FEM 11087 - Applied Microeconometrics

Assignment 2: Panel Data Analysis

Empirical Application

Group 33

Alejandra Betancourt 778686
Yinli Hu 744678
Teun Mulder 565586
Andrés Piñón 775387

30 September 2025

Question 1 [0.7 points]

A central question in labor economics is: ***How much more do individuals earn with higher levels of education?*** *Economists often estimate the* returns to education*—that is, the increase in earnings associated with completing high school, college, or additional years of schooling.*

Using the panel data provided, begin by constructing a ***bar chart*** *showing* ***mean income by education group****. Group individuals based on their* ***highest level of educational attainment*** *(e.g., less than high school, high school graduate, some college, college degree or more), and plot the* ***average income*** *for each category.*

```
bysort pid: egen max_edyears = max(edyears)

gen edyears_cat = .
replace edyears_cat = 1 if max_edyears <= 11 & !missing(edyears)
replace edyears_cat = 2 if max_edyears == 12 & !missing(edyears)
replace edyears_cat = 3 if max_edyears >= 13 & max_edyears <= 15 ///
        & !missing(edyears)
replace edyears_cat = 4 if max_edyears >= 16 & !missing(edyears)

graph bar (mean) income, over(edyears_cat)
```

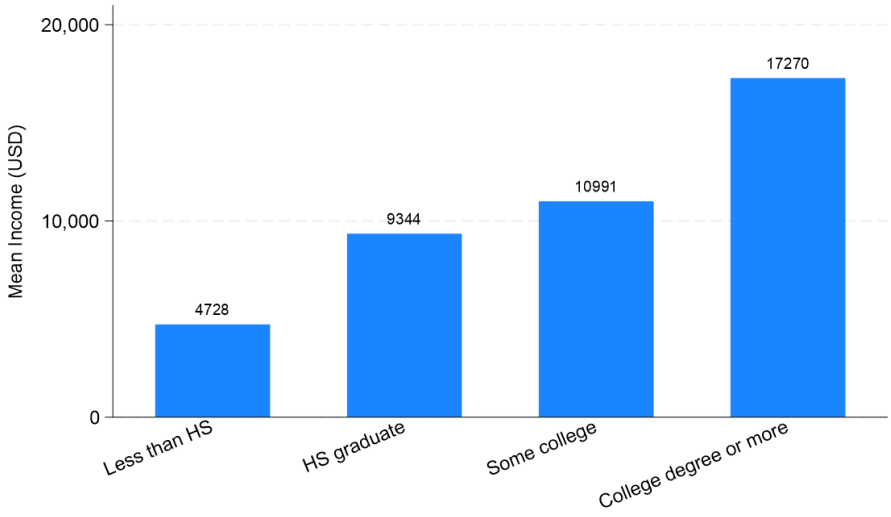

Figure 1.1: Mean income by education level

Recent debates around student debt and the value of higher education often assume that education "pays off" equally for everyone. ***Does your analysis support that assumption?*** *To explore this, create* ***separate plots by gender*** *to highlight any differences in the relationship between education and earnings. Discuss your findings.*

```
graph bar (mean) income, over(edyears_cat) by(male)
```

As seen in the resulting graphs (**Figure 1.2**, below), there is a stark contrast in average mean income across genders at every education level.

Men consistently earn higher average incomes than women across every category. Male individuals with a college degree stand to earn roughly $13,000 more on average than women with comparable education, while men belonging to the first three categories earn, on average, roughly three and a half times as much as their female counterparts.

While the difference in amount of dollars earned appears to widen as education level increases, the largest relative gap occurs at the first three levels; this suggests a significant gender gap in average income across the board. These findings would challenge the assumption that education pays off for everyone. While both genders benefit from higher education, men benefit substantially more at every level.

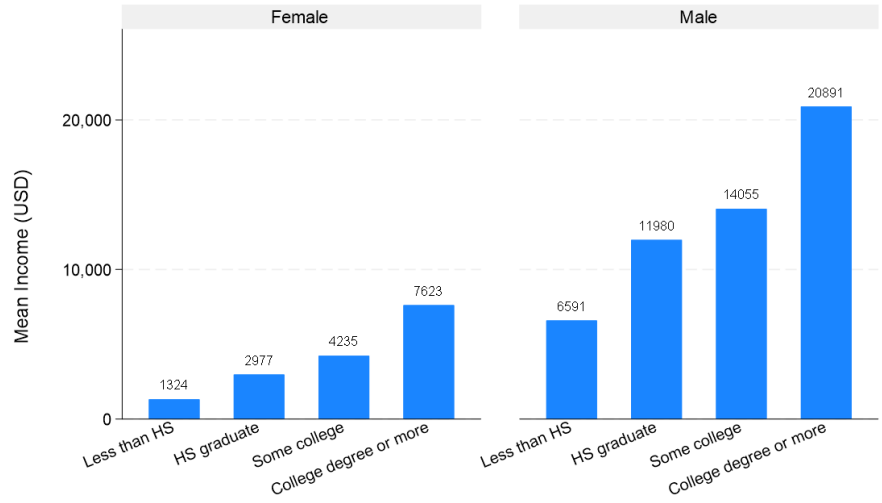

Figure 1.2: Mean income by education level, by gender

Question 2 [1 point]

Now, we turn to formally estimating the effect of years of education on income. Use ***pooled OLS*** *to examine the impact of years of education (edyears) on* ***log(income)****, controlling for age, gender (male), marital status categories, ethnicity categories, and childbirth.*

```
gen log_income = log(income)
reg log_income edyears age i.male ib0.mstatus ib4.ethnicity ///
    i.child_birth
```

a) What is the estimated return to an additional year of education? Interpret the coefficient on years of education in terms of its ***sign****,* ***magnitude****, and* ***statistical significance****.*

For each additional year of schooling, individuals receive a 15.70% higher average income, ceteris paribus. This difference is statistically significant at the 1% significance level.

