



Norwegian University of
Science and Technology

A guardian's effect on pupil's reading test scores

A comparative analysis on New Zealand

Bacheloroppgave i Samfunnsøkonomi SØK2901

Student: Eghe Jeffrey Avent-Umamwen

Supervisor: Bjarne Strøm

Mai 2020

CONTENTS

1. INTRODUCTION	3
1.1 motivation	
1.2 the thesis	
1.3 our hypothesis and former studies.	
1.4 the summary	
2. THE ANALYTICAL THEORY AND METHOD OF OUR ANALYSES.	7
2.1 the function form	
2.2 The educational production Function.	
- Its relationship to the method of a linear function.	
2.3 A regression model –	
a refined version of the educational production function.	
2.4 The analytical importance of our “x” inputs and the sum of least squares.	
2.5 the analytical weight of our model – R^2	
2.6 Hypothesis testing.	
3. REVIEWING AND CLASSIFYING OUR COLLECTED DATA.	15
3.1 data descriptive analysis	
4. THE REGRESSION ANALYSIS.	20
4.1 an overview of the chosen regression model for testing	
4.2 regression table	
4.3 Empirical result	
5. CONCLUSION	26
6. REFERENCES	28
7 . APPENDIX	30

1. INTRODUCTION

1.1 Motivation.

We look to research, to how big an effect, the guardian of a pupil has on the test performance of a pupil at school, specifically at reading. A guardian of an individual who is yet of legal age is normally their parent. However, someone who has adopted a young child, as well a schoolteacher from a school that a young child is officially registered as a pupil, can all both be classified as a guardian for that pupil. It is widely assumed that the role of a guardian of a young person plays a role in what type of individual that person may become as an adult. Therefore, we wish to examine how much of an effect that a guardian may have on how well a pupil does at their reading test score. Specifically, reading performances of pupils, between the third and fifth grade of school. The country of choice for this research is New Zealand. New Zealand is the chosen country of choice due to Its similar government style to Norway. Both New Zealand and Norway operate with a “constitutional monarchy”. Therefore, similarities between the two nations. However, the official spoken language in New Zealand is British English. We hope that this will help narrow our research and eliminate a few deterministic and stochastic factors related dialects. Because, Norway has four geographical Norwegian dialects: Vestlands, Østlandsk, Trøndersk and Nordnorsk.

1.2 The thesis.

In New Zealand, how does the educational pedigree of the pupil’s guardians affect the pupil’s test scores in reading?

- Educational pedigree refers to how well educated a person is. Normally, the word considers the performance level of the schools attended by a person - The difficulty of the studies attended, how well the individual performed at school, and how high up the educational degree system that the individual got – a university bachelor degree, a PhD, and so forth. However, for this paper, we only consider the person – parent or teacher – to have either and or a university degree, teacher’s certificate and a high income, since most well-educated individuals are assumed to have high incomes.

1.3 Our hypothesis and former studies.





(Hans Bonesrønning, 2004, article, 14) It is highlighted two particularly important models. The first model is Coleman’s report from the year 1966 (Coleman and co. 1966). Coleman had an assignment to find empirical proof on if the educational system in the USA could help strengthened the educational pedigree of the black (racial) community in the USA. It is stated in

Hans article, that Colemans conclusion what that it was all in the family. This conclusion meant that educational institutions had little effect on the educational achievement of pupils at school. This conclusion by Coleman was not considered a particularly promising conclusion.

At a later period, and by Rivkin (Rivkin and co. 2001) a new model, like Colemans, but with the addition of a new “input”. The teacher’s factor. The model highlighted that Colemans model is potentially wrong because the teachers at various school have great effect on the student’s performances as well.

The parents and teachers are both considered to play an important role in pupils’ performance (Hans B, 2004, article 14). Therefore, and as a leading question to our coming hypothesis: can we conclude that a guardian with a higher educational pedigree - meaning at minimum a university degree - would lead to a pupil performing better in their educational achievements - like in their reading test score? We do not know the empirical answer to this question. However, there have been qualitative assessments, that may be considered an equivalent to our coming hypothesis. One, mentioned in this paper of these assessments is from the Norwegian “Senter for leseforskning” regarding educational resources at home, parent’s attitude towards reading and schoolteacher’s attitude toward reading.

The assessment of the PIRL “progress in international reading literacy study” from 2001 by the Norwegian “Senter for leseforskning” (Solheim Tønnesen, report on PIRLS, 2001) are as below:

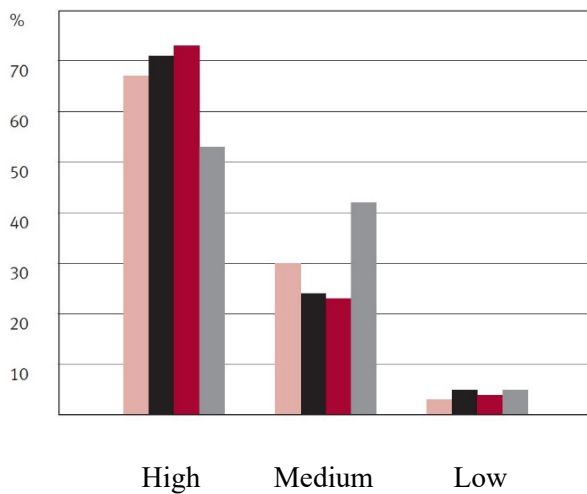
	NORWAY
	SWEDEN
	ISLAND
	INTERNATIONAL AVERAGE.

