Panel Data: Homework

Mai-Anh Dang | Student ID: 21608631 | TSE M2 EEE 2017-18

November 16, 2017

1 Data Manipulation

We work on the data *wagedata.dta*, which has 545 male individuals from 1980 to 1987

Procedure of Data Manipulation

- The original data is in wide format. It is transformed to long format, by `reshape` command. The column `year` is created from 1980-1987.
- In the long format, each row corresponds to one period of one individual.
- Create the `numobs` for the numbers of observed periods for each individual
- By `tab year` and `tab numobs`, we check the balance of the panel data. In fact, it is balanced with **4,360 obs.**, **545 individual obs. for each year**, and **8 time obs. for each individuals**.
- This is more about the technical limitations of `xtoverid`, which would be used to test and compare models. Currently, `xtoverid` does not automatically support factor regressors in models (i.e. `year` when we use `i.year`). The dummies of year is created manually, to obtain `year1980` to `year1987` binary variables.

Inspection for Panel Data

The distribution of $log(wage)$ variables is presented in **Figure 1** and **Figure 2**. The summary statistic of interest variables is also reported in **Table 1**:

Table 1: **Summary Statistics for Interest Variables**

		Mean	Std. Dev	Min	Max	Observations
$log(wage)$	overall	1.6491	0.5326	-3.5790	4.0518	$N = 4360$
	between		0.3907	0.3333	3.1742	$n = 545$
	within		0.3622	-2.4672	3.2047	$T = 8$
$educ$	overall	11.767	1.7461	3	16	$N = 4360$
	between		1.7476	3	16	$n = 545$
	within		0	11.767	11.767	$T = 8$
$exper$	overall	6.5147	2.8258	0	18	$N = 4360$
	between		1.7476	3.5	14.5	$n = 545$
	within		2.2916	3.0147	10.0146	$T = 8$
$expersq$	overall	50.4247	40.7820	0	324	$N = 4360$
	between		26.351	17.5	215.5	$n = 545$
	within		31.1431	-44.0752	158.9248	$T = 8$

We are interested in the effect of eduction ($educ$) on wage ($lwage$), i.e. the return of schooling. There are other factors which might affect the wage, as regressors involved in the model. The key issue of this estimation is that there are potential unobserved individual characteristics correlating with both the dependent variable (*lwage*) and interest regressor (*educ*), such as the individual ability. That varies across individuals and constant over time, assumed to be captured in α_i, we have the following equation:

$$\begin{aligned} log(wage_{it}) = \beta_1 + \beta_2 educ_i &+ \beta_3 black_i + \beta_4 hispan_i + \beta_5 exper_{it} + \beta_6 exper_{it}^2 \\ &+ \beta_7 married_{it} + \beta_8 union_{it} + \lambda_t + \alpha_i + \epsilon_{it} \end{aligned}$$

2 Pooled OLS

First, we ignore α_i, or in other words, we assume that the individual-specific unobserved effect is insignificant. The results of OLS estimators (in standard form) is in **Column (1), Table 3**. *educ* has the positive effect on *lwage*, which is significant at 1%. *exper* also has positive statistically significant effect, while *expersq* coefficient is negative. It is reasonable and expectable that the experience has diminishing effect on wage. This OLS estimates are based on the assumptions that:

1. $E[\epsilon_{it}|x_{it}] = 0$, ignoring α_i
2. ϵ_i and ϵ_j are independent for $i \neq j$
3. E[X'X] is full rank (K)
4. $E[\epsilon_i \epsilon_i' | X_i] = \sigma^2 I_T$

Under these assumptions, the asymptotic distribution of pooled OLS estimate is:

$$\sqrt{N}(\hat{\beta}_{POLS} - \beta) \xrightarrow[n \to \infty]{d} N(0, \sigma^2 E[X_i' X_i]^{-1})$$

In which, the asymptotic variance-covariance matrix of pooled OLS estimate, which is used to construct the standard errors for test statistics, relies on **Assumption 4**. However, this assumption is very likely violated, as the error terms are hardly *i.i.d* in reality. First of all, the error terms are not likely homoskedastic. We might assume that these errors are independent between individuals, but within a given individual (i.e. comparing this individual over time), errors are likely correlated. Thus, the usual OLS standard errors (SE) tends to report the POLS to be more accurate that it actually is, specifically with smaller standard error.