Table 2.1: Pooled OLS model

	Coefficient	Robust std. err.	P>|t|	Conf. int.
Explanatory variable				
Education years	0.1458	(0.0031)	0.000	[0.140,0.152]
Control variables				
Age	0.1263	(0.0017)	0.000	[0.123,0.130]
Male	0.6000	(0.0115)	0.000	[0.577,0.623]
Childbirth in past year	-0.0795	(0.0198)	0.000	[-0.118,-0.041]
Married	0.6189	(0.0211)	0.000	[0.578,0.660]
Separated or divorced	0.0767	(0.0472)	0.104	[-0.016,0.169]
Widowed	0.0647	(0.2690)	0.810	[-0.462,0.592]
Black	-0.4286	(0.0139)	0.000	[-0.456,-0.401]
Hispanic	-0.0825	(0.0147)	0.000	[-0.111,-0.054]
Mixed Race (Non-Hispanic)	-0.3744	(0.0627)	0.000	[-0.497,-0.252]
Constant	3.0156	(0.0318)	0.000	[2.953,3.078]
Number of obs	55874			
F-statistic	3875.31			
Prob > F	0.0000			
R-squared	0.4089			
Adj. R-squared	0.4088			

Dependent variable: log(income)

b) Differences in returns to schooling by gender are sometimes interpreted as potential evidence of ***labor market discrimination****. Test whether the effect of years of education using the categorical variable created in Question 1 on log(income) is the* ***same for men and women****. Based on your results, do you find any evidence consistent with discrimination?*

```
reg log_income i.edyears_cat##i.male age ib0.mstatus ib4.ethnicity ///
    i.child_birth

test 2.edyears_cat#1.male 3.edyears_cat#1.male 4.edyears_cat#1.male
```

Table 2.2: Joint significance test of interaction terms

F-statistic	23.86
Prob > F	0.0000

(1) HS graduate × Male = 0
(2) Some college × Male = 0
(3) College degree or more × Male = 0

The interaction terms are highly statistically significant at the 1% significance level, and we can therefore reject the null hypothesis that the interaction effects coefficients are equal to zero. In other words, men and women see significantly different income effects from additional education levels.

Table 2.3: Pooled OLS model with interaction effects

	Coefficient	Robust std. err.	P>|t|	Conf. int.
Explanatory variable				
HS graduate	0.1611	(0.0242)	0.000	[0.114,0.209]
Some college	0.3593	(0.0241)	0.000	[0.312,0.407]
College degree or more	0.4756	(0.0270)	0.000	[0.423,0.529]
Interaction terms				
HS graduate × Male	0.2498	(0.0325)	0.000	[0.186,0.314]
Some college × Male	0.2201	(0.0322)	0.000	[0.157,0.283]
College degree or more × Male	0.1979	(0.0338)	0.000	[0.132,0.264]
Control variables				
Age	0.1591	(0.0015)	0.000	[0.156,0.162]
Male	0.3647	(0.0235)	0.000	[0.319,0.411]
Childbirth in past year	-0.0902	(0.0200)	0.000	[-0.130,-0.051]
Married	0.5959	(0.0213)	0.000	[0.554,0.638]
Separated or divorced	-0.0539	(0.0481)	0.262	[-0.148,0.040]
Widowed	-0.1767	(0.2636)	0.503	[-0.693,0.340]
Black	-0.4637	(0.0142)	0.000	[-0.491,-0.436]
Hispanic	-0.1137	(0.0150)	0.000	[-0.143,-0.084]
Mixed Race (Non-Hispanic)	-0.3563	(0.0625)	0.000	[-0.479,-0.234]
Constant	3.8188	(0.0307)	0.000	[3.759,3.879]
Number of obs	55874			
F-statistic	2380.33			
Prob > F	0.0000			
R-squared	0.3920			
Adj. R-squared	0.3918			

Dependent variable: log(income)

c) Under what conditions is the pooled OLS estimate of the effect of years of education ***unbiased and efficient****? Do you believe these conditions are likely to hold in this context?*

For a pooled OLS estimate to be unbiased, the explanatory variable must be uncorrelated with the error term. In our case, this would require that all relevant variables affecting income are either included in the model or are uncorrelated with education; otherwise, the model would suffer from endogeneity due to factors like individual ability, family background or motivation, which are not present in our

model.

For the model to be efficient, serial correlation must not be present. Since individuals possess unobserved characteristics that remain constant across time and which impact their income across all time periods, this is highly unlikely. As a result, the error term for each individual is correlated across time periods. This does not create bias in the estimators, but it leads to unreliable significance tests and confidence intervals.

In the context of our model, these conditions are unlikely to hold. There are many personal factors that influence both education levels and income, not all of which are accounted for in our model. Additionally, it is likely that unobserved individual characteristics which remain constant across time impact income across all time periods, which means that serial correlation is present and our estimate is inefficient.

Question 3 [0.5 points]

So far, the panel structure of the data has been largely unexploited. Random effects (RE) estimation can improve the efficiency of the estimates compared to pooled OLS.

a) Estimate the effect of years of education (edyears) on ***log(income)*** *using the* ***random effects*** *(RE) model, controlling for age, gender (male), marital status categories, ethnicity categories, and childbirth. Interpret the estimated coefficient for years of education in terms of its* ***sign***, ***magnitude***, *and* ***statistical significance***. *Then, compare the RE estimate and standard error of the education coefficient with those obtained from the* ***pooled OLS model***.

```
xtset pid wave
xtreg log_income edyears age i.male ib0.mstatus ib4.ethnicity ///
    i.child_birth, re
estimates store random
```

Table 3.1: Random effects (RE) model

	Coefficient	Std. err.	P>|t|	Conf. int.
Explanatory variable				
Education years	0.1509	(0.0031)	0.000	[0.145,0.157]
Control variables				
Age	0.1273	(0.0015)	0.000	[0.124,0.130]
Male	0.5698	(0.0167)	0.000	[0.537,0.603]
Childbirth in past year	-0.0599	(0.0174)	0.001	[-0.094,-0.026]
Married	0.5107	(0.0197)	0.000	[0.472,0.549]
Separated or divorced	0.0454	(0.0392)	0.246	[-0.031,0.122]
Widowed	0.0855	(0.2619)	0.744	[-0.428,0.599]
Black	-0.3833	(0.0192)	0.000	[-0.421,-0.346]
Hispanic	-0.0655	(0.0203)	0.001	[-0.105,-0.026]
Mixed Race (Non-Hispanic)	-0.3244	(0.0813)	0.000	[-0.484,-0.165]
Constant	2.9423	(0.0341)	0.000	[2.875,3.009]
Number of obs	55874			
Number of groups	7126			
R-squared within	0.3380			
R-squared between	0.5546			
R-squared overall	0.4084			
Wald χ^2	34633.21			
Prob > χ^2	0.0000			
σ_u	0.4412			
σ_e	1.1766			
ρ	0.1233			

Dependent variable: log(income)

According to the RE model, for each additional year of schooling individuals receive a 16.29% higher average income, ceteris paribus. This difference is statistically significant at the 1% significance level.