Note.

Solheim assessed three Scandinavian countries, then compared her findings to the international average from PIRL, 2001. More on PIRL, 2001 in our data review, later in this paper.

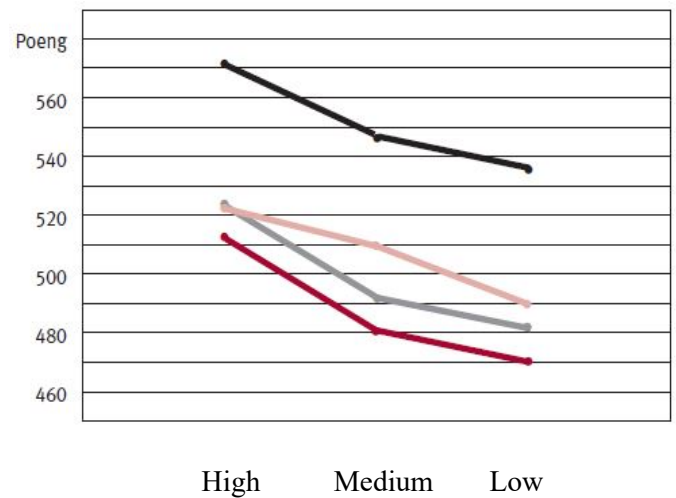
Figur 1.

Parent's responded attitude to reading



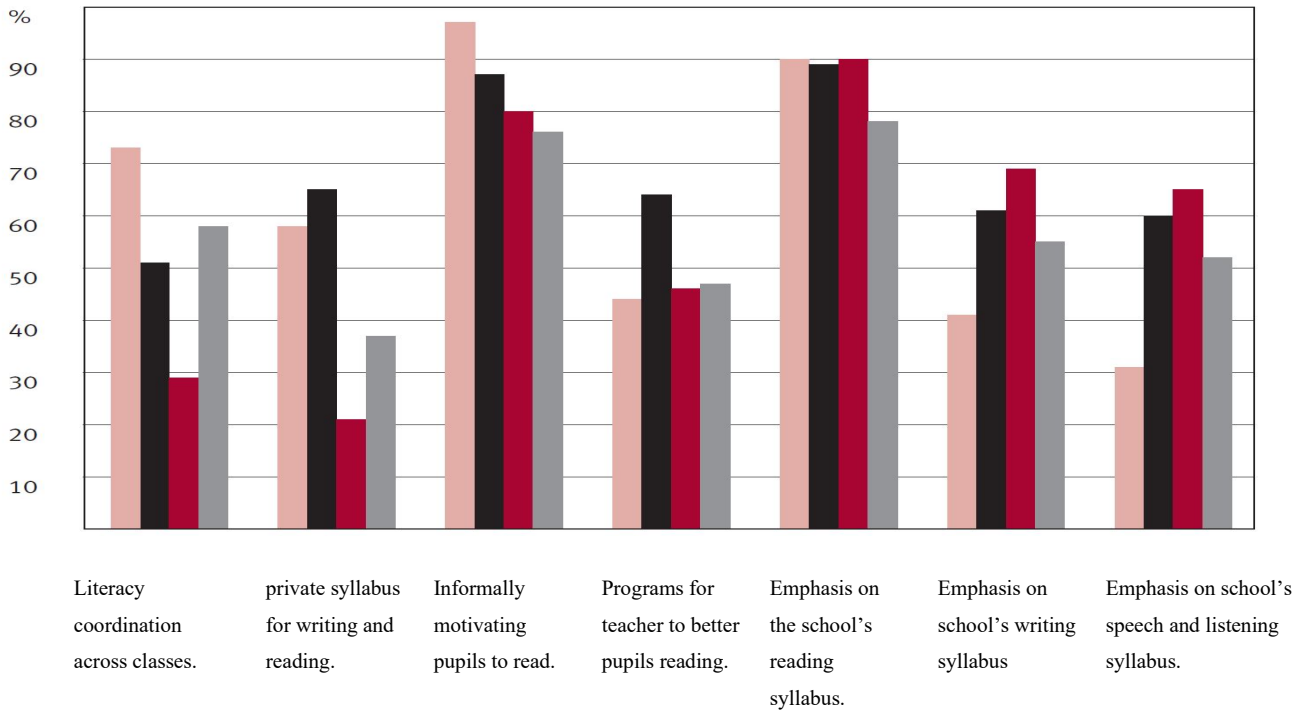
Figur 2

Parent's attitude and actual pupil score



Parent responded to these following questions: I read when I must, I like to discuss with others about books, I like to use my spare time to read books, I read only when I need information, reading is an important activity at home.

Figur 3. Percent of pupils that have schools with more emphases on reading.



1.4 The summary.

As observed in figure 1 and 2, there is correlation between homes, where parents have high attitude towards reading and pupils with high test scores in reading. However, there is a significant test score gap between Sweden and the international average. There is also a significant gap between Sweden, and Norway. Norway, a Scandinavian country like Sweden, and yet a noticeable educational gap between their test score.

In figure 3, we try to understand the cause of this gap. Schools in Sweden put as much emphasis on reading as the international average. The areas with big differences between Sweden and the international average in figure 3 is at: 1) A private syllabus for writing and reading, 2) School based program for teachers on how to better pupil's literacy. The observation in figure 3 brings us back to the importance of our thesis:

Thesis:

In New Zealand, how does the educational pedigree of the pupil's guardians affect the pupil's test scores in reading?