More robust and accurate standard errors could be obtained by clustering by individuals. By this approach, the heteroskedasticity is adjusted, (yet it does not necessarily account for auto-correlation). The result with robust SE is reported in **Column (2), Table 3**. Indeed, the SE is higher as it takes into account that error terms are not *i.i.d.* The coefficients are same. As when assumption 4 is violated, it only causes the misleading SE and results of test statistics, while the POLS coefficients are still consistent.

3 Random Effects

In random effects model, the individual-specific effect is taken into account, but assuming that random factors $\alpha_i \sim N(0, \sigma_\alpha^2)$. The error terms of the model is:

$$u_{it} = \alpha_i + \epsilon_{it}$$

It is based on the assumptions that:

1. Random effect α_i are *iid*:
   $E[\alpha_i | educ_i, \cdots, union_i] = E[\alpha_i] = 0, i = 1, \cdots, N$
   $E[\alpha_i^2 | educ_i, \cdots, union_i] = \sigma_\alpha^2$, and α_i and α_j are independent for $i \neq j$

2. Errors ϵ_{it} are *iid*:
   $E[\epsilon_i | educ_i, \cdots, union_i] = E[\epsilon_i] = 0, i = 1, \cdots, N$
   $E[\epsilon_i \epsilon_i' | educ_i, \cdots, union_i] = \sigma_\epsilon^2 I_T$, and ϵ_i and ϵ_j are independent for $i \neq j$

Under these assumptions, we can define: $E[u_{it}|educ_i, \cdots, union_i] = 0$, so the estimated coefficients are consistent. With $\hat{\sigma}_\epsilon^2$ and $\hat{\sigma}_\alpha^2$ from between and within models, it is feasible to construct the $\hat{\Omega} = V\hat{ar}[u|X]$. Random Effects is the FGLS with $\hat{\Omega}$ *(together with the Full Rank Condition)*. The results of Random Effects is in **Column (3), Table 3**. The robust SE is applied to adjust further potential within heteroskedasticity.

Comparing to the counter OLS model, the magnitude of coefficients are quite close, the sign and significance are same. As under $E[u_{it}|X_i] = 0$, both estimators are unbiased and consistent. Yet, the random effects (FGLS version) would be more efficient. In fact, the robust SE of RE is smaller for most of coefficients. The RE is better, except the case that $Var(\alpha_i) = 0$. To test if the random effects is better than POLS, the **Breusch and Pagan Lagrangian multiplier test** is applied, the $\bar{\chi}^2(1) = 3203.64, p-val = 0.000$, we reject the $H_0 : Var(\alpha_i) = 0$. The RE is preferred.

4 Fixed Effects

Comparing to RE, FE has less strong assumptions. It only relies on the assumptions about ϵ_{it} that: $E[\epsilon_{it}|educ_i, \cdots, union_i] = E[\epsilon_i] = 0, i = 1, \cdots, N, t = 1, \cdots, T$. The heteroskedasticity of $E[\epsilon_i \epsilon_i'|x_i]$ is adjusted by clustering. The unobserved α_i is treated by *within model*:

$$(y_{it} - \bar{y}_i) = (\bar{x}_{it} - \bar{x}_i)'\beta + (\epsilon_{it} - \bar{\epsilon}_i)$$

The FE estimator $\hat{\beta}_{FE}$ is unbiased and consistent. The FE estimator is presented in **Column (4), Table 3**. The size of coefficients are relatively different from RE model, yet the sign and the significance of them are similar.

There are several regressors omitted from the model, namely *educ, black, hisp, year1987*. By subtracting the original model by the time averaged of variables, the FE model gets rid of time-invariant unobserved α_i, yet it also eliminates other regressors *black, hisp, educ*, which are unchanged by time.

The question is that why $exper_{it}$ is redundant from the model. In fact, my STATA results eliminate *year1987* (reference dummies is *year1980*, already eliminated from the model), instead of $exper_{it}$. If I change the order of input with $exper_{it}$ after *year1987*, the $exper_{it}$ is redundant. It is because of the collinearity between the $exper_{it}$ and year dummies.

5 Hausman Test: Compare FE vs. RE

If the individual effect is actually random (as the assumption of RE), both $\hat{\beta}_{FE}$ and $\hat{\beta}_{RE}$ is consistent, $\hat{\beta}_{RE}$ is at least as efficient as $\hat{\beta}_{FE}$. Otherwise, $\hat{\beta}_{FE}$ is consistent and $\hat{\beta}_{RE}$ is not.

The Hausman test is used to compare FE and RE models. It will test if there is any systematic difference between two models (on the time-varying variables only). If there is, FE model is preferred (as RE estimator is inconsistent).

