ECOM90003: Applied Microeconometric Modelling – Assignment 2

Q	Response
1	The four assumptions are: • The relationship between the dependent (student test scores) and independent (school characteristics and class size) variables must be linear in the parameters β and α. Practically, this means a linear function can correctly specify the functional form of the relationship and there are no omitted variables with an additive error term. • The cross-sectional data use to construct this model must be drawn randomly from the population of students, classes and schools. Doing this ensures a representative sample which is unaffected by selection bias. Practically, this means the test score of each student, their class size and school characteristics are randomly observed across different schools and classes. • The independent variables $\boldsymbol{X}_s$ and n_{sc} cannot exhibit perfect multicollinearity. This implies in the sample (and therefore the population) there are no constant independent variables, nor are there any exact linear relationships between them. Ensuring this means each variable provides unique information about how variation in the dependent variable occurs, ensuring coefficient estimates are precise. • Finally, the error term must have a mean of zero conditional on the independent variables $\boldsymbol{X}_s$ and n_{sc}. Ensuring this means there is no correlation between the independent variables and the error term, implying these independent variables are not systematically related to other unobserved factors affecting test scores.
2	There are four potential sources of endogeneity in this regression: • Depending on the exact number and nature of variables included in $\boldsymbol{X}_s$, its possible this regression could suffer from omitted variable bias. If these omitted variables are correlated with class size, this introduces endogeneity. Without more information about $\boldsymbol{X}_s$, its impossible to rule this out. • Simultaneity bias could also be an issue for the endogeneity of this regression model. If the class sizes at schools are adjusted according to previous test scores for students (i.e. poorly performing students might be put into smaller class sizes) then class size would influence both test scores an class size itself. • Measurement error in the independent variables can also cause endogeneity, which would be caused by inaccurately reporting class sizes for example (but could apply to anything in $\boldsymbol{X}_s$,). This is particularly problematic if the issue is systematic. • Selection bias could also be a problem. If higher performing students tend to be placed into smaller classes or in schools with certain characteristics then the correlation between scores and classes size would also be impacted by some other unobserved variable that influences placement into certain sized classes.

Josh Copeland (SID: 1444772)

3	No, having access to panel data would certainly be helpful, but it would not be sufficient by itself to deal with the sources of endogeneity identified above. Panel data would help with endogeneity for the following reason: • Panel data (via a fixed effects model) would allow us to control for unobserved time-invariant individual-specific effects (α_i). This would address omitted variable bias where the omitted variables are constant over time for each individual (e.g. student motivation/ability or socioeconomic background, etc). However, panel data would not be able to fully resolve this issue for the following reasons: • Panel data cannot address time-varying omitted variables, where these variables would continue to cause endogeneity in this model. For example, its reasonable to expect the engagement of parents' in their child's education is an important determinant of their performance, but is likely also time varying as its driven by a range of factors (stress, income, employment status). Unless this is directly measured by another variables we can observe, it would not be captured by α_i. • It cannot fundamentally resolve simultaneity. While including some lagged independent variables in the model could help, there is a fundamental problem here in that there is a two-way relationship between the dependent and independent variables that can only be resolved with the use of an IV to address this feedback loop. • If there is a systematic issue with how class sizes (or some other independent variable) are measured, then this issue will persist regardless of what kind of data format is used. • Similarly, the use of panel data would not resolve selection bias, as this issue is fundamental to data collection itself and not the format it is presented in.
4	This specification of measurement error is known as classical measurement error and will result in a biased α estimate. Where an independent variable suffers from classical measurement error, it suffers from attenuation bias. This can be shown by substituting $\widetilde{n_{sc}}$ into our equation: $$y_{isc} = \boldsymbol{X}_s\beta + \widetilde{n_{sc}}\,\alpha + \epsilon_{isc}$$ $$y_{isc} = \boldsymbol{X}_s\beta + (n_{sc} + v_{sc})\alpha + \epsilon_{isc}$$ $$y_{isc} = \boldsymbol{X}_s\beta + n_{sc}\alpha + (v_{cs}\alpha + \epsilon_{isc})$$ Practically, this adds noise to the independent variable and reduces the absolute value of the estimator towards zero as the error now incorporates part of it.
5	To be a valid instrument, Maimonides' rule needs to be both relevant to class sizes and exogenous to test scores. Where Maimonides' rule establishes a threshold for the size of school classes, this means this rule needs to be strongly correlated with class sizing decisions by schools. To be exogenous with the error terms, Maimonides' rule should not directly affect test scores, except through its impact on class size. Implicitly, this

