Causal Analysis between Income and performance in Basic education: Empirical Studies from Ontario in 2021-2022

Edward Sen

[1006719598]

[edward.sen@mail.utoronto.ca]

Steven Chung

[1007765526]

[hochunsteven.chung@mail.utoronto.ca]

ECO375: Applied Econometrics

University of Toronto

Department of Economics

Abstract:

Studies in the UK show a rising correlation between household income and educational attainment. This study aims to reveal whether similar results hold in Ontario.. Using data sourced from the government of Ontario, we studied how percentage of low-income households affected Grade 10 Ontario Secondary School Literacy Test pass rates for 717 Ontario secondary schools in years 2017-2022. We found that the percentage of low-income households has a economically insignificant but statistically significant negative effect on Grade 10 OSSLT pass rate, which vanishes when accounting for school board fixed effects.

1. Introduction

It is widely recognized that children having poorer financial backgrounds have worse educational outcomes than their wealthy peers. Research in the UK in recent decades suggests that correlation between household income and educational level has risen continually for children born post WW2 (Blandon and Paul, 2004).A report by Brookings (2011) corroborates this claim, stating that student household income is a significant factor influencing educational attainment.

The aim of this study is to test if similar results hold in Ontario. More specifically, we aim to determine how percentages of students from low-income households affect Grade 10 Ontario secondary school literacy test (OSSLT) first-time pass rates for 717 Ontario Secondary schools for school years 2017-2018 through 2021-2022, with data sourced from the Government of Ontario's data catalogue.

Using pooled ordinary least squares method, regressing OSSLT first-time pass rates on percentages of students from low-income households reveals a small, but statistically significant negative relationship between the two, implying schools with a higher percentage of low-income households have,on average, lower OSSLT pass rates. However, an analogous relationship cannot be established when panels for school boards are constructed.

2. Context and Data

Our data follows a panel structure and contains 3585 observations. Each observation represents one of 717 Ontario secondary schools, our population of interest, during a school year from September 2017 to June 2022. The main variable of interest is each secondary school's percentage of students from low-income households during a given school year, where low-income is defined as having household income less than 50,000 CAD for most family sizes. The dependent variable is each secondary school's Grade 10 OSSLT first-time pass rate during a given school year, which measures whether students have reached the minimum standard for literacy among all subjects in the elementary level.

Among all secondary schools, OSSLT first-time pass rates ranged from 0 to 1, with a mean of 0.782, and percentage of students from low-income households ranged from 0 to 0.460, with a mean of 0.151. These summary statistics can be found in table 1.

Also present in our data are various school demographic measures that we will analyze as covariates in our multiple linear regression model. These include continuous variables for each school's percentage of students whose parents forwent post-secondary education, percentage of students whose native tongue is not the instruction language, percentage of students who are new to Canada from countries that do not speak the school's instruction language, and percentage of students with special educational needs. Summary statistics for these variables can also be found in table 1.

3. Regression Analysis

3.1. Simple Linear Regression

Our baseline model for the relationship between OSSLT first-time pass rate and percentage of students from low-income households for Ontario secondary schools is given by the following equation:

$$ossltftpr_{it} = \beta_0 + \beta_1 * lowincpct_{it} + u_{it}$$

such that $lowincpct_{i,t}$ is the percentage of students from low income households for a given secondary school i during the given year t, and $OSSLTFTPR_{i,t}$ is the OSSLT first-time pass rate for the given secondary school i during the given year t.The coefficients β_0 and β_1 were estimated using pooled ordinary least squares, and robust standard errors were used to account for heteroskedasticity.

Table 2 describes the results of the baseline model. Our estimates yielded a β_1 of 0.256. This implies a school that has an additional 1% of students from low-income households has, on average, an OSSLT first-time pass rate that is 0.256% lower in any given year, which is economically insignificant. The confidence interval for average change in OSSLT first-time pass rate is [-0.322, -0.190], which does not include 0, rejecting the null hypothesis that percentage of students in low-income households does not affect OSSLT first-time pass rates. The t-statistic of the test was less than 0.001, indicating that this result is statistically significant at a 0.1% level

This simple regression model has several pitfalls concerning Least Squares Assumptions (LSAs).. Firstly, LSA 1 (E(u|X) =0) is violated as the model fails to account for covariates, causing β_1 to be biased. Secondly, LSA2 (Sample is independently and identically distributed) is violated as pooled OLS method was used to estimate the coefficients, implicitly assuming no school-specific effects, while the panel structure of the dataset implies differences in schools' underlying characteristics

Additionally, the simple regression model only has an R^2 of 0.0187, therefore only 1.87% of the variation in OSSLT first-time pass rate is explained by the model. The differences in observed and predicted values are visualized in table 3, in which there is a high amount of deviation from the line of best fit between observed and predicted OSSLT first time pass rates.

3.2. Multiple Linear Regression

Due to scarcity of literature on factors influencing literacy test performance in secondary schools, a trial-and-error approach was employed to determine the functional form of the multivariate linear model. OSSLT first-time pass rate was regressed on each relevant continuous variable individually. Each regression comprised multiple specifications corresponding to different functional forms, including linear, quadratic, cubic, lin-log, log-lin and log-log. Then, the best performing specifications were selected based on fit and parsimony with Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), for which lower relative values indicate better fit and additional parameters are penalized. RMSE was also used to evaluate goodness of fit, and specifications with statistically insignificant coefficients were not considered. To further narrow down the most ideal specifications, predicted values from best performing specifications were plotted to visually inspect for uninterpretable behavior.

Combining the best performing specifications for each variable into one expression, we obtain the maximal functional form:

$$ossltftpr_{it} = \beta_0 + \beta_1 * lowincpct_{it} + \beta_2 * lowincpct^2_{it} + \beta_3 * ln(slpct_{it}) +$$
$$\beta_4 * slcpct_{it} + \beta_5 * slcpct^2_{it} + \beta_6 * spedpct_{it} + \beta_7 * spedpct^2_{it} + \beta_8 * parnopse_{it} +$$
$$\beta_9 * eng_{it} + \beta_1 0 * pub_{it} + \delta_2 * D2018 + ...\delta_5 * D2021 + u_{it}$$

where for given secondary school i and given time t, $slpct$ denotes the percentage of students whose native tongue was not the school's language of instruction, $slcpct$ denotes the percentage of students whose country's most used language was not the school's language of instruction, $spedpct$ denotes the percentage of students with special educational needs, $parnopse$ denotes percentage of students whose parents forwent post-secondary education and $D2018$ to $D2021$ denote dummy variables for school years 2018-2019 through 2021-2022.