The hypotheses are that:

1. $H_0 : E[x_{it}\alpha_i] = 0$, No systematic difference between FE and RE
2. $H_1 : E[x_{it}\alpha_i] \neq 0$

In fact, for this case, we want to compare the FE and RE model in robust version. In STATA, the `xoverid` is applied, which is the robust version of Hausman test. It reports the Sargan-Hansen statistic, and conducts the **test of overidentifying restrictions** for FE and RE models. The FE model only relies on the orthogonality conditions for regressors that $E[x_{it}\epsilon_{it}] = 0$, while the RE model has additional orthogonality conditions that $E[x_{it}\alpha_i] = 0$, that are overidentifying restrictions. The null hypothesis is: H_0: Overidentifying restriction is valid. The t-statistics is reported in **Table 3**, with chi-square = 44.53, p-val=0.000. We reject the null hypothesis (RE model is rejected).

It is concluded that FE model is preferred, and unobserved individual effect α_i is not random.

6 Effects of Time-invariant Variable: *Educ*

As the previous discussion, FE has less strong assumption about the α_i than RE, hence its estimator is more reliable to be consistent. Yet, we face the situation that the interest variable (*educ*) is time-invariant.

Table 2: Advantages and Disadvantages of FE and RE Panel Model

	Advantages	Disadvantages
Fixed Effects	Weaker assumptions, it only relies on the orthogonality condition of X_{it} and ϵ_{it} without any assumption about α_i	Unable to estimate the impact of time-invariant variables Coefficient estimates of time-variant variables not reliable when most of the variation of regressors is cross-sectional rather than over time
Random Effects	If the RE estimator is consistent, it will be more efficient than FE estimator The RE model enables estimating the effect of time-invariant variables	Strong assumptions of α_i (i.e. the distribution of unobserved individual-specific effect is random. If this assumption is not true, the RE estimator is not consistent)

7 Hausman-Taylor IV

The FE model is not able to estimate the effect of interest variable *educ* (which is time-invariant), while the RE model is rejected by the Robust Hausman Test. That is the motivation for the Hausman-Taylor approach, where IV is used to overcome the issue of potential correlation between $educ_i$ and α_i.

Consider the model (without time dummies):

$$\begin{aligned} log(wage_{it}) = \beta_1 + \beta_2 educ_i &+ \beta_3 black_i + \beta_4 hispan_i + \beta_5 exper_{it} + \beta_6 exper_{it}^2 \\ &+ \beta_7 married_{it} + \beta_8 union_{it} + \alpha_i + \epsilon_{it} \end{aligned}$$

Then, we present it as:

$$log(wage_{it}) = x'_{1it}\beta_1 + w'_{1i}\gamma_1 + \gamma_2 educ_i + \alpha_i + \epsilon_{it}$$

where:

- $educ_i$ is assumed to be the only endogenous regressors (i.e. correlated with α_i)
- $x_{1it} = (exper_{it}, exper_{it}^2, married_{it}, union_{it})'$, time-variant and uncorrelated with α_i and ϵ_{it}
- $w_{1i} = (black_i, hispan_i)'$, time-invariant and uncorrelated with α_i and ϵ_{it}

By Hausman-Taylor Approach, the below could be IV for the model:

- x_{1it} for x_{1it}
- w_{1i} for w_{1i}
- $\bar{x}_{1i}$ for $educ_i$

8 Hausman-Taylor IV Estimates

The results of Hausman-Taylor IV model is presented in **Column (6), Table 3**. Comparing with the counter within estimates, even though the sign and significance of estimated coefficients are same, the magnitude is slightly different.

9 Hausman Test: Hausman-Taylor IV Estimates vs. FE

The validity of the IV is tested by the Hausman test (**Test of Overidentifying Restrictions**) for Hausman-Taylor model and the counter within model. The FE model only relies on the orthogonality conditions for regressors that $E[x_{it}\epsilon_{it}] = 0$, while the Hausman-Taylor model has additional assumptions about the exogenous regressors (i.e. independent from fixed effect), that are overidentifying restrictions. The null hypothesis is: H_0: Overidentifying restriction is valid. The t-statistics is reported in **Table 3**, with chi-square = 25.18, p-val=0.000. The degree of freedom is 3, equals to the difference between the number of exogenous time varying variables (4) and one endogenous $educ_i$. We reject the null hypothesis **(Hausman-Taylor model is rejected)**.

The instruments are not valid.