	requires that the application of this rule only influences class size adjustment, and not any other facet of school characteristics. Their key rationale is that Maimonides' rule represents a natural experiment in the Israeli education system which allows them to leverage variation on class size due to the rule. To substantiate this as a relevant instrument they draw several pieces of evidence: • Figure I helps to indicate the satisfaction of the relevance assumption for this instrument. By demonstrating a close relationship between the functional form of Maimonides' rule and actual class sizes across Israel the authors clearly indicate the instrument is relevant. • Figure III does the same thing by demonstrating relevance between class size and Maimonides' rule by showing there is correlation between these variables after controlling for other relevant factors (control variables). • The descriptive statistics in Table I also help indicate this relevance assumption by showing Maimonides' rule can be used to identify the effects of class size as it induces a discontinuity in the relationship between enrolment and class size at enrolment multiples of 40. Specifically, by comparing Panels A and B of Table I this is shown by the very small difference in average characteristic between the overall and discontinuity samples, despite fewer than one quarter of enrolment being in the discontinuity sample. • The authors' discussion of Campbell (1969) is also informative for rationalising the choice of the instrument, which exploits the fact a deterministic function can be used to assign the causal effect of a treatment. That is, if the methods of assignment are discontinuous, then you can control for the smooth part of the function and still estimate the impact at the point of the discontinuity.
6	The effect of Maimonides' rule on class size ranges from .54 to .77, which is far and away the largest effect reported in these regression tables at around at least ten times greater then per cent disadvantages or enrolment in any specification. Due to endogeneity issues with this regression, this effect is not causal, but it does indicate correlation between this predicted and actual class sizes is very strong relative to the other variables shown.
7	Refer to Appendix A.
8	Refer to Appendix A.
9	These estimates indicate higher predicted class sizes are associated with larger classes and lower test scores. The negative association between Maimonides' rule and test scores is strongest for fifth graders, but also applies to fourth graders. Importantly, the relationship between Maimonides' rule and reading scores in both grades are largely insensitive to enrolment size. However, there's no evidence of a relationship between math scores and predicted class size for fourth graders. These estimates are reduced form, which simplifies the relationship between variables by omitting the full structural form of the relationship. Practically, this means these estimates omit the endogenous variable (class size) from our regression. Therefore, these tell us that there is a relationship between our instrument (Maimonides rule) and test scores, but this is not causal. This is because, structurally, the authors think the instrument does not have a direct impact on test scores. It only impacts it indirectly through class size.

	Therefore, to make this relationship causal, we need to use the IV approach to estimating this relationship between class size and scores, using Maimonides' rule as an instrument using two stage least squares estimation.
10	Refer to Appendix A.
11	I will only be comparing comparable models where the same independent variables are using (class size, per cent disadvantaged and enrolment). In Table 2, using OLS estimates, class size does not have a statistically significant impact on fifth grader reading comprehension or math test scores (but decreases scores by 0.025 and increases them by 0.019 respectively. Each additional unit of a school's per cent disadvantaged value decreases fifth grader reading comprehension and math scores by 0.351 and 0.332 respectively significant at the 1 per cent level. Each additional student enrolled increases math scores by 0.017 significant at the 10 per cent level (impact on reading comprehension is insignificant). In Table 4, each additional student in the class reduces fifth grader reading comprehension and math scores by 0.277 and 0.231 respectively (significant at the 1 and 10 per cent levels respectively). Each additional unit of a school's per cent disadvantaged value decreases fifth grader reading comprehension and math scores by 0.369 and 0.350 respectively significant at the 1 per cent level. Each additional student enrolled in fifth grader classes increases reading comprehension and math scores by 0.022 and 0.041 respectively (significant at the 10 and 1 per cent levels respectively). In the naïve equations, the relationship between test scores and class size is biased upwards as the IV estimates are much smaller. For per cent disadvantaged, this bias is also upwards but much smaller, adding evidence to this variable being a true independent variable. For enrolments, this biased downwards, as the IV estimates are larger than their OLS counterparts.

Word count: 1600 words

Appendix A: requested tables

Table 1: Reduced form class size estimates for 1991 (full sample)

	5th Graders		4th Graders	
	(1)	(2)	(7)	(8)
Mean	29.9		30.3	
(s.d.)	(6.5)		(6.3)	
f_{sc}	0.704 (0.025)	0.544 (0.037)	0.773 (0.017)	0.673 (0.033)
Percent disadvantaged	-0.076 (0.011)	-0.054 (0.010)	-0.054 (0.008)	-0.040 (0.009)
Enrolment		0.043 (0.006)		0.026 (0.005)
Root MSE	4.551	4.380	4.193	4.128
R^2	0.517	0.553	0.564	0.578
N	2024		2053	

Note: standard errors in parentheses.

