Microeconometrics Assignment 3

Jia Ru, 27720141152749

March 21, 2016

1 Treatment Effect

Consider the treatment-outcome model

$$y = x\beta + \alpha d + \epsilon$$

(a) Is randomized treatment a sufficient condition for identification of α and β ?

Random assignment of treatment allow the identification of the ATT(Average Treatment Effect of the Treated,or TOT in some books),i.e. the parameter α, but without further assumptions do not necessarily allow one to identify the underlying structural parameters of β.
To interprete it in detail, let Y_1 and Y_0 denote the outcome if the individual is treated/not treated, respectly. So only one of Y_1 and Y_0 can be observed, the other is "counterfactual". And the observed outcome Y_i, can be write as

$$Y_i = d_i Y_{1i} + (1 - d_i) Y_{0i}$$

Define the treatment effect

$$\alpha_i = Y_{1i} - Y_{0i}$$

And Average Treatment Effect(ATE) as well as Average Treatment Effect of the Treated(ATT)

$$E[\alpha_i] = E[Y_{1i} - Y_{0i}]$$

$$E[\alpha_i | d_i = 1] = E[Y_{1i} - Y_{0i} | d_i = 1]$$

Since

$$\begin{aligned}
&E[Y_i|d_i = 1] - E[Y_i|d_i = 0] \\
=&E[Y_{1i}|d_i = 1] - E[Y_{0i}|d_i = 1] + E[Y_{0i}|d_i = 1] - E[Y_{0i}|d_i = 0] \\
=&ATT + E[Y_{0i}|d_i = 1] - E[Y_{0i}|d_i = 0]
\end{aligned}$$

The first term is the parameter of interest, α, the remainder is the selection bias term. In the case of random assignment of treatment, it is zero . So we can

identify parameter α.
But we could not identify parameter β without the information of the correlation between x, d and ϵ. If

$$E[u|x] = 0$$

then we can identify parameter β.

(b)If participation of treatment (d = 1 or 0) is correlated with y, suggest a method to obtain α.

We can use IV method to estimate α: Find an inturment variable Z, such that Z is correlated with d but not correlated with X or ϵ, i.e. Z affacts the outcome Y only thouth treatment d.

(c) Suppose the outcome equation is unknown, eligibility of treatment is exogenously determined, suggest a method to obtain the average treatment effect.

If eligibility of treatment (instead of treatment per se) is randomized, It can be proved that(the proof is omitted here):

$$TOT \equiv E[Y_1 - Y_0|d = 1] = \frac{E[Y|e = 1] - E[Y|e = 0]}{Prob(d = 1|e = 1)} = \frac{ITT}{program\ participation\ rate.}$$

where dummy e denote "Eligibility", ITT means "Intent-to-Treat" effect
So given observed data, ATE can be estimated as: difference of sample mean of subsamples of eligible and not eligible, divided by program paricipation rate.

2 Empirical Analysis

2.1 Explore the data set.Identify which variables are time variant, individual variant or both.

Group the observations by id and time, respectively. Then see the summary statistic of all the variables. Analysis the tables (especially the Std.Dev. collomn) I find:

- OCC, IND, SOUTH, SMSA, FEM, UNION, ED, BLK is invariant accross time.
- No variable is invariance accross id
- Other variables are both time variant and individual variant.

```
sort id
by id: summ
sort time
by time:summ
```

(results ommited, since it' too much)

2.2 Set panel Structure. Compute summary statistics for all the variables.Are there any suspicious values?

```
xtset, clear
xtset id time
xtdescribe
xtsum
```

Consider the "Std.Dev." column of ‘xtsum‘ table. For id, FEM, UNION, ED, BLK the within standard deviation is 0. Only time's between Standard deviation is 0.