- And if yes, that it does affect the pupil's test score, how particularly significant is this effect?

Hypothesis:

My assumption is that we can conclude that a guardian with a higher educational pedigree - in other words, a university degree or more, would lead to a pupil performing significantly better in their educational achievements - like in their reading test score. And that a significant negative effect can be expected when the guardian has a low educational pedigree.

Questions corresponding to the above thesis and hypothesis are what we will hope to answer through the course of this paper. It is important to take note of shortcomings relating to factors not included in our data for this paper. An example is the factor of geographical dialects. High correlation issues like parents with higher education have higher income, therefore can afford better schools, with better teacher, better peer groups, teaching methods etc. A collection of data that efficiently separates all these factors would be particularly expensive and considered socially immoral to a community.

2. THE ANALYTICAL THEORY AND METHOD OF OUR ANALYSES.

For this research, we would expand the educational production function into a multiple regression formula of our own. This formula would take in the effects of underlying factors from our data set, as deemed relevant for the thesis of this paper.

The educational production function in its standard form is like every other production function that is rooted from the economic theory of production and demand. A process that follows the path of input, into the production function, then output that comes out from the other end of the function. An analytical showing of the education production function and its analytical relationship with the multiple regression method is shown below.

2.1 Function form:

$Y = f(x)$ - A normal production function

$Y = f'(x) > 0$ - Its function form when solved.

A linear variant of its function form can be written as below:

$$E(y) = b_0 + \beta x$$

$E(y)$ – Expected output from the function.

B_0 – A constant coefficient in the function.

βx – The marginal output with relationship to data (x) aka - The functions independent input factors.

Important Note.

The actual outputs, in relationship to real life events differs from the expected outputs. This is because human behavior is not deterministic, but stochastic. Therefore, there is always to be expected a margin of error that represents the stochastic factor in our inputs.

The definition of stochastic is as follow: a random distribution of probability that can be observed and analyzed statistically. However, can not be precisely predicted by any means due to factors that cannot be controlled.

The linear function rewritten, and with the addition of the error factor, represented as ε .

$$Y = E(y) + \varepsilon \quad (1)$$

ε – Represent the influence from very randomized inputs that are not included in the model. However, they may affect the results from the model. Either in a negative way, or in a positive way.

The formula (1) can be rewritten as:

$$Y = b_0 + \beta x + \varepsilon$$

Y – In this new formula the sign “y” now represents, not the expected output, but the actual output from our function. It is important to remember that, although the formula represents the actual output, we still would not get a precise output. This is due to the error factor. What we get is an output from within a margin of a predicted expectation.

How does the error factor affect our output?

If $\varepsilon > 0$ this means that our actual output “y” > is greater than our expected output E(y).

If $\varepsilon < 0$ this means that our actual output “y” < is lesser than our expected output E(y).

2.2 The educational production Function.

It’s relationship to the method of a linear function.

The educational function below is a model first created by Coleman in his report from 1966. The relationship between his model and the production function discuss in (2.1) is shown below. It is also important to take note on how his model can either be expanded upon or reduced.

$$O_{it} - O_{it^*} = f(F_i^{(t-t^*)}, P_i^{(t-t^*)}, S_i^{(t-t^*)}) + \varepsilon_{it}$$

$O_{it} - O_{it^*}$ - Represents “Y”. It is a continuous output that represents the students output between the time period of O_{it} and O_{it^*} .

$f(F_i^{(t-t^*)}, P_i^{(t-t^*)}, S_i^{(t-t^*)})$ – Represents f(x). It is continuous as well, between the time period t to t* across all three “x” factors.

ε_{it} – Represents the error factor.

Important note

Within $f(F_i^{(t-t^*)}, P_i^{(t-t^*)}, S_i^{(t-t^*)})$ are a series of different factors joined together to represent a single “x” input in the production function. Why it is made so, can be because the various factors are highly correlated. Therefore, is more productive that correlated inputs be made into one factor. This gives us room to add other factors that are correlated to our output. But are non-correlated to the other (x) input.

This can be observed in the education production function by Rivkin and co. From their report from 2001:

$$O_{it} - O_{it^*} = f(F_i^{(t-t^*)}, P_i^{(t-t^*)}) + \sum t_j T_{ij} + \varepsilon_{it}$$

$\sum t_j T_{ij}$ - In the production function, “ $\sum t_j T_{ij}$ ” is an indicator for an uncorrelated “x” input for Rivkin’s production function model. Taking the above formula into consideration, one should take a mental note on how we can choose to expand upon or reduce the education production function in accordance to the output that we may hoped to statistically observe in any given situation.

2.3 A regression model – a refined version of the educational production function.

$$O_{it} - O_{it^*} = f(F_i^{(t-t^*)}, P_i^{(t-t^*)}) + \sum t_j T_{ij} + \varepsilon_{it}$$

The production function above can be mathematically refined and rewritten as below:

$$Y_i = b_0 + \beta X_1 + \beta X_2 + \varepsilon \quad - \text{A regression model/formula.}$$

i – representing the continuous time aspect of both “y” and the “x” inputs that may be added to the regression model $I = 1, 2, 3, \dots, n$

$\beta X_1 + \beta X_2$ - Representing a series of unknown “x” inputs that can be single factors or a series of correlated “x” inputs that are statistically grouped together, into a single “x” input by a given parameter. Example: a time span can be a given parameter that help group a series of other inputs into one continuous “x” input for a regression model.

