University of Warsaw
Faculty of Economic Sciences

Family Background in Returns to Education

An Empirical Analysis of Family's Impact on Personal Achievement

Submitted to

Professor Rafał Woźniak

for

19/20 Advanced Econometrics Project

Author:

Zimin Luo (417124)

August 29, 2020

Contents

Abstract

This report estimates the effects of family background in returns to education using panel data collected by the National Longitudinal Survey of Youth (NLSY) and processed by Koop and Tobias. Two empirical models are constructed, the first one consists of all linear relationships between the individual's wage after logarithm transform and all explanatory variables, the second one with an additional quadratic term of the individual's potential, which resembles the Mincer wage equation. The result shows that the latter one fit the data better. Meanwhile, OLS, fixed effects model and random effects model are used and the random effects model is eventually chosen as the final model as opposed to the Hausman test. The result suggests that living in a broken residence has negative impact on one's earnings. Furthermore, parent's education is used as instruments to account for the endogeneity in education and returns to education is estimated using the 2SLS approach. Both empirical results indicate that family background has a significant role in one's earnings.

Chapter 1

Introduction

Intergenerational mobility, returns to education and inequalities have become increasingly popular in the research field in recent years. Intergenerational mobility refers to changes in socioeconomic status from one generation to another, while returns to education is the increase in wage from one additional year of schooling. Conceptually, all three influence each other. It has been widely acknowledged that intergenerational mobility is strongly correlated with inequality, in other words, in countries with higher inequality, children are more likely to have the same socioeconomic background as their parents. In 2012, American economist Alan Krueger introduced the Great Gatsby Curve and showed the positive correlation between intergenerational immobility and inequality across different countries[15]. His idea was later supported by Jerrim et al[13], whom also discovered that educational attainment is the key linkage between inequality and intergenerational. Such finding is consistent with the human capital theory: education improves one's skills and productivity, hence the wages and socioeconomic status. Moreover, if the educational attainment has a strong correlation with an individual's family socioeconomic background, the inequality is more likely to pass down to the next generation. Studies concerning intergenerational mobility have spiked in recent decades thanks to the availability of the data, which provided grounds for policymakers to review and improve current redistributive policies. For instance, researchers from Taiwan and Argentina have found that the current university subsidiary programs are in fact exacerbating inequalities, as individuals attending universities are from the most privileged families [16][20]. Similarly, estimating returns to education can help us allocate resources more effectively. Some studies have shown that the returns to return is not linear across different cohorts: having a higher education level tends to increase the returns to education, meaning that an additional year at university have a higher increase in wage than an additional year in high school[17], whereas others found contradictory results [1]. In this project, my focus is not on the returns to education itself, rather, the intergenerational transmission in returns to education.

1.1 Objective

The main objective of this project is to verify the hypothesis that family background plays a significant role in children's achievement. If the hypothesis is proven to be true, I would like to further investigate to what extend family's background affect children's achievement; conversely, if the alternative hypothesis is true, that the family background have insignificant impact on children's achievements, I would like to know how much individual's personal characteristics affect his or her achievements. The sub-tasks of this project include: construction of empirical models for the main objective, selection of an appropriate regression model, and interpretation of the results.

1.2 Methodology

One's achievements can be in many forms. In this project, the dependent variable is restricted to be the individual's wage in natural log form for robust results. First of all, two empirical models are estimated using (1) a linear combination of all the available estimators including an interaction term and (2) an additional quadratic explanatory variable which is selected via exploratory analysis; next, ordinary least squares (OLS), fixed effects and random effects models are regressed on both models for comparison purpose. R^2 values are used to judge the fitness of empirical models and Breusch-Pagan Lagrangian test, F-tests are used to measure the individual effects and Hasuman tests are used to compare the random effects model and fixed effects model. After the final model is decided, instrumental variable approach is also utilized as an attempt to address the endogeneity bias.

1.3 Structure

The report is organized in the following way: the second chapter discusses some of the current findings in intergenerational mobility and returns to education; the third chapter describes the basic statistics of the dataset and representations of all the variables; the fourth chapter is devoted to explaining how the empirical models are built and how the final model is chosen; followed by an instrumental variable approach, the last chapter summaizes the findings and conclusions.

Chapter 2

Literature Review

Previous studies have already confirmed the existence of intergenerational transmission in various forms: Liu et al [16] found that in Taiwan family income is positively significant in the children's attendance of universities, individuals attending the most prestigious universities are generally from the wealthiest families, meanwhile, mother's educational level has stronger effect than the father's; Rozada et al [20] found that in Argentina individuals that attending higher education are from much wealthier families than those who do not, although no significant differences in socio-economic background are found between those who attend private universities and those who public universities; in rural China, significant intergenerational transmission was found in groups who were born after the 1980s but not before that, and the transmission effect is less so than in urban areas [9]; in Norway, father's education was found insignificant in children's education, but mother's education is only significant in the son's education not the daughters [4]. One of the main challenges in estimating intergenerational mobility and returns to education is that the results can hardly be generalized, as such, studies of different subjects or in different countries can have very different results. Pronzato argued that with relatively smaller sample size, a stronger father's effect is likely to be discovered in the study of intergenerational transmission of education [18].

The other main challenge in estimating returns to education is the endogeneity problem, particularly in error terms correlated with education, and the omitted abilities, causing the estimators to be biased. To tackle this issue, instrumental variables estimations are frequently used. Brunello and [6], Ismail [12] and Black [4] used educational system reform, a dummy variable indicating whether the individual was born before or after the reform, the researchers argue that such change is exogenous to the wage but correlated with individual's educational attainment, which should lead to a consistent estimates of returns to education; Chen et al [7] used spouses' education as instruments because couples share similar, if not the same, educations; Bhatti et al divided the individuals into sub-groups (by gender, age, residential area) and used average years of schooling of the sub-groups as instruments[11][10] , which combines some conventional instruments such as the distance to the schools.

Chapter 3

The Data

The original dataset was obtained from the National Longitudinal Survey of Youth (NLSY) in the United States and was preprocessed by Koop and Tobias[14]. A total number of 17919 observations were sampled from the national data, which consists of 2178 individuals' information including educational years, hourly wage and family background from 1979 to 1993. The Koop and Tobias dataset selected only white males aged between 16 and 22 when being surveyed to maximize consistency, and worked at least 30 weeks and 800 hours a year to ensure they have already been actively involved in the workforce. Koop and Tobias have filtered out misreported data such as hourly wage less than $1 or larger than $100.

The panel data consists of 10 variables: each individual is assigned an unique `PERSONID`, four time-varying variables that include schooling years `EDUC`, hourly wage in natural logarithm form `LOGWAGE`, potential experience `POTEXPER` is constructed as $\texttt{POTEXPER} = \texttt{age} - \texttt{EDUC} - 5$, and a time trend `TIMETRND`; five time-invariant variables include the `ABILITY`, which is a score that is measured by cognitive studies, mother's education `MOTHERED` and father's education `FATHERED` measured in years, a dummy variable `BROKNHOME` that is 1 if the individual is residing in a broken

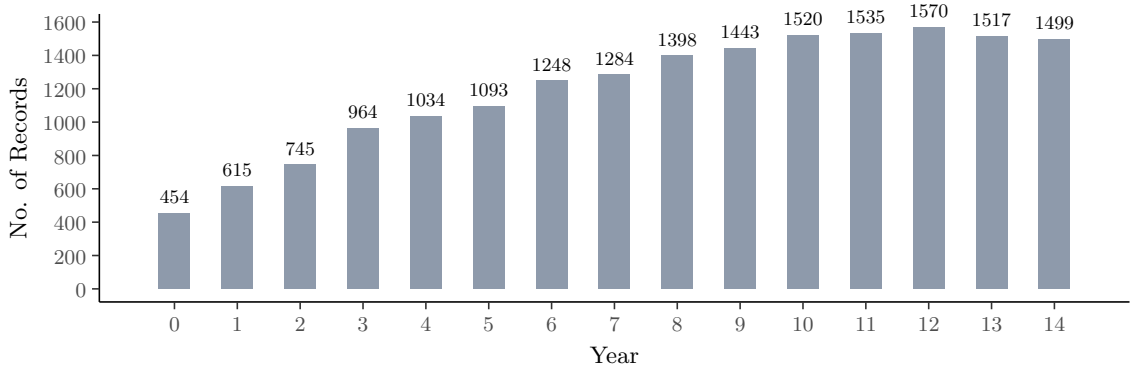

Figure 3.1. Distribution of the yearly record shows that panel data is heavily unbalanced.