Table 2: OLS estimates for 1991

	5th Grade						4th Grade					
	Reading comprehension			Math			Reading comprehension			Math		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Mean score	74.4			67.3			72.5			68.9		
(s.d.)	(7.7)			(9.6)			(8.0)			(8.8)		
Class size	0.221 (0.034)	-0.031 (0.026)	-0.025 (0.033)	0.322 (0.040)	0.076 (0.036)	0.019 (0.042)	0.141 (0.035)	-0.053 (0.028)	-0.040 (0.032)	0.221 (0.039)	0.055 (0.036)	0.009 (0.040)
Per cent disadvantaged		-0.350 (0.014)	-0.351 (0.015)		-0.340 (0.018)	-0.332 (0.019)		-0.339 (0.015)	-0.341 (0.016)		-0.289 (0.017)	-0.281 (0.017)
Enrolment			-0.002 (0.006)			0.017 (0.008)			-0.004 (0.006)			0.014 (0.007)
Root MSE	7.54	6.10	6.10	9.36	8.32	8.31	7.94	6.65	6.65	8.66	7.82	7.81
R^2	0.035	0.367	0.367	0.048	0.248	0.251	0.013	0.309	0.309	0.025	0.204	0.207
N	2019			2018			2049			2049		

Note: standard errors in parentheses.

Table 3: Reduced form class size estimates for 1991 (full sample)

	5th Graders				4th Graders			
	Reading comprehension		Math		Reading comprehension		Math	
	(3)	(4)	(5)	(6)	(9)	(10)	(11)	(12)
Mean	74.4		67.3		72.5		68.9	
(s.d.)	(7.7)		(9.6)		(8.0)		(8.8)	
f_{sc}	-0.111 (0.029)	-0.150 (0.039)	-0.009 (0.041)	-0.125 (0.051)	-0.085 (0.031)	-0.089 (0.040)	0.038 (0.040)	-0.033 (0.050)
Percent disadvantaged	-0.359 (0.014)	-0.354 (0.015)	-0.354 (0.018)	-0.337 (0.019)	-0.340 (0.015)	-0.340 (0.016)	-0.292 (0.017)	-0.282 (0.017)
Enrolment		0.010 (0.006)		0.031 (0.008)		0.001 (0.007)		0.019 (0.008)
Root MSE	6.085	6.079	8.340	8.294	6.635	6.637	7.829	7.813
R^2	0.374	0.375	0.246	0.254	0.311	0.311	0.203	0.207
N	2019				2049			

Note: standard errors in parentheses.

Table 4: Full sample 2SLS estimates for 1991 (5th Graders)

	Reading comprehension			Math		
	(1)	(2)	(3)	(7)	(8)	(9)
Mean score	74.4			67.3		
(s.d.)	(7.7)			(9.6)		
Class size	-0.158 (0.042)	-0.277 (0.076)	-0.263 (0.094)	-0.013 (0.058)	-0.231 (0.098)	-0.264 (0.123)
Per cent disadvantaged	-0.371 (0.016)	-0.369 (0.016)	-0.369 (0.016)	-0.355 (0.020)	-0.350 (0.020)	-0.350 (0.020)
Enrolment		0.022* (0.009)	0.013 (0.026)		0.041 (0.012)	0.063 (0.036)
Enrolment squared/100			0.004 (0.010)			-0.010 (0.014)
Root MSE	6.161	6.242	6.228	8.339	8.400	8.425
N	2019			2018		

Note: standard errors in parentheses.

Appendix B: Stata code

```
*******************************************************************************
************** ECOM90003 - Applied Microeconometric Modelling  *********
****************************** ASSIGNMENT 2 ****************************
*******************************************************************************

*******************************************************************************
****************************** SET-UP **********************************
*******************************************************************************

clear // clear all data from memory
log close

set more off // pause more message

cd "C:\Users\joshc\OneDrive\Desktop\git\mae_unimelb\2024\S2\ECOM90003 - Applied Microeconometric Modelling\Assignments\Assignment 2\"

log using "assignment2.log", replace // if you get log file open error message, insert: "log close" into command


frame create temp1
frame create temp2

frame change temp1
use grade5_final, clear

frame change temp2
use grade4_final, clear

frame rename temp1 grade5
frame rename temp2 grade4
```

********** TABLE 1 **********

// Using the data provided, replicate Panel A (columns 1-2 and 7-8) of Table 3 and label your replication "Table 1".