The RE estimate of 0.1509 is 0.0051 points higher than the pooled OLS estimate of 0.1458, while yielding marginally higher standard error at 0.0031, compared to 0.0027 for the POLS model. This suggests that the POLS estimate is overly optimistic, arising from its failure to account for within-individual correlation across time.

Table 3.2: POLS, RE comparison

	POLS	RE
Explanatory variable		
Education years	0.1458***	0.1509***
	(0.0027)	(0.0031)
Control variables		
Age	0.1263***	0.1273***
	(0.0014)	(0.0015)
Male	0.6000***	0.5698***
	(0.0126)	(0.0167)
Childbirth in past year	-0.0795***	-0.0599***
	(0.0184)	(0.0174)
Married	0.6189***	0.5107***
	(0.0185)	(0.0197)
Separated or divorced	0.0767*	0.0454
	(0.0375)	(0.0392)
Widowed	0.0647	0.0855
	(0.2531)	(0.2619)
Black	-0.4286***	-0.3833***
	(0.0139)	(0.0192)
Hispanic	-0.0825***	-0.0655**
	(0.0146)	(0.0203)
Mixed Race (Non-Hispanic)	-0.3744***	-0.3244***
	(0.0583)	(0.0813)
Constant	3.0156***	2.9423***
	(0.0319)	(0.0341)
Number of obs	55874	55874
Number of groups		7126
F-statistic	3864.88	
Wald χ^2		34633.21
P-value	0.0000	0.0000
R-squared	0.4089	
Adj. R-squared	0.4088	
R-squared within		0.3380
R-squared between		0.5546
R-squared overall		0.4084
σ_u		0.4412
σ_e		1.1766
ρ		0.1233

Standard errors in parentheses.
* p < 0.05, ** p < 0.01, *** p < 0.001

*b) Under which conditions and why can the random effects estimator be **more efficient** than pooled OLS?*

Random effects can be more efficient than pooled OLS when there is serial correlation in the POLS estimate. RE can be useful in this situation since it takes into account the correlation within individuals across different time periods with the use of GLS.

Question 4 [1.55 points]

*Alternatively, the panel structure of the data can be used to perform **fixed effects (FE)** estimation.*

*a) Based on theoretical considerations, would you **prefer** fixed effects or random effects estimation? Justify your answer.*

The choice between fixed effects or random effects hinges on whether or not individual effects are correlated with the explanatory variable. In our model, this is extremely likely, due to the fact that many individual factors, such as ability, motivation, background and socio-economic status are correlated with further education and also with higher income for factors other than higher education. Hence, we would prefer fixed effects since even though it cannot estimate the effect of time invariant characteristics on income, it can provide us with an unbiased estimator.

*b) Use a **fixed effects estimator** to examine the impact of years of education (edyears) on **log(income)**, controlling for age, gender (male), marital status categories, ethnicity categories, and childbirth. Interpret the coefficient on years of education in terms of its **sign**, **magnitude**, and **statistical significance**. Compare your results with those from the **pooled OLS** and **random effects** models.*

```
xtreg log_income edyears age i.male ib0.mstatus ib4.    ethnicity ///
    i.child_birth, fe
estimates store fixed
```

Table 4.1: Fixed effects (FE) model

	Coefficient	Std. err.	P>|t|	Conf. int.
Explanatory variable				
Education years	0.1610	(0.0038)	0.000	[0.153,0.168]
Control variables				
Age	0.1259	(0.0017)	0.000	[0.123,0.129]
Childbirth in past year	-0.0545	(0.0180)	0.002	[-0.090,-0.019]
Married	0.4178	(0.0223)	0.000	[0.374,0.461]
Separated or divorced	-0.0112	(0.0432)	0.796	[-0.096,0.073]
Widowed	0.1575	(0.2861)	0.582	[-0.403,0.718]
Constant	3.1536	(0.0378)	0.000	[3.080,3.228]
Number of obs	55874			
Number of groups	7126			
R-squared within	0.3383			
R-squared between	0.4945			
R-squared overall	0.3713			
Wald χ^2				
Prob > χ^2	0.0000			
σ_u	0.8174			
σ_e	1.1766			
ρ	0.3255			

Dependent variable: log(income)

According to the FE model, for each additional year of schooling individuals receive a 17.47% higher average income, ceteris paribus. This difference is statistically significant at the 1% significance level.

Compared to the POLS and RE models, FE has both a higher coefficient (0.1610 compared to 0.1458 and 0.1509, respectively) as well as higher standard error (0.0038 compared to 0.0027 and 0.0031, respectively). This implies that individual-specific effects were negatively correlated with education, meaning the POLS and RE estimates had a downward bias.

The higher standard error can also be expected, since FE trades some amount of efficiency in exchange for an unbiased estimate.

Table 4.2: POLS, RE, FE comparison

	POLS	RE	FE
Explanatory variable			
Education years	0.1458***	0.1509***	0.1610***
	(0.0027)	(0.0031)	(0.0038)
Control variables			
Age	0.1263***	0.1273***	0.1259***
	(0.0014)	(0.0015)	(0.0017)
Male	0.6000***	0.5698***	0.0000
	(0.0126)	(0.0167)	(.)
Childbirth in past year	-0.0795***	-0.0599***	-0.0545**
	(0.0184)	(0.0174)	(0.0180)
Married	0.6189***	0.5107***	0.4178***
	(0.0185)	(0.0197)	(0.0223)
Separated or divorced	0.0767*	0.0454	-0.0112
	(0.0375)	(0.0392)	(0.0432)
Widowed	0.0647	0.0855	0.1575
	(0.2531)	(0.2619)	(0.2861)
Black	-0.4286***	-0.3833***	0.0000
	(0.0139)	(0.0192)	(.)
Hispanic	-0.0825***	-0.0655**	0.0000
	(0.0146)	(0.0203)	(.)
Mixed Race (Non-Hispanic)	-0.3744***	-0.3244***	0.0000
	(0.0583)	(0.0813)	(.)
Constant	3.0156***	2.9423***	3.1536***
	(0.0319)	(0.0341)	(0.0378)
Number of obs	55874	55874	55874
Number of groups		7126	7126
F-statistic	3864.88		4154.09
Wald χ^2		34633.21	
P-value	0.0000	0.0000	0.0000
R-squared	0.4089		0.3383
Adj. R-squared	0.4088		0.2415
R-squared within		0.3380	0.3383
R-squared between		0.5546	0.4945
R-squared overall		0.4084	0.3713
σ_u		0.4412	0.8174
σ_e		1.1766	1.1766
ρ		0.1233	0.3255

Standard errors in parentheses.
* p < 0.05, ** p < 0.01, *** p < 0.001

*c) Perform the **Hausman** test. What do the results indicate? Based on the test outcome, **which estimator** (RE or FE) is more appropriate in this context?*

```
hausman fixed random
```

Table 4.3: Hausman test

χ^2	141.94
Prob > χ^2	0.0000

H0: Difference in β not systematic

The Hausman test checks the null hypothesis that the individual-specific effects are uncorrelated with the explanatory variable. These results indicate that we can reject the null hypothesis at the 1% significance level, meaning that the test indicates that individual-specific effects are correlated with education years.

