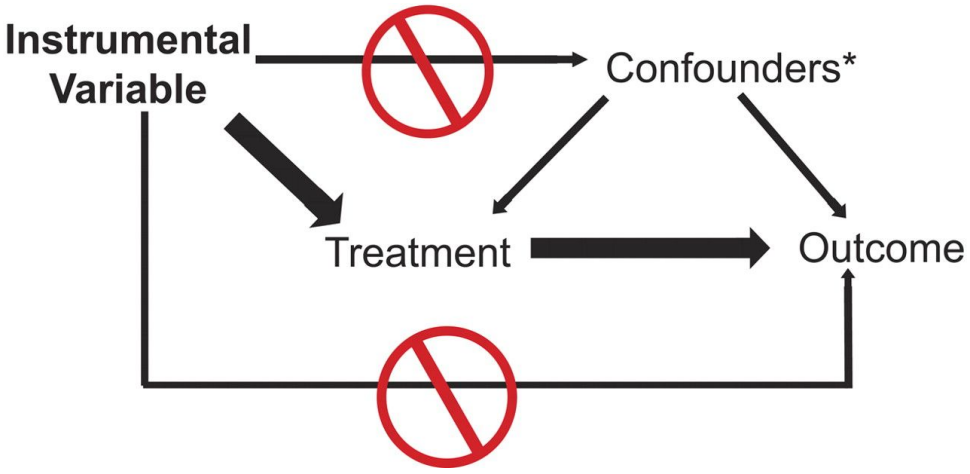
LATE TO THE PARTY: INVESTIGATING INSTRUMENT VARIABLES FOR EDUCATION

Cassandra Pengelly



20346212

Econometrics 871: Cross Section Project

Stellenbosch University

July 2021

Table of Contents

1 Introduction **3**

2 Data **3**

3 Methodology **6**

3.1 Specification Tests and LATE Assumptions 6

4 Results **8**

5 Conclusion **8**

References **10**

Appendix A: Specification Tests **12**

Appendix B: Regression Tables **13**

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, if certain assumptions are met. 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 investigates whether the LATE assumptions for the instrumental variables hold. Following this, section 4 discusses 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 its implementation (Brophy, Branson, Daniels, Leibbrandt, Mlatsheni & Woolard (2018)). NIDS was started in 2008, with over 28,000 people interviewed. 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

[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

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

Villiers, Leibbrandt & Woolard (2009) outlay 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 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 are 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 Asian/Indian 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.