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
PERSONID	17,855	1,081.736	629.827	1	534	1,633	2,178
EDUC	17,855	12.687	1.916	9	12	14	20
LOGWAGE	17,855	2.299	0.527	0.070	1.960	2.660	4.570
POTEXPER	17,855	8.363	4.127	0	5	11	22
TIMETRND	17,855	8.204	3.954	0	5	12	14
ABILITY	17,855	0.057	0.923	−4	−0.5	0.8	2
MOTHERED	17,855	11.478	2.985	0	11	12	20
FATHERED	17,855	11.716	3.764	0	10	14	20
BRKNHOME	17,855	0.153	0.360	0	0	0	1
SIBLINGS	17,855	3.155	2.121	0	2	4	18

Table 3.1. Basic statistics of the final dataset. 42 PERSONIDs are removed from the original Koops and Tobias dataset.

home and 0 otherwise, as well as the number of `SIBLINGS` of the individual. In this project, `EDUC`, `POTEXPER` and `ABILITY` are deemed as an individual's abilities and `MOTHERED`, `FATHERED`, `BROKNHOME` and `SIBLINGS` are deemed as the individual's background. The hypothesis is that at least one of the family background variables is significant variables that affects one's `LOGWAGE`.

3.1 Exploratory Analysis

Figure 3.1 shows the distribution of the number of observations per year. It is clear that the panel data is unbalanced: having only 454 observations in the first year of the survey and 1,499 in the last year, and the measured unbalancedness is $\gamma = 0.655$ and $\nu = 0.827$. The main concern with the panel data being unbalanced is that it can lead to biased estimations when the observations are systematically missing, therefore, it is prominent to identify the reasons why they are missing. To deal with this issue, one approach is to simply drop all subjects that have missing observations, however, this could result in survival bias. Considering that students who are working while they are still studying (as part-timers) earns systematically lower than those who have graduated (can work as full-time employees), I decided to remove individuals that have very few observations as well as low highest educations. By doing this, I eliminated individuals who may have not graduated by the end of the survey. The thresholds set for minimum number of records is 3 and 10 for maximum schooling years. Moreover, for simplicity purpose, it is assumed that all remaining observations are missing completely at random (MCAR) thus the estimators are consistent and unbiased in this regard. The final data summary is listed in Table 3.1 after 42 individual records are dropped from the Koop and Tobias dataset.

Visualizations are used to help with the establishment of the empirical models, each explanatory variable is plotted against the dependent variable `LOGWAGE`. In most cases, a rather linear rela-

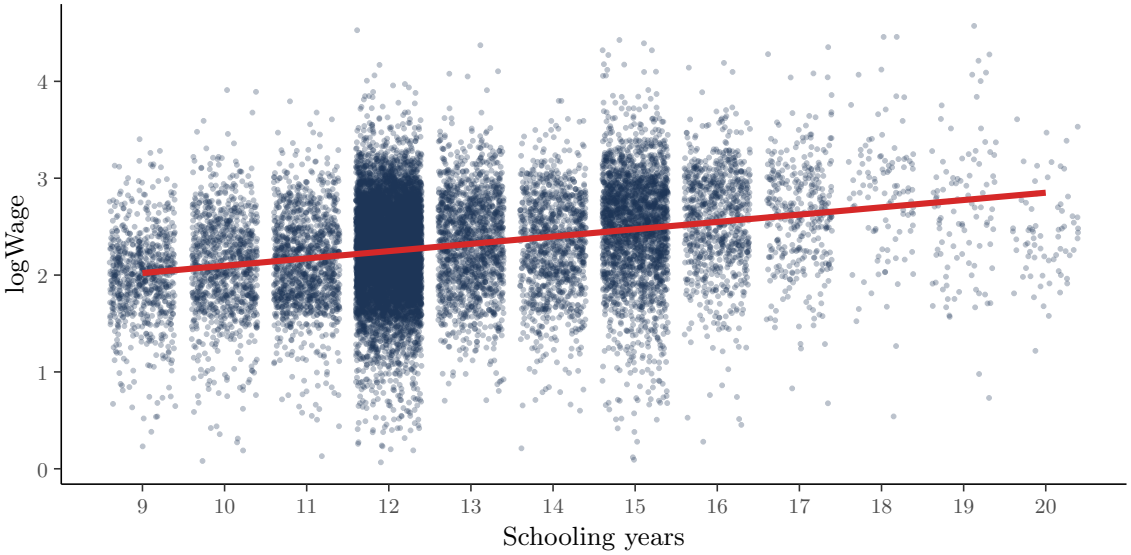

Figure 3.2. Scatter plot of logWage by Schooling Years of the entire panel data. A positive relationship can be observed. The red solid line represents the slope of a linear function of schooling that explains log wage: $\log wage = \beta schooling\_years + u$

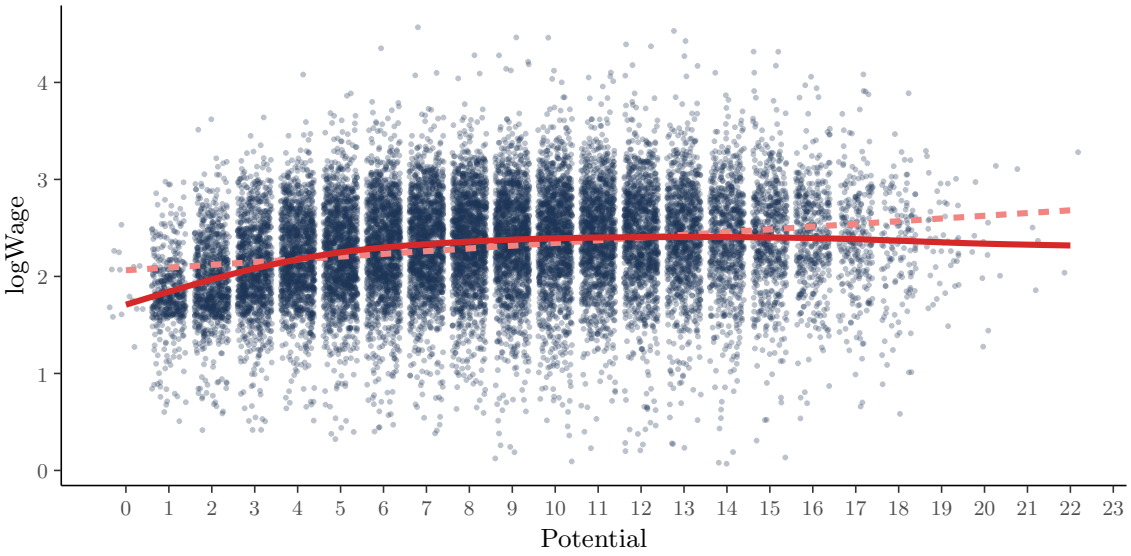

Figure 3.3. Scatter plot of logWage by Potential. The red dashed line represents the slope of a linear function of potential that explains log wage: $\log wage = \beta potential + u$, the solid line represents a non-linear relationship which fits better than the linear relationship.

tionship can be observed between the explanatory variable and dependent variable, including the schooling years `EDUC` and the log wage `LOGWAGE`, as shown in Figure 3.2; yet a non-linear relationship tends to fit the log wage `LOGWAGE` and potential `POTEXPER` plot better than a linear one does (Figure 3.3). This observation is consistent with the Mincer equation, nonetheless, two separate empirical models are constructed in order to compare the performances in the next section.