Question 5 [0.9 points]

*Next, estimate a **Correlated Random Effects (CRE)** model to examine the effect of years of education (edyears) on **log(income)**.*

```
by pid: egen edyears_mean = mean(edyears)
by pid: egen age_mean = mean(age)
by pid: egen male_mean = mean(male)
by pid: egen mstatus_mean = mean(mstatus)
by pid: egen ethnicity_mean = mean(ethnicity)
by pid: egen child_birth_mean = mean(child_birth)

xtreg log_income edyears age i.male ib0.mstatus ib4.ethnicity ///
    i.child_birth age_mean mstatus_mean child_birth_mean ///
    edyears_mean male_mean ethnicity_mean, re
```

According to the CRE model, for each additional year of schooling individuals receive a 17.43% higher average income, ceteris paribus. This difference is statistically significant at the 1% significance level.

*a) What is one advantage of the **CRE estimator** compared to the **random effects (RE)** estimator?*

While RE requires that individual effects be uncorrelated with the explanatory variables, CRE models the correlation between individual effects and time-varying

explanatory variables, making it more robust to violations of the random effects assumption.

Table 5.1: Correlated Random Effects (CRE) model

	Coefficient	Std. err.	P>|t|	Conf. int.
Explanatory variable				
Education years	0.1607	(0.0039)	0.000	[0.153,0.168]
Control variables				
Age	0.1255	(0.0017)	0.000	[0.122,0.129]
Male	0.5505	(0.0180)	0.000	[0.515,0.586]
Childbirth in past year	-0.0577	(0.0183)	0.002	[-0.094,-0.022]
Married	0.4504	(0.0221)	0.000	[0.407,0.494]
Separated or divorced	-0.0627	(0.0430)	0.145	[-0.147,0.022]
Widowed	-0.0767	(0.2632)	0.771	[-0.592,0.439]
Black	-0.3805	(0.0196)	0.000	[-0.419,-0.342]
Hispanic	-0.0769	(0.0205)	0.000	[-0.117,-0.037]
Mixed Race (Non-Hispanic)	-0.3089	(0.0813)	0.000	[-0.468,-0.150]
edyears_mean	-0.0285	(0.0064)	0.000	[-0.041,-0.016]
CRE variables				
age_mean	0.0109	(0.0038)	0.004	[0.003,0.018]
child_birth_mean	-0.0879	(0.0579)	0.129	[-0.201,0.026]
mstatus_mean	0.2105	(0.0347)	0.000	[0.142,0.278]
Constant	2.9644	(0.0614)	0.000	[2.844,3.085]
Number of obs	55874			
Number of groups	7126			
R-squared within	0.3383			
R-squared between	0.5569			
R-squared overall	0.4096			
Wald χ^2	34746.58			
Prob > χ^2	0.0000			
σ_u	0.4412			
σ_e	1.1766			
ρ	0.1233			

Dependent variable: log(income)

*b) What is one advantage of the **CRE estimator** compared to the **fixed effects (FE)** estimator?*

While FE models remove all time-invariant characteristics through the within transformation, CRE retains the ability to estimate effects of variables that don't change over time, such as gender, race, or place of birth. By modelling the correlation

structure between individual effects and explanatory variables, CRE becomes useful when estimating the effects of time-constant characteristics.

*c) Compare the estimated coefficient for years of education from the **CRE model** with those from the **RE** and **FE** models. Are the coefficients similar or different? Explain why this is the case.*

Compared to the previous models, CRE has a coefficient similar to the FE estimate (0.1607 compared to 0.1610) as well as the similar standard error (0.0039 compared to 0.0038). This suggests that CRE succesfully replicates the FE estimate's ability to handle endogeneity without having to omit time-invariant variables.

Compared to the RE estimate, the CRE coefficient is notably higher than the RE coefficient (0.1607 compared to 0.1509), which is similar to the difference between RE and FE. The higher FE coefficient indicates that there is indeed correlation between individual effects and education, confirming the Hausman test results. While CRE and FE take this correlation into account, the RE estimate's lower coefficient reflects downward bias from failing to account for this correlation.

All in all, the CRE estimate is very similar to the FE estimate, addressing the correlation between individual effects and regressors detected by the Hausman test. The key advantage of CRE over FE is that it can estimate time-invariant characteristics, making it appropriate for handling endogeneity when time-invariant variables cannot be omitted.

*d) Based on your CRE estimates, does the assumption of **exogeneity** appear to hold? Which estimator would you consider most appropriate in this context?*

According to the model, the `mstatus_mean` CRE estimator is statistically significant at the 1% significance level. This suggests that the exogeneity assumption does not hold.

```
test edyears_mean age_mean male_mean child_birth_mean ///
    mstatus_mean ethnicity_mean
```

Table 5.2: Joint significance test

χ^2	71.51
Prob > χ^2	0.0000

(0) Education mean × Age mean ×
Male mean × Childbirth mean ×
Marriage Status mean × Ethnicity mean = 0

The joint significance test of the CRE time-average variables rejects the null hypothesis that all time averages equal zero with a 1% significance level, confirming that the exogeneity assumption does not hold.

To most accurately estimate *returns on education*, the FE model would make most sense as it is unambiguously robust; however, if time-invariant effects had to be estimated, CRE would provide unbiased estimates while allowing for time-invariant variables.