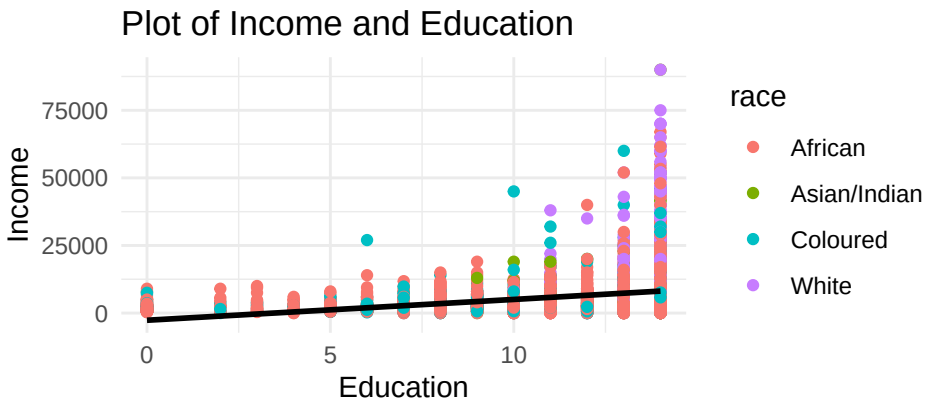


Figure 2.1: Income and Education Relationship

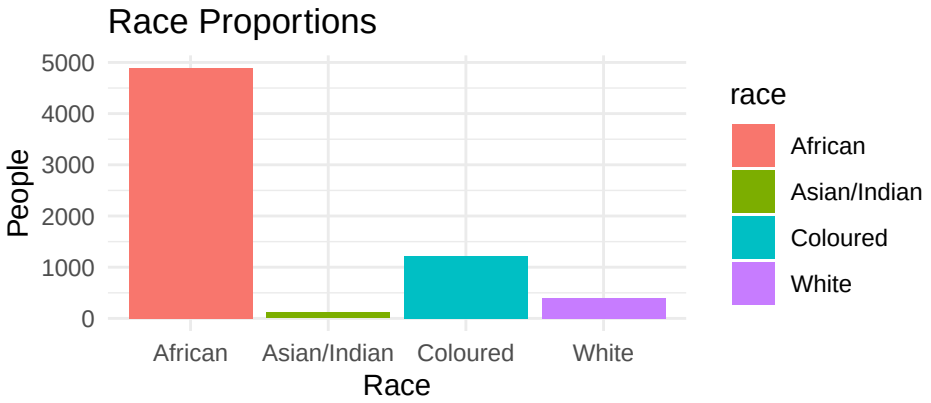


Figure 2.2: Race Proportions of Sample

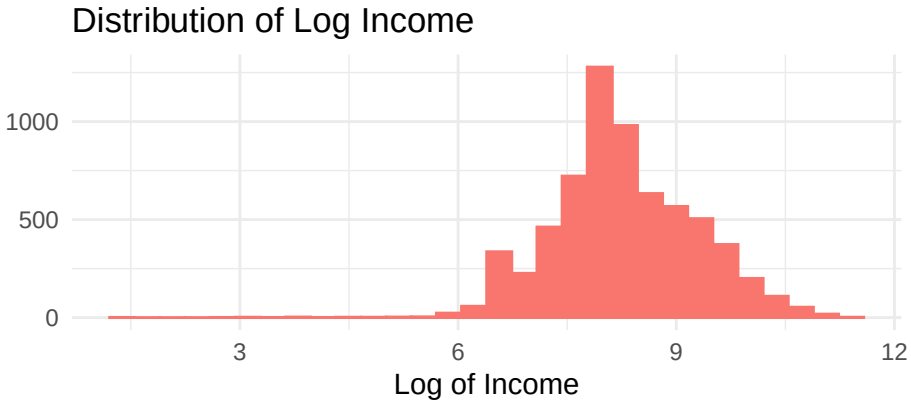


Figure 2.3: Log Income Distribution

3. Methodology

Following the earnings function proposed by Mincer (1974), to model the relationship between income and education in 5.5, 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: father's 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: relevance, monotonicity, random assignment and exclusion restriction. The section below, 3.1, interprets the diagnostic tests for the instrument variables and evaluates whether the LATE assumptions hold.

3.1. Specification Tests and LATE Assumptions

The relevance assumption is directly testable on the data, for which we can use the 'Weak Instrument' test. The results of the specification tests in the appendix, 5, show that all four instruments are not weak instruments. The null hypothesis for the 'Weak Instrument' test is that the instrument is weak. The p-values for each of the individual tests are zero, which rejects the null hypothesis. This supports the relevance assumption and these instruments affect treatment for at least some individuals. This makes intuitive sense as well: if an individual was on the fence about studying further, and his/her parents studied further, the parents are likely to convince that individual to follow in their footsteps/take the "safe" option of studying further.

The Wu-Hausman test evaluates whether the OLS and instrumental variable regressions give significantly different estimates. The null hypothesis is that the estimated coefficients of OLS and IV regression are not statistically different from one another. For all four instruments, the null hypothesis

is rejected, which means the OLS and IV estimates are sufficiently different from one another. The Sargan test is applicable when the number of instruments exceeds the number of endogenous variables. Rejecting the null of the Sargan test implies that the instruments are invalid. For father's occupation and mother's occupation, the Sargan test fails to reject the null, thus all the instrumental variables are valid.

