Cross Section Research Assignment

22876634

2023-06-19

Motivation for Research

There is a belief in economics that an educated person does not only improve their own circumstances, but in large groups these individuals can help their entire community, even the uneducated. This is known as the *social return* to schooling (Angrist & Pischke, 2008:144). Thus, it is important for governments to investigate the drivers of education outcomes. One would hope that education affects all equally capable individuals in a homogeneous manner. This is not the case. There are socioeconomic divides between children, even once they are in the same school, with the same opportunities. This study seeks to argue, and prove, that a powerful driver of the divide in academic outcomes is due to the status of one's home as urban or rural, even once many traditionally influential variables are controlled for. With the awareness of potentially biased OLS estimates discussed at a later juncture, this process will make use of an Instrumental Variable (IV) approach. An unbiased estimate would be of particular interest to policy makers because it can indicate that while funding the creation of adequate schools is important, it is as important to create an environment where all students are equally able to make use of the opportunities a good school might offer.

Data

The data set used for this study is a survey of students from two Portuguese schools. The sampling design employed is Simple Random Sampling. However, it is important to acknowledge that the selection process to enter the schooling system itself is not random. There is a discrepancy between the frequencies of urban and rural children in the sample relative to their true proportion in the Portuguese population. As a result, the target population for this study is specifically school children who are already part of the education system. Any generalizations or extrapolations of the study's findings should be cautiously made in reference to this target population of school children. Furthermore, the two selected schools are considered relatively representative of middle-income schools, suggesting that the sampling frame of these schools can serve as a reasonable microcosm of Portugal as a whole. Therefore, I argue that the target population and sampling frame exhibit similar characteristics, and an econometrician can make valid inferences from this survey.

The measure of academic performance in this survey consists of three separate test scores for mathematics. While there is a time element to these tests, with one being written before the other, these are the only variables exhibiting time dimensions. To account for this, I have chosen to use the average standardised mathematics test score as my measure of academic performance. This decision allows me to focus on the overall performance of students without placing emphasis on specific test instances. I aim to minimize the impact of random errors that a student might experience in a particular test, especially given the relatively small sample size (n = 397). It is important to note that my goal is not to measure 'intelligence' but rather to assess academic performance, which I believe the chosen measurement exhibits construct validity for. One would expect a correlation between mathematics scores and academic performance.

One quarter of the participants in this study reside in rural areas. This is a significant portion that allows for meaningful inferences to be drawn. Though it is worth noting that this proportion is lower than the one third of Portuguese citizens who currently reside in rural regions (European Commission, 2022). The study reveals that more than 80% of the students in the sample have access to the internet, and approximately two thirds of them live within a one-hour travel distance from their respective schools. The gender distribution within the school is balanced, with roughly 50% of

students identifying as male and the remaining 50% as female. To ensure the validity of the findings, tutoring has been considered as a control variable in this study. However, it is worth mentioning that only 13% of students have stated that they receive tutoring.

The primary motivation for this investigation stems from the statistic that children residing in rural areas are 55% more likely to experience academic failure compared to their urban counterparts. It is also noteworthy that urban mothers are 71% more likely to have received tertiary education, implying that this might be a generational trend. These academic disparities between rural and urban contexts have necessitated this investigation, aiming to utilize inferential statistics to gain deeper insights into the factors contributing to these divisions.