Backwards stepwise selection was used to simplify the maximal model, Variables resulting in the least reduction in R^2 were removed in a stepwise manner. Table 10 compares models obtained in each step based on AIC and BIC. Specifications (1) and (2) have relatively low AIC and contain more parameters, while specification (4) has the lowest BIC and less parameters, therefore we select specification (4) on the basis of parsimony and lower BIC. This yields the simplified model

$$osslt ftpr_{it} = \beta_0 + \beta_1 * lowincpct_{it}^2 + \beta_2 * ln(slpct_{it}) + \beta_3 * slcpct_{it} + \\ \beta_4 * spedpct_{it} + \beta_5 * spedpct_{it}^2 + \beta_6 * parnopse_{it} + \beta_7 * eng_{it} + \beta_8 * pub_{it} + u_{it}$$

Controlling for factors such as language spoken by student or student's country in relation to school's language of instruction, proportion of school population with special educational needs, parental educational level and type of school yields a coefficient of -0.374 for $lowincpct^2$, our variable of interest. To interpret this coefficient, we take partial derivative of pass rate with respect to the percentage of low-income household students, yielding a marginal effect of $2 * (-0.374) * lowincpct$, which indicates an increasingly large negative marginal effect of percentage of low-income students on OSSLT first-time pass rates, as illustrated in table 10.

However, despite the increasing marginal effect of percentage of students from low-income households on OSSLT first-time pass rates, the overall effect is economically insignificant for the observed range of values - at the mean percentage of low-income students, a 1% increase in percentage of low-income students is only associated with a 0.11% decrease in OSSLT first-time pass rates.

The 95% confidence interval for the effect of squared percentage of students from low-income households on OSSLT first-time pass rates is [-0.547, -0.201], thus rejecting the null hypothesis ($H_0 : \beta_1 = 0$) that percentage of students from low-income households has no effect on OSSLT first-time pass rates. The t-statistic of the test was less than 0.001, indicating that this result is statistically significant at a 0.1% level.

3.3. Extension - School Board Fixed Effects

As an extension to our analysis, we analyzed our dataset with panel data methods to account for fixed effects of policy differences between school boards and eliminate unobserved heterogeneity. School boards were chosen as the panel members to increase within variation as our dataset only had observations for five years, and school board parameters were generated by taking the mean values of school demographic parameters for individual school boards. This effectively divided Ontario secondary schools into 71 groups, and for each school board there were 355 observations across five years.

For our analysis of fixed effects, we constructed a simplified multivariate linear model given by the following equation:

$$\begin{gathered} m\_ossltftpr_{it} = m\_lowincpct_{it} + m\_slpct_{it} + m\_slcpct_{it} + \\ m\_spedpct_{it} + m\_parnopse_{it} + u_{it} \end{gathered}$$

where the $m\_$ prefix denotes the mean of previously defined school parameters, taken over individual school boards.

Table 11 describes the results of a multiple linear regression. The F-test for the model yields a F-statistic of 3.65, rejecting at a 0.1% significance level the null hypothesis that none of the variables in the model affect OSSLT first-time pass rates. However, all variable coefficients, with the exception of the dummy variable for the year 2021 and the constant term, have t-stats greater than 0.05, making them statistically insignificant at a 5% level. Given that within R^2 is relatively high at 0.159, compared to between R^2 which is 0.100, high standard errors cannot be explained by insufficient within variation. Instead, the correlation matrix for the panel data model (table) reveals high levels of correlation between covariates, indicating that the cause of the inflated standard errors is due to a case of near-multicollinearity.

4. Limitations of results

Omitted variable bias presents a major threat to the study's internal validity. Running ovtest, STATA's inbuilt version of the Ramsey RESET test on the multiple regression model specified in section 3.2, yields a F-statistic of 5.37, therefore the null hypothesis that the model has no omitted variables is rejected at a 0.11% significance level. Plausible omitted variables include laziness, intellectual quotient, and other unmeasurable factors.

Simultaneous causality bias also threatens the internal validity of our model. It assumes that a higher percentage of low-income households one-sidedly induced a lower OSSLT pass rate for secondary schools, but it is also possible that lower OSSLT pass rates negatively affected secondary school reputations, which led to higher enrollment of students from low-income households.

In addition, the study may suffer from errors-in-variables bias. Given that some schools reported no change in the values of certain demographic variables for several years, it is reasonable to suspect inaccuracies in data voluntarily reported by the schools. Also, the absence of clear incentives for student families to provide accurate income data to secondary schools presents another possible source of bias.

Finally, our model also showed limited external validity, as we failed to find datasets with the same parameters for regions similar to Ontario.

5. Conclusion

This report suggests a negative correlation between low-income household students and the Grade 10 OSSLT Test first-time pass rate, which is economically insignificant but statistically significant. We developed our multivariate model through backwise stepwise selection on a maximal model, comparing specifications based on AIC, BIC, RMSE, R^2. Then, we adopted fixed effect estimation to extend our model by transforming our dataset into panel data with school boards as panel members. The extension analysis yielded statistically insignificant results due to near-multicollinearity between the explanatory variable and the covariates.

Our findings contradict established literature on the subject, which states that income has a economically significant causal effect on education level (Blandon and Paul, 2004; Brookings, 2011). This contradiction may be attributed to the Ontario government's success in promoting equity in education, which may have minimized the effect of student household income on educational outcomes.

