A STATISTICAL ENQUIRY INTO THE FACTORS AFFECTING IRISH SECONDARY SCHOOL PERFORMANCE.

ALEXANDER REDMOND
Senior Sophister

Through the use of a comprehensive dataset provided by The Irish Times, Alexander Redmond analyses the impact of attending a fee paying school on the prospects of third level education. A further econometric analysis on the relationship between fee size and third level attendance is then conducted, with surprising results. This project is particularly timely given the current debate surrounding the government funding of fee-paying schools.

Introduction

The Leaving Certificate is Ireland's national final exam for the secondary school system. Students' leaving certificate results determine their matriculation into third level education in Ireland. Many believe that the type of school a student attends will affect whether or not they continue into third level education. In this paper I will attempt to examine the characteristics of Irish schools which determine the number of students continuing on to third level education. I have considered many variables that could affect matriculation, and have chosen a select few based on the availability of data and their economic implications. My research is motivated by my own experience with the Leaving Certificate as a pupil in a private school. There is much public speculation about whether or not attending a private school improves students' academic performance. I wish to examine potential statistical correlations that may be present regarding school performance. Finally I think that this topic is of great relevance in Ireland today. There is currently a debate about whether or not public finances should be given to private schools. During times of tightening government budgets, questions of equity arise, questions which are certainly relevant in the context of education. I will attempt to identify correlations, with an emphasis on fee-paying schools in particular.

Theoretical Model

I have decided to use a multiple regression model to describe the correlated effects of secondary school characteristics on school performance. I have taken "performance" to

mean academic performance of the schools' pupils. The continuation of pupils on to third level education is hence my dependent variable. For my independent variables I had originally chosen a variety of characteristics that I considered important, but due to difficulty in acquiring data, some are omitted. This will be discussed in section 3. Due to the presence of fee-paying schools in Ireland I have decided to create two separate models. Both contain similar dependent variables as mentioned above. I will outline the differences in my models below:

A. Complete Model:

$$thirdlevel = \beta 0 + \beta 1\ fee + \beta 2\ class + \beta 3\ class2 + \beta 4\ urban + ut$$

Thirdlevel: This variable equals the percentage of the senior class to follow on to third level education after completing their Leaving Certificate.

$\beta 0$: This is the intercept value of the regression.

fee: A dummy variable to classify a school as fee-paying. Its associated parameter is $\beta 1$.

class: This is the number of students in the 6th (final) year class at the school. Its associated parameter is $\beta 2$.

class2: This is the number of students in the 6th year class squared. I suspect there may be a non-linear relationship and so I want to investigate this. Its associated parameter is $\beta 3$.

urban: This is a dummy variable identifying whether or not the school is located within a major city. Its associated parameter is $\beta 4$.

u: This is the error term/disturbance as is typical standard with in regression models.

B. Fee-Paying Model:

$$thirdlevel = \beta\_0 + \beta\_1\ fee + \beta\_2\ class + \beta\_3\ class2 + \beta\_4\ urban + u$$

thirdlevel: See above.
fee: This is the annual fee charged by the school. Its associated paramter is $\beta 1$.
class: See above.
class2: See above.
urban: See above.
u: See above.

Dataset

In order to conduct my analysis I used already existing data, along with my own research, to order create the dataset used. A sample of 678 fee-paying and public schools in the country has been included for Model A's regression. A sample of 44 fee-paying schools has been used for Model B's regression. These were reduced from an original data set of 685 secondary schools, representing the entire country. This complete data set included 630 public schools, 54 fee-paying schools, and a single grind school. The reduction in the sample size is due to difficulties in obtaining information from a handfull of fee-paying, private schools that refused to disclose their annual cost for my research. I decided to remove these schools from both Model A and Model B in order to make comparisons more accurate.

The dataset is cross-sectional, with data relevant to the 2011 academic year. After explaining my research to Liam Stebbings at The Irish Times, he graciously agreed to provide me their annual Secondary School League Table data. The author thanks Sean Flynn, Peter McGuire, and Mick Crowley for compiling an invaluable dataset. This data included a list of all secondary schools in the country, along with the number of students in the secondary school class, and the percentage of the senior class that went onto third level. These preliminary numbers allowed me to build my own dataset on top of this.