Chapter 4

Empirical Models

Based on the explanatory data analysis, the empirical models with multiple regressions are proposed as follows:

$$\begin{aligned}\texttt{LOGWAGE}_{it} =& \beta_0 + \beta_1 \texttt{EDUC}_{it} + \beta_2 \texttt{POTEXPER}_{it} + \beta_3 \texttt{ABILITY}_i + \beta_4 \texttt{MOTHERED}_i + \beta_5 \texttt{FATHERED}_i \\ &+ \beta_6 \texttt{BROKNHOME}_i + \beta_7 \texttt{SIBLINGS} + \beta_8 \texttt{MOTHERED}_i * \texttt{FATHERED}_i + u_i + \varepsilon_{it}\end{aligned} \tag{4.1}$$

where $\texttt{MOTHERED}_i * \texttt{FATHERED}_i$ is an interaction term of father's education and mother's education, β_i are the estimators for each variable, u_i contains all unobservable time-invariant individual characteristics, such as intelligence, persistence, resourcefulness, etc., and ε_{it} are random disturbances. The subscript it denotes time-variant variables and i denotes time-invariant variables.

The first model consists only linear relationships between all seven explanatory variables and the dependent variable. In the second model, an quadratic term $\texttt{POTEXPER}^2$ is introduced as explained in the previous section:

$$\begin{aligned}\texttt{LOGWAGE}_{it} =& \beta_0 + \beta_1 \texttt{EDUC}_{it} + \beta_2 \texttt{POTEXPER}_{it} + \beta_3 \texttt{POTEXPER}_{it}^2 + \beta_4 \texttt{ABILITY}_i \\ &+ \beta_5 \texttt{MOTHERED}_i + \beta_6 \texttt{FATHERED}_i + \beta_7 \texttt{BROKNHOME}_i + \beta_8 \texttt{SIBLINGS} \\ &+ \beta_9 \texttt{MOTHERED}_i * \texttt{FATHERED}_i + u_i + \varepsilon_{it}\end{aligned} \tag{4.2}$$

To find out which model has a better explanatory power, OLS, fixed effects and random effects regressions are applied on both models and the regression results are shown in Table 4.1.

4.1 Empirical Results

OLS1, FE1 and RE1 are the corresponding regressions applied on Model 1 (Equation 4.1) and OLS2, FE2, and RE2 on Model 2 (Equation 4.2). RE1s and RE2s are derived from RE1 and RE2 with all insignificant variables removed via the general-to-specific approach. For Model 1, the returns to education estimator β_1 for OLS is 0.07, meaning that one additional year of schooling will increase individual's wage by 7 percent; it increased to 12.4 percent when fixed effects model is used and 9.4 percent when using random effects regression. The estimator for POTEXPER is fairly consistent across three regressions, meaning that one additional unit of POTEXPER increases individual's wage by around 4 percent. The remaining variables are all time-invariant, which fixed effects models ignored. The OLS regression estimates that one additional score in ABILITY increases wage by 7.8 percent, living in a broken home reduces individual's wage by 5.3 percentage as compared to those who don't, and having one additional sibling increases the individual's wage by 0.5 percent, the remaining variables are all insignificant. Meanwhile, the RE regression estimates that one unit additional ABILITY increases the individual's wage by 6.3 percent, and living in a broken home reduces the individual's wage by 4.8 percent, the remaining variables including SIBLINGS are insignificant.

For Model 2, after introducing $\texttt{POTEXPER}^2$ to the supply side, the estimations become much closer for all three regressions and nearly identical for FE2 and RE2 regressions. The returns to education has generally reduced from that of the first model, reaching at 6.5 percent for OLS2, and around 7.5 percent for FE2 and RE2. The effect of ABILITY have increased to around 8 percent. The most drastic change comes from POTEXPER's estimator β_2, which nearly doubled for all three regressions compared with their counterpart, making it the most dominate variable in the model. The quadratic term becomes significant despite having a small negative value -0.004, and the negative effect of BROKNHOME have slightly increased to -0.055 for all three regressions. SIBLINGS remains significant for OLS2 only and its value decreased from 0.005 to 0.004.

Given the significance level at 0.05, the F-tests report significant individual effects ($F_{2129,15717} = 9.745$, $p < 0.0001$ for Model 1, and $F_{2129,15716} = 10.182$, $p < 0.0001$ for Model 2), likewise, the Breusch-Pagan Lagrangian test report significant differences across individuals ($\chi_1^2 = 17606$, $p < 0.0001$ for Model 1 and $\chi_1^2 = 18586$, $p < 0.0001$ for Model 2), hence, the OLS is worse than both fixed effects models and random effects models; the Hausman specification tests report the random effects models are inconsistent ($\chi_2^2 = 52.323$, $p < 0.0001$ for Model 1 and $\chi_3^2 = 32.096$, $p < 0.0001$ for Model 2), therefore, fixed effects models are preferred over the random effects models. On the other hand, Model 2 generally has better explanatory power than Model 1 for having higher R^2 values, it thereby confirms the previous assumption that the relationship between an individual's potential POTEXPER and LOGWAGE is non-linear.

	Dependent variable:							
	LOGWAGE							
	OLS1	FE1	RE1	RE1s	OLS2	FE2	RE2	RE2s
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
EDUC	0.070***	0.124***	0.094***	0.094***	0.065***	0.076***	0.073***	0.074***
	(0.002)	(0.006)	(0.004)	(0.004)	(0.002)	(0.006)	(0.004)	(0.004)
POTEXPER	0.040***	0.039***	0.040***	0.040***	0.095***	0.107***	0.106***	0.106***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.003)	(0.003)	(0.003)	(0.003)
I(POTEXPER^2)					−0.003***	−0.004***	−0.004***	−0.004***
					(0.0002)	(0.0001)	(0.0001)	(0.0001)
ABILITY	0.078***		0.060***	0.063***	0.080***		0.080***	0.086***
	(0.005)		(0.011)	(0.010)	(0.005)		(0.011)	(0.010)
MOTHERED	−0.003		0.003		−0.003		0.001	
	(0.003)		(0.006)		(0.003)		(0.006)	
FATHERED	0.002		0.006		0.002		0.002	
	(0.003)		(0.006)		(0.003)		(0.006)	
BRKNHOME	−0.053***		−0.048**	−0.048**	−0.055***		−0.055**	−0.055**
	(0.010)		(0.022)	(0.022)	(0.010)		(0.022)	(0.022)
SIBLINGS	0.005**		0.003		0.004**		0.002	
	(0.002)		(0.004)		(0.002)		(0.004)	
MOTHERED:FATHERED	0.0003		−0.0003		0.0003		0.0001	
	(0.0002)		(0.0005)		(0.0002)		(0.0005)	
Constant	1.035***		0.671***	0.743***	0.903***		0.732***	0.773***
	(0.044)		(0.081)	(0.048)	(0.044)		(0.081)	(0.047)
Observations	17,855	17,855	17,855	17,855	17,855	17,855	17,855	17,855
R^2	0.175	0.197	0.252	0.252	0.188	0.229	0.278	0.278
Adjusted R^2	0.175	0.088	0.252	0.252	0.187	0.124	0.278	0.278
F Statistic	473.157*** (df = 8; 17846)	1,930.178*** (df = 2; 15717)	5,918.765***	5,917.002***	458.552*** (df = 9; 17845)	1,558.957*** (df = 3; 15716)	6,799.768***	6,796.731***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4.1. Regression Results

4.2 The Final Model

More often than not, the Hausman specification test has been used as a convention to decide between the fixed effects model and the random effects model, with its null hypothesis being "both models are consistent but random effects model is more efficient", and the alternative hypothesis "only the fixed effects model is consistent". The idea of Hausman test is that when the explanatory variables are all exogenous, both estimates are consistent but random effects have a smaller variance hence it is more efficient; when there exists endogeneity, the random effects estimates are biased hence can not be adopted. Yet, its validity has been challenged [8]. In addition, rejecting the random effects model based solely on the Hausman specification test is not helpful as all the potential effects caused by time-invariant variables are filtered out by the fixed effects model, such as parental education, number of siblings, etc., which are of interest in this project. The aim of this section is to justify why the random effects model is a more appropriate choice for this project, despite the Hausman test result.