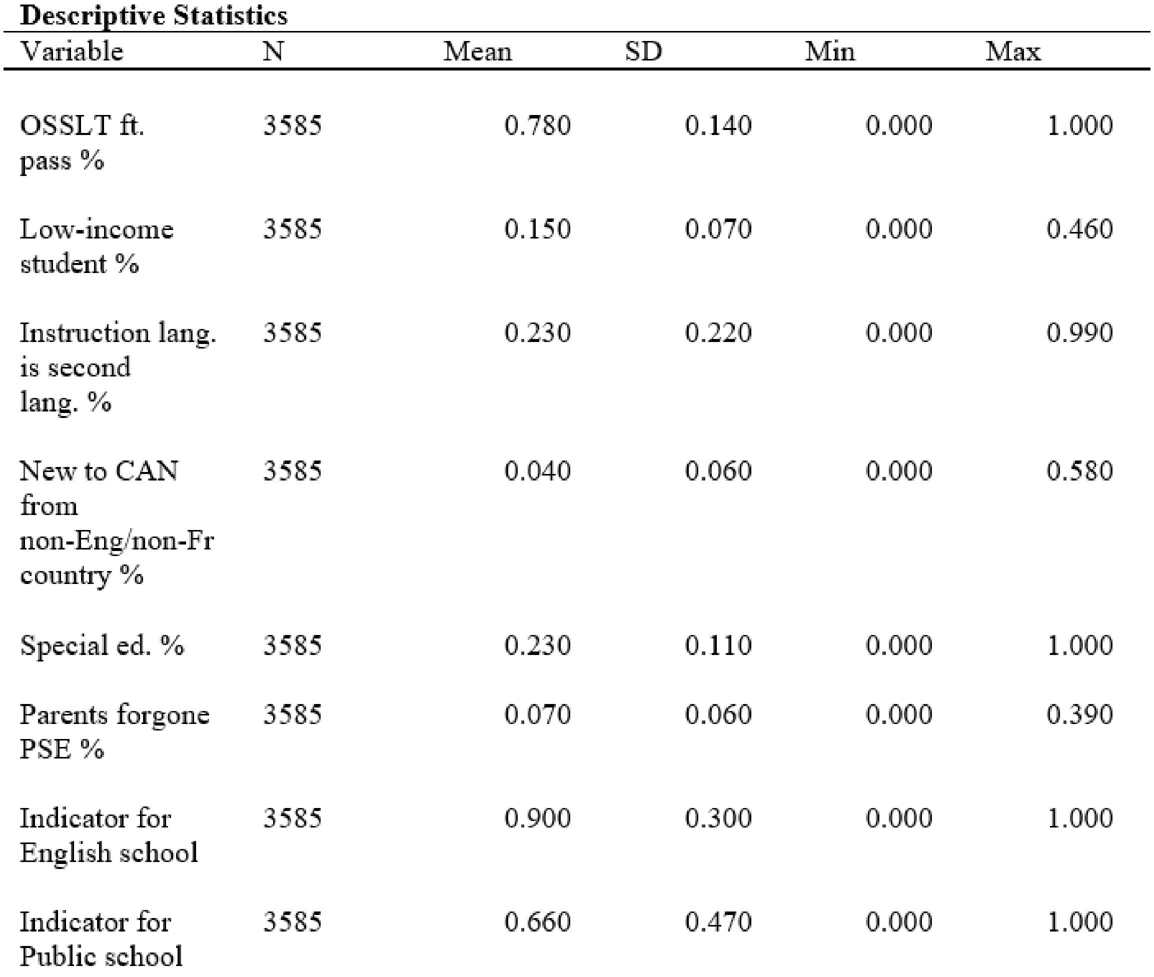
However, the study is plagued by omitted variable bias, simultaneous causality bias, errors-in-variables bias and limited external validity.

References

1. BLANDEN, JO, and PAUL GREGG. "FAMILY INCOME AND EDUCATIONAL ATTAINMENT: A REVIEW OF APPROACHES AND EVIDENCE FOR BRITAIN." Oxford Review of Economic Policy 20, no. 2 (2004): 245–63. http://www.jstor.org/stable/23606627
2. Isaacs, J. and Magnuson, K. (2011). Income and Education as Predictors of Children's School Readiness. Washington, DC: Brookings Institution. https://www.brookings.edu/articles/income-and-education-as-predictors-of-childrens-school-readiness/
3. Education Quality and Accountability Office. 'Ontario Secondary School Literacy Test (OSSLT)'. https://www.eqao.com/the-assessments/osslt/

Appendix

Table 1: Summary Statistics



Descriptive Statistics

Variable	N	Mean	SD	Min	Max
OSSLT ft. pass %	3585	0.780	0.140	0.000	1.000
Low-income student %	3585	0.150	0.070	0.000	0.460
Instruction lang. is second lang. %	3585	0.230	0.220	0.000	0.990
New to CAN from non-Eng/non-Fr country %	3585	0.040	0.060	0.000	0.580
Special ed. %	3585	0.230	0.110	0.000	1.000
Parents forgone PSE %	3585	0.070	0.060	0.000	0.390
Indicator for English school	3585	0.900	0.300	0.000	1.000
Indicator for Public school	3585	0.660	0.470	0.000	1.000

Table 2: Results from Simple Linear Regression

Simple Linear Regression

Number of obs	= 3585
F(1, 715)	= 57.90
Prob > F	= 0.0000
R-squared	= 0.0187
Root MSE	= 0.13518

Variable

OSSLT test first time pass rate	ossltftpr	Coefficient	Robust Std. Error	T-stat	P>|t|
Low-income household percentage	lowincpct	-0.256111	0.0336591	-7.61	0
intercept	_cons	0.820366	0.0051188	160.27	0

Table 3: Observed vs Predicted OSSLT first-time pass rates for simple linear model