I conducted my research into both fee-paying schools and public schools. The majority of the fee-paying schools had a website, with approximately 60 per cent of them providing the public with their fee policy, including annual cost. I contacted the other 40 per cent of schools directly, explained my research, and requested their fee policy. This led to a reduction in my sample size, as explained above. In order to get the data for the urban variable I used the addresses of schools located in Dublin, Cork, and Limerick. Within this subset I had to differentiate between the county and the city. If the school was located in the city limits it was considered Urban and was given a binary value of 1. In Model A, I simply assigned the fee-paying schools a value of 1. I also scaled the thirdlevel variable for ease of interpretation of coefficients.

Descriptive Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
thirdlevel2	678	7.530206	2.171703	.63	10
fee	678	.0663717	.2491145	0	1
class	678	71.72566	40.84663	5	257
class2	678	6810.558	7509.744	25	66049
urban	678	.2507375	.4337576	0	1

Table presents the output of Model A

Variable	Obs	Mean	Std. Dev.	Min	Max
thirdlevel2	44	9.284318	1.149357	4.56	10
fee	44	4743.159	1703.358	2550	12456
class	44	80.65909	41.59939	6	196
class2	44	8197.068	7889.397	36	38416
urban	44	.5681818	.501056	0	1

Table presents the output of Model B

Empirical Model

In my multiple regressions of both models, I started with a simple OLS estimation as a baseline. I then decided to conduct a Tobit regression, due to the nature of my dependent variable. I transformed my dependent variable from a percentage into a decimalized interger between 0 and 10. This allowed me to run a Tobit regression with a lower limit of value 0 and an upper limit of value 10.

Model A

Variable	Obs	Mean	Std. Dev.	Min	Max
thirdlevel2	44	9.284318	1.149357	4.56	10
fee	44	4743.159	1703.358	2550	12456
class	44	80.65909	41.59939	6	196
class2	44	8197.068	7889.397	36	38416
urban	44	.5681818	.501056	0	1

Table presents the output of Model A (augmented)

fee: Our dummy variable fee is positively related to our dependent variable thirdlevel. This agrees with economic intuition. Fee paying schools are generally expected to perform better than their public counterparts. A strongly significant t-score of 7.44 is welcomed.

class: Our class independent variable is also positively correlated with thirdlevel. Again this can be argued to agree with conventional wisdom. Very small senior classes are generally located in remote areas that may be lacking in resources and labour supply of teachers. Of course this is all speculative. It should be noted that there exist some studies which find the opposite relationship between class size and performance (Leithwood, 2009). It is also statistically significant, with a t-score of 4.52.

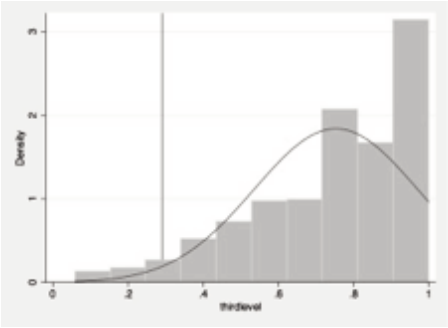
class2: This is a worrying result. The estimation suggests that a non-linear relationship between class size and performance is negative.

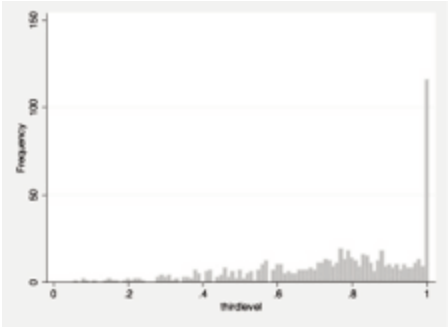
urban: Our dummy variable urban is also statistically significant, with a t-score of -9.85. It has a negative relationship with thirdlevel, suggesting an urban environment may not be conducive to strong academic results.

R2: Our R-squared value of 0.2051 tells us that 20% of the variation in thirdlevel is explained by our model.

The estimation and results above are a simple OLS regression and, as discussed earlier, are not suitable for the data used, but good for an initial test of the data. I will now give with my empirical reasoning for choosing a Tobit regression and the results for Model A.