Table 5.3: POLS, RE, FE, CRE comparison

	POLS	RE	FE	CRE
Explanatory variable				
Education years	0.1458***	0.1509***	0.1610***	0.1607***
	(0.0027)	(0.0031)	(0.0038)	(0.0039)
Control variables				
Age	0.1263***	0.1273***	0.1259***	0.1255***
	(0.0014)	(0.0015)	(0.0017)	(0.0017)
Male	0.6000***	0.5698***	0.0000	0.5505***
	(0.0126)	(0.0167)	(.)	(0.0180)
Childbirth in past year	-0.0795***	-0.0599***	-0.0545**	-0.0577**
	(0.0184)	(0.0174)	(0.0180)	(0.0183)
Married	0.6189***	0.5107***	0.4178***	0.4504***
	(0.0185)	(0.0197)	(0.0223)	(0.0221)
Separated or divorced	0.0767*	0.0454	-0.0112	-0.0627
	(0.0375)	(0.0392)	(0.0432)	(0.0430)
Widowed	0.0647	0.0855	0.1575	-0.0767
	(0.2531)	(0.2619)	(0.2861)	(0.2632)
Black	-0.4286***	-0.3833***	0.0000	-0.3805***
	(0.0139)	(0.0192)	(.)	(0.0196)
Hispanic	-0.0825***	-0.0655**	0.0000	-0.0769***
	(0.0146)	(0.0203)	(.)	(0.0205)
Mixed Race (Non-Hispanic)	-0.3744***	-0.3244***	0.0000	-0.3089***
	(0.0583)	(0.0813)	(.)	(0.0813)
CRE variables				
age_mean				0.0109**
				(0.0038)
mstatus_mean				0.2105***
				(0.0347)
child_birth_mean				-0.0879
				(0.0579)
edyears_mean				-0.0285***
				(0.0064)
Constant	3.0156***	2.9423***	3.1536***	2.9644***
	(0.0319)	(0.0341)	(0.0378)	(0.0614)
Number of obs	55874	55874	55874	55874
Number of groups		7126	7126	7126
F-statistic	3864.88		4154.09	
Wald χ^2		34633.21		34746.58
P-value	0.0000	0.0000	0.0000	0.0000
R-squared	0.4089		0.3383	
Adj. R-squared	0.4088		0.2415	
R-squared within		0.3380	0.3383	0.3383
R-squared between		0.5546	0.4945	0.5569
R-squared overall		0.4084	0.3713	0.4096
σ_u		0.4412	0.8174	0.4412
σ_e		1.1766	1.1766	1.1766
ρ		0.1233	0.3255	0.1233

Standard errors in parentheses.
* p < 0.05, ** p < 0.01, *** p < 0.001

Question 6 [0.9 points]

Recent research provides compelling evidence that after the birth of a first child, women's earnings decline sharply and remain persistently lower, while men's earnings remain largely unaffected.

a) Estimate the effect of childbirth on ***log(income)*** *using the* ***most appropriate model****. Control for age, gender (male), marital status categories, ethnicity categories, and years of education (edyears). Interpret the estimated coefficient for childbirth in terms of its* ***sign****,* ***magnitude****, and* ***statistical significance****.*

In order to choose an estimator, both endogeneity and serial correlation must be taken into account, as well as the ability to estimate time-invariant variables. Intuition points to CRE as the model of choice, but we should first formally test these assumptions.

First, the pooled OLS estimate can be rejected because it fails to account for the panel structure of the data. POLS incorrectly assumes observations are independent across time, leading to inefficient estimates and incorrect standard errors when individual-specific effects create serial correlation.

To decide between the RE and FE estimates, we can conduct a Hausman test to check whether or not the exogeneity assumption holds:

```
xtreg log_income i.child_birth age i.male ib0.mstatus ib4.ethnicity ///
    edyears, re
estimates store random

xtreg log_income i.child_birth age i.male ib0.mstatus ib4.ethnicity ///
    edyears, fe
estimates store fixed

hausman fixed random
```

Table 6.1: Hausman test

χ^2	141.94
Prob > χ^2	0.0000

H0: Difference in β not systematic

We can reject the null hypothesis that individual-specific effects are uncorrelated with the explanatory variable, which suggests that in order to estimate the effect

of childbirth on income, the FE estimate would stand as the most robust approach; however, this would omit the *gender* variable due to its time-invariant nature. In this situation, the CRE estimate stands as the most appropriate model, since it can deal with endogeneity while still accounting for time-invariant effects.

```
by pid: egen edyears_mean = mean(edyears)
xtreg log_income i.child_birth age i.male ib0.mstatus ib4.ethnicity ///
    edyears age_mean mstatus_mean edyears_mean, re
xtreg log_income i.child_birth##i.male age ib0.mstatus ib4.ethnicity ///
    edyears age_mean mstatus_mean edyears_mean, re
```

According to our model, an individual who has had a child during the past year will receive a 5.94% lower average income, ceteris paribus. This difference is statistically significant at the 1% significance level.

b) Test whether the effect of childbirth on log(income) ***differs*** *between males and females. What conclusions can you draw from your results?*

```
xtreg log_income i.child_birth##i.male age ib0.mstatus ib4.ethnicity ///
    edyears age_mean mstatus_mean, re

test 1.child_birth#1.male
```

Table 6.2: Significance test

χ^2	114.33
Prob > χ^2	0.0000

(1) Childbirth × Male = 0

We can therefore reject the null hypothesis that the interaction effect of childbirth and male equals zero, meaning there is indeed difference in the effect of childbirth across genders.

Table 6.3: Correlated Random Effects (CRE)

	Coefficient	Std. err.	P>|t|	Conf. int.
Explanatory variable				
Childbirth in past year	-0.0577	(0.0183)	0.002	[-0.094,-0.022]
Control variables				
Age	0.1255	(0.0017)	0.000	[0.122,0.129]
Male	0.5505	(0.0180)	0.000	[0.515,0.586]
Education years	0.1607	(0.0039)	0.000	[0.153,0.168]
Married	0.4504	(0.0221)	0.000	[0.407,0.494]
Separated or divorced	-0.0627	(0.0430)	0.145	[-0.147,0.022]
Widowed	-0.0767	(0.2632)	0.771	[-0.592,0.439]
Black	-0.3805	(0.0196)	0.000	[-0.419,-0.342]
Hispanic	-0.0769	(0.0205)	0.000	[-0.117,-0.037]
Mixed Race (Non-Hispanic)	-0.3089	(0.0813)	0.000	[-0.468,-0.150]
child_birth_mean	-0.0879	(0.0579)	0.129	[-0.201,0.026]
CRE variables				
age_mean	0.0109	(0.0038)	0.004	[0.003,0.018]
edyears_mean	-0.0285	(0.0064)	0.000	[-0.041,-0.016]
mstatus_mean	0.2105	(0.0347)	0.000	[0.142,0.278]
Constant	2.9644	(0.0614)	0.000	[2.844,3.085]
Number of obs	55874			
Number of groups	7126			
R-squared within	0.3383			
R-squared between	0.5569			
R-squared overall	0.4096			
Wald χ^2	34746.58			
Prob > χ^2	0.0000			
σ_u	0.4412			
σ_e	1.1766			
ρ	0.1233			