Regarding the assumption of monotonicity: it seems unlikely that someone who would have pursued education further would decide not to because either of his/her parents pursued education further. Thus, it is reasonable to conclude that the monotonicity assumption holds for father's and mother's education (i.e. there are no defiers). Similarly, it seems unreasonable that someone who would have pursued education further would decide not to due to his/her parents' occupation. Although, one could argue that: if an individual would have liked to study further, but doesn't want the pressure of living up to his/her parents' academic achievements and so opts out, the monotonicity assumption is invalid. However, this is unlikely to occur for the majority of the sample and can be ignored.

The random assignment assumption is difficult to argue for, since the distribution of potential outcomes and potential take-up values should look same for those whose parents studied further and for those whose didn't, when in reality the distribution doesn't look the same. We would expect individuals whose parents have lower levels of education to have different potential earnings distributions or potential take-up probabilities than individuals whose parents have higher levels of education. Similarly for individual's whose parents have different occupations. In South Africa, networks play a large role in employment opportunities and therefore play a role in earning potentials. The earnings' potential of an individual changes depending on their connections and their parents' connections, which is correlated to parental occupation; this is a violation of the exclusion restriction.

The exclusion restriction assumption is broken if an individual's parents' education affects his/her potential earnings through another variable other than education. This is possible, for example, through socioeconomic standing. If an individual's parents' education/occupation is correlated with their being wealthy and well-connected, and having wealthy and well-connected parents increases earning potential (they could provide financial and social capital for business ventures) then the exclusion restriction assumption no longer holds. However, according to Hoogerheide, Block & Thurik (2012), violations of the strict validity assumption does not necessarily lead to results which differ significantly from those of the strict validity case. Hoogerheide, Block & Thurik (2012) make the case that family background variables can be used as instruments for regressions involving income. The next section 4 discusses the results of the ordinary least squares regression (OLS) compared to the instrumental variable (IV)regressions.

4. Results

Table 5.1 in the appendix (5) presents the regression results, and the figure below it plots the regression coefficients with their 95% confidence intervals. The OLS regression regresses log income on education, controlling for age, race, and gender. The coefficients of the control variables have the expected signs and most are statistically significant at 1%. The coefficient for education is 0.194, which implies that a 1 unit (a year) increase in education is associated with a 21,4% increase in income, ceteris paribus. However, in order to show there exists a causal relationship between education and income, and that there exists bias in the OLS estimate, two-stage least squares (2SLS) was applied using the 4 different instrumental variables.

The first 2SLS regression employs father's education as an instrument for education (2SLS FE in 5.5). The coefficient for this regression is 0.309, which is significantly higher than the OLS estimate. This implies that the OLS estimate is bias downwards. The second IV regression uses mother's education as an instrument for education (2SLS ME in 5.5). The coefficient for education here is 0.280; again this is a lot higher than the OLS estimate. The third and fourth 2SLS regressions that exploit father's and mother's occupation as instrument variables for education (2SLS FO, 2SLS MO in 5.5) show similar results. They estimate the coefficient for education to be 0.327 and 0.387 respectively. However, the sample size diminishes when using occupation as a variable, which makes them less statistically trustworthy than the other estimates. This is illustrated clearly by the large confidence bands for 'FO' and 'MO' in the figure in the appendix 5.5. Overall, the IV regression results show that the OLS estimate is substantially downwards biased. These findings are similar to those of Biyase & Zwane (2015), who also show that returns to education are underreported by OLS regressions.

However, it is important to recognize that the 2SLS estimators recover local average treatment effects (LATE) whereas OLS estimators recover average treatment effects. In the IV regressions, the effect of further education is only captured for those individuals who are more likely to further their education because their parents pursued education further. This empirical stance does not provide information about the effect of studying further among people would have always studied further or never studied further, regardless of their parents' education/occupations.