// Generating means

```
frame change grade5

summarize classize

frame change grade4

summarize classize
```

// Generating actual table content

```
// Clear previous estimates
eststo clear

// Run regressions for grade5
frame change grade5
eststo: reg classize func1 tipuach, cl(schlcode)
estimates store model_grade5_1  // Store first model for grade5

eststo: reg classize func1 tipuach c_size, cl(schlcode)
estimates store model_grade5_2  // Store second model for grade5

// Run regressions for grade4
frame change grade4
eststo: reg classize func1 tipuach, r
estimates store model_grade4_1  // Store first model for grade4

eststo: reg classize func1 tipuach c_size, cl(schlcode)
estimates store model_grade4_2  // Store second model for grade4

// Switch back to the default frame to produce the combined table
frame change default

// Output the table with all four models
esttab model_grade5_1 model_grade5_2 model_grade4_1 model_grade4_2 using
"table_1.csv", ///
b(%5.3f) se obslast r2 star(* 0.1 ** 0.05 *** 0.01) replace ///
title("Model 1 (Grade 5)" "Model 2 (Grade 5)" "Model 1 (Grade 4)" "Model 2 (Grade 4)") ///
stats(rmse r2 N)
```

******************************* TABLE 2 *******************************

// Replicate Columns 1 to 12 of Table 2 in the paper. Include it in your assignment appendix, labelled as "Table 2".

// Generating means

```
frame change grade5

summarize avgverb
summarize avgmath

frame change grade4

summarize avgverb
summarize avgmath
```

//Generating actual table output

```
local outcome_var avgverb avgmath

// Clear previous estimates
eststo clear

// List of dataframes to loop through
local dataframes grade5 grade4

// Loop through each dataframe
foreach df in `dataframes' {
// Change to the current dataframe
frame change `df'

// Loop through each outcome variable
foreach y in `outcome_var' {
// Run regressions and store results
eststo: reg `y' classize, cl(schlcode)
eststo: reg `y' classize tipuach, cl(schlcode)
eststo: reg `y' classize tipuach c_size, cl(schlcode)
}
}

// Switch back to the default frame to produce the combined table
frame change default

// Output the table with all models into "table_2.csv"
```

```
esttab using "table_2.csv", ///
b(%5.3f) se obslast replace ///
title("Regression Results for Grade 5 and Grade 4") ///
stats(rmse r2 N) ///
```

********** TABLE 3 **********

```
// Replicate Columns 3 - 6 and 9 – 12 in Panel A of Table 3 in the paper. Include it in your assignment appendix, labelled as "Table 3".

//Generating actual table output

local outcome_var avgverb avgmath

// Clear previous estimates
eststo clear

// List of dataframes to loop through
local dataframes grade5 grade4

// Loop through each dataframe
foreach df in `dataframes' {
// Change to the current dataframe
frame change `df'

// Loop through each outcome variable
foreach y in `outcome_var' {
// Run regressions and store results
eststo: reg `y' func1 tipuach, cl(schlcode)
eststo: reg `y' func1 tipuach c_size, cl(schlcode)
}
}

// Switch back to the default frame to produce the combined table
frame change default

// Output the table with all models
esttab using "table_3.csv", ///
b(%5.3f) se obslast replace ///
title("Regression Results for Grade 5 and Grade 4") ///
stats(rmse r2 N) ///
```

```
******************************** TABLE 4 ********************************

// Replicate Columns 1-3 and 7-9 of Table 4 in the paper. Include it in your assignment
appendix, labelled as "Table 4".


local outcome_var avgverb avgmath

// Clear previous estimates
eststo clear

// List of dataframes to loop through
local dataframes grade5

// Loop through each dataframe
foreach df in `dataframes' {
// Change to the current dataframe
frame change `df'

// Loop through each outcome variable
foreach y in `outcome_var' {
// Run 2SLS regressions and store results
eststo: ivregress 2sls `y' (classize = func1) tipuach, cl(schlcode)
eststo: ivregress 2sls `y' (classize = func1) tipuach c_size, cl(schlcode)
eststo: ivregress 2sls `y' (classize = func1) tipuach c_size c_size2, cl(schlcode)
}
}

// Switch back to the default frame to produce the combined table
frame change default

// Output the table with all models
esttab using "table_4.csv", ///
b(%5.3f) se obslast replace ///
title("Regression Results for Grade 5") ///
stats(rmse r2 N)
```