Fixed effects models assume homogeneity whereas the random effects models allow for heterogeneity, nonetheless, fixed effects models are almost always preferred in the research field due to the underlying endogeneity that violates the random effects model assumption. A general random effects model takes the following form:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \beta_2 z_i + (u_i + \varepsilon_{it}) \tag{4.3}$$

which satisfies the following assumptions[1]:

$$E(u_i|x_{it}, z_i) = E(\varepsilon_{it}|x_{it}, z_i) = 0 \tag{4.4}$$

y_{it} denotes the dependent variable, x_{it} is a vector that contains all time-varying explanatory variables, z_i is the time-invariant individual characteristics, and ε_{it} is random disturbance. When u_i is correlated with x_{it},, the random effects estimates suffers from the endogenous bias. According to Bell et al[3][2], the panel data has a hierarchical structure: Level 1 contains all time-varying observations and Level 2 contains all time-invariant observations[2], and the source of the endogeneity can be explained via Equation 4.5:

$$x_{it} = x_i^B + x_{it}^W \tag{4.5}$$

where x_i^B denotes the higher-level between effects, x_{it}^W denotes the lower-level within effects, x_{it} denotes the total effect at Level 1. By default, x_i^B and x_{it}^W are assumed to be equal at β_1 is consistent; if such condition is not met, β_1 is unable to account for both effects simultaneously and inevitably causes the unexplained residuals to be absorbed by $u_i + \varepsilon_{it}$, hence the correlation and endogenous bias (to be more specific, heterogeneity bias). The authors indeed proposed a

[1] Apart from the normality assumptions

[2] Throughout this project, Level 1 is deemed as the lower level and Level 2 the higher level.

hybrid model that combines the two models that named the within-between RE model (REWB), which unfortunately is beyond the scope of this project.

By and large, taken into consideration that there is no drastic difference between the random effects estimates RE2s and the FE2 estimates, more importantly, its ability to address the effects of the time-invariant variables, we conclude that the random effects model RE2s is favored despite the Hasuman test result. Certainly, it's probable that there is still endogeneity in the model, but ultimately, "all models are wrong; the practical question is how wrong do they have to be to not be useful"[5]. The final model is presented in Equation 4.6:

$$\begin{aligned}\texttt{LOGWAGE}_{it} =& 0.773 + 0.074 \times \texttt{EDUC}_{it} + 0.106 \times \texttt{POTEXPER}_{it} - 0.004 \times \texttt{POTEXPER}_{it}^2 \\ &+ 0.086 \times \texttt{ABILITY}_i - 0.055 \times \texttt{BROKNHOME}_i + u_i + \varepsilon_{it}\end{aligned} \tag{4.6}$$

With the model being finalized, it is hereby confirmed that the hypothesis that family background indeed impacts an individual's achievements is true, to be more specific, living in a broken household will decrease the individual's income by 5.5%, compared to those who do not. Other variables that constitutes family background, such as parents' schooling years, number of siblings have no significant impact on one's wage. Significant variables include the individual's schooling years: one additinoal year in schools increases the hourly wage by 7.4%; one additinoal unit score in the cognitive study increases the hourly wage by 8.6%; the potentional has a non-linear effect, an additinoal x unit increase in one's potential result in 10.6%x - 0.4%x^2 change in the hourly wage.

Multiple diagnostics have been run on the final model in Equation 4.6, unfortunately, the model suffers from (1) cross-sectional dependence which could lead to biased estimates (Breusch-Pagan LM test reports $\chi^2_{1906878} = 3114220$, $p < 0.0001$ and Pesaran's CD test reports $z = 55.188$, $p < 0.0001$), (2) heteroskedasticity, which may cause the variances estimates to be biased and inconsistent (Breusch-Pagan Test reports $\chi^2_5 = 231.86$, $p < 0.0001$), (3) serial correlation (Breusch-Godfrey/Wooldridge test reports $\chi^2_1 = 1338.5$, $p < 0.0001$), which may not only lead to biased standard errors but also a seemingly higher R^2 value. Coefficient test is performed subsequently to account for the potential biases, and the results are presented in Table 4.2.

4.3 Instrumental Variables

The final model was selected based on the similarities of the estimators between different models, and the flexibility to estimate the time-invariant variables, but the existence of endogeneity was not addressed. In the previous section, the Hausman test indicates that the explanatory variables are correlated with the error term ($Cov(u_i, x_{it}) \neq 0$), which can cause the estimators to be biased. The goal of this section is to apply 2SLS regression to tackle this problem.

	Dependent variable:				
	LOGWAGE				
	panel linear		*coefficient test*		
	(1)	(2)	(3)	(4)	(5)
EDUC	0.074***	0.074***	0.074***	0.074***	0.074***
	(0.004)	(0.005)	(0.005)	(0.004)	(0.006)
POTEXPER	0.106***	0.106***	0.106***	0.106***	0.106***
	(0.003)	(0.004)	(0.004)	(0.003)	(0.010)
I(POTEXPERˆ2)	−0.004***	−0.004***	−0.004***	−0.004***	−0.004***
	(0.0001)	(0.0002)	(0.0002)	(0.0002)	(0.0004)
ABILITY	0.086***	0.086***	0.086***	0.086***	0.086***
	(0.010)	(0.011)	(0.011)	(0.010)	(0.029)
BRKNHOME	−0.055**	−0.055**	−0.055**	−0.055**	−0.055**
	(0.022)	(0.022)	(0.022)	(0.024)	(0.022)
Constant	0.773***	0.773***	0.773***	0.773***	0.773***
	(0.047)	(0.065)	(0.065)	(0.052)	(0.089)
Observations	17,855				
R^2	0.278				
Adjusted R^2	0.278				
F Statistic	6,796.731***				
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 4.2. Coefficient tests for (1) original model, (2) cluster-robust estimator, (3) heteroskedasticity-robust estimator, (4) autocorrelation-robust estimator, (5) cross-sectional and serial correlation estimator.

