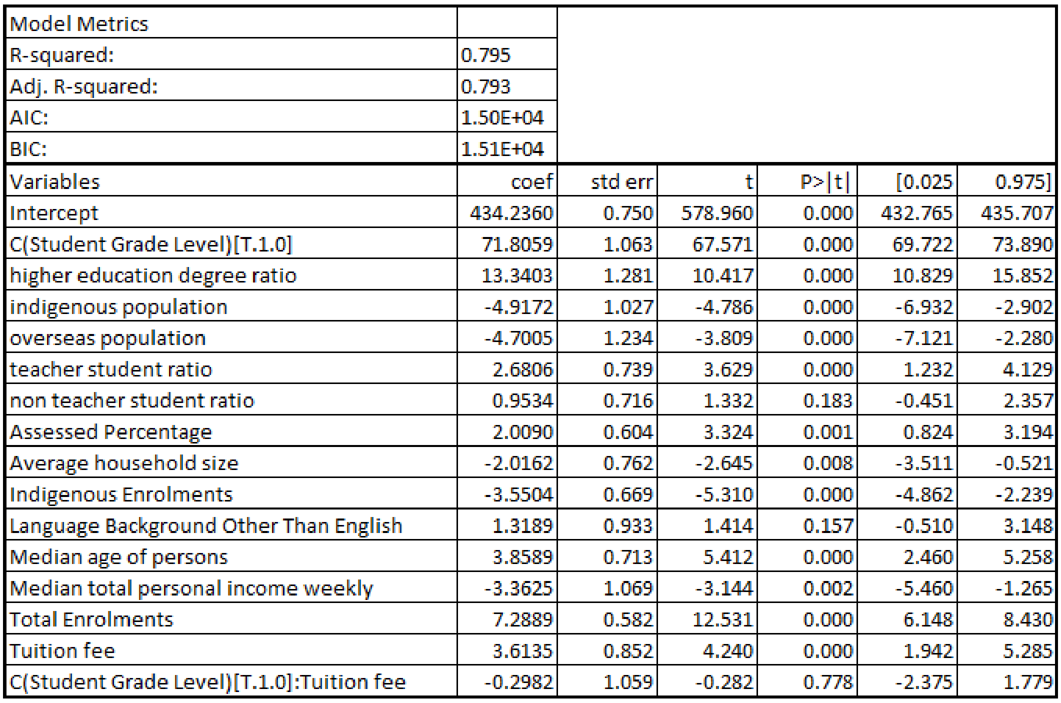
Figure 5.4: The change of AIC and BIC for ridge and lasso regression when increase the penalty factor alpha (also called lambda). From the graph we can see that with the increase of alpha, the AIC and BIC of the model increases.

5.2 Result in the second stage

In order to reduce the effect of over-fitting, regularisation was also used in constructing the model, and the grid search method was used in parameter tuning for regularisation. The grid search process was done on multiple models for determine the effect of changing penalty factors, and one of the result was shown here as an example in figure 5.4. From the graph we can see that with the increase of alpha, the AIC and BIC of the model increases, so the regularization in this model is not appropriate.

The addition of penalty factors actually increase the error of the model instead of decrease, so the penalty factor added in lasso and ridge regression is not supportive in our model.

In the second stage, instead of adjusting the model manually, we used stepwise feature selection to select appropriate features. As stepwise feature selection iterates through variables automatically. Though the result is still a multiple linear regression model that looks similar to the former ones we had, the feature selection method actually iterated through much more different combinations of variables compared to the former stage when we were



Model Metrics						
R-squared:	0.795					
Adj. R-squared:	0.793					
AIC:	1.50E+04					
BIC:	1.51E+04					
Variables	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	434.2360	0.750	578.960	0.000	432.765	435.707
C(Student Grade Level)[T.1.0]	71.8059	1.063	67.571	0.000	69.722	73.890
higher education degree ratio	13.3403	1.281	10.417	0.000	10.829	15.852
indigenous population	-4.9172	1.027	-4.786	0.000	-6.932	-2.902
overseas population	-4.7005	1.234	-3.809	0.000	-7.121	-2.280
teacher student ratio	2.6806	0.739	3.629	0.000	1.232	4.129
non teacher student ratio	0.9534	0.716	1.332	0.183	-0.451	2.357
Assessed Percentage	2.0090	0.604	3.324	0.001	0.824	3.194
Average household size	-2.0162	0.762	-2.645	0.008	-3.511	-0.521
Indigenous Enrolments	-3.5504	0.669	-5.310	0.000	-4.862	-2.239
Language Background Other Than English	1.3189	0.933	1.414	0.157	-0.510	3.148
Median age of persons	3.8589	0.713	5.412	0.000	2.460	5.258
Median total personal income weekly	-3.3625	1.069	-3.144	0.002	-5.460	-1.265
Total Enrolments	7.2889	0.582	12.531	0.000	6.148	8.430
Tuition fee	3.6135	0.852	4.240	0.000	1.942	5.285
C(Student Grade Level)[T.1.0]:Tuition fee	-0.2982	1.059	-0.282	0.778	-2.375	1.779

Figure 5.5: The result of using stepwise feature selection method to choose variables.

adding or removing variables by hand. Therefore, the result in this stage is more convincing. A sample result of this stage is shown below in 9.1.