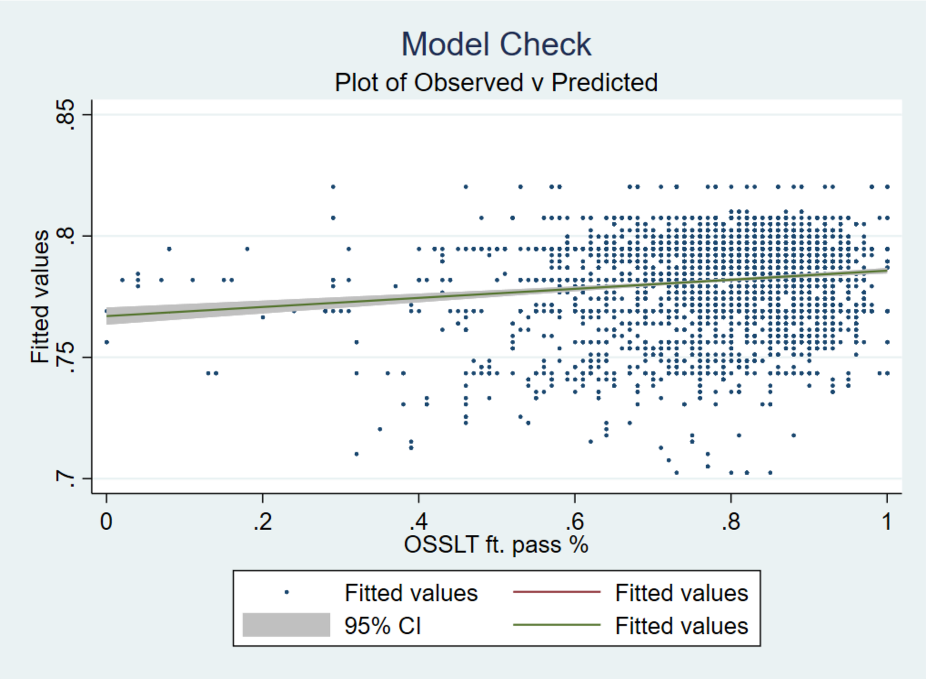


Table 4: Simple linear regression line-fit

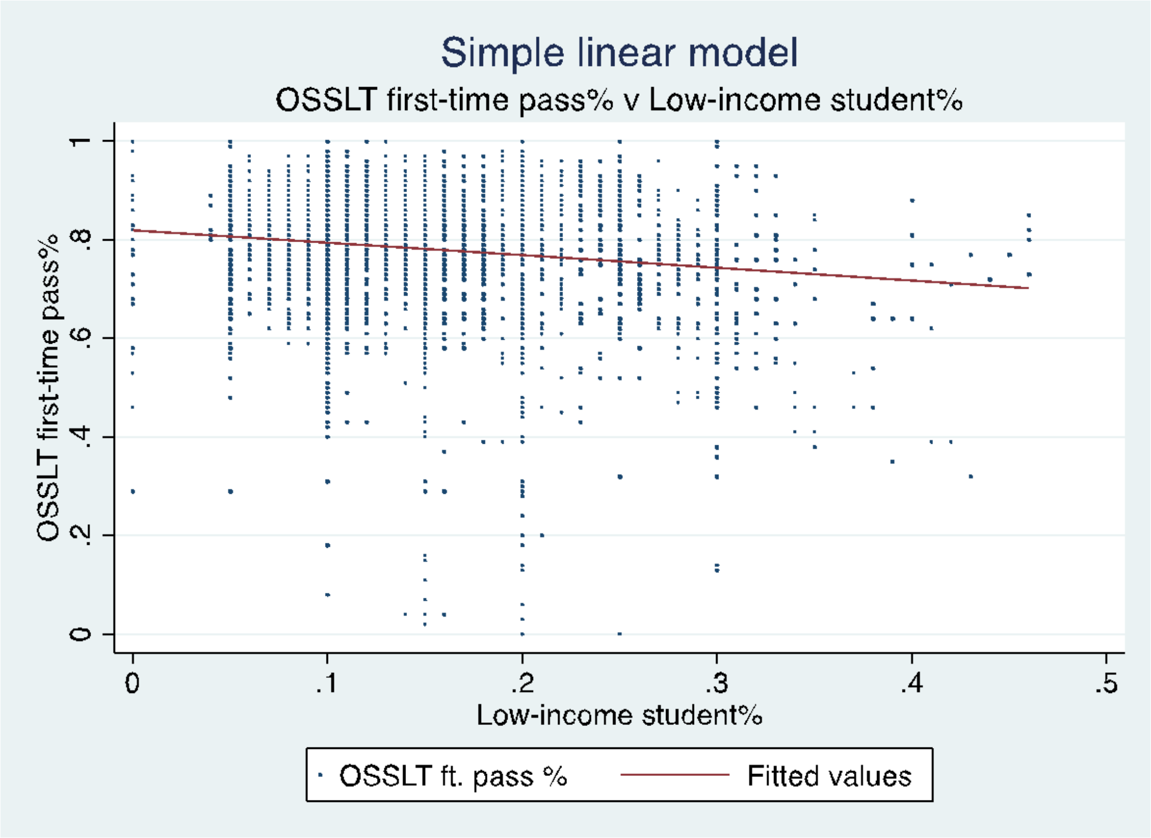


Table 5: Plots of different fits for OSSLT first-time pass rate v low-income student %

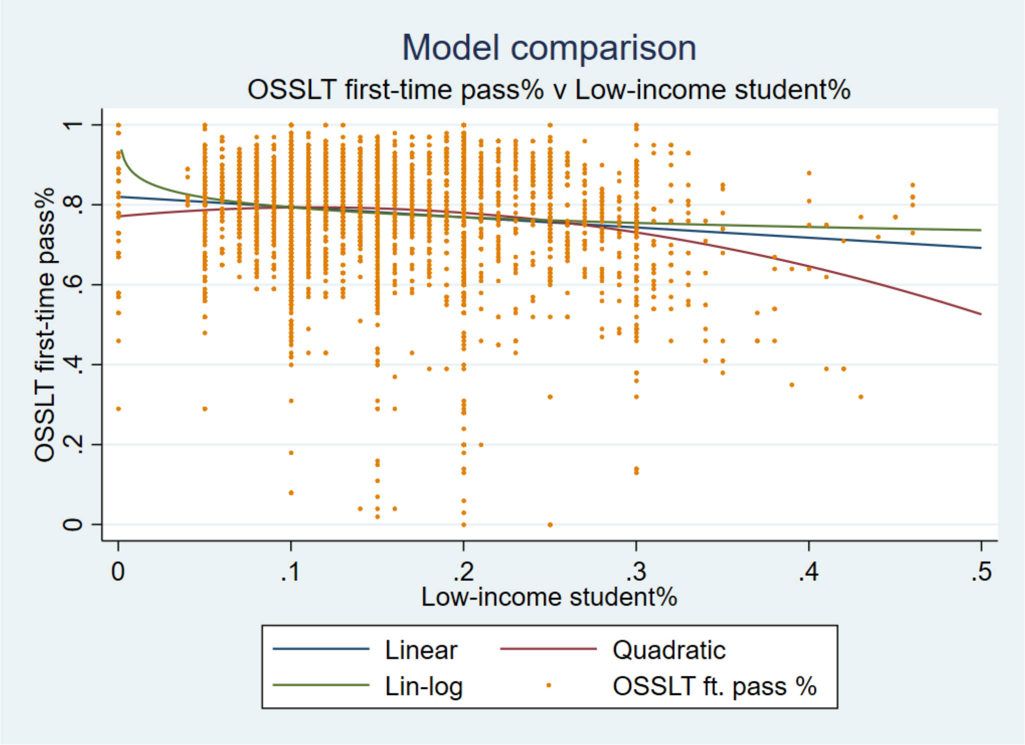


Table 6: Plots of different fits for OSSLT first-time pass rate v Non-Eng/non-French first lang %