Dependent variable: log(income)

Table 6.4: Correlated Random Effects (CRE) with interaction effects

	Coefficient	Std. err.	P>|t|	Conf. int.
Explanatory variable				
Childbirth in past year	-0.3400	(0.0321)	0.000	[-0.403,-0.277]
Childbirth in past year × Male	0.3967	(0.0371)	0.000	[0.324,0.469]
Control variables				
Age	0.1245	(0.0017)	0.000	[0.121,0.128]
Male	0.5074	(0.0184)	0.000	[0.471,0.543]
Education years	0.1619	(0.0039)	0.000	[0.154,0.170]
Married	0.4503	(0.0221)	0.000	[0.407,0.494]
Separated or divorced	-0.0637	(0.0430)	0.138	[-0.148,0.020]
Widowed	-0.0725	(0.2629)	0.783	[-0.588,0.443]
Black	-0.3792	(0.0196)	0.000	[-0.418,-0.341]
Hispanic	-0.0777	(0.0205)	0.000	[-0.118,-0.038]
Mixed Race (Non-Hispanic)	-0.3023	(0.0813)	0.000	[-0.462,-0.143]
child_birth_mean	-0.0446	(0.0580)	0.442	[-0.158,0.069]
CRE variables				
age_mean	0.0114	(0.0038)	0.003	[0.004,0.019]
edyears_mean	-0.0289	(0.0064)	0.000	[-0.041,-0.016]
mstatus_mean	0.2121	(0.0347)	0.000	[0.144,0.280]
Constant	2.9913	(0.0614)	0.000	[2.871,3.112]
Number of obs	55874			
Number of groups	7126			
R-squared within	0.3397			
R-squared between	0.5568			
R-squared overall	0.4109			
Wald χ^2	34919.38			
Prob > χ^2	0.0000			
σ_u	0.4418			
σ_e	1.1753			
ρ	0.1238			

Dependent variable: log(income)

Looking at the regression table for the CRE model with interaction effects, we can see that a female individual who has had a child during the past year will receive a 40.49% lower average income, ceteris paribus; however, a male individual who has had a child during the past year will receive a 5.83% higher income, ceteris paribus. Both estimates are statistically significant at the 1% significance level. The astounding 46.32 percentage point gap in the effect of childbirth on average income reveals significant gender disparities, with women experiencing a severe child penalty, while men seem to receive a small child bonus.

Question 7 [1.2 points]

Without conducting any empirical analysis:

a) Compare the key assumptions underlying ***pooled OLS***, ***fixed effects (FE)***, *and* ***random effects (RE)*** *estimators. Discuss theoretically in which scenarios you would prefer to use each method.*

Pooled OLS assumes exogeneity and no serial correlation within individuals, as well as the standard MLR assumptions. It ignores the panel structure entirely, treating all observations as independent.

Random Effects assumes exogeneity (same as POLS), but uses GLS to account for within-individual correlation, making it more efficient than POLS when exogeneity holds.

Fixed Effects requires only that the idiosyncratic error be uncorrelated to the explanatory variables, allowing the individual effects to correlate with the explanatory variables. It eliminates individual effects through within-transformation.

With panel data, POLS is rarely appropriate as it produces inefficient estimates and incorrect standard errors due to serial correlation. It could theoretically be used in cases in which there is only one observation per person and exogeneity holds. RE can be used when individual effects exist but are uncorrelated with the explanatory variables, while FE is necessary to produce unbiased estimators when individual characteristics are correlated with the explanatory variables. A Hausman test can verify whether the RE assumption holds, in which case RE is preferred over FE. The main limitations of the FE model are its lower efficiency and its inability to estimate time-invariant variables.

b) Within the practical context of this assignment (effect of education on earnings), provide an example situation for each estimator in the form of a ***Directed Acyclic Graph (DAG)****. For each case (Pooled OLS, FE, and RE), explain why the assumptions required for the respective method hold in that example, and why that method would be preferred.*

An example situation where the POLS estimator would be preferred could be a short-term study tracking education levels and income, where graduate programs are repeatedly randomly assigned through a scholarship lottery. It would control for age, gender, marital status and ethnicity. The exogeneity assumption would hold since the lottery mechanism ensures that education is uncorrelated with individual effects, and serial correlation would be minimized through the small time frame, mitigating the emergence of individual trends (in practice, RE would most likely be preferred since it would account for any residual within-person correlation).

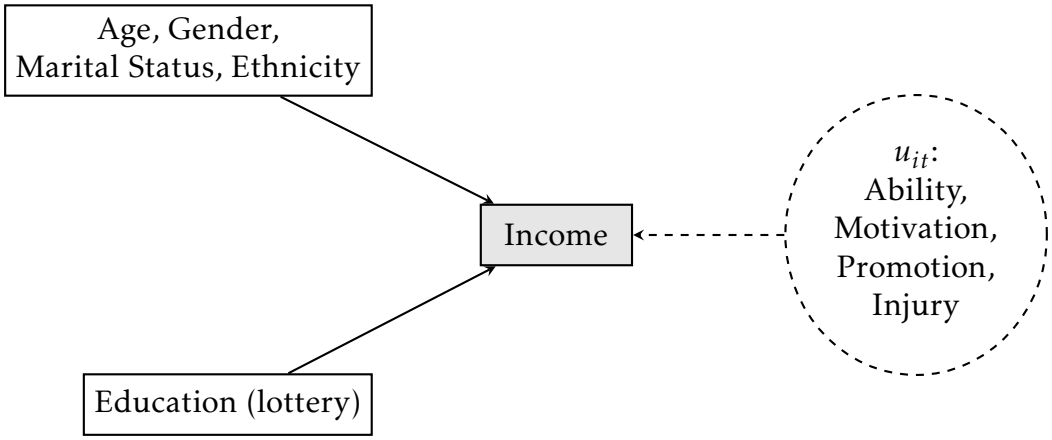

Figure 7.1: DAG POLS example

A situation in which the RE estimator would be preferred could be a multi-year study following workers across regions with different education policies, where said policy changes create variation in educational attainment. Said study would track education levels and income, controlling for age, gender, marital status and ethnicity. Individual effects exist and create within-person correlation over time, but these characteristics are uncorrelated with the regional policy changes affecting education access. The exogeneity assumption would hold because education variation stems from external policies rather than individual choices. RE would be preferred over POLS because it efficiently accounts for within-person correlation using GLS, and preferred over FE because it maintains efficiency when exogeneity holds while also allowing the estimation of time-invariant characteristics like gender or ethnicity.