As we can see there is a large proportion of data points skewed to the right in this normalized histogram of thirdlevel. One can see that the natural censoring of data above the value 1.





If we view the histogram with all values represented we get a similar result. Again we see a large number of thirdlevel values being 1. Both the normally distributed and discrete histograms highlight the need to use a Tobit regression. We can see the results of the regression below.

Tobit regression — Number of obs = 678; LR chi2(4) = 141.03; Prob > chi2 = 0.0000; Log likelihood = −1368.2123; Pseudo R2 = 0.0490

thirdlevel2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fee	3.014685	.3867372	7.80	0.000	2.255331	3.77404
class	.0222193	.0067819	3.28	0.001	.008903	.0355355
class2	−.0000764	.0000368	−2.08	0.038	−.0001487	−4.14e−06
urban	−1.9011	.2057895	−9.24	0.000	−2.305165	−1.497034
_cons	6.955577	.2835246	24.53	0.000	6.39888	7.512275
/sigma	2.24089	.0695396			2.10435	2.37743

Obs. summary: 0 left-censored observations
562 uncensored observations
116 right-censored observations at thirdlevel2>=10

As the table above shows us, the Tobit regression censored 116 observations at 100% third-level attendance.

fee: This states that if a school is fee-paying then it will have three times as many students going on to third level as a public school. This result is also strongly statistically significant at the 1 per cent level.

class: This value states that a one unit increase in the final year class size will lead to a 2 per cent increase in the predicted value of thirdlevel attendace. Again this is statistcially significant.

class2: Just like the OLS, this variable is throwing out strange results. It is negatively correlated, and is not as statistically significant as the other variables.

urban: This variable is negatively correlated with thirdlevel, and is highly statistcially significant. It states that if a school is located in an urban area, it is going to have 1.9 times fewer students attending third level. This is a worrying result for economic policy in an increasingly urbanized world.

Model B

The results for Model B will be displayed and examined below. The dramatically lower sample size and inclusion of specific costs within fee-paying schools are to be noted. Our OLS gives us back some strange and conflicting results compared to our full-sample regression including public schools.

Source	SS	df	MS	Number of obs =	44
				F(4, 39) =	6.51
Model	22.7490801	4	5.68727004	Prob > F =	0.0004
Residual	34.0548013	39	.873200033	R-squared =	0.4005
				Adj R-squared =	0.3390
Total	56.8038814	43	1.3210205	Root MSE =	.93445

thirdlevel2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fee	-.0004112	.0000847	-4.85	0.000	-.0005826	-.0002399
class	-.0036048	.0112112	-0.32	0.750	-.0262816	.0190719
class2	.0000578	.0000594	0.97	0.336	-.0000622	.0001779
urban	.014689	.288496	0.05	0.960	-.5688492	.5982271
_cons	11.0431	.622616	17.74	0.000	9.783736	12.30246

fee: Interestingly, there is a statistically significant negative correlation between the annual cost and the thirdlevel attendance. It has a t-score of -4.85.

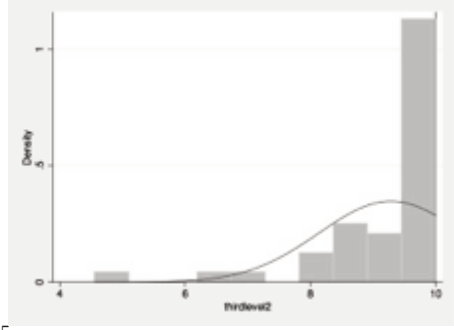
class: This is result is not statistically significant, and we can conclude that the coefficient is not statistically different from zero.

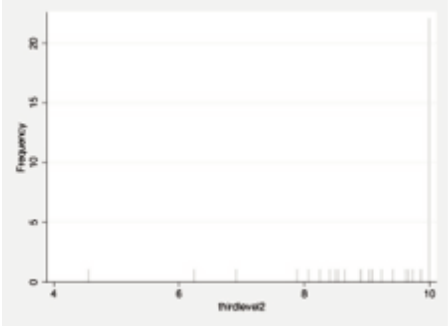
class2: This result is the same as that for class, although not as strongly statistcially insignificant.

urban: Again this is highly statistically insignificant.