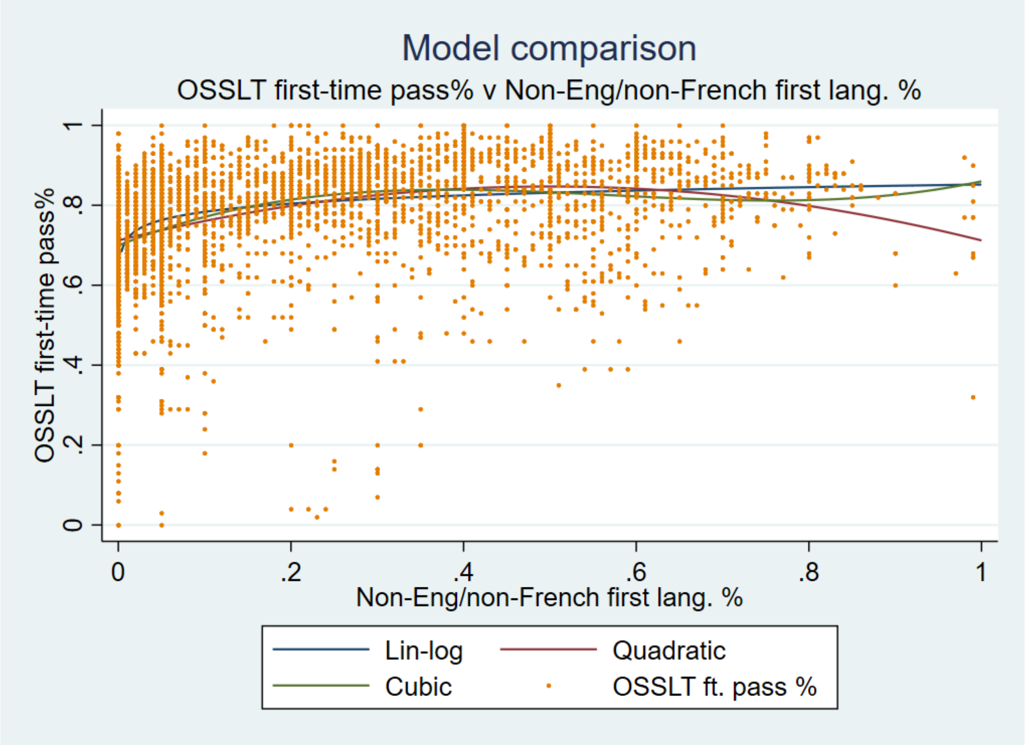


Table 7: Plots of different fits for OSSLT first-time pass rate v Parents forwent postsecondary %

