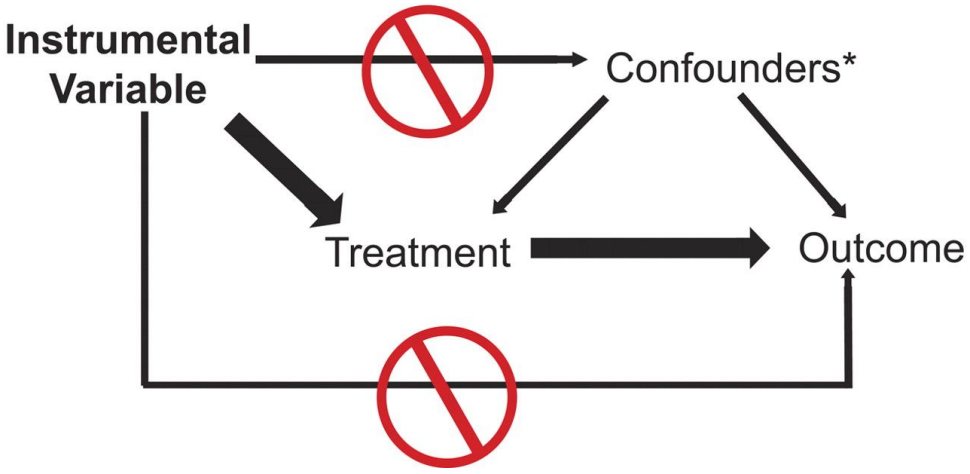
LATE TO THE PARTY: INVESTIGATING INSTRUMENT VARIABLES FOR EDUCATION

Cassandra Pengelly



20346212

Econometrics 871: Cross Section Project

Stellenbosch University

July 2021

Table of Contents

1. Introduction

The return to education is a cornerstone topic in labour economics and is of particular interest to policymakers. This is no different in South Africa, especially given the significant income inequality, which is often claimed to be strongly linked to differences in education. However, wages cannot simply be regressed on education because there is likely endogeneity present. This arises from the problem of there being an omitted variable, where education and wages are both correlated with a variable in the error term. One variable of this nature that has been extensively studied is innate ability (Hertz (2003)). Returns to schooling could be biased upwards if ability is positively correlated with both income and education. However, as Lang (1993) assesses, the overall findings of research on the impact of ability bias are inconclusive. In fact, Lang (1993: 1) notes that several papers find that returns to education are *downwardly* biased.

This paper will investigate whether OLS estimates are biased in the South African case by estimating the return to education using four different instrumental variable estimators: the first two exploit parents' education, and the other two use parents' occupations. These instruments estimate a 'local average treatment effect', which - this paper will argue - is more appropriate than OLS estimators for analysing the returns to education for South Africa. These instruments are tested for strength and validity and then implemented on the NIDS Wave 5 data set. This essay[1] is structured as follows: section 2 details the data set used and discusses the descriptive statistics. Section 3 outlines the methodology and argues that the LATE assumptions for the instrumental variables hold. Following this, section 4 presents the regression results and evaluates the robustness of the estimators used to obtain a causal effect. The final section, 5, concludes[2].

2. Data

The data used for this paper is sourced from the National Income Dynamics Survey (NIDS) (*National income dynamics study 2017, wave 5 dataset* (2018)), which was the first national household panel study in South Africa. NIDS is an initiative of the Department of Planning, Monitoring and Evaluation, and the Southern Africa Labour and Development Research Unit (SALDRU) is tasked with implementation (Brophy, Branson, Daniels, Leibbrandt, Mlatsheni & Woolard (2018)). NIDS was started in 2008, interviewing over 28,000 people. These same people are then interviewed every two years. The latest survey is Wave 5 (2017), which is the data set used for the regression analysis, where the individual is the unit of observation. Wave 5 consists of 37,368 individuals, where the high rate of attrition among high-income, Indian/Asian, and White respondents has led to the sample be-

[1] This essay was written in R using the Texevier package by Katzke (2017)
[2] The code and write up for this project can be found on Github https://github.com/cass-code/Cross-Section.git

ing increased by 2,775 to maintain sample representativeness (Brophy, Branson, Daniels, Leibbrandt, Mlatsheni & Woolard (2018)).