Methodology

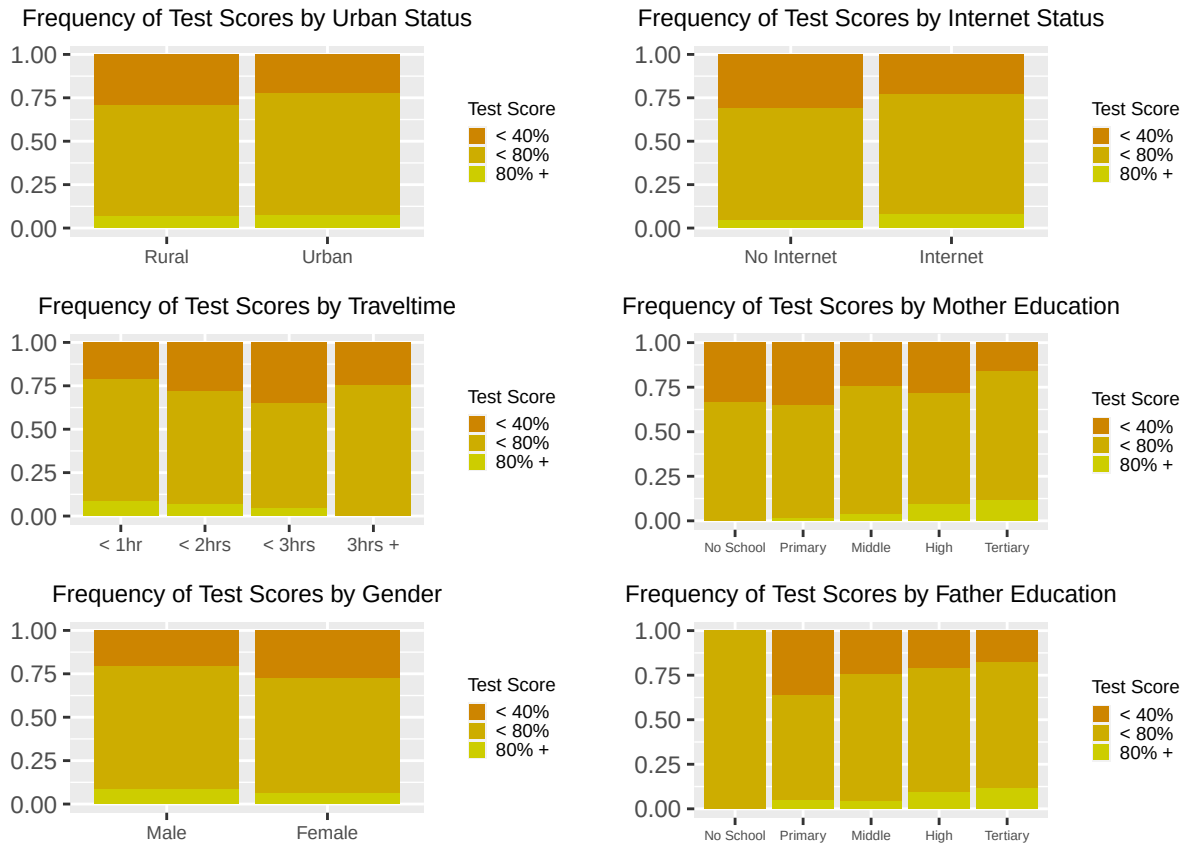
Instrumental Variable Approach

OLS estimations will yield biased results if the condition of exogeneity is not upheld. Therefore, if the treatment effect "Urban Status" is correlated with the error term, it is considered endogenous (Angrist, 1990:318). Endogeneity can arise due to various factors, but in this case, I posit that the endogeneity of this treatment stems from the omission of certain variables in the survey, which may have been impossible to include. These omitted variables exhibit correlations with both the treatment variable and the dependent variable. For example, it could be argued that individuals residing in rural areas may have limited capacity to take advantage of academic opportunities. If the variable "Ability" is not included in the OLS regression, it becomes impossible to ascertain the true effect of "Urban Status". In the absence of "Ability" or any other correlated variables, the measurement of "Urban Status" actually captures a composite of factors that collectively influence the treatment effect, rather than solely reflecting the treatment effect itself (Angrist, 1990:318). Econometricians employ various techniques to address endogeneity. In this study the Two Stage Least Squares (TSLS) method, has been chosen. Internet Access and Travel Time (to school) have been chosen as the instruments for the first stage of the process.

The effectiveness of an instrument relies on two fundamental rules. Firstly, the instrument must be uncorrelated with the error term and have no direct effect on the dependent variable when controlling for the treatment (Angrist, 1990:319). Secondly, the instrument should be strongly correlated with the treatment variable. Fitted values derived from the first stage of the TSLS procedure would lack meaningful interpretation if this correlation were weak. These conditions are known as the Exclusion Restriction and the Rank Condition, respectively. Longer travel time does not inherently affect a child's motivation or ambition, nor does it directly influence test scores. Internet access may be more questionable if one believes online resources significantly impact test scores. However, since the tests in question appear to be standardised and administered in person rather than online, it is less likely that internet access plays a direct role in academic performance. Furthermore, both instruments are strongly correlated with Urban status (at the 99% level). Travel time has a negative correlation, while access to the internet is positively correlated with living in an urban area.

Control Variables

The selection of control variables in this study was meticulously designed to mitigate potential biases that may affect the results. The vector of controls includes fundamental academic indicators, such as study time and the number of prior instances of academic failure before undertaking the standardised assessments. Additionally, the vector contains relevant demographic information, specifically the self-identified gender of the student and the educational attainment level of the student's mother. Moreover, the inclusion of whether the student received tutoring during their educational journey is considered to further enhance the precision of the estimates.