As the actual “ $\beta X_1 + \beta X_2$ ” and ϵ are unknown, we would have to estimate these inputs out from our collected data set. In addition, it would need to be estimated in accordance to the thesis that we hoped to observe.

As mentioned before, it is productively efficient that the “x” factors separated by the plus sign has little to no correlation between its selves. Although, they should all be significantly correlated to the output in the regression model. An analytical demonstration on why this is important is written below:

2.4 The analytical importance of our “x” inputs and the sum of least squares.

What happens when we fail to estimate and fail to include an important “x” input from our data set? What happens when we include an “x” input that is highly correlated to another “x” input in our regression model? The answers to these questions are part of what helps decide how efficient the regression model that we use to estimate our output is.

Let us assume:

(1) $Y_i = b_0 + \beta X_1 + \beta X_2 + \epsilon$ (1) is our most appropriate regression model.

However, we used a wrong regression model, by leaving out an “x” input. This “x” input is also highly correlated to the other “x” input in our regression model.

So, we get (2) $Y_i = b_0 + \beta X_1 + \epsilon$

To estimate our β_1 constant we use the formula below:

$$(3) \quad b_1 = \frac{\sum (X_1 - \bar{X}_1) (Y_i - \bar{Y}_1)}{\sum (X_1 - \bar{X}_1)^2}$$

Important note.

The linear fraction (3) is the accepted formula for estimating the constant term β . We use a lower-case b and the number “1” to signify that it is an estimate of the actual constant term for our first “x” input. The actual constant term β remains an unknown, so our aim is the best possible estimate of every term and “x” inputs given the parameters that we set. This is as mentioned about the regression model.

This mathematical method of estimating our terms, from our collected data and for the use on our regression model is called the least square method, abbreviated as “LSM”. Another name for the same method is the ordinary least square method abbreviated as “OLS”.

The linear fraction (3) is the same formula that would be used when there are multiple “x” inputs in our regression model. It is important to remember that these would all be terms collected from the same data sets. Therefore, all the relevant estimated “x” inputs should be highly correlated to our “y” output.

This means:

$$(4) (Y_i - \bar{Y}) = (X_n - \bar{X}_n) + (\epsilon_n - \bar{\epsilon})$$

That the difference between our actual “Y” output and its mean is equal to the difference between all relevant “x” input and their mean [added] all relevant stochastic error terms and their mean – We use a software for this calculation because it gets very complicated.

The least square method helps us estimate a linear function that is remarkably close to our expected linear function. It does this by dividing the difference of all relevant “x” inputs in our data by the sum of least squared of a specific “x” input against its mean.

One of many reasons as to why this method works is because of the law of large numbers. The law of large numbers in probability states that the mean narrows down to the actual mean as a sample size grows - By squaring, we double the numbers. By setting parameters, we can narrow to specific intervals in our collected data, that we wish to observe independently.

The risk.

Our parameter becomes meaningless if multiple “x” inputs are highly correlated with each other. This would lead to a skewed display of our data. This is because we would be observing one specific interval in our collected data instead of multiple independent “x” input intervals in our collected data.

In other words, an “x” input that is not significantly correlated with our “y” output can be misinterpreted as significant because it is highly correlated to another “x” input that is significantly correlated to “y”, but not included in the our regression model.

Therefore, an “x” input is highly significant if the other “x” inputs does not correlate with our “y” output, or the “x” inputs are independent (not correlated) to each other.

2.5 The analytical weight of our model – R²

The analytical weight of our model is called the R² of our regression mode. It tells us how much of our output “y” (dependent variable), is explained by the regression model that we have constructed.

Remember: $(Y_i - \bar{Y}) = (X_n - \bar{X}_n) + (\epsilon_n - \bar{\epsilon})$

In order to fine the R2 we square the sum of the difference of one or more of our independent “x” inputs $(X_n - \bar{X}_n)^2$ and divide it by the squared sum of the difference of our “y” output $(Y_i - \bar{Y}_1)^2$

$$(1.1) \quad R^2 = \frac{SSE}{SST} \left(= \frac{b^2 \sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} \right)$$

Note: $0 \leq R^2 \leq 1$

SSE – stands sum of squared errors - errors is our difference as mentioned before.

SST – stands for total sum of squares.

For this model to be consistent, some assumptions about our data must be made and checked for.

- (1) $E(\varepsilon_n) = 0$ - The expected coefficient of our error term must be zero. This indicates little difference between our predicted linear function and the actual linear function that is unknown.
- (2) $\text{Var}(\varepsilon_n) = \text{Standard deviation} < \infty$ - The inconsistency in our data should be constant and never ending for all the data. Remember: “Var” stands for the variance in our data and the standard deviation is the variance squared.
- (3) $\text{Cov}(\varepsilon_1), (\varepsilon_2) = 0$ - Covariance between the errors must be non-existent. In other words, the joint variability between errors should not correlate with each other in any ways.
- (4) $\varepsilon_n \approx \text{normal}$ – The error terms should be normally distributed in our data. This is important because it lets us perform statistical hypothesis testing. Something we will do at the final stages of our regression model.

Take note that these assumptions are also congruent to our analytical theory and method analysis.