Villiers, Leibbrandt & Woolard (2009) outlays the design of the survey. To sample the households used in Wave 1, two-stage cluster sample design with stratification was employed. Stage 1 involved selecting 400 Primary Sampling Units, based on Statistics South Africa's Master Sample of 300 Primary Sampling Units (2003). Private households in all of South Africa's 9 provinces are the target population for NIDS. The 53 disctrict councils make up the explicit strata in the Master sample. Based on the allocation of the district councils in the Primary Sampling Units in the Master Sample, the sample was proportionally allocated and the Primary Sampling Units were randomly chosen within the strata (Villiers, Leibbrandt & Woolard (2009) p.9). Fieldworkers are assigned to the selected addresses and are instructed to interview all households living at the dwelling unit.

The NIDS wave 5 data set contains 30,110 observations of 1,144 variables. As a part of the data cleaning process, I selected 14 variables: year of birth, gender, race, income, marital status, highest level of schooling, highest level of tertiary education, union status, father's schooling, father's tertiary education, father's occupation, mother's schooling, mother's tertiary education, and mother's occupation. The occupation variables have been re-coded with the appropriate profession names from the *International standard classification of occupations* (2012), and are therefore categorical variables. I constructed the variables age and age-squared from the year of birth, and I constructed an education variable, which represents the number of years of education for an individual (i.e. summing the years of schooling and years of tertiary education). I constructed a similar variable for mother and father's education. Unfortunately, there are a significant number of missing observations for the variables income, parents' education, marital status and union status. Because of this, the sample size for the regressions is reduced to 3,300 and under.

As a part of the initial data exploration, figure 2.1 gives a quick snapshot of the relationship between income and education. As shown by the black regression line, income and education are positively correlated, which is what we expect. Section 4 will argue that this is a causal relationship. We can also glimpse how education and income is distributed in the sample. To get a clearer view of the race distribution, figure 2.2 shows a bar plot of the four main race groups in South Africa: African, Asian/Indian, Coloured and White. The graph shows the sample is roughly representative of the South African population, with a slightly smaller sample of Coloured and White individuals (despite the top up sample). We also want to check that income is normally distributed to combat heteroskedasticity in the regression analysis. We find that the income distribution is skewed but log of income is fairly normal, as figure 2.3 shows. This indicates that log of income should be the dependent variable in the regressions (rather the linear form of income), which is also supported by the literature.