R2: Our R-squared value is 0.40, which tells us 40 per cent of the variation in our dependent variable is explained by our independent variables.

As with Model A, I will now conduct a Tobit regression with upper and lower limits for my dependent variables, and present the histograms illustrating the censoring.





This too shows a highly skewed dataset, with a majority of my dependent variable results being at 10. We can see this using a discrete frequency as with Model A. The small sample of fee paying schools may be throwing off the results in this case. There are clearly many more values of thirdlevel equalling 10 than anywhere else. Coupling this fact with the sample size of 44 may help to explain our strange estimates. The Tobit results are outlined below.

```
Tobit regression                                  Number of obs     =         44
                                                  LR chi2(4)        =      14.82
                                                  Prob > chi2       =     0.0051
Log likelihood = -53.147534                       Pseudo R2         =     0.1224
```

thirdlevel2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fee	-.0005213	.0001407	-3.70	0.001	-.0008057	-.0002369
class	-.0239604	.0230203	-1.04	0.304	-.0704862	.0225654
class2	.0001922	.000132	1.46	0.153	-.0000746	.000459
urban	.0539591	.5029278	0.11	0.915	-.9624959	1.070414
_cons	12.73769	1.185536	10.74	0.000	10.34163	15.13375
/sigma	1.454409	.2411399			.9670472	1.941771

```
  Obs. summary:          0  left-censored observations
                        22     uncensored observations
                        22 right-censored observations at thirdlevel2>=10
```

Our results show us that 22 observations were censored at thirdlevel value of 10 – which corresponds with 100 per cent third level throughput.

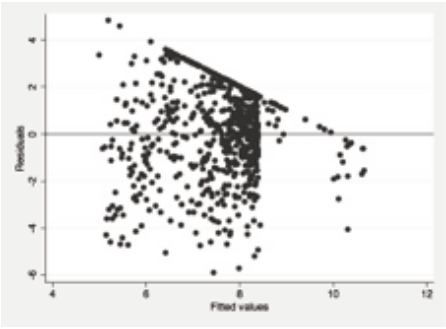
fee: We see that fee, as with the OLS estimate, is negatively correlated and statistically significant with our dependent variable.

class: class is again not statistically significant in this model. There is a 30 per cent chance that our coefficient is equal to zero and thus fails to reject our null hypothesis.

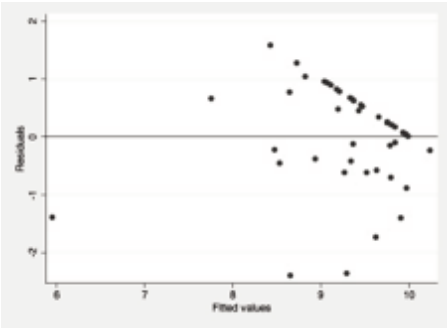
class2: class2 is the same as class however is less insignificant.
urban: This variable is highly insignificant in our model, with an extremely low t-score.

Diagnosis check

A major assumption of OLS is that of homoscedasticity of variance in the residuals. I feel a test for this is necessary in identifying possible problems in applying these statistical models onto data that may or may not fit the assumptions. In order to test for constant variance between the residuals we can plot them against the fitted values. We see this for both models below.



Model A



Model B

These results are very worrying for our assumption of homoscedasticity, in our OLS estimates at least. It is clear that the residuals in both Model A and Model B are not scattered randomly and evenly; there is a clear pattern. This means that the residuals are not homoscedastic and highlights a possible heteroscedastic problem for our regression. This is especially worrying in my Tobit regressions, as the results attained are achieved by using a false estimate of the distribution of residuals – which determine whether or not to censor a value. Our coefficients, therefore, may be biased. This highlights the caution needed in running OLS and Tobit regressions, and a caveat must be stated about interpretting possibly incorrect results.