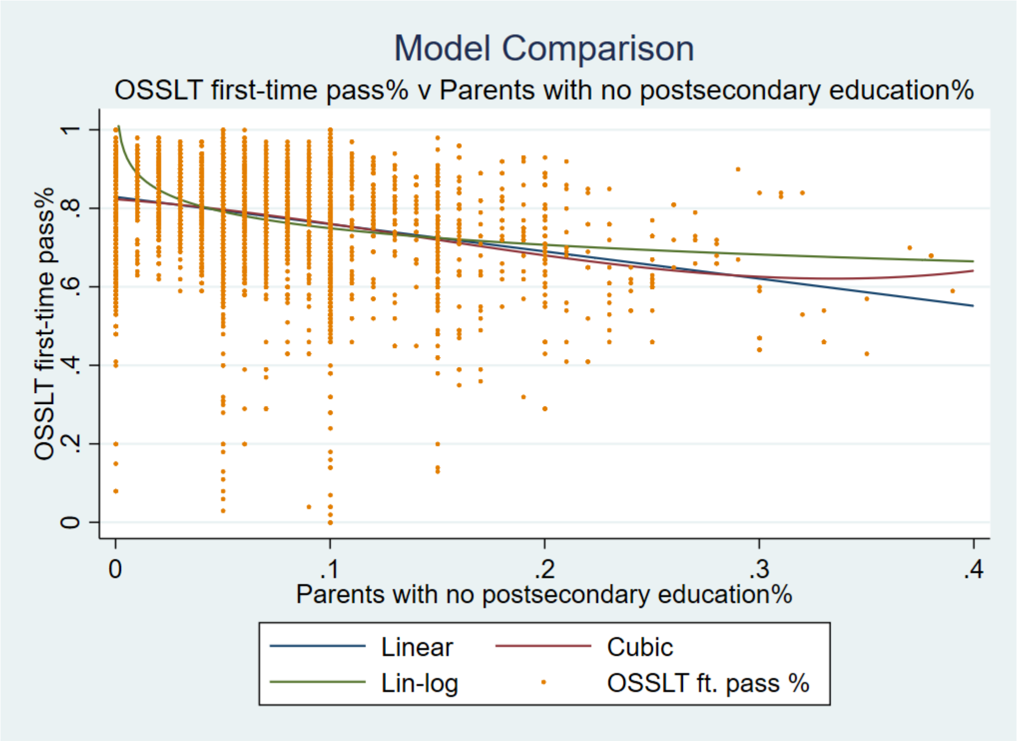


Table 8: Plots of different fits for OSSLT first-time pass rate v Special ed. %

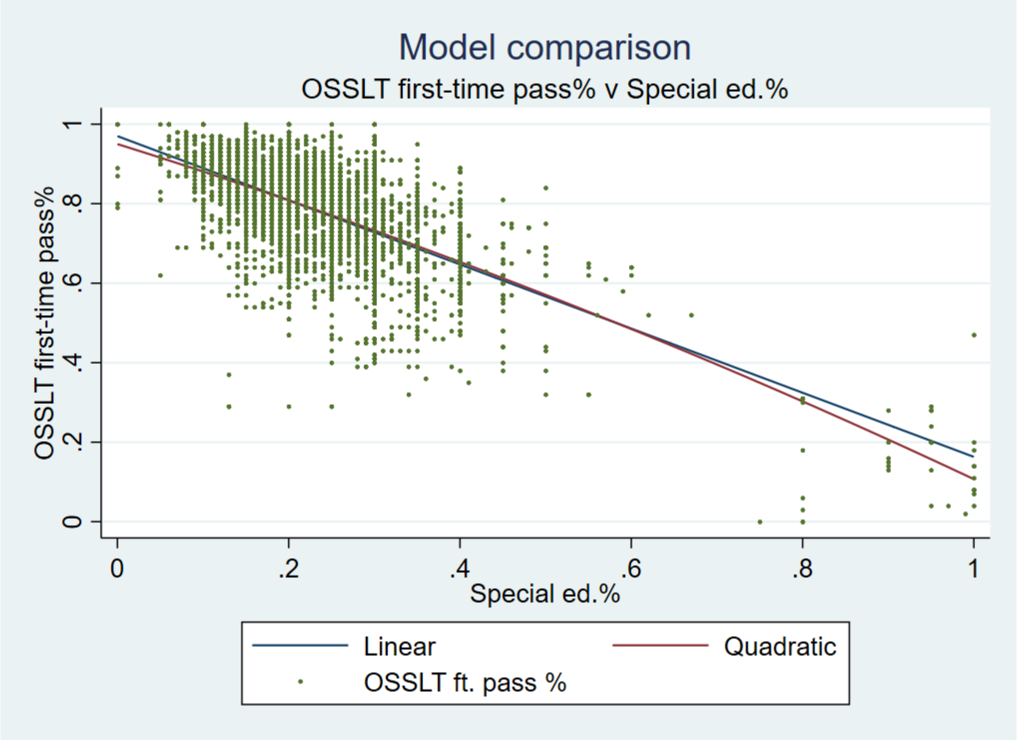
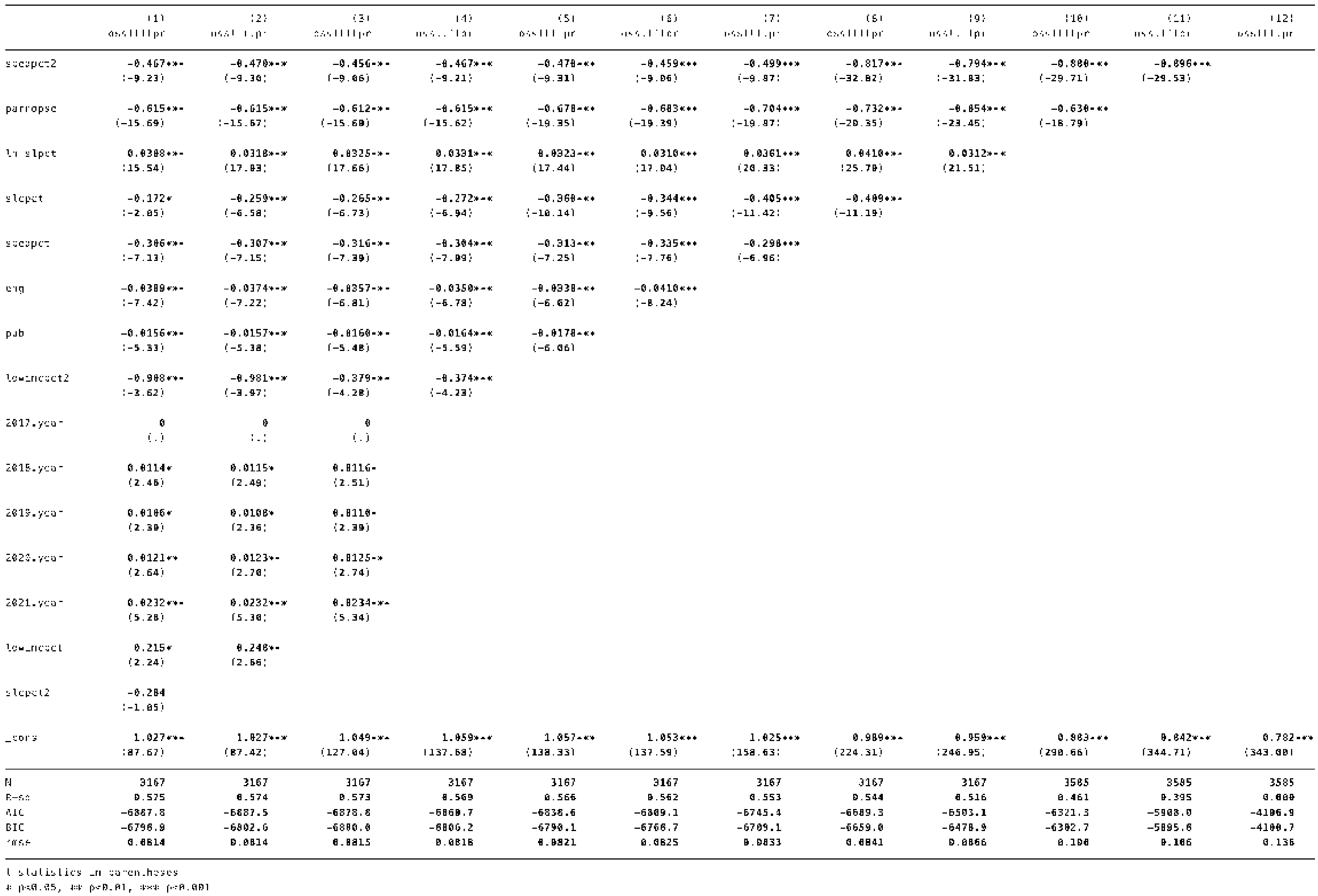


Table 9: Marginal effect of selected percentage changes in low-income student %

$\Delta lowincpct$	$\Delta ossltftpr$
0.1 to 0.2	$(-0.374) * 0.2^2 - (-0.374) * 0.1^2 = -0.112$
0.15 to 0.16	$(-0.374) * 0.16^2 - (-0.374) * 0.15^2 = -0.00116$
0.2 to 0.3	$(-0.374) * 0.3^2 - (-0.374) * 0.2^2 = -0.0187$
0.3 to 0.4	$(-0.374) * 0.4^2 - (-0.374) * 0.3^2 = -0.0262$