2.6 Hypothesis testing.

Sometimes, we are more concerned about whether some parameter is similar or is different from a given value (R L Thomas, 2005). For concerns such as these, we would need benchmarks. Statistical hypothesis testing plays the role of a confirmatory benchmark, where we start with a null hypothesis, often represented as “ H_0 ”. The null hypothesis is a representation of a given value that is known and accepted by us, or other researchers. This null hypothesis can only be rejected, if the probability from our statistical analysis falls above, below, or outside a predetermined significance value.

There are multiple predetermined significance levels, as there are often written on significance tables like the T-table, F-table and many more. It is also possible for researcher to create their own specific predetermined significant values. However, we will be using the standard T-table and F-table for this paper.

When a statistical analysis differs from our H_0 hypothesis, we represent the alternate value as H_A .

3. REVIEWING AND CLASSIFYING OUR COLLECTED DATA.

Our data set is taken from the “PIRLS” data set. PIRLS stands for Progress in the International Reading Literacy Study. This is data gotten from evaluating the educational achievements of pupils and students. And has been a regular evaluation since 1958. Norway participated in its very first evaluation in the year 1991. However, our country of choice for our thesis is New Zealand.

The group of countries available from our current data, PIRLS 2001 (International Reading Literacy Study, 2001) is as observed in the picture to the right.

Argentina	Italia	Russland
Belize	Kuwait	Singapore
Bulgaria	Kypros	Skottland
Canada	Latvia	Slovakia
Colombia	Litauen	Slovenia
England	Makedonia	Sverige
Frankrike	Marokko	Tsjekia
Hellas	Moldova	Tyrkia
Hong Kong	Nederland	Tyskland
Iran	New Zeeland	Ungarn
Island	Norge	USA
Israel	Romania	

The data contains around 150 000 pupils from an approximate number of 5777 schools from 35 countries. In addition, parents, schools and teachers are given forms to fill. These forms

includes questions regarding hobbies, activities, and few information about teachers educational, parents income and education.

IEA – International association for the evaluation of educational achievement is an international and aims to maintain fairness across the various participating countries and schools. They administer these tests for major subjects like mathematics, science and reading. Reading score is our “y” dependent output and the output that we hope to analyze and estimate for this paper’s thesis. And the country New Zealand as our primary and unchanging parameter for all our independent “x” inputs. For the “x” inputs, there will be more parameters, secondary parameters, and tertiary parameters. These two other parameters will be showcases below as primary and secondary to make classification easier.

Classifying our data.

Test score – our “y” output – in other words, our dependent variable.

Read	Test score	Continuous variable.
------	------------	----------------------

Our main “x” inputs – In other words, our independent variables.

Parents education	Parent have a university degree = 1 all else = 0
Teachers education	Teacher has a university degree = 1 all else = 0

Teachers have a university degree shows high collinearity. As this is almost obligatory at every decent school. Therefore, it is omitted, and parent education is hence our main “x” input variable. Teachers educational will be replace with control “x” variable “teacher’s certificate” for the remainder of the paper.

Control “x” input variables rated as primary – meaning potentially important in our regression test. Rated secondary – meaning potentially less important in our regression test. Remember that our conclusion may change post our regression test. Control variables help narrow our regression model, while reducing the negative effect of omitted “x” variables. Remember the importance and risk of omitted variable from (2.4)

Primary “x” Inputs (independent variables)	Secondary “x” inputs (independent variables)
Born in country - Not_born	Same teacher for 4 years – samteacher_4plus
Annual income - income	Parents employment habits – Par_emp
Parents education – Par_edu	Teacher certification - teacher_cert
Teacher experience – teacher_exp	Parents not born in country – par_not_born
Students economical state – pct_disadv	Teachers sex – teacher_fem

Teachers education level – teacher_educ	Teachers age – teacher_age
Class size – clsz	Girl – sex of the pupils.

Important note:

The classification of our independent variables into primary and secondary are pre-regression analysis. This means that some of our primary variables may prove to be insignificant for our thesis. And some of our secondary variables may prove to be particularly important after our regression analysis.

Some of these variables also may prove to be highly correlated with each other, and therefore, will be omitted from our regression analysis. Potential possibilities are annual income and students economical state, teacher’s experience and teacher’s age, or same teacher for 4 years.

3.1 Data descriptive analysis.

Revision - from the analytical theory and explanation.

Remember:

$$(1) E(y) = b_0 + \beta x_1 + \dots + \beta x_n$$

....., n – stands for all potential “x” input (independent variable) that we may use.

E(y) - stands for our expected function line, not our actual line. This expected function line is created from the regression software’s calculated “mean” of dependent and independent variables divided by each chosen dependent variable mean.

Summary variables:

Variable	Obs	Mean	Std. Dev.	Min	Max
read	2,504	530.3829	91.20145	206.7324	763.358
teacher_age	2,427	3.487845	1.224979	1	6
not_born	2,385	.1832285	.3869349	0	1
par_not_born	2,286	.1951006	.3963648	0	1
par_edu	2,037	2.08002	1.029515	1	5
income	1,890	4.428571	1.698387	1	6
teacher_exp	2,386	13.79799	10.04958	1	40
teacher_cert	2,409	.9983396	.0407231	0	1
teacher_edu	2,405	0	0	0	0
clsiz	2,399	28.02209	4.911469	2	48
pct_disadv	2,372	1.932125	1.09792	1	4
par_emp	1,910	1.871728	.754096	1	4
sameteach	2,379	.0163934	.1270098	0	1
teacher_fem	2,431	.7445496	.4362038	0	1
girl	2,476	.4830372	.4998131	0	1

- We have over a thousand observations on all our chosen variables. It is important to pay close attention to the means and standard deviations of our chosen variables in relation to it selves. The mean above does not tell us the mean of the “y” and “x” variables in relation to every other variable, but to itself only – **Check appendix for relation to others.** The standard deviation stands for the average deviation of our observations from its own mean.