Urban children in the sample exhibit a lower occurrence of receiving marks below 40% relative to their representation in the total sample. The selected instrumental variables, travel time and internet status, demonstrate patterns consistent with the treatment effect. Mother's education is included instead of father's education due to global educational access disparities, so it is more telling when a student's mother has been educated. Furthermore, there is limited variation in lower levels of father education, as seen in the stacked bar charts. The gender disparity in academic performance is evident, with fewer female students excelling and a higher proportion performing poorly in the sample. It will be important to identify whether these relationships remain once the entire vector of controls, along with the treatment, are all estimated together.

Estimation

As can been seen below, all control variables return their expected effect on academic performance. Mother's Education has a statistically significant positive effect on test scores, along with tutoring. The gender education gap is also exemplified by this study, with those identifying as female performing less well than their male counterparts. As expected, having failed a class before, and studying less, would adversely affect test performance. These are all statistically significant at the 95% level or more. The controls are also remarkably consistent between IV and OLS. This could imply that these are their true parameter values. Nevertheless, these factors are not the focus, and as such, more attention in this section is placed on the treatment effect.

Instrumental Variable Regression Results

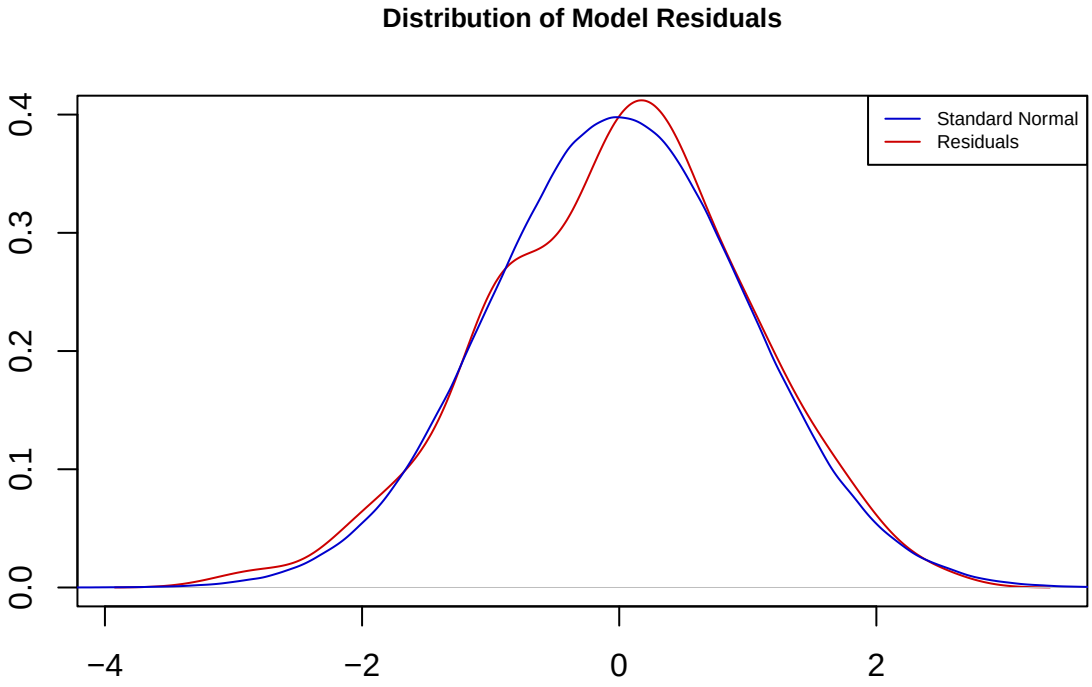
	Test Score (%)			
	Specification 1 (1)	Specification 2 (2)	IV (3)	OLS (4)
Urban	19.746***	11.915**	10.343*	3.438*
	(6.229)	(5.857)	(5.628)	(2.004)
Failures		-7.905***	-7.959***	-8.170***
		(1.229)	(1.190)	(1.162)
Study Time		2.872**	3.540***	3.397***
		(1.111)	(1.145)	(1.123)
Female		-5.805***	-5.021***	-4.716***
		(1.848)	(1.803)	(1.762)
Mother Education		1.524*	2.421***	2.786***
		(0.874)	(0.890)	(0.833)
Tutoring			25.762***	26.440***
			(9.236)	(9.083)
Mother Education x Tutoring			-7.379***	-7.332***
			(2.345)	(2.310)
Study Time x Tutoring			-6.323**	-6.594**
			(3.012)	(2.960)
Constant	37.925***	39.679***	37.642***	42.168***
	(4.930)	(5.132)	(5.013)	(3.592)
Observations	397	397	397	397
R2	-0.097	0.153	0.212	0.235
Adjusted R2	-0.100	0.142	0.195	0.219
Residual Std. Error	19.466 (df = 395)	17.196 (df = 391)	16.651 (df = 388)	16.402 (df = 388)
F Statistic				14.903*** (df = 8; 388)
Note:				*p<0.1; **p<0.05; ***p<0.01