Assuming that in Equation 4.6 schooling years EDUC is endogenous, and POTEXPER, ABILITY and BROKNHOME are all exogenous, we can use parent's education as the instruments to estimate the returns to education. I argue that the parents' education are valid instruments because (1) they have no direct effect on children's hour wage, which was proven in earlier estimates (2) empirical studies have found that they are correlated with children's education. The equation is formulated as follows:

$$\begin{aligned}\texttt{LOGWAGE}_{it} =& \beta_0 + \beta_1 \texttt{EDUC}_i + \beta_2 \texttt{POTEXPER}_i + \beta_3 \texttt{POTEXPER}_i^2 \\ &+ \beta_4 \texttt{ABILITY}_i + \beta_5 \texttt{BROKNHOME}_i + \beta_6 \texttt{SIBLINGS} + u\end{aligned} \tag{4.7}$$

where

$$\begin{aligned}\texttt{EDUC}_i =& \pi_0 + \pi_1 \texttt{MOTHERED}_i + \pi_2 \texttt{FATHERED}_i + \pi_3 \texttt{POTEXPER}_i \\ &+ \pi_4 \texttt{POTEXPER}_i^2 + \pi_5 \texttt{ABILITY}_i + \pi_6 \texttt{BROKNHOME}_i + \pi_7 \texttt{SIBLINGS} + v\end{aligned} \tag{4.8}$$

The 2SLS result is presented in Table 4.3. Compared to results from Model 2, the returns to school estimated by 2SLS rises up to 12.8%, 41% higher than returns to school estimated by FE2 and RE2s and nearly doubled than in OLS2, becoming comparable to FE1, all the other significant estimators remain significant but their absolute values become smaller: POTEXPER drops by about 11% for OLS2 and 18% for FE2 and RE2s to 8.7%, ABILITY drops by around 69% to 2.7%, the negative effect of BRKNHOME reduces from -5.5% to -3.2%. SIBLINGS becomes significant, and has a positive impact on hourly wage: one additional siblings increases the individual's hourly wage by 0.6%. The weak instruments diagnose reports that we have strong instruments ($F_{2,17847} = 270.541$, $p < 0.001$), the Wu-Hasuman test reports that the 2SLS is preferred over OLS ($F_{1,17847} = 21.781$, $p < 0.001$), and Sargan test reports invalid instruments ($\chi_1^2 = 0.923$, $p = 0.337$). The result is consistent with the previous study that the 2SLS estimation of returns to education is higher than OLS [7].

	Dependent variable:			
	LOGWAGE			
		panel linear		*instrumental variable*
	OLS2	FE2	RE2s	parent's education
	(1)	(2)	(3)	(4)
EDUC	0.065***	0.076***	0.074***	0.128***
	(0.002)	(0.006)	(0.004)	(0.013)
POTEXPER	0.095***	0.107***	0.106***	0.087***
	(0.003)	(0.003)	(0.003)	(0.004)
I(POTEXPERˆ2)	−0.003***	−0.004***	−0.004***	−0.003***
	(0.0002)	(0.0001)	(0.0001)	(0.0002)
ABILITY	0.080***		0.086***	0.027*
	(0.005)		(0.010)	(0.014)
MOTHERED	−0.003			
	(0.003)			
FATHERED	0.002			
	(0.003)			
BRKNHOME	−0.055***		−0.055**	−0.032***
	(0.010)		(0.022)	(0.011)
SIBLINGS	0.004**			0.006***
	(0.002)			(0.002)
MOTHERED:FATHERED	0.0003			
	(0.0002)			
Constant	0.903***		0.773***	0.153
	(0.044)		(0.047)	(0.168)
Observations	17,855	17,855	17,855	17,855
R^2	0.188	0.229	0.278	0.154
Adjusted R^2	0.187	0.124	0.278	0.154
Residual Std. Error				0.485 (df = 17848)
F Statistic	458.552*** (df = 9; 17845)	1,558.957*** (df = 3; 15716)	6,796.731***	
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 4.3

Chapter 5

Conclusions

In this project, the effect of family background on individual's earnings using the panel data provided by Koop and Tobias was estimated using random effects model and instrumental variables approach. Both models imply that family background played a significant role in individual's earnings. The random effects model indicated that living in a broken reduces one's wage by 5.5%; in the meantime, parent's education are found to have no direct impact on the individual's earnings. Using parent's education as instruments, it is found that the effect of living in a broken household reduces to by 3.2% whereas the effect of siblings becomes significant, having one more sibling increases the individual's wage by 0.6%. The reason for such positive relationship is unclear as the "siblings effect" can have mixed influences on one's life. The returns to education estimated by the random effects model was 7.4%, using instrumental approach substantially increases the return rate to 12.8%, yet both values still greatly differ from the US benchmark (10%) set by Psacharopoulos et al [19]. It is worth noting that the random effects model suffers from cross-sectional dependence, heteroskedasticity, and serial correlation, and despite a high F-test value for the weak instruments test, the validity of using family background as instruments remains an empirical question [21].

Bibliography

[1] Harry and Patrinos. Estimating the return to schooling using the Mincer equation. *IZA World of Labor*, 2016.

[2] Andrew Bell, Malcolm Fairbrother, and Kelvyn Jones. Fixed and random effects models: making an informed choice. *Quality and Quantity*, 2019.

[3] Andrew Bell and Kelvyn Jones. Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science Research and Methods*, 2015.

[4] Sandra E. Black, Paul J. Devereux, and Kjell G. Salvanes. Why the apple doesn't fall far: Understanding intergenerational transmission of human capital, 2005.

[5] G Box and N Draper. Empirical Model Building and Response Surfaces, 1st Edition. *Wiley series in probability and mathematical statics, ISBN 978-0471810339*, 1987.

[6] Giorgio Brunello and Raffaele Miniaci. The economic returns to schooling for Italian men. An evaluation based on instrumental variables. *Labour Economics*, 1999.

[7] Guifu CHEN and Shigeyuki HAMORI. Economic returns to schooling in urban China: OLS and the instrumental variables approach. *China Economic Review*, 2009.

[8] Tom S. Clark and Drew A. Linzer. Should I Use Fixed or Random Effects? *Political Science Research and Methods*, 2015.

[9] Yongqing Dong, Renfu Luo, Linxiu Zhang, Chengfang Liu, and Yunli Bai. Intergenerational transmission of education: The case of rural China. *China Economic Review*, 2019.

[10] Sajjad Haider Bhatti, Muhammad Aslam, and Jean Bourdon. Market Returns to Education in Pakistan, Corrected for Endogeneity Bias. *THE LAHORE JOURNAL OF ECONOMICS*, 2018.

[11] Sajjad Haider Bhatti, Jean Bourdon, and Muhammad Aslam. Economic Returns to Education in France: OLS and Instrumental Variable Estimations. *THE LAHORE JOURNAL OF ECONOMICS*, 2013.

[12] Ismail and Ramlee. The Impact of Schooling Reform on Returns to Education in Malaysia. *MPRA Paper*, 2012.

[13] John Jerrim and Lindsey Macmillan. Income Inequality, Intergenerational Mobility, and the Great Gatsby Curve: Is Education the Key? *Social Forces*, 94(2):505–533, 06 2015.

[14] Gary Koop and Justin L. Tobias. Learning about heterogeneity in returns to schooling. *Journal of Applied Econometrics*, 2004.

[15] Alan Krueger. The rise and consequences of inequality in the united states, 01 2012.

[16] Jin Tan Liu, Shin Yi Chou, and Jin Long Liu. Asymmetries in progression in higher education in Taiwan: Parental education and income effects. *Economics of Education Review*, 25(6):647–658, 2006.

[17] Claudio E. Montenegro and Harry Anthony Patrinos. Comparable estimates of returns to schooling around the world. *World Bank Group: Education Global Practice Group*, 2014.

[18] Chiara Pronzato. An examination of paternal and maternal intergenerational transmission of schooling. *Journal of Population Economics*, 2012.

[19] George Psacharopoulos and Harry Anthony Patrinos. Returns to investment in education: A further update. *Education Economics*, 2004.

[20] Martín Rozada and Alicia Menendez. Public university in argentina: Subsidizing the rich? *Economics of Education Review*, 21:341–351, 02 2002.

[21] Philip Trostel, Ian Walker, and Paul Woolley. Estimates of the economic return to schooling for 28 countries. *Labour Economics*, 2002.