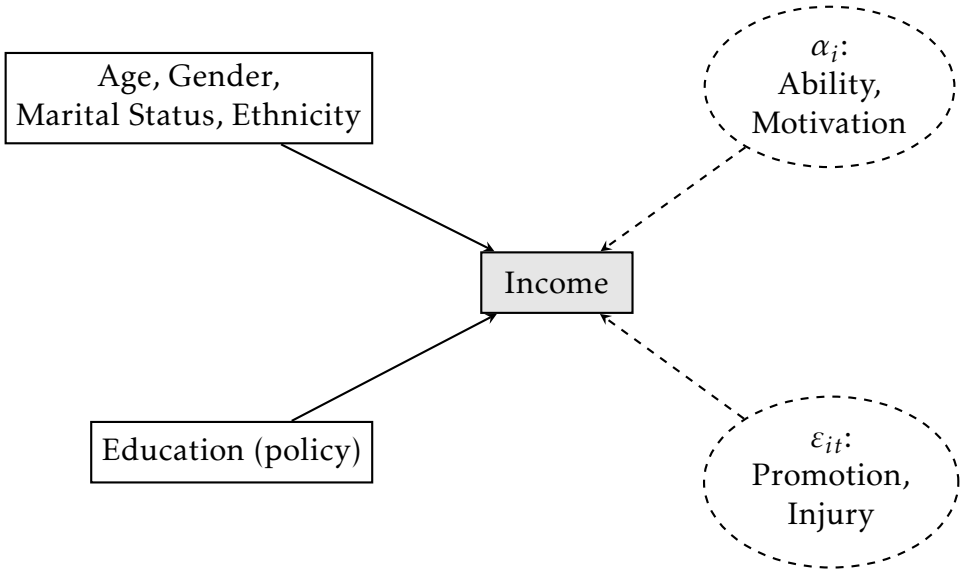

Figure 7.2: DAG RE example

A situation in which the FE estimator would be preferred could be a standard observational panel study following individuals over several years as they make their own education decisions. This study would track education levels and income while controlling for age and marital status. Since unobserved individual characteristics influence both educational choices and income, FE would be necessary to eliminate endogeneity through the use of within-transformation. While this approach has lower efficiency than RE and cannot estimate time-invariant variables like gender or ethnicity, it would be able to provide unbiased estimates of the causal effect of education on income despite endogeneity, hence making it the preferred estimator.

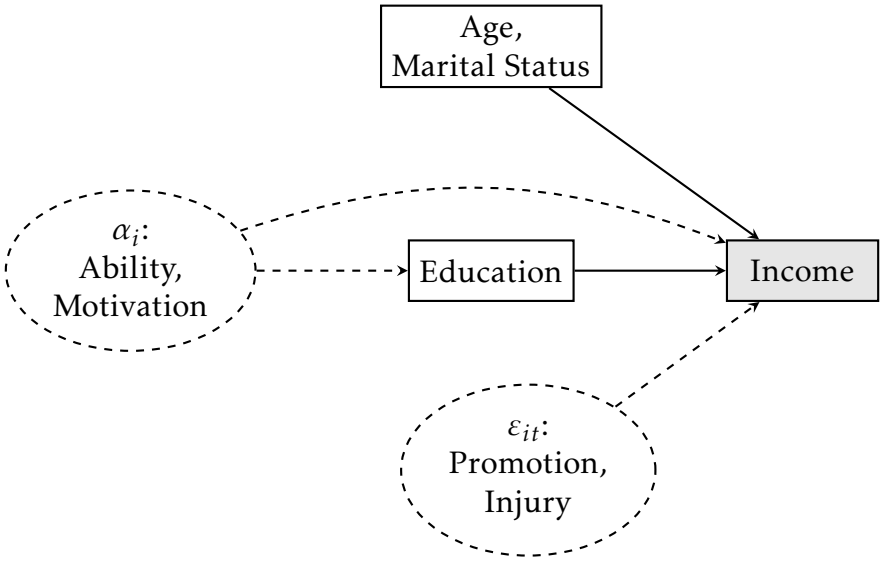

Figure 7.3: FE DAG: Individual effects correlated with education

Question 8 [0.75 points]

Finally, revisit your data and evaluate whether ***attrition*** *is present in your sample. Based on your preferred model, discuss the likelihood of* ***attrition bias****. What conclusions can you draw regarding its presence, and how might it affect the validity of your results?*

```
bysort pid (wave): gen n_waves = _N
gen all_waves = n_waves == 17

xtreg log_income i.child_birth##i.male age ib0.mstatus ib4.ethnicity ///
    edyears age_mean mstatus_mean all_waves, re
```

Table 8.1: Attrition bias: *all waves* indicator

	Coefficient	Std. err.	P>|t|	Conf. int.
Explanatory variable				
Childbirth in past year	-0.3415	(0.0321)	0.000	[-0.405,-0.279]
Childbirth in past year × Male	0.3989	(0.0371)	0.000	[0.326,0.472]
Control variables				
Age	0.1245	(0.0017)	0.000	[0.121,0.128]
Male	0.5057	(0.0184)	0.000	[0.470,0.542]
Education years	0.1619	(0.0039)	0.000	[0.154,0.170]
Married	0.4498	(0.0221)	0.000	[0.406,0.493]
Separated or divorced	-0.0626	(0.0430)	0.145	[-0.147,0.022]
Widowed	-0.0766	(0.2629)	0.771	[-0.592,0.439]
Black	-0.3783	(0.0196)	0.000	[-0.417,-0.340]
Hispanic	-0.0776	(0.0205)	0.000	[-0.118,-0.037]
Mixed Race (Non-Hispanic)	-0.3042	(0.0813)	0.000	[-0.463,-0.145]
child_birth_mean	-0.0231	(0.0587)	0.694	[-0.138,0.092]
CRE variables				
age_mean	0.0091	(0.0039)	0.022	[0.001,0.017]
edyears_mean	-0.0288	(0.0064)	0.000	[-0.041,-0.016]
mstatus_mean	0.2106	(0.0347)	0.000	[0.143,0.279]
Bias indicator				
all_waves	0.0591	(0.0256)	0.021	[0.009,0.109]
Constant	3.0308	(0.0637)	0.000	[2.906,3.156]
Number of obs	55874			
Number of groups	7126			
R-squared within	0.3397			
R-squared between	0.5571			
R-squared overall	0.4110			
Wald χ^2	34929.90			
Prob > χ^2	0.0000			
σ_u	0.4416			
σ_e	1.1753			
ρ	0.1237			