Important note:

Variables with low min and max are often categorical variables or dummy variables. Categorical variables will all be converted into dummy variables. Dummy variables are continuous variables, like our “Y” variable test score “read”.

Correlation table:

Remember that all dependent variables should be correlated to the independent “y” output, but less correlated to each other. Therefore, observe the correlation table below:

	read	teacher_age	not_born	par_not_born	par_edu	income	teacher_exp	teacher_cert	teacher_edu	cls_size	pct_disadv	par_emp	same_teacher_us	teacher_fem	girl
read	1.0000														
teacher_age	-0.0545	1.0000													
not_born	-0.0058	-0.0419	1.0000												
par_not_born	0.0478	-0.0117	0.4692	1.0000											
par_edu	-0.3391	0.0592	-0.1507	-0.1565	1.0000										
income	0.2970	-0.0824	0.0014	-0.0149	-0.3782	1.0000									
teacher_exp	-0.0193	0.7982	-0.0462	-0.0217	0.0198	-0.0388	1.0000								
teacher_cert	1.0000							
teacher_edu	1.0000						
cls_size	0.0556	-0.0855	0.0772	0.0417	-0.0373	0.0094	-0.0148	.	.	1.0000					
pct_disadv	-0.3083	0.1778	-0.0507	-0.0231	0.2670	-0.3044	0.0488	.	.	-0.1120	1.0000				
par_emp	-0.0519	0.0152	0.0069	-0.0055	0.0676	-0.3248	0.0015	.	.	-0.0095	0.0462	1.0000			
same_teacher_us	-0.0701	0.0094	-0.0318	-0.0341	-0.0204	-0.0150	0.0110	.	.	-0.0086	0.0709	-0.0041	1.0000		
teacher_fem	0.0595	-0.0830	0.0717	0.0827	-0.0137	0.0376	-0.1139	.	.	-0.0753	-0.0138	0.0047	0.0039	1.0000	
girl	0.1856	-0.0090	-0.0053	0.0120	0.0022	-0.0358	-0.0254	.	.	0.0240	0.0115	0.0330	0.0010	0.0305	1.0000

0.3 and below - Signifies a weak correlation

0.5 – signifies a moderate correlation

0.7 and above – signifies a strong correlation

The numbers with a negative symbol (-) in front signifies a downward correlation. We aim to pick for the same regression models, independent “x” variables that are moderately or have little correlation with each other. However, when using dummy “x” variables, the correlation table does not properly demonstrate the relationship between variables. Therefore, the correlation table above, has no significant value for our thesis.

Important note:

We hope to predict a line that is remarkably close to our expected line. But with the addition our error margin “ε”

$$(2) Y = b_0 + \beta x_i + \dots + n + \epsilon$$

Our constant terms, b_0 , $\beta(x_i)$ and ϵ are all estimates of our true population terms, which are unknown. However, this means that the distributions of our “y” variables in relation to our predictors (“x” variables) taken from our sample (Reading test score, New Zealand, PIRLS data) is particularly important. **Check appendix for a visual assessment.**

4. THE REGRESSION ANALYSIS.

Categorical variables converted to dummy variable for our regression analysis:

The old variables	The new dummy variables.
income	Income30000plus - All parent annual income above 30000 dollars.
Par_edu	Par_eduhigh – All parent with university level education
Par_emp	Par_emphalf – One parent works full time.
Teacher_age	Teacher_agehigh – All teacher above the age of 40.
Pct_disadv	Pct_disadvhigh – Percent of homes that are considered poor.

Revision – Our thesis:

In New Zealand, how does the educational pedigree of the pupil’s guardians affect the pupil’s test scores in reading?

- We consider the pupils teacher and parents as guardians for the pupil.

4.1 An overview of our chosen regression models for testing.

$$(i) \text{Read} = b_0 + b_1 \text{par_eduhigh} + b_2 \text{teacher_cert} + b_3 \text{par_not_born} + b_4 \text{not_born} + b_5 \text{income30000plus} + b_6 \text{par_emphalf} + b_7 \text{teachers_agehigh} + b_8 \text{pct_disadvhigh} + b_9 \text{teacher_exp} + b_{10} \text{teacher_edu} + b_{11} \text{sameteacher4_plus} + b_{12} \text{clsiz} + b_{12} \text{teacher_fem} + b_{13} \text{girl}$$

This has all the independent “x” variables that we hope to test for on our regression analysis.

When we do the regression on the Stata software, the error term is registered as zero. It is still in the data, but not visually displayed.

To decide on the most significant “x” inputs to use for further analysis, we do a T- statistical test and then check for its probability value. The probability value is often abbreviated as the P-value”.

Formula:

$$(6.1) \quad t\text{-statistic} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

(R.L Thomas, 2005, s. 178)

Note:

$$\bar{X} = \sum X_n/n$$

μ – Is the predicted population mean.

n – Number of observations – Often, we write n minus the regression's number of degrees of freedom, when the predicted population mean is considered unknown for a t-test.