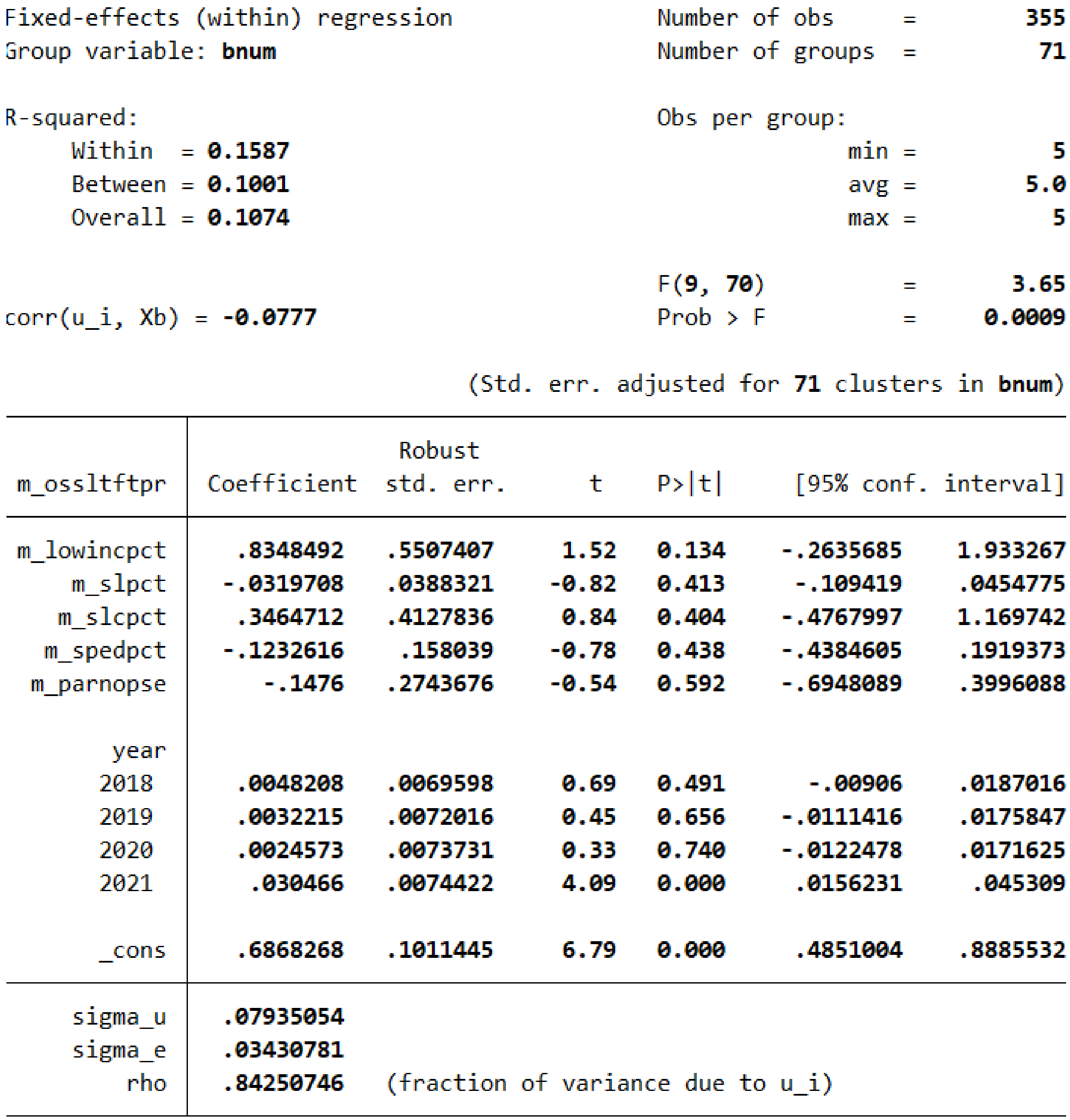
Table 5: Rotated table for different specifications in each step of the Backward stepwise selection process, referencing Table 1.