Dependent variable: log(income)

The *all waves* indicator has a positive coefficient of 0.0591, statistically significant at the 5% significance level. This indicates that individuals who stayed for all waves had a 6.09% higher income than those who dropped out at some point, ceteris paribus.

```
bysort pid (wave): gen next_wave = (wave[_n+1] == wave + 1)

xtreg log_income i.child_birth##i.male age ib0.mstatus ib4.ethnicity ///
    edyears age_mean mstatus_mean next_wave, re
```

Table 8.2: Attrition bias: *next wave* indicator

	Coefficient	Std. err.	P>|t|	Conf. int.
Explanatory variable				
Childbirth in past year	-0.3419	(0.0321)	0.000	[-0.405,-0.279]
Childbirth in past year × Male	0.3965	(0.0371)	0.000	[0.324,0.469]
Control variables				
Age	0.1263	(0.0017)	0.000	[0.123,0.130]
Male	0.5075	(0.0184)	0.000	[0.471,0.544]
Education years	0.1620	(0.0039)	0.000	[0.154,0.170]
Married	0.4484	(0.0221)	0.000	[0.405,0.492]
Separated or divorced	-0.0639	(0.0430)	0.137	[-0.148,0.020]
Widowed	-0.0785	(0.2629)	0.765	[-0.594,0.437]
Black	-0.3800	(0.0196)	0.000	[-0.418,-0.342]
Hispanic	-0.0788	(0.0205)	0.000	[-0.119,-0.039]
Mixed Race (Non-Hispanic)	-0.3048	(0.0812)	0.000	[-0.464,-0.146]
child_birth_mean	-0.0100	(0.0584)	0.864	[-0.125,0.104]
CRE variables				
age_mean	0.0078	(0.0039)	0.046	[0.000,0.015]
edyears_mean	-0.0295	(0.0064)	0.000	[-0.042,-0.017]
mstatus_mean	0.2094	(0.0347)	0.000	[0.141,0.277]
Bias indicator				
next_wave	0.0665	(0.0140)	0.000	[0.039,0.094]
Constant	2.9844	(0.0614)	0.000	[2.864,3.105]
Number of obs	55874			
Number of groups	7126			
R-squared within	0.3399			
R-squared between	0.5573			
R-squared overall	0.4112			
Wald χ^2	34959.21			
Prob > χ^2	0.0000			
σ_u	0.4414			
σ_e	1.1752			
ρ	0.1236			

Dependent variable: log(income)

The *next wave* indicator has a positive coefficient of 0.0591, statistically significant at the 1% significance level. This indicates that individuals who appear in the subsequent wave have 6.87% higher income in the current period compared to those who drop out before the next wave, ceteris paribus.

```
xtreg log_income i.child_birth##i.male age ib0.mstatus ib4.ethnicity ///
    edyears age_mean mstatus_mean n_waves, re
```

The *number of waves* indicator has a positive coefficient of 0.0070, statistically significant at the 5% significance level. This indicates that for each additional wave an individual has been part of, he or she will have 0.7% higher income than the baseline, ceteris paribus.

All three attrition indicators are statistically significant, providing strong evidence that attrition bias is present in the sample. The consistent pattern across all measures shows that higher-income individuals are more likely to remain in the panel, meaning the sample becomes progressively less representative of lower-income populations over time.

If lower-income individuals experience larger income penalties from childbirth and are also more likely to drop out, our estimates may underestimate the true magnitude of the child penalty, and hence of gender differences.

The existence of non-random attrition threatens the validity of our estimates, since the CRE cannot account for time-varying causes of attrition that correlate with both childbirth and income; therefore, the results must be interpreted cautiously, as they are more likely to represent the effects of higher-income panel participants rather than the broader population of interest.

Table 8.3: Attrition bias: *number of waves* indicator

	Coefficient	Std. err.	P>|t|	Conf. int.
Explanatory variable				
Childbirth in past year	-0.3431	(0.0321)	0.000	[-0.406,-0.280]
Childbirth in past year × Male	0.4011	(0.0371)	0.000	[0.328,0.474]
Control variables				
Age	0.1245	(0.0017)	0.000	[0.121,0.128]
Male	0.5012	(0.0186)	0.000	[0.465,0.538]
Education years	0.1620	(0.0039)	0.000	[0.154,0.170]
Married	0.4500	(0.0221)	0.000	[0.407,0.493]
Separated or divorced	-0.0632	(0.0430)	0.142	[-0.147,0.021]
Widowed	-0.0722	(0.2629)	0.784	[-0.588,0.443]
Black	-0.3797	(0.0196)	0.000	[-0.418,-0.341]
Hispanic	-0.0785	(0.0205)	0.000	[-0.119,-0.038]
Mixed Race (Non-Hispanic)	-0.3048	(0.0812)	0.000	[-0.464,-0.146]
child_birth_mean	0.0040	(0.0612)	0.947	[-0.116,0.124]
CRE variables				
age_mean	0.0024	(0.0052)	0.642	[-0.008,0.013]
edyears_mean	-0.0286	(0.0064)	0.000	[-0.041,-0.016]
mstatus_mean	0.2064	(0.0347)	0.000	[0.138,0.275]
Bias indicator				
n_waves	0.0070	(0.0028)	0.012	[0.002,0.013]
Constant	3.1084	(0.0772)	0.000	[2.957,3.260]
Number of obs	55874			
Number of groups	7126			
R-squared within	0.3397			
R-squared between	0.5574			
R-squared overall	0.4110			
Wald χ^2	34935.12			
Prob > χ^2	0.0000			
σ_u	0.4412			
σ_e	1.1753			
ρ	0.1235			

Dependent variable: log(income)