Appendix A

R code

```
# load libraries
library(tidyverse)    # for pipe operations
library(stargazer)    # for latex formatting
library(plm)          # for panel data regression
library(aod)          # for wald test
library(AER)          # for IV estimates ivreg
library(lmtest)       # model test
library(ggplot2)      # for vidualizations
library(ggExtra)
library(ggthemes)
library(tikzDevice)   # for latex figures
#For some reason, Rstudio needs to know the time zone for tikzDevice
options(tz="PL")
library(reshape2)     # melt function
# set customized plot colors for visualizations
col_b = "#1d3557"
col_r = "#d62828"
col_g = "#2a9d8f"
col_i = "#264653"

# ------------------------------------------------------------------
# variables:
#
# PERSONID  -   personal ID
# EDUC      -   educational level in years
# LOGWAGE   -   per hour logWage
# POTEXPER  -   individual's potential = age - education - 5
# TIMETRND  -   time trend
# ABILITY   -   individual's cognitive score
# MOTHERED  -   mother's educational level in years
# FATHERED  -   father's educational level in years
# BRKNHOME  -   a dummy variable indicating if the individual is from a
    broken home
# SIBLINGS  -   number of siblings in the household
# ------------------------------------------------------------------
```

```r
# ------------------------------------------------------------------
# IMPORT DATA
# ------------------------------------------------------------------

# load data
df <- read.csv("data.csv")
# take a glance at the dataframe
tibble::glimpse(df)

# print summary in a nice format
stargazer(df, type="text", summary = TRUE)

# check if the panel data is balanced
# count the number of observations by year
table(df$TIMETRND)
# the result suggests that it is unbalanced dataset
# visualization
df %>%
  group_by(TIMETRND) %>%
  summarise(CT = n()) %>%
  ggplot(., aes(x=TIMETRND, y=CT)) +
  geom_col(fill=col_b, width=0.5, alpha=0.5) +
  geom_text(aes(label=CT), position=position_dodge(width=0.9), vjust=-0.5) +
  ylim(0, 1800) +
  labs(
    x = "Year",
    y = "Counts"
  ) +
  scale_x_continuous(breaks = seq(0, 14, by = 1)) +
  scale_y_continuous(breaks = seq(0, 1600, by = 200)) +
  theme_classic()

# unbalancedness of the data
punbalancedness(df, index=c("PERSONID", "TIMETRND"))

# for latex plot only
# ----------------------
# tikz(file = "logWageEduc.tex", width = 6, height = 3, bg = "transparent",
    pointsize = 10)
# plot <- ggplot(df, aes(x=as.factor(EDUC), y=LOGWAGE)) +
#   geom_boxplot(fill=col_b, width=0.5, alpha=0.5, size=1, color=col_b,
    outlier.size=0.5, outlier.alpha=0.1) +
#   # geom_smooth(method = "lm", se=FALSE, color = col_r, size=2) +
#   labs(
#     y = "logWage",
#     x = "Schooling Years"
#   ) +
#   # scale_x_continuous(breaks = seq(0, 20, by = 1)) +
#   theme_classic()
#
```

```
# #This line is only necessary if you want to preview the plot right after
    compiling
# print(plot)
# #Necessary to close or the tikxDevice .tex file will not be written
# dev.off()
#
# # for latex plot only
# tikz(file = "timetrnd.tex", width = 6, height = 2, bg = "transparent",
    pointsize = 10)
# plot <- df %>%
#   group_by(TIMETRND) %>%
#   summarise(CT = n()) %>%
#   ggplot(., aes(x=TIMETRND, y=CT)) +
#   geom_col(fill=col_b, width=0.5, alpha=0.5) +
#   geom_text(aes(label=CT), position=position_dodge(width=0.9), vjust=-0.5)
    +
#   ylim(0, 1800) +
#   labs(
#     x = "Year",
#     y = "Counts"
#   ) +
#   scale_x_continuous(breaks = seq(0, 14, by = 1)) +
#   scale_y_continuous(breaks = seq(0, 1600, by = 200)) +
#   theme_classic()
# #This line is only necessary if you want to preview the plot right after
    compiling
# print(plot)
# #Necessary to close or the tikxDevice .tex file will not be written
# dev.off()


# ------------------------------------------------------------------
# DATA CLEANING
# ------------------------------------------------------------------

# create a new dataframe 'df_edu' that contains each individual's highest
# educational attainment only,
# use TIMETRND == max(TIMETRND) to filter the latest record
df_edu <- df %>%
  group_by(PERSONID) %>%
  filter(EDUC == max(EDUC), TIMETRND == max(TIMETRND))  %>%
  mutate(FAMED = (MOTHERED+FATHERED)/2)

head(df_edu)

# count the number of occurences for each individual
n_TIME <- df %>%
  group_by(PERSONID) %>%
  select(TIMETRND) %>%
  summarize(n())
```

```r
# assign the new colnames
colnames(n_TIME) <- c("PERSONID", "TIMECTN")

# inner join df_edu & n_TIME on PERSONID, we have a new dataframe
# that contains information of each individual's highest
# educational attainment and his/her number of records in the dataset.
# The thresholds are set to be:
# TIMECTN <= 3, and EDUC < 9
# for these records are likely to be
IDs_to_drop <- df_edu %>%
  inner_join(n_TIME, by=("PERSONID" = "PERSONID")) %>%
  filter(TIMECTN < 3 & EDUC <= 10)

# 42 individual records will be removed
dim(IDs_to_drop)[1]

# print IDs_to_drop
IDs_to_drop

# drop them
df <- subset(df,!PERSONID %in% IDs_to_drop$PERSONID)
stargazer(df, type="text", summary = TRUE)

# ----------------------------------------------------------------
# DATA EXPLORATION
# ----------------------------------------------------------------

# basic distribution plots

# density plot of dependent var: LOGWAGE
ggplot(df, aes(LOGWAGE)) +
  geom_density(fill=col_b, color="white", alpha=0.5) +
  labs(
    title="Density plot of LOGWAGE",
    x = "LogWage",
    y = "Density"
  ) +
  theme_classic()

# for latex plot only
# ---------------------
# tikz(file = "../report/figures/logWageDensity.pdf", width = 6, height = 3,
    bg = "transparent",  pointsize = 10)
# ggplot(df, aes(LOGWAGE)) +
#   geom_density(fill=col_b, color="white", alpha=0.5) +
#   labs(
#     x = "LogWage",
#     y = "Density"
#   ) +
#   theme_classic()
# dev.off()
```

```
# distribution of father's, mother's and children's education
df_edu %>%
  melt(id.vars=c("PERSONID", "LOGWAGE","POTEXPER", "TIMETRND", "ABILITY",
                 "BRKNHOME","SIBLINGS"), measure.vars=c("FATHERED", "MOTHERED
                    ", "EDUC"),
       variable.name = "FAMEMBER",
       value.name = "HIEDUC") %>%
  group_by(FAMEMBER) %>%
  ggplot(aes(HIEDUC, fill=FAMEMBER)) +
  geom_bar(color="white", alpha=0.7, width=0.75, position="dodge") +
  scale_fill_manual(values = c(col_b, col_r, col_g)) +
  scale_x_continuous(breaks=seq(0, 20, 1)) +
  scale_y_continuous(breaks=seq(0, 1000, 200)) +
  labs(
    title = "Counts per each family member by schooling years",
    x = "Schooling Years (highest)",
    y = "Counts"
  ) +
  theme_classic()