	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	ossllllpr	ossllllpr	ossllllpr	ossllllpr	ossllllpr	ossllllpr	ossllllpr	ossllllpr	ossllllpr	ossllllpr	ossllllpr	ossllllpr
sscapct2	-0.467***	-0.470***	-0.456***	-0.467***	-0.478***	-0.459***	-0.499***	-0.817***	-0.794***	-0.880***	-0.896***	
	(-9.23)	(-9.30)	(-9.06)	(-9.21)	(-9.31)	(-9.06)	(-9.87)	(-32.82)	(-31.83)	(-29.71)	(-29.53)	
parropsc	-0.615***	-0.615***	-0.612***	-0.615***	-0.678***	-0.683***	-0.704***	-0.732***	-0.854***	-0.630***		
	(-15.69)	(-15.67)	(-15.60)	(-15.62)	(-19.35)	(-19.39)	(-19.87)	(-20.35)	(-23.45)	(-18.79)		
ln slpct	0.0308***	0.0318***	0.0325***	0.0331***	0.0323***	0.0310***	0.0361***	0.0410***	0.0312***			
	(15.54)	(17.83)	(17.66)	(17.85)	(17.44)	(17.04)	(20.33)	(25.70)	(21.51)			
slpct	-0.172*	-0.259***	-0.265***	-0.272***	-0.360***	-0.344***	-0.405***	-0.409***				
	(-2.05)	(-6.58)	(-6.73)	(-6.94)	(-10.14)	(-9.56)	(-11.42)	(-11.19)				
sscapct	-0.306***	-0.307***	-0.316***	-0.304***	-0.313***	-0.335***	-0.298***					
	(-7.13)	(-7.15)	(-7.39)	(-7.09)	(-7.25)	(-7.76)	(-6.96)					
eng	-0.0389***	-0.0374***	-0.0357***	-0.0350***	-0.0338***	-0.0410***						
	(-7.42)	(-7.22)	(-6.81)	(-6.78)	(-6.62)	(-8.24)						
pub	-0.0156***	-0.0157***	-0.0160***	-0.0164***	-0.0178***							
	(-5.33)	(-5.38)	(-5.48)	(-5.59)	(-6.06)							
lowincpct2	-0.908***	-0.981***	-0.379***	-0.374***								
	(-3.62)	(-3.97)	(-4.28)	(-4.23)								
2017.year	0	0	0									
	(.)	(.)	(.)									
2018.year	0.0114*	0.0115*	0.0116*									
	(2.46)	(2.49)	(2.51)									
2019.year	0.0106*	0.0108*	0.0110*									
	(2.30)	(2.36)	(2.39)									
2020.year	0.0121**	0.0123**	0.0125**									
	(2.64)	(2.70)	(2.74)									
2021.year	0.0232***	0.0232***	0.0234***									
	(5.28)	(5.30)	(5.34)									
low_incpct	0.215*	0.240**										
	(2.24)	(2.66)										
slcpct2	-0.284											
	(-1.05)											
_cons	1.027***	1.027***	1.049***	1.059***	1.057***	1.053***	1.025***	0.989***	0.959***	0.903***	0.842***	0.782***
	(87.67)	(87.42)	(127.04)	(137.68)	(138.33)	(137.59)	(158.63)	(224.31)	(246.95)	(290.66)	(344.71)	(343.00)
N	3167	3167	3167	3167	3167	3167	3167	3167	3167	3585	3585	3585
R-sq	0.575	0.574	0.573	0.569	0.566	0.562	0.553	0.544	0.516	0.461	0.395	0.000
AIC	-6887.8	-6887.5	-6878.8	-6860.7	-6838.6	-6809.1	-6745.4	-6689.3	-6503.1	-6321.3	-5908.0	-4106.9
BIC	-6796.9	-6802.6	-6800.0	-6806.2	-6790.1	-6766.7	-6709.1	-6659.0	-6478.9	-6302.7	-5895.6	-4100.7
rmse	0.0814	0.0814	0.0815	0.0818	0.0821	0.0825	0.0833	0.0841	0.0866	0.100	0.106	0.136

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Table 11: Panel data regression results



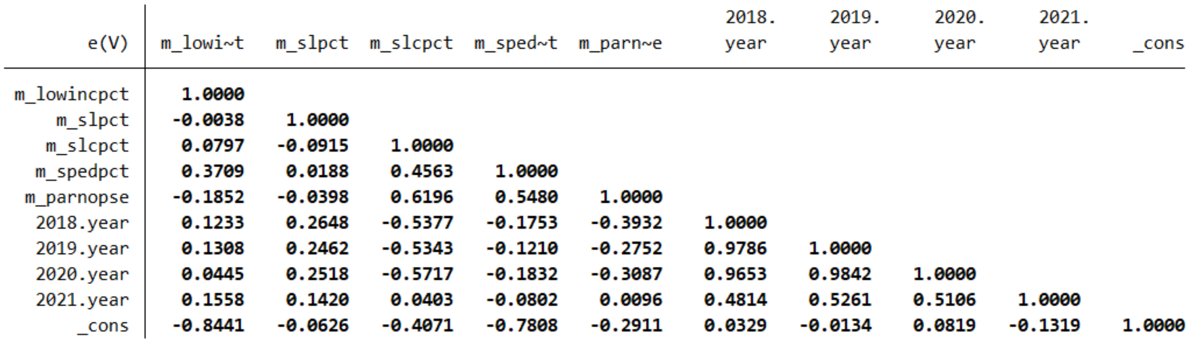
```
Fixed-effects (within) regression               Number of obs     =        355
Group variable: bnum                            Number of groups  =         71

R-squared:                                      Obs per group:
     Within  = 0.1587                                         min =          5
     Between = 0.1001                                         avg =        5.0
     Overall = 0.1074                                         max =          5

                                                F(9, 70)          =       3.65
corr(u_i, Xb) = -0.0777                         Prob > F          =     0.0009

                                 (Std. err. adjusted for 71 clusters in bnum)
------------------------------------------------------------------------------
             |               Robust
  m_ossltftpr| Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
  m_lowincpct|   .8348492   .5507407     1.52   0.134    -.2635685    1.933267
      m_slpct|  -.0319708   .0388321    -0.82   0.413     -.109419    .0454775
     m_slcpct|   .3464712   .4127836     0.84   0.404    -.4767997    1.169742
    m_spedpct|  -.1232616    .158039    -0.78   0.438    -.4384605    .1919373
   m_parnopse|     -.1476   .2743676    -0.54   0.592    -.6948089    .3996088
             |
         year|
        2018 |   .0048208   .0069598     0.69   0.491      -.00906    .0187016
        2019 |   .0032215   .0072016     0.45   0.656    -.0111416    .0175847
        2020 |   .0024573   .0073731     0.33   0.740    -.0122478    .0171625
        2021 |    .030466   .0074422     4.09   0.000     .0156231     .045309
             |
        _cons|   .6868268   .1011445     6.79   0.000     .4851004    .8885532
-------------+----------------------------------------------------------------
      sigma_u|  .07935054
      sigma_e|  .03430781
          rho|  .84250746   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Table 12: Correlation matrix for panel data model



```
                                                                  2018.     2019.     2020.     2021.
        e(V)  m_lowi~t    m_slpct   m_slcpct  m_sped~t  m_parn~e       year      year      year      year     _cons

 m_lowincpct    1.0000
     m_slpct   -0.0038    1.0000
    m_slcpct    0.0797   -0.0915    1.0000
   m_spedpct    0.3709    0.0188    0.4563    1.0000
  m_parnopse   -0.1852   -0.0398    0.6196    0.5480    1.0000
   2018.year    0.1233    0.2648   -0.5377   -0.1753   -0.3932    1.0000
   2019.year    0.1308    0.2462   -0.5343   -0.1210   -0.2752    0.9786    1.0000
   2020.year    0.0445    0.2518   -0.5717   -0.1832   -0.3087    0.9653    0.9842    1.0000
   2021.year    0.1558    0.1420    0.0403   -0.0802    0.0096    0.4814    0.5261    0.5106    1.0000
       _cons   -0.8441   -0.0626   -0.4071   -0.7808   -0.2911    0.0329   -0.0134    0.0819   -0.1319    1.0000
```