S - sample variance

Degrees of freedom:

Example: $(Y_i - \bar{Y})$

The formula also means $-(Y_1 - \bar{Y}) + (Y_2 - \bar{Y}) + (Y_3 - \bar{Y}) = 0$

Therefore, if $(Y_1 - \bar{Y}) = 2$ and $(Y_2 - \bar{Y}) = -6$ Then $(Y_3 - \bar{Y})$ must be $+4$. (Thomas, 2005, s 175)

Important note:

We use the above method. However, this is done in the statistical software program, Stata. When the t-value is calculated, we check for the probability value of our t-test on a t-table. We are using the significance level of 0.05 percent. What this means is that: given that our null hypothesis is true – In this case, our null hypothesis is always that the estimated coefficient obtained as the parameter for a given “x” independent input is likely due to random probability. And not because the x independent input is significantly correlated with our “y” dependent output - Therefore, on a “x” input were our null hypothesis is true, we accept that the “x”

independent input is insignificant. In a situation, were our null hypothesis is false – we accept the estimated parameter to be bot due to random probability. But, due to the significance of our “x” independent input been correlated with our “y” dependent output – Therefore, we reject the null Hypothesis and conclude that the “x” independent variable is important for our regression model.

Using the above explanation, we have our second regression model, were “x” variables with low p-value are omitted. Also, the “x” variable “teacher_edu” - teachers education was automatically omitted due to its high collinearity with other variables in our regression.

$$(ii) \text{ Read} = b_0 + b_1\text{par_eduhigh} + b_2\text{teacher_cert} + b_4\text{not_born} + b_5\text{income30000plus} + b_6\text{par_emphalf} + b_8\text{pct_disadvhigh} + b_{11}\text{sameteacher4_plus} + b_{12}\text{clsiz} + b_{13}\text{girl}$$

Teacher is female was deemed insignificant due to low p-value. As well as teacher’s age, parents not born in country, and teacher’s experience.

Remember that we are analyzing how a guardian’s pedigree affects pupil’s performance. And currently, experience, age and guardian’s sex are proven to be insignificant. Whether the pupil is a girl or boy proves to be significant. However, I would like to know how much is explained by parent’s education and teacher’s certification, only. Therefore, our third regression for this paper. Check (4.2) for result from the regressions.

$$(iii) \text{ Read} = b_0 + b_1\text{par_eduhigh} + b_2\text{teacher_cert}$$

Both independent variable, parent education and teachers’ certificate, both proves to be significant with a high f-value, t-value, and significant p-value. However, the model has a R^2 at 0.09 compared to my current best at 0.19 R^2 , regression model (ii). I would like to check for

how much more significant my two main “x” independent variables are, especially when in conjunction with some of my control variables. Therefore, I multiply Parents education with income3000plus – This tells me how much more effect I can get if the guardian has a high educational pedigree and is financially successful as well. I also multiply Teacher certificate with the negative impairing control variable “clsiz” – This tells me the effect of a certified teacher in relation to a class size. Both these multiplications generates an interaction term in Stata – New “x” independent variable inputs – parent’s success and teacher’s influence and therefore, my fourth regression model below:

Note: The F-value in Stata tells the joint significance of a regression model and the T-value tells the singular significance of my “x” inputs.

$$(iv) \text{ Read} = b_0 + b_1\text{par_eduhigh} + b_2\text{teacher_cert} + b_4\text{not_born} + b_5\text{income30000plus} + b_6\text{par_emphalf} + b_8\text{pct_disadvhigh} + b_{11}\text{sameteacher4_plus} + b_{12}\text{clsiz} + b_{13}\text{girl} + b_{14}\text{par_succ} + b_{15}\text{teacher_influnce}$$

4.2 The regression table:

VARIABLES	(1) m1 read	(2) m2 read	(3) m3 read	(4) m4 read
par_eduhigh	40.578 (4.150)	42.943 (3.999)	52.776 (3.889)	59.366 (12.757)
teacher_cert	199.022 (38.518)	191.786 (38.926)	178.106 (41.920)	
par_not_born	9.527 (5.705)			
not_born	-16.390 (5.803)	-16.808 (5.056)		-16.665 (5.056)
income30000plus	28.348 (5.838)	24.602 (5.747)		28.917 (6.569)
par_emphalf	11.524 (4.024)	12.336 (3.910)		12.364 (3.909)
teachers_agehigh	-2.407 (5.871)			
pct_disadvhigh	-42.303 (4.683)	-39.357 (4.573)		-39.135 (4.575)

teacher_exp	0.188			
	(0.288)			
o.teacher_edu	-			
sameteacher4_plus	-36.150	-40.173		-40.344
	(17.522)	(17.380)		(17.376)
clsize	0.722	1.208		-9.508
	(0.391)	(0.379)		(2.231)
teacher_fem	7.571			
	(4.546)			
girl	32.343	31.486		31.358
	(3.881)	(3.790)		(3.791)
o.teacher_cert				-
par_succ				-18.130
				(13.374)
teacher_influnce				10.698
				(2.162)
Constant	270.432	271.975	341.939	460.654
	(39.638)	(39.688)	(41.851)	(12.758)
Observations	1,526	1,659	1,966	1,659
R-squared	0.215	0.199	0.095	0.200

Standard errors in parentheses

4.3 Empirical result

All four regressions have an f-probability value greater than the f-table's predetermined benchmark level. And are therefore significant. Regression m1 takes all our independent "x" variables and regression m2 removes all "x" variables that are shown to singularly contribute insignificantly to the overall regression model. Regression m3 shows our two main "x" variables.