From the results of feature selection in multiple experiments, we can conclude the significant variables(with p value lower than 0.1) with positive impact on student's NAPLAN score include Tuition fee, total enrollments (of a school), higher education degree ratio, etc. Significant variables like indigenous population and school indigenous enrollment percentage have a negative impact on students' NAPLAN score. Noticing the variables in the model were transformed and therefore the scale of the variables are changed. Using sklearn package, zero-mean and unit-variance normalization is applied to the transformed data. Because of the transformation, we can better see the influence of tuition fee on student NAPLAN grade. That is, with a difference of tuition fee between highest(around \$ 31,000) tuition fee school and the lowest (0 \$) school, the expected score difference is only about 3 points (using Table in 9.1 as an example). Refer back to (figure 2.2), the grade difference between government and independent schools are roughly 60 points. So the influence of tuition fee is actually very small, comparing to other factors like total enrollments, which has a coefficient of 7, and higher education degree ratio with coefficient of 13.

5.3 Following work

In this section, I will be introducing the result of Instrumental Variable analysis performed by my partner Nicholas. Instrumental Variable (Sander, 2000) is commonly used in studying the casual relationship. Through we have already studied on a large amount of variables that correlates to the NAPLAN score variable, there are a lot more that are not considered in the model. Some are not considered because we have no data to capture the influential factor, while some are factors that only affect the NAPLAN score by influencing some other factors, which means they do not have influence on NAPLAN score directly. These are called instrumental variables. Nicholas used the family marital statuses and parent education level to capture the factors related to the care of children at an early age, and obtained models with better estimation. Similar result that the tuition fee factor is having small coefficient towards the NAPLAN score was obtained: roughly \$1000 of increase in tuition fee is correlated with 1 point of increase of NAPLAN score, which is too small to be an important factor.

6 Shortcomings and further work

6.1 Scope of the project

The score of this project could be enlarged in future studies. With the NAPLAN score information and school information of all schools across Australia, the study was able to be done in a large scope. However, due to the limitation of time and resources in data collection, we were unable to collect the school fee data for all schools across Australia. Therefore, the research was limited to the scope of schools in Melbourne. Further study including more data will reduce the possible effect of over fitting, and generate more convincing result at a larger scale.

What is more, more independent variables may be introduced to this project like the teacher's satisfactory level, extra-curricular data of students, student physical health information, etc. In this study, the data set we were studying on was limited, and therefore cannot reflect on the whole picture of the children's development problem.

6.2 Limitation in correlation

In this project, though we were using school fee as an independent variable and the grade as dependent variable, we actually cannot assert the relationship between these two variables. We can only tell that these two variables are correlated with each other, but cannot tell whether one of them is the result or cause of the other. Using common knowledge we are making the judgement that an increase of education input would result in the improvement of students grade, but the causality is not proved. Further studies may touch on the problem of causality and correlation between tuition fee and student academic performance.

7 Conclusion

To summarise, this project explored on the factors that influences students' academic performance quantitatively, especially focus on the relationship between tuition fee and students' NAPLAN grade for primary school students in the Melbourne area. We used multiple regression to model the multi factor problem, using tuition fee and other factors as independent variables and NAPLAN result as dependent variable. Different techniques were used in this study. In the preprocessing stage, the transformation of variables improves the model performance. In the modeling stage, the variables that are not significant or highly correlated with other variables are removed from the model. Through the data of tuition fees varies in a large range, and the number of observations we used is only over one thousand, the regularization of the model like lasso and ridge regression did not really improve the result of the model. With the help of stepwise feature selection, we converged to a model with the best performance. In the model, tuition fee was significant as the p-value of it is very low. However, the coefficient of the tuition fee is too small to be influential, namely tuition fees is not really the reason that leads to the large NAPLAN score difference between independent schools and government schools. Other factors, like school indigenous enrollment, higher education degree ratio are actually more influential.

8 Reference

Abs.gov.au. 2020. Australian Statistical Geography Standard (ASGS). [online] Available at: <https://www.abs.gov.au>

A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation of Nonorthogonal problems, Technometrics, 12(1970), 55-67.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2011). NAPLAN: National assessment program.