Table 3: Results by Different Estimation Approaches

	Dependent variable: log(wage)					
	OLS	*OLS Robust SE*	*RE Robust SE*	*FE Robust SE*	*FE Robust SE*	*HTaylor IV*
	(1)	(2)	(3)	(4)	(5)	(6)
educ	0.091^{***}	0.091^{***}	0.092^{***}	-	-	0.114^{***}
	(0.005)	(0.011)	(0.011)			(0.016)
exper	0.067^{***}	0.067^{***}	0.106^{***}	0.132^{***}	0.116^{***}	0.111^{***}
	(0.014)	(0.020)	(0.016)	(0.012)	(0.010)	(0.008)
expersq	-0.002^{***}	-0.002^{**}	-0.005^{***}	-0.005^{***}	-0.004	-0.004^{***}
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
black	-0.139^{***}	-0.139^{***}	-0.139^{***}	-	-	-0.140
	(0.024)	(0.050)	(0.051)			(0.049)
hisp	0.016	0.016	0.022	-	-	0.033
	(0.021)	(0.039)	(0.040)			(0.045)
married	0.108^{***}	0.108^{***}	0.064^{***}	0.047^{**}	0.045^{**}	0.061^{***}
	(0.016)	(0.026)	(0.019)	(0.021)	(0.017)	
union	0.182^{***}	0.182^{***}	0.106^{***}	0.080^{***}	0.082^{***}	0.106^{***}
	(0.017)	(0.027)	(0.021)	(0.023)	(0.022)	(0.018)
Constant	0.092	0.092	0.024	1.027^{***}	1.064^{***}	-0.264
	(0.078)	(0.161)	(0.160)	(0.040)	(0.189)	
Year Dummies	Yes	Yes	Yes	Yes	No	No
Clustering	-	Yes	Yes	Yes	Yes	-
Obs.	4,360	4,360	4,360	4,360	4,360	4,360
Groups	-	545	545	545	545	545
R^2 within	-	-	0.180	0.180	0.179	
R^2 between	-	-	0.186	0.001	0.186	
R^2 overall	0.189	0.189	0.183	0.064	0.183	
Breusch-Pagan LM	-	-	3203.6^{***} p-val =0.000	-	-	-
$Sigma_\alpha$	-	-	0.324	0.400	0.325	0.332
$Sigma_\epsilon$	-	-	0.351	0.351	0.351	0.351
rho	-	-	0.461	0.566	0.461	0.473
Hausman test	-	-	$\chi^2(9) = 44.53$	-	-	$\chi^2(3) = 25.18$
	-	-	p-val = 0.000	-	-	p-val = 0.000

Note: The coefficients of time period dummies are not reported $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