5. Conclusion

This essay set out to show that there is a causal link between education and income using four different instrumental variables. The OLS regression finds that education has a positive coefficient, as expected, but the four 2SLS regressions indicate that OLS may be underestimating the returns to education. The fact that all four 2SLS regressions show higher education estimates suggests the results are robust, and the instrumental variables shift the estimate enough to indicate that OLS would be

inconsistent. Although some of the LATE assumptions may be violated by selecting parents' education and occupations as instrumental variables, they may still be viable tools to introduce exogeneity into the model (Hoogerheide, Block & Thurik (2012)). Further robustness checks could be conducted by including more controls in the regressions (such as marital status and union status[3]), using fixed effects to control for time invariant unobservables or using education lagged as an instrument variable.

[3]I ran these regressions and found similar results; however the sample sizes were very small

References

10 Biyase, M. & Zwane, T. 2015. Does education pay in south africa? Estimating returns to education using two stage least squares approach. *International Business & Economics Research Journal (IBER)*. 14(6):807–814.

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.

Hoogerheide, L., Block, J. & Thurik, R. 2012. Family background variables as instruments for education in income regressions: A bayesian analysis. *Economics of Education Review*. 31(5):515–523.

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 A: Specification Tests

Table 5.1: 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 5.2: 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 5.3: 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 5.4: Specification Tests for Mother's Occupation

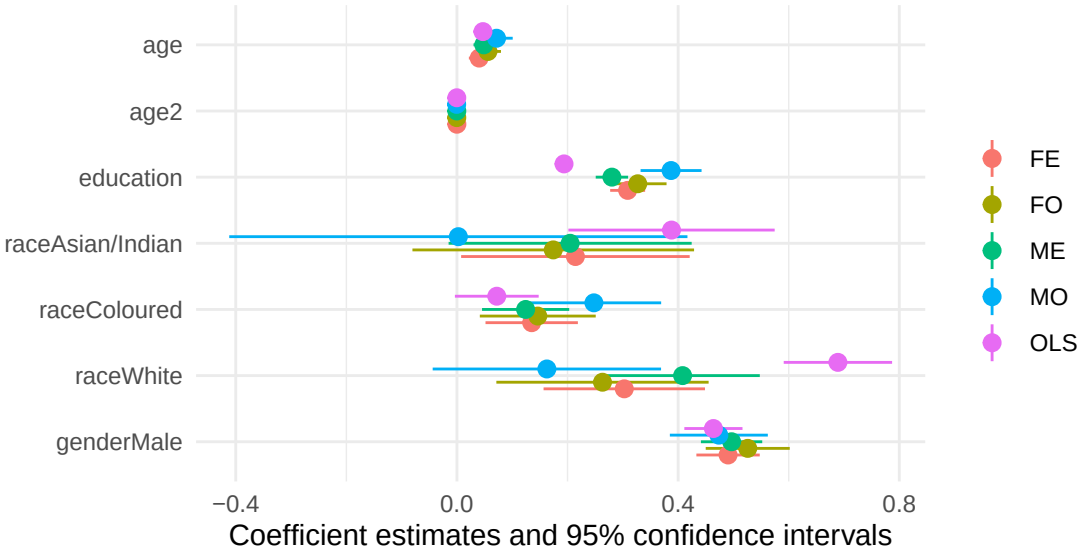
	df1	df2	statistic	p-value
Weak instruments	9	1731	38.61704	0.000000
Wu-Hausman	1	1738	64.04297	0.000000
Sargan	8	NA	4.50194	0.809239

Appendix B: Regression Tables

Table 5.5: 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

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


age
age2
education
raceAsian/Indian
raceColoured
raceWhite
genderMale
−0.4
0.0
0.4
0.8
Coefficient estimates and 95% confidence intervals
FE
FO
ME
MO
OLS