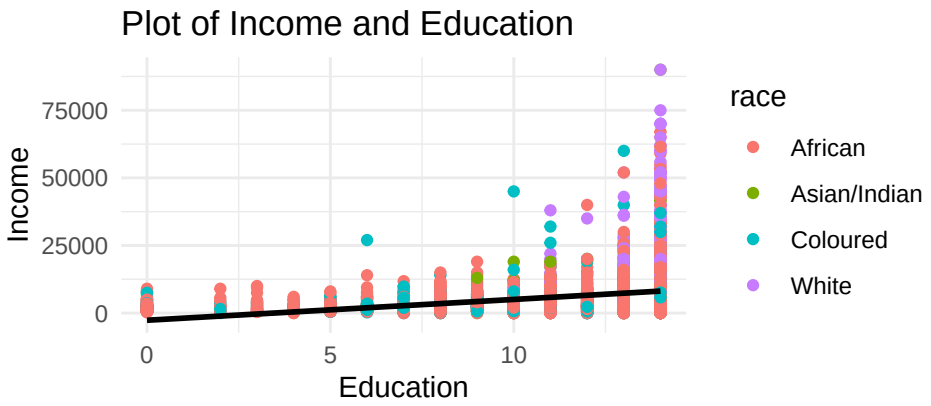


Figure 2.1: Income and Education Relationship

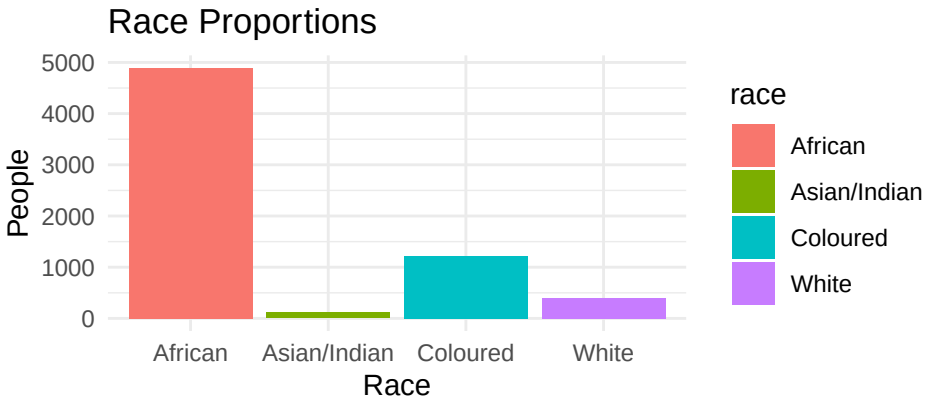


Figure 2.2: Race Proportions of Sample

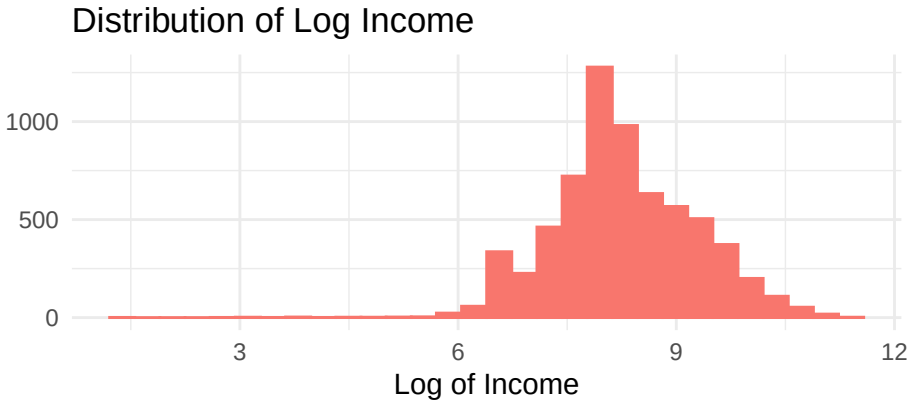


Figure 2.3: Log Income Distribution

3. Methodology

Following the earnings function proposed by Mincer (1974), to model the relationship between income and education in 4.1, the following general OLS regression, 3.1, is used:

$$Log(income) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u \tag{3.1}$$

where X_1 is education and $\ldots, X_k$ includes the other regressors: age, age-squared, race and gender, $\beta_0, \ldots, \beta_k$ are the regression coefficients, and u is the error term. Race is a categorical variable with the reference group as African, and gender is a dummy variable that takes a value of 1 if an individual is male and 0 otherwise.

If education is an endogenous regressor - i.e. correlated with the error term - then the OLS coefficient for education (β_1) will be biased (Colonescu (2016)). If we can find an instrumental variable for education, then we can remove this endogeneity and get an unbiased coefficient for education (Stock & Watson (2015) & Venables & Smith (2010)). 4 different variables are used as an instrument for an individual's education: fathers education, mother's education, father's occupation and mother's occupation. In order to use these variables as instruments, we have to check their validity(Imbens & Angrist (1994)). Economic theory and the literature suggests that returns to schooling are heterogeneous (Koop & Tobias (2004: 827–828)), which implies that for the instruments to be valid they have to comply with the Local Average Treatment Effects (LATE) identifying assumptions. The four LATE assumptions are: independence, random assignment, exclusion restriction, and monotonicity. The first section below, ref{spec}, interprets the diagnostic tests for instrument variables and the second part, ref{late}, evaluates whether the LATE assumptions hold for the instrument variables.

3.1. Specification Tests

Table 3.1: Specification Tests for Father's Education

	df1	df2	statistic	p-value
Weak instruments	1	3368	406.13620	0
Wu-Hausman	1	3367	70.06385	0
Sargan	0	NA	NA	NA

Table 3.2: Specification Tests for Mother's Education

	df1	df2	statistic	p-value
Weak instruments	1	3605	400.15525	0
Wu-Hausman	1	3604	46.19299	0
Sargan	0	NA	NA	NA

Table 3.3: Specification Tests for Father's Occupation

	df1	df2	statistic	p-value
Weak instruments	9	2162	19.86956	0.0000000
Wu-Hausman	1	2169	32.61388	0.0000000
Sargan	8	NA	12.65711	0.1242046

Table 3.4: Specification Tests for Mother's Occupation

	df1	df2	statistic	p-value
Weak instruments	9	1731	38.61704	0.000000
Wu-Hausman	1	1738	64.04297	0.000000
Sargan	8	NA	4.50194	0.809239

3.2. LATE Assumptions

If our instrument is: 1. randomly assigned 2. only affects our outcome through its effect on our endogenous variable 3. has a significant impact on take-up of treatment, and 4. does not deter anyone from treatment

Random Assignment: Random assignment of draft eligibility means independence assumption is very convincing. Would not expect individuals who were born on low-lottery days to have different potential earnings distributions or potential take-up probabilities than high-lottery individuals.