# explanatory variables vs dependent variables
# the intention is to explore the relationship between various explanatory
# variables and the dependent variables

# EDUC vs. LOGWAGE
# according to Mincer, log-wage is linear in schooling
# from the visualization, it is clear that there is a positive relationship
    between
# log wage and schooling years. And the relationship is plausibly linear.
ggplot(df, aes(x=EDUC, y=LOGWAGE)) +
  geom_point(color=col_b, alpha=0.3, position = "jitter", size=0.5) +
  geom_smooth(method="lm",se=FALSE, color=col_r, size=2) +
  labs(
    title = "Education years vs. Log Wage",
    x = "Schooling years",
    y = "log Wage"
  ) +
  scale_x_continuous(breaks=seq(0,20,by=1)) +
  theme_classic()

# for latex plot only
# ----------------------
# tikz(file = "../report/figures/logWageSchooling.pdf", width = 6, height =
    3, bg = "transparent",  pointsize = 10)
# ggplot(df, aes(x=EDUC, y=LOGWAGE)) +
#   geom_point(color=col_b, alpha=0.3, position = "jitter", size=0.5) +
#   geom_smooth(method="lm",se=FALSE, color=col_r, size=2) +
#   labs(
#     x = "Schooling years",
```

```
#     y = "log Wage"
#   ) +
#   scale_x_continuous(breaks=seq(0,20,by=1)) +
#   theme_classic()
# dev.off()

# POTEXPER vs LOGWAGE
# from the visualization that a quadratic relationship fits better
# than a linear relationship
ggplot(df, aes(x=POTEXPER, y=LOGWAGE)) +
  geom_point(color=col_b, alpha=0.3, position = "jitter", size=1) +
  geom_smooth(method="lm",se=FALSE, color="#f28482", size=2, linetype = "
      dashed") +
  geom_smooth(se=FALSE, color=col_r, alpha=0.5, size=2) +
  labs(
    title = "Potential vs logWage",
    x = "Potential",
    y = "logWage"
  ) +
  scale_x_continuous(breaks=seq(0,24,by=1)) +
  theme_classic()


# # for latex only
# tikz(file = "../report/figures/potexper.pdf", width = 6, height = 3, bg = "
    transparent",  pointsize = 10)
# ggplot(df, aes(x=POTEXPER, y=LOGWAGE)) +
#   geom_point(color=col_b, alpha=0.3, position = "jitter", size=0.5) +
#   geom_smooth(method="lm",se=FALSE, color="#f28482", size=2, linetype = "
    dashed") +
#   geom_smooth(se=FALSE, color=col_r, alpha=0.5, size=2) +
#   labs(
#     x = "Potential",
#     y = "logWage"
#   ) +
#   scale_x_continuous(breaks=seq(0,24,by=1)) +
#   theme_classic()
# dev.off()

# ABILITY vs LOGWAGE
# a linear relationship seems fit
ggplot(df, aes(x=ABILITY, y=LOGWAGE)) +
  geom_point(color=col_b, alpha=0.3, size=1) +
  geom_smooth(method="lm", se=FALSE, color=col_r, size=1.5) +
  labs(
    title = "Ability vs logWage",
    x = "Ability",
    y = "logWage"
  ) +
  theme_classic()
```

```r
# BRKNHOME vs LOGWAGE
# residing in a broken (BROKNHOME = 1) household seems to have a negative
# impact on the LOGWAGE: having a lower mean wage than BROKNHOME = 0
ggplot(df, aes(x=factor(BRKNHOME), y=LOGWAGE)) +
  geom_boxplot(fill=c(col_b, col_r), width=0.5, size=1.5, alpha=0.5) +
  labs(
    title = "Broken Home vs logWage",
    x = "Broken Home",
    y = "logWage"
  ) +
  theme_classic()

# SIBLINGS vs LOGWAGE
# a linear relationship seems fit
ggplot(df, aes(x=SIBLINGS, y=LOGWAGE)) +
  geom_point(color=col_b, alpha=0.3, size=1, position = "jitter") +
  geom_smooth(method="lm", se=FALSE, color=col_r, size=1.5) +
  labs(
    title = "Number of siblings vs logWage",
    x = "siblings",
    y = "logWage"
  ) +
  theme_classic()

# FATHERED vs. LOGWAGE
# a linear relationship seems fit
ggplot(df, aes(x=FATHERED, y=LOGWAGE)) +
  geom_point(color=col_b, alpha=0.3, position = "jitter") +
  geom_smooth(method="lm", se=FALSE, color=col_r, size=1.5) +
  labs(
    title = "Father's education vs. children's Wage",
    x = "Father's Education Level (Years)",
    y = "Children's logWage"
  ) +
  scale_x_continuous(breaks=seq(0, 20, 1)) +
  theme_classic()

# MOTHERED vs. LOGWAGE
# a linear relationship seems fit
ggplot(df, aes(x=MOTHERED, y=LOGWAGE)) +
  geom_point(color=col_b, alpha=0.3, position = "jitter") +
  geom_smooth(method="lm", se=FALSE, color=col_r, size=1.5) +
  labs(
    title = "Mother's education vs. children's Wage",
    x = "Mother's Education Level (Years)",
    y = "Children's logWage"
  ) +
  scale_x_continuous(breaks=seq(0, 20, 1)) +
  theme_classic()

# from the visualizations above, I decided to construct a linear model
```

```
# between all the explanatory variables and the dependent variable LOGWAGE.
# A quadratic term of POTEXPER will be added to a second model for comparison

# ------------------------------------------------------------------
# PANEL MODELS
# ------------------------------------------------------------------

# test the model using ‘ols‘, ‘fixed effects‘‘, and ‘random effects‘
# respectively.
# model: LOGWAGE~EDUC+POTEXPER+I(POTEXPER^2)+ABILITY+MOTHERED+FATHERED+
# BRKNHOME+SIBLINGS+MOTHERED:FATHERED

# Model 1:
# LOGWAGE ~ EDUC+POTEXPER+ABILITY+MOTHERED+FATHERED+BRKNHOME+SIBLINGS+
    MOTHERED:FATHERED,
# ----------
# ols model
ols.1 <- plm(LOGWAGE~EDUC+POTEXPER+ABILITY+MOTHERED+FATHERED+
             BRKNHOME+SIBLINGS+MOTHERED:FATHERED,
           data=df, index=c("PERSONID", "TIMETRND"), model="pooling")

# check model summary
summary(ols.1)

# fixed model
fixed.1 <-plm(LOGWAGE~EDUC+POTEXPER+ABILITY+MOTHERED+FATHERED+
              BRKNHOME+SIBLINGS+MOTHERED:FATHERED,
            data=df, index=c("PERSONID", "TIMETRND"), model="within")

# check model summary
summary(fixed.1)

# random model
random.1 <- plm(LOGWAGE~EDUC+POTEXPER+ABILITY+MOTHERED+FATHERED+
                BRKNHOME+SIBLINGS+MOTHERED:FATHERED,
              data=df, index=c("PERSONID", "TIMETRND"), model="random")

# check model summary
summary(random.1)

# compare the models to find the most appropriate one
# poolability test
pFtest(fixed.1, ols.1) # fixed effect is more appropriate

# plm test
plmtest(ols.1, type=c("bp")) # random effect is more appropriate

# Hausman test
phtest(fixed.1, random.1) # fixed effect is more appropriate

# compare the results
```

```
stargazer(ols.1, fixed.1, random.1, type ="text",
          column.labels = c("OLS", "FE","RE"))


# ------------------------------------------------------------------
# General to Specific Approach: RANDOM MODEL 1
# ------------------------------------------------------------------

# look at the model summary again
summary(random.1)

# check joint significanc - can we remove all insignificant variables at once
    ?
# - MOTHERED
# - FATHERED
# - SIBLINGS
# - MOTHERED:FATHERED

random.1$coefficients



# use Wald test to test the hypothesis that all insignificant var are jointly
# insignificant
H <- rbind(
  c(0, 0, 0, 0, 1, 0, 0, 0, 0),
  c(0, 0, 0, 0, 0, 1, 0, 0, 0),
  c(0, 0, 0, 0, 0, 0, 0, 1, 0),
  c(0, 0, 0, 0, 0, 0, 0, 0, 1)
)

# wald test to check if the smaller model performs better
wald.test(b = coef(random.1), Sigma = vcov(random.1), L = H)

# P(> X2) = 0.77 -> we can't reject the Null Hypothesis that these variables
    are all