The IV estimate exhibits a notable increase of approximately 7% in comparison to the OLS regression. This change is attributed to the elimination of bias. The coefficient is statistically significant at the 90% level in both OLS and IV, and in both regressions urban status has a positive effect on a student's average test score. The direction of the treatment effect is not of particular interest, as many econometricians would have expected a positive relationship prior to examining the results. Rather, the significance lies in the impact of unobservable factors, such as mindset, ability to seize opportunities, and motivation. These may differ between rural and urban children. These unobserved variables lead to an underestimation of the true treatment effect by the OLS model. The gap is in fact more than three-times larger than what OLS predicts.

This can be attributed to the challenges faced by rural children when entering the education system. These individuals are less likely to enter into the system in the first place. This means that only rural students who, either themselves or within their family, have a strong commitment to the academic journey tend to overcome difficulties and access education. Consequently, this self-selected group of rural students displays more ambition compared to their urban counterparts, on average. By isolating the treatment effect from the influence of ambition, it becomes evident that the actual gap would be even more concerning if rural children in middle-income schools were not inherently driven.

I believe that this effect would not apply universally. Rather, it would apply to situations where rural children enroll in schools that they perceive as providing an improvement over their usual circumstances. In such cases, where rural children do not feel disempowered, they might become more ambitious. Conversely, rural children attending rural schools may exhibit lower motivation due to the realisation that their living conditions are likely to remain unfavorable, as has been the case for generations. However, when rural children are exposed to realistic opportunities, such as in middle-income schools (where they represent a minority), some individuals may become more hopeful, and ultimately driven, relative to their urban peers.

Finally, as seen below, the residuals of this estimation are relatively normally distributed. A Breusch-Pagan test for heteroskedasticity returned a p-value of 0.2193, which implies that we cannot reject the null hypothesis of homoskedasticity, which along with mean zero residuals could imply that the coefficients are efficient, and exhibit little bias (Angrist & Pischke, 2008:230). Perhaps more importantly, given that the treatment effect is not significant at the desired 95% level, a normal distribution in the residuals enhances the validity of t-tests. This allows an econometrician to be more certain of the confidence interval provided for the treatment effect.



Note: Model residuals are standardised for the purposes of this comparison, but are mean zero regardless of standardisation.

Discussion and Critique

The change in the treatment effect between OLS and IV is certainly counter intuitive. Some economists may think that once all controls and bias has been accounted for, simply living in one area compared to another should not have a statistically significant effect on academic outcomes. It could be because bias has been removed, but it is built on a few untestable assumptions. Firstly, "rural status" in Portugal is unlikely to resemble what it means to be rural in a country like South Africa, because Portugal is a far less socioeconomically divided country. Children from a rural upbringing and children from urban living might not be as fundamentally different in Portugal as they are in more divided countries. This would make the integration process and support structure in Portuguese middle-income schools better suited to the needs of the relatively few underprivileged children. Portuguese middle-income schools might represent a real opportunity to these children, whereas a South African child might face many more obstacles at the South African equivalent of middle-income schools. Therefore, my ambition hypothesis might not be externally valid.

This study is also founded on the assumption that these are representative middle-income schools for Portugal. The sample size is quite small, and it is comprised of the students at only two schools. These schools were also assumed to be of the same quality. In order to make government policy from this, one would need repeated experiments. I do believe the estimation to be consistent, and statistically sound, but all hypotheses and findings need further testing in order to become widely accepted fact.

In conclusion, a robust IV approach has yielded causal effect between rural and urban children in these two Portuguese schools - a large academic performance gap. The quality of the school does not change from student to student, which would indicate that government policy should not only focus on improving school quality, but also focus on the living situation of students. The government cannot only invest in the opportunities provided to school children. There is a fundamental difference in their initial conditions, before entering the schooling system. It would seem that an incredibly important part of educating the youth is to invest in the lives they live when not at school. Rural living, or rather, what it entails, has effects on a child's ability to succeed long before they enter the classroom.

Bibliography

Angrist, J.D. 1990. Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. *The American Economic Review*, 80(3):313-336.

Angrist, J.D. & Pischke, J. 2008. *Mostly Harmless Econometrics*. Princeton University Press: Oxford.

Factsheet on 2014-2022 Rural Development Programme for Mainland, Portugal [Online]. [n.d.]. Available: https://agriculture.ec.europa.eu/system/files/2023-01/rdp-factsheet-portugal-continente_en.pdf [10 June, 2023].

Student Performance [Online]. [n.d.]. Available: https://www.kaggle.com/datasets/whenamancodes/student-performance [10 June, 2023]