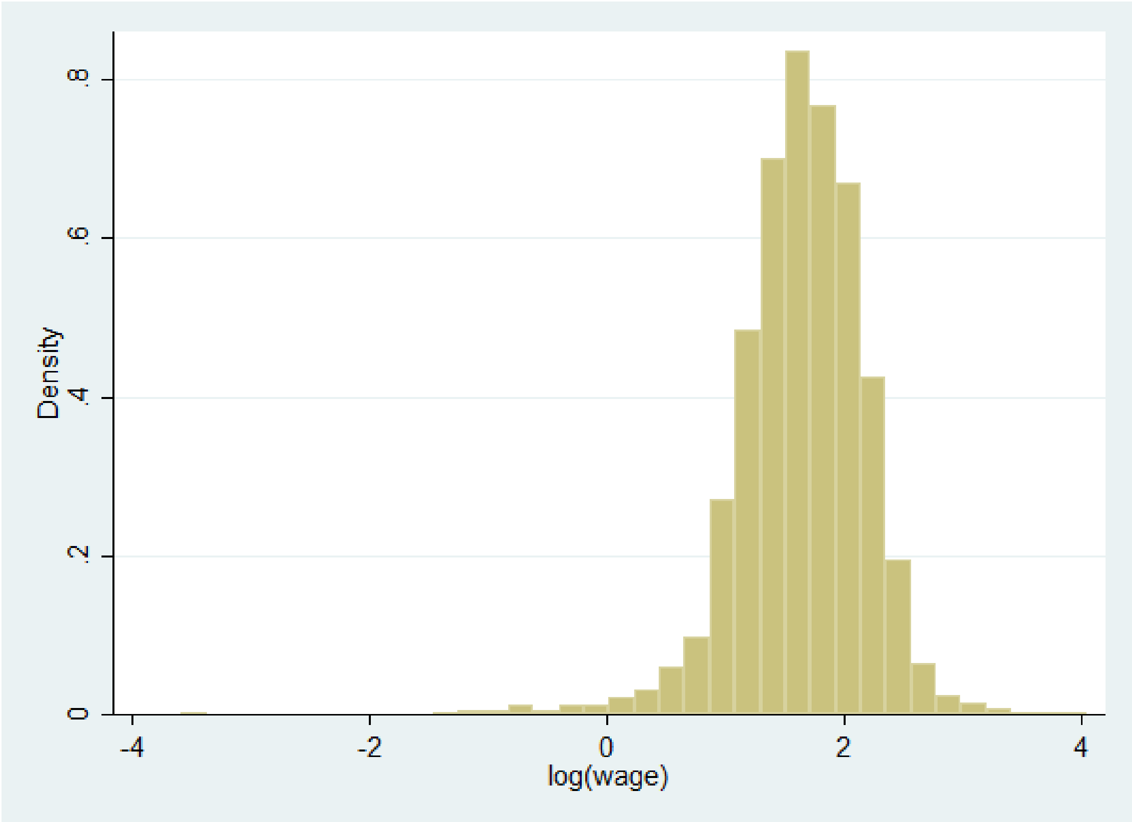

Figure 1: Distribution of $Log(Wage)$

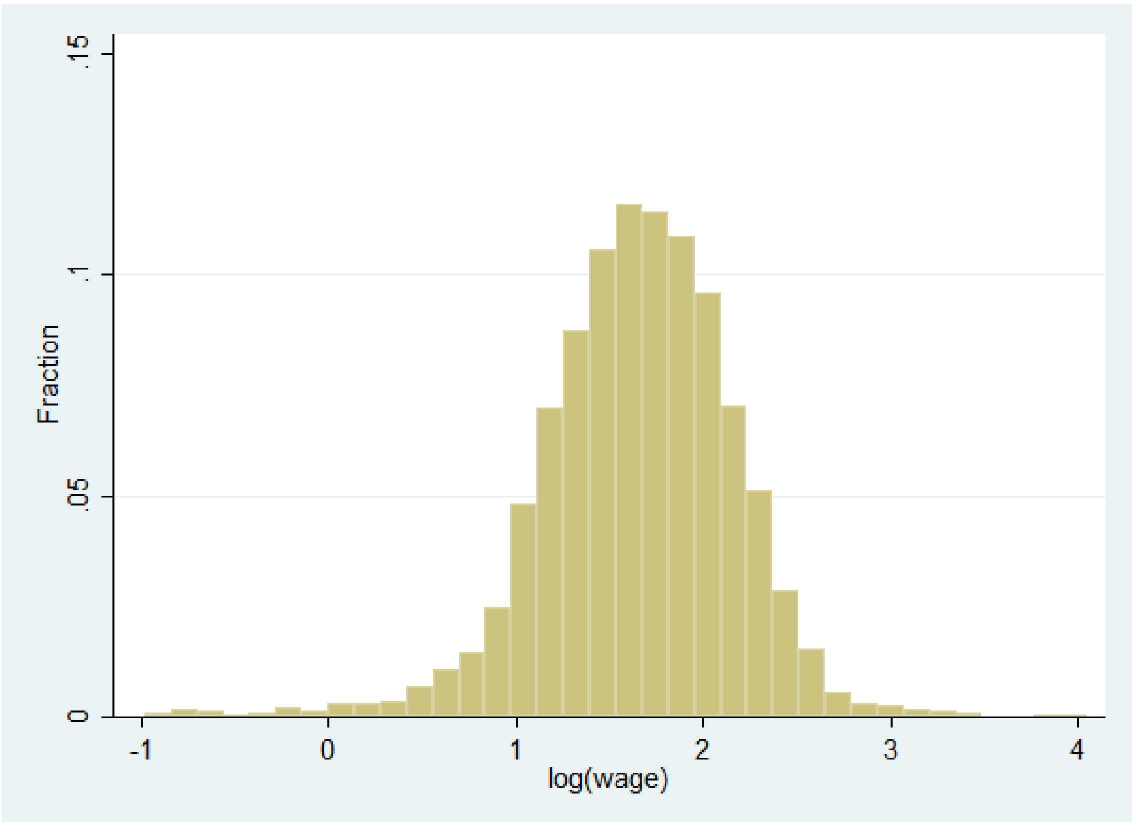

Figure 2: Distribution of $Log(Wage)$ by Fraction