Exclusion restriction: More reason to be concerned. Draft eligible men were allowed to defer their military service if they wanted to proceed with their studies. If some men decided to get more schooling because they had low lottery numbers and did not want to serve in the military, then this could have affected earnings via education attainment rather than via military service.

Monotonicity: It seemse unlikely that someone who would have pursued education further would decide not to because either of his/her parent pursued education further. Thus, it's reasonable to conclude that the monotonicity assumption holds (i.e. there are no defiers). Similarly

Is it possible that someone would have volunteered for the army (if draft ineligible) but decided not to because of draft eligibility? Seems unlikely.

The relevance assumption is testable and table ref{reg5} shows that both parents' education is correlated with education (and the coefficients are statistically significant at a 1% level). This means these instruments affect treatment for at least some individuals. The regression of education on parents' occupation shows that joint effect of the occupation types are significant. The p-values associated with the F-test for each regression are presented in 3.6. The p-values are essentially 0 and thus reject the null hypothesis that the joint effect is zero. From an intuitive stand point, if an individual's parents pursued more years of education,

Are there individuals who would not have gone to war if they didn't get picked, but ended up going just because they were eligible? Seems plausible.

Table 3.5: Regressions: Instrument Relevance

	Education	Education
Father's Education	0.234 ***	
	(0.008)	
Mother's Education		0.247 ***
		(0.008)
N	3376	3613
R2	0.190	0.197

*** p < 0.001; ** p < 0.01; * p < 0.05.

Table 3.6: p-values of F statistics

Regression	p.value
Regress Education on Father's Occupation	4.748773e-54
Regress Education on Mother's Occupation	2.356427e-61

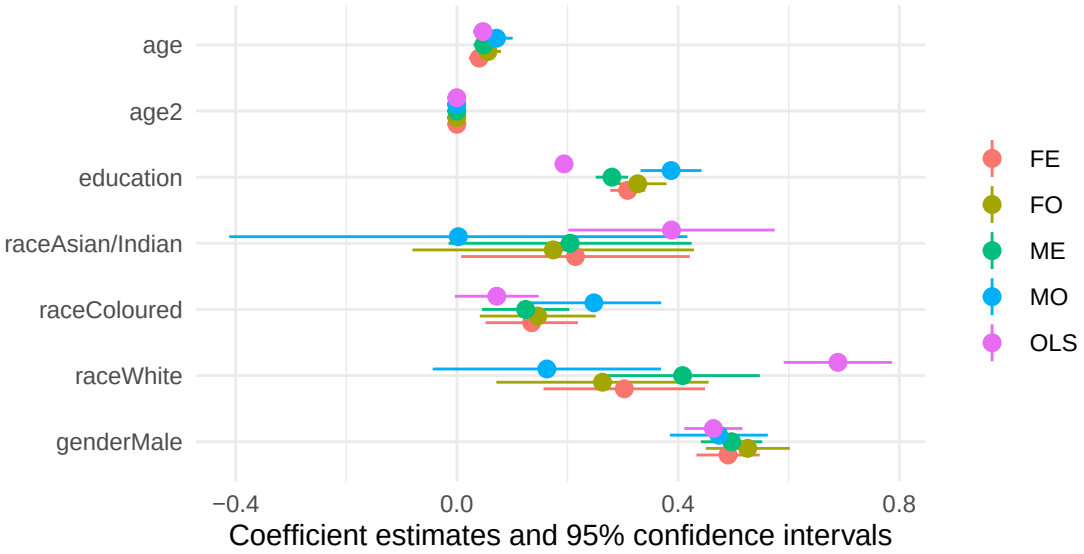
4. Results

Table 4.1 presents the regressions results and figure **??** plots the regression coefficients with their 95% confidence intervals.