# jointly insignificant based on 0.05 significance level

# remove all insignificant variables:

random.1s <- plm(LOGWAGE~EDUC+POTEXPER+ABILITY+BRKNHOME,
                 data=df, index=c("PERSONID", "TIMETRND"), model="random")

# compare random.1 and random.1s
stargazer(random.1, random.1s, type ="text",
          column.labels = c("RE1", "RE1s"))

# Model 2:
# LOGWAGE ~ EDUC+POTEXPER+I(POTEXPER^2)+ABILITY+MOTHERED+FATHERED+BRKNHOME+
    SIBLINGS+MOTHERED:FATHERED,
# ----------
```

```
# ols model
ols.2 <- plm(LOGWAGE~EDUC+POTEXPER+I(POTEXPER^2)+ABILITY+MOTHERED+FATHERED+
             BRKNHOME+SIBLINGS+MOTHERED:FATHERED,
           data=df, index=c("PERSONID", "TIMETRND"), model="pooling")

# check model summary
summary(ols.2)

# fixed model
fixed.2 <-plm(LOGWAGE~EDUC+POTEXPER+I(POTEXPER^2)+ABILITY+MOTHERED+FATHERED+
              BRKNHOME+SIBLINGS+MOTHERED:FATHERED,
            data=df, index=c("PERSONID", "TIMETRND"), model="within")

# check model summary
summary(fixed.2)

# random model
random.2 <- plm(LOGWAGE~EDUC+POTEXPER+I(POTEXPER^2)+ABILITY+MOTHERED+FATHERED
    +
                BRKNHOME+SIBLINGS+MOTHERED:FATHERED,
              data=df, index=c("PERSONID", "TIMETRND"), model="random")

# check model summary
summary(random.2)

# compare the models to find the most appropriate one
# poolability test
pFtest(fixed.2, ols.2) # fixed effect is more appropriate

# plm test
plmtest(ols.2, type=c("bp")) # random effect is more appropriate

# Hausman test
phtest(fixed.2, random.2) # fixed effect is more appropriate

# compare the results
stargazer(ols.2, fixed.2, random.2, type ="text",
          column.labels = c("OLS", "FE","RE"))

# ------------------------------------------------------------------
# General to Specific Approach: RANDOM MODEL 2
# ------------------------------------------------------------------

# look at the model summary again
summary(random.2)

# check joint significanc - can we remove all insignificant variables at once
    ?
# - MOTHERED
# - FATHERED
```

```
# - SIBLINGS
# - MOTHERED:FATHERED

random.2$coefficients

# use Wald test to test the hypothesis that all insignificant var are jointly
# insignificant
H <- rbind(
  c(0, 0, 0, 0, 0, 1, 0, 0, 0, 0),
  c(0, 0, 0, 0, 0, 0, 1, 0, 0, 0),
  c(0, 0, 0, 0, 0, 0, 0, 0, 1, 0),
  c(0, 0, 0, 0, 0, 0, 0, 0, 0, 1)
)

# wald test to check if the smaller model performs better
wald.test(b = coef(random.2), Sigma = vcov(random.2), L = H)
# P(> X2) = 0.58 -> we can't reject the Null Hypothesis that these variables
    are all
# jointly insignificant based on 0.05 significance level

# remove all insignificant variables:
random.2s <- plm(LOGWAGE~EDUC+POTEXPER+I(POTEXPER^2)+ABILITY+BRKNHOME,
                 data=df, index=c("PERSONID", "TIMETRND"), model="random")
summary(random.2s)

# compare the results again
stargazer(ols.1, fixed.1, random.1, random.1s, ols.2, fixed.2, random.2,
    random.2s,
          type ="text", no.space=TRUE, title="Regression Results",
          column.labels = c("OLS1", "FE1","RE1", "RE1s", "OLS2", "FE2","RE2",
              "RE2s"))


# FE model is more appropriate as suggested by the Hausman Test.
# One possible explanation is the engogeneity lies in the residuals,
# causing RE model estimators to be biased. However, FE model ignores
# all time-invariant variables which are of interest for this project.
# Therefore, I chose to proceed with the RE model despite the Hausman
# test result.

# Final model: REs2

# ------------------------------------------------------------------
# Diagnostic tests for the final model: RE2s
# ------------------------------------------------------------------

# Testing for cross-sectional dependence/contemporaneous correlation:
# using Breusch-Pagan LM test of independence and Pasaran CD test
# H0: residuals across entities are not correlated
# H1: there is contemporaneous correlation (we have it)
# Breusch-Pagan's LM: bad for N > T panels
```

```
pcdtest(random.2s, test = c("lm"))
# Pesaran's CD performs well even for small T and large N
pcdtest(random.2s, test = c("cd"))
# p-value < 2.2e-16 -> reject the null hypothesis that there is no
# contemporaneous correlation.
# there is contemporaneous correlation (can lead to bias in tests results)

# heteroskedasticity
# Breusch-Pagan Test for heteroskedasticity
# H0: homoskedasticity
# H1: heteroskedasticity
bptest(LOGWAGE~EDUC+POTEXPER+I(POTEXPER^2)+ABILITY+BRKNHOME,
       data=df, studentize=F)
# p < 2.2e-16 -> reject the null hypothesis that there is homoskedasticity.
# there is heteroskedasticity (normally distributed accrdoing to the original
    paper)
# need to use a robust covariance matrix to account for it

# serial correlation
# Breusch--Godfrey test for serial correlation
# H0: there is no serial correlation
# H1: there is serial correlation
pbgtest(random.2s)
# p < 2.2e-16 -> reject the null hypothesis that there is no serial
    correlation.
# there is serial correlation
# Serial correlation causes the standard errors of the coefficients to be
# smaller than they actually are and higher R-squared

# Robust Covariance Matrix Estimators
# cluster-robust estimator
diagnose.1 = coeftest(random.2s, vcov.=vcovHC(random.2s, type="HC0"))

# heteroskedasticity-robust estimator
diagnose.2 = coeftest(random.2s, vcov.=vcovHC(random.2s, type="HC0", cluster=
    "group"))

# autocorrelation-robust estimator
diagnose.3 = coeftest(random.2s, vcov.=vcovNW(random.2s, type="HC0", cluster=
    "group"))

# robust for cross-sectional and serial correlation estimator
diagnose.4 = coeftest(random.2s, vcov.=vcovSCC(random.2s, type="HC0", cluster
    ="time"))

# put them together
stargazer(random.2s, diagnose.1, diagnose.2, diagnose.3, diagnose.4, no.space
    =TRUE,
          title="Diagnostic tests", type="text")

# ------------------------------------------------------------------
```

```
# Instrumental variables approach
# ------------------------------------------------------------------

# instrumental variables for comparison:
# 2SLS: ivreg(second stage | instrument, data)

# iv: both father and mother's educational level as instruments

iv <- ivreg(LOGWAGE ~ EDUC+POTEXPER+I(POTEXPER^2)+ABILITY+BRKNHOME
                +SIBLINGS | .-EDUC+MOTHERED+FATHERED, data=df)

summary(iv, diagnostics = TRUE)

# Diagnostic tests:
#                     df1   df2 statistic  p-value
# Weak instruments      2 17847   270.541  < 2e-16 *** (stats > 10, we have
    strong instruments)
# Wu-Hausman            1 17847    21.781 3.08e-06 *** (iv is preferred over
    OLS)
# Sargan                1    NA     0.923    0.337     (instrument is valid)

stargazer(ols.2, fixed.2, random.2s, iv, type = "latex", no.space=TRUE,
          column.labels = c("OLS2", "FE2", "RE2s", "parent's education"))
```