As hypothesized, there is a positive increase of 40 and above in test score when the parent of the pupil has a university degree. Moreover, an increase at 178 and above when the teacher has a teacher certificate, specific for teaching. The reason why the effect of parent's education on pupil's test score has variations from 40 to 52.776 is because of omitted bias variables. This effect is likewise on the variable teacher's certificate as well. The workings of omitted bias variable is explained at (2.4). In regression m4, we test for interaction terms by multiplying "x" variables together.

To calculate the effect from these interaction terms, we follow this pattern of calculation:

$$\text{Par_eduhigh} + \text{income30000plus} - \text{Par_succ} = 59 + 28 - 18 = 69.$$

This means that the extra parameter effect from a parent having a university degree while also been financially successful equates to a 69 increase in test score for the pupil who has that parent as a guardian. However, the p-value on this result is deemed insignificant, and due to random chance by the Stata software. This can be because our sample data is not the most appropriate for this research, or maybe it is insignificant.

For teacher's influence

$$\text{Teacher certificate} - \text{class size} + \text{Teacher influence} = 0 - 9 + 10 = 1$$

Every unit increase of the class size leads to a - 9 decrease in test score. However, the influence from a teacher with a teacher certificate adds an additional 1 unit increase in test score per every unit increase of a class size. Parent with high education and or incomes have a significant positive effect on pupil's test score in reading. And pupils from economical disadvantageous background have a significant negative test score in reading. An addition, and as well an interesting observation, is the negative effect that come from pupils having the same teacher for 4 or more years. As observed earlier, there is also a negative effect from age and experience. The reason for this is not something we can deduct from this paper's data analysis, thesis, and hypothesis.

Important note:

When we create interaction terms by multiplying variables, some data are omitted from the original data. This means when we multiply parent education with parent with high income, then only parent with a university degree but low income remains in our original "parent_education_high" variable. Therefore, we have a higher coefficient 59 in regression m4 than in regression m2 "42" and m1 "40". This the same for every other variables included in an interaction term.

5. CONCLUSION

Writing this paper has been challenging. Not necessarily from a theoretical approach. But much more from a practical approach. The practical approach on the regression analysis had on several occasions made me rethink my theoretical understanding of the subject. Hence leading me to conclude that there is more that could be done from a practical approach to better understand the sample data. This fault may not only stem from my limited understanding of the method of regression analysis, but also in the way the data may have been collected. As mentioned earlier in (1.1) and (1.4).

Moreover, more theoretical understanding may be beyond my current education's syllabus. However, I can at the very least conclude my hypothesis (1.4) to have been true at a 0.05 percent significance level.

$$H_0 = 0$$

$$H_A \neq 0$$

Answer = we can safely reject our null hypothesis "H₀" at a 5 % significance level.

Regarding my thesis - In New Zealand, how does the educational pedigree of the pupil's guardians affect the pupil's test scores in reading?

- And if yes, that it does affect the pupil's test score, how particularly significant is this effect?

The answer is yes, it is highly significant and at an explanation weight of minimum 9 percent to a maximum at 20 percent of the entire pupil's test score in reading – given our sample data from PIRL, 2001. I considered it a major effect to think that we can predict about 20 percent of the reason most pupils do considerably well at their reading test scores is because their guardians are either highly well-educated and or economically successful in life.

I have taken an expository approach to writing this paper. And from a deductive perspective based from my own individual understanding of the sum of least squares regression method of analysis.

REFERENCES

Assessment - PIRL "progress in international reading literacy study" from 2001 by the Norwegian "Senter for leseforskning" (Solheim Tønnesen, report on PIRLS, 2001)

Rivkin, S.G, E.A. Hanushek og J.F. Kain (2001): «Teachers, schools and academic achievement. » NBER Working Paper no 6691

R L Thomas, using statistics in economics, (2005). Mcgraw-Hill.

Bonesronning, H (2004): utforming av utdanningspolitikken – kva kan økonomene bidra med? Økonomisk forum 58 (3), 14-23 - <https://samfunnsokonomene.no/wp-content/uploads/2019/05/Trykkutgave-4-2012.pdf>

IEA - International Association for the Evaluation of Educational Achievement - <https://www.iea.nl/about>

Hanushek, E. A. (2020): Education production functions. I Bradley, S. og Green, C. (red): Economics of Education, 2nd Edition, London: Academic Press, 161-170.
<http://hanushek.stanford.edu/sites/default/files/publications/Hanushek%202020%20Education%20Production%20Functions.pdf>

Coleman, J.S., E.Q. Campbell, C.J. Hobson, J. McPartland, A.M.Mood, F.D. Weinfeld og R.L. York (1966): «Equality of Educational Opportunity.» U.S. Government Printing Office, Washington D.C.

New Zealand wealth distribution.
<http://archive.stats.govt.nz/Census/2006CensusHomePage/QuickStats/quickstats-about-a-subject/incomes/personal-income.aspx>

Norway wealth distribution - <https://www.ssb.no/en/inntekt-og-forbruk/artikler-og-publikasjoner/wealth-distribution-in-norway>

APPENDIX