Table 4.1: Regressions: OLS and 2SLS with Various Instruments

	OLS	2SLS FE	2SLS ME	2SLS FO	2SLS MO
Age	0.047 ***	0.040 ***	0.049 ***	0.056 ***	0.072 ***
	(0.009)	(0.009)	(0.009)	(0.012)	(0.015)
Age2	-0.000 **	-0.000	-0.000 *	-0.000 *	-0.000 **
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Education	0.194 ***	0.309 ***	0.280 ***	0.327 ***	0.387 ***
	(0.005)	(0.016)	(0.015)	(0.027)	(0.028)
Coloured	0.072	0.135 **	0.124 **	0.146 **	0.248 ***
	(0.039)	(0.043)	(0.040)	(0.053)	(0.062)
Asian/Indian	0.388 ***	0.214 *	0.205	0.174	0.003
	(0.095)	(0.105)	(0.112)	(0.130)	(0.211)
White	0.689 ***	0.303 ***	0.408 ***	0.263 **	0.163
	(0.050)	(0.074)	(0.071)	(0.098)	(0.105)
Male	0.464 ***	0.490 ***	0.497 ***	0.526 ***	0.474 ***
	(0.027)	(0.029)	(0.028)	(0.039)	(0.045)
N	3376	3376	3613	2178	1747
R2	0.456	0.364	0.368	0.337	0.247

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.



wn possibly long tables. Note that the following will fit on one page if it can, but cleanly spreads over multiple pages:

5. Conclusion

You should NOT write an extensive essay about the topic of interest, nor should you conduct an extensive literature review behind it: you should only reference articles that support the motivation for your econometric strategy (in the quest of finding a causal interpretation of a relationship that you are modeling). Provide descriptive statistics and graphs to aid in your discussion. The assignment should be between 3-6 pages, including your tables and figures.

You are required to go beyond simply estimating and presenting your results, but to convince the reader of their robustness by presenting alternative specifications. You should apply different estimators and specifications where possible. Discuss the shortcomings of the estimators in obtaining a causal effect and argue why your strategy is the best available to obtain a causal effect that satisfies relevant assumptions.

Determine an effect of interest and find instrumental variables to estimate it causally. If possible, use more than one instrument. Given the LATE assumptions, try and explain why your results differ and which is likely to represent the causal effect you are looking for. Conduct sufficient specification tests to establish whether you are overidentifying your instrument set or whether the IVs shift the estimates enough to indicate that OLS would be inconsistent.

References

10 Brophy, T., Branson, N., Daniels, R.C., Leibbrandt, M., Mlatsheni, C. & Woolard, I. 2018. *National income dynamics study panel user manual*. (Release 2018. Version 1). Cape Town: Southern Africa Labour; Development Research Unit; Southern Africa Labour; Development Research Unit.

Colonescu, C. 2016. Principles of econometrics with r. 11. [Online], Available: https://bookdown.org/ccolonescu/RPoE4/.

Hertz, T. 2003. Upward bias in the estimated returns to education: Evidence from south africa. *American Economic Review*. 93(4):1354–1368.

Imbens, G.W. & Angrist, J.D. 1994. Identification and estimation of local average treatment effects. *Econometrica*. 62(2):467–475. [Online], Available: http://www.jstor.org/stable/2951620.

International standard classification of occupations: Structure, group definitions and correspondence tables. 2012. (ISCO - 08). International Labour Office, Geneva; International Labour Organization.

Katzke, N.F. 2017. *Texevier: Package to create elsevier templates for rmarkdown*. Stellenbosch, South Africa: Bureau for Economic Research.

Koop, G. & Tobias, J.L. 2004. Learning about heterogeneity in returns to schooling. *Journal of Applied Econometrics*. 19(7):827–849. [Online], Available: http://www.jstor.org/stable/25146329.

Lang, K. 1993. *Ability bias, discount rate bias and the return to education*. (MPRA Paper). University Library of Munich, Germany. [Online], Available: https://EconPapers.repec.org/RePEc:pra:mprapa:24651.

Mincer, J. 1974. *Schooling, experience and earnings*. National Bureau of Economics, New York: Columbia University Press.

National income dynamics study 2017, wave 5 dataset. 2018. Cape Town, South Africa: Department of Planning, Monitoring,; Evaluation [funding agency] & DataFirst [distributor]. [Online], Available: https://doi.org/10.25828/fw3h-v708.

Stock, J. H. & Watson, M.W. 2015. *Introduction to econometrics, third update, global edition*. Pearson Education Limited.

Venables, W.N. & Smith, D.M. 2010. *R-Intro: An introduction to r*. [Online], Available: https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf.

Villiers, L. de, Leibbrandt, M. & Woolard, I. 2009. *Methodology: Report on NIDS wave 1.* (Technical Paper no. 1). Cape Town: Southern Africa Labour; Development Research Unit; Southern Africa Labour; Development Research Unit.

Appendix

Appendix A

Some appendix information here

Appendix B