Davis-Kean, P. E. (2005). The Influence of Parent Education and Family Income on Child Achievement: The Indirect Role of Parental Expectations and the Home Environment. Journal of Family Psychology, 19(2), 294-304. doi:10.1037/0893-3200.19.2.294

Denis Napthin, Peter Lee, Caroline Graham & Meredith Wills.(2018). National Regional, Rural and Remote Education Strategy. Australian Government, Department of Education, Skills and Employment.

Frank A. Wolak (1987) An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model, Journal of the American Statistical Association, 82:399, 782-793.

Goss, P., Emslie, O., and Sonnemann, J. (2018). Measuring student progress: A state-by-state report card – Technical Report. Grattan Institute. ISBN: 978-0-6483311-7-9

IQBAL, A., AZIZ, F., FAROOQI, T., & ALI, S. (2016). Relationship between Teachers' Job Satisfaction and Students' Academic Performance. Eurasian Journal Of Educational Research, 16(64), 1-35. doi: 10.14689/ejer.2016.64.19

Jenny Chesters & Louise Watson (2013) Understanding the persistence of inequality in higher education: evidence from Australia, Journal of Education Policy, 28:2, 198-215, DOI: 10.1080/02680939.2012.694481

Karmel, P. (1973). Schools in Australia: Report of the Interim Committee of the Australian Schools Commission. Canberra: Australian Government Publishing Service.

Kuhn, Max, and Kjell Johnson. Feature Engineering And Selection: A Practical Approach For Predictive Models (Chapman Hall/CRC Data Science Series). 1st ed., Chapman And Hall, 2019.

Lamb, S. (2014). Educational Disadvantage in Regional and Rural Schools. Research Conference, 2014

McNiece, R., & Jolliffe, F. (1998). An investigation into regional differences in educational performance in the National Child Development Study. Educational Research, 40(1), 17-30. doi: 10.1080/0013188980400102

Myers, R. (1990). CLASSICAL AND MODERN REGRESSION WITH APPLICATIONS (2nd ed., pp. 82-85). Boston: PWS-Kent.

Sander Greenland (2000). An introduction to instrumental variables for epidemiologists, International Journal of Epidemiology, Volume 29, Issue 4, Pages 722–729

Seber, G., & Lee, A. (2003). Linear Regression Analysis (2nd ed., p. 4). Auckland: John Wiley & Sons. Copyright.

Sheldon Rothman (2003) "The changing influence of socioeconomic status on student achievement: Recent evidence from Australia". LSAY CONFERENCE PAPERS, 4-2003

Tan, P.N. & Steinbach, Michael & Kumar, Vipin. (2005). Cluster Analysis: Basic Concepts and Algorithms. Introduction to Data Mining. 487-568.

Win, R., & Miller, P. (2005). The Effects of Individual and School Factors on University Students' Academic Performance. The Australian Economic Review, 38(1), 1-18. doi: 10.1111/j.1467-8462.2005.00349.x

9 Appendix

Graph 9.1: How do the variables affect student academic achievement in different NAPLAN tests?

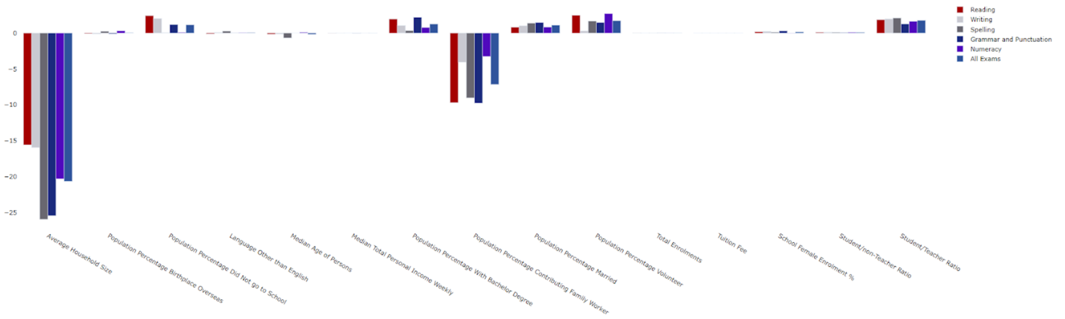


Figure 9.1: Results we see here for the averaged NAPLAN results are seen consistently across each exam area: The strength and direction of correlation between each of our model's variables and school achievement within different NAPLAN subjects are similar to each other. That is, the effect of variables are similar on different NAPLAN test, no matter it is on literacy or numeracy. Also, the effect is similar to the result of averaging the grades on different tests. This graphic result credit to my partner Nicholas Thompson.