Results

The statistical insignificance in Model B compared to Model A is disappointing. However, there are still significant results from both OLS and Tobit regressions for both models. The strong correlation between fee and thirdlevel throughput in Model A is not surprising, but it is a welcome result. This reinforces the notion of fee-paying schools achieving generally better academic results then their public counterparts. Supporting evidence of this relationship has also been found in the Australian secondary school market (Vella, 1999). This may be due to selection-bias of students rather than the school themselves, as fee-paying schools sometimes require an entrance examination to attend. This is called 'creaming', and empirical research exists on the subject (West, 2006). One also cannot deny the additional resources available to fee-paying school students due to the larger budgets expected. The ability to afford a fee-paying secondary school will also be correlated with other supports like grinds and weekend classes. The opposing results for class and class2 are worrying. On the one hand we see a positive relationship for class; however on the other hand class2 is negative. I believe that this attempt to analysis a non-linear relationship has failed, and a more appropriate examination would use a variable of student/teacher ratio. This data was, unfortunately unavailable. An exciting result is the negative correlation

between thirdlevel and urban. Its strong statistical correlation is welcome. Empirical results for underperforming schools in urban areas have been found by other researchers (Lankford, 2002). This is a worrying result for schools located in cities, and may lead to questions of equitable budgeting between urban and non-urban schools.

As stated earlier, the results for Model B were disappointing. The negative correlation between fee and thirdlevel was surprising, particularly the statistical significance. There are a number of possible explanations for this. One is the religious nature of fee-paying schools. Many fee-paying schools give as their motivation for their fees the religious ethos of the school, rather than its academic performance. The schools are not justifying their high costs with better academic performance, and so a positive relationship is perhaps not expected.

Other studies have been carried out in order to identify how school characteristics affect performance, and these should be acknowledged for their results. In particular, there have been conflicting results with regards to private school performance, as well as supporting results with regards to class size (Shulruf, 2008).

Extensions

There is much scope available for future research on this topic. An increase in sample size for the fee-paying schools model is a possible improvement, although obtaining the missing data may be difficult. A panel data approach combining previous years' results would be of most use; however this would require further cooperation by the generous Irish Times staff.

A major disappointment for my own study was the lack of a gender variable due to time constraints of collecting gender information from 650+ schools. The results with such a variable included would have been quite interesting. It is anecdotally said that females perform better than males during secondary school; I would like to see the statistical relationship, if any, and would welcome further research in this area. An analysis of pupil/teacher ratio is another possible avenue of study with education policy implications.

Conclusion

The goal of this study was to identify statistical relationships between the characteristics of secondary schools and their student continuation rates into third level education. I also wanted to examine the differences in performance between fee-paying and public schools, as well as the differences in performance between fee-paying schools themselves, treating them as a subset of the larger sample size. Statistically significant results were found for both models, which tend to agree with economic and intuitive thinking. The results shed light on the apropriateness of sending public-sector money to fee-paying, private schools when they already out-perform their public counterparts. The issue of school location is

also raised. Urban schools underperform compared to non-urban schools. More governmnet spending on public schools in urban areas seems appropriate from an equity viewpoint. This study may help to guide policymakers during this time of debate around school funding.

Acknowledgements

I would like to thank The Irish Times Team: Peter McGuire, Mick Crowley, and in particular the late Sean Flynn, the former education editor, who allowed me to use their data for my project. I would also like to thank the secondary schools I contacted for their cooperation and generosity.

References

McGuire, P., and Crowley M., 2011. Irish Times Feeder School List, November 2011.

Leithwood, K., and Jantzi, D., 2009. A Review of Empirical Evidence about School Size Effects: A Policy Perspective. Review of Educational Research 79, pp.464-490.

Francis, V., 1999. Do Catholic Schools Make a Difference? Evidence from Australia. The Journal of Human Resources 34, pp.208-224.

West, A., 2006. School Choice, Equity and Social Justice: The Case for More Control. British Journal of Education Studies 54, pp.15-33.

Lankford, H, Loeb S., and Wychoff, J., 2002. Teacher Sorting and the Plight of Urban Schools: A Descriptive Analysis. Education Evaluation and Policy Analysis 24, pp.15-32.

Shulruf, B, Hattie, J., and Tumen, S., 2008. Individual and School Factors Affecting Students' Participation and Success in Higher Education. Higher Education 56, pp.613-632.