The 'suspicious values' appears for the 4 variables: OCC, IND, SOUTH, SMSA. The conclusion seems contradict in 1) and 2) After checking the meannning of these variables, I speculate that it's becuase, take OCC for example, a few people transfers form blue-collar occupation to other occupation type during the survey years, consisting a small percent of the total sample however. So I didn't perceive these pattern in question 1), (Since this data have 595 id, I couldn't check it one by one).

Variable		Mean	Std. Dev.	Min	Max	Observations	
EXP	overall	19.85378	10.96637	1	51	N =	4165
	between		10.79018	4	48	n =	595
	within		2.00024	16.85378	22.85378	T =	7
WKS	overall	46.81152	5.129098	5	52	N =	4165
	between		3.284016	31.57143	51.57143	n =	595
	within		3.941881	12.2401	63.66867	T =	7
OCC	overall	.5111645	.4999354	0	1	N =	4165
	between		.469327	0	1	n =	595
	within		.1731615	-.3459784	1.368307	T =	7
IND	overall	.3954382	.4890033	0	1	N =	4165
	between		.4648725	0	1	n =	595
	within		.152739	-.4617047	1.252581	T =	7
SOUTH	overall	.2902761	.4539442	0	1	N =	4165
	between		.4489462	0	1	n =	595
	within		.0693042	-.5668667	1.147419	T =	7
SMSA	overall	.6537815	.475821	0	1	N =	4165
	between		.4601658	0	1	n =	595
	within		.1223035	-.2033613	1.510924	T =	7
MS	overall	.8144058	.3888256	0	1	N =	4165
	between		.3686109	0	1	n =	595
	within		.1245274	-.0427371	1.671549	T =	7
FEM	overall	.112605	.3161473	0	1	N =	4165
	between		.3163754	0	1	n =	595
	within		0	.112605	.112605	T =	7
UNION	overall	.3639856	.4812023	0	1	N =	4165
	between		.4543848	0	1	n =	595
	within		.1593351	-.4931573	1.221128	T =	7
ED	overall	12.84538	2.787995	4	17	N =	4165
	between		2.790006	4	17	n =	595
	within		0	12.84538	12.84538	T =	7
BLK	overall	.0722689	.2589637	0	1	N =	4165
	between		.2591505	0	1	n =	595
	within		0	.0722689	.0722689	T =	7
LWAGE	overall	6.676346	.4615122	4.60517	8.537	N =	4165
	between		.3942387	5.3364	7.813596	n =	595
	within		.2404023	4.781808	8.621092	T =	7
id	overall	298	171.7821	1	595	N =	4165
	between		171.906	1	595	n =	595
	within		0	298	298	T =	7
time	overall	4	2.00024	1	7	N =	4165
	between		0	4	4	n =	595
	within		2.00024	1	7	T =	7

2.3 GLS regression

```
gen EXP2 = EXP*EXP

global Y "LWAGE"
global X "EXP EXP2 WKS OCC IND SOUTH SMSA MS FEM
   UNION ED BLK"

est clear
set matsize 595, permanently

xtgls $Y $X, i(id) corr(ind)
est store GLS
```

I. Do the signs of the parameters agree with your intuition?

For variable EXP: as one accumulate his working experience, his wages will increase (coef. of EXP is positive), however this marginal effects decrease (coef. of EXP2 is negative).
coefficients of other variables all agree with my intuition: The people who have more working hours and high education, being in in a manufacturing industry, married, being in Union tends to have more wage, while people who are black female, blue-collar worker, live in south ends to have less wage.

II. What variables are significant at 5% level?

All the variables.

III.Is this an appropriate procedure?

No, this pooled model do not consider heterogenity, that is to say, it implictly assumes that different individuals have same constant term. There are some factors, say, people's ability, can affact WAGE, which is unobservable, since different people have different ability, it's unreasonable to use pooled model.

2.4 Implement the following methods for the wage equation:

- Fixed effects (within)
- Random effects (GLS)
- Between
- Random effects (ML)

```
xtreg $Y $X , fe
est store fe

xtreg $Y $X , re
est store re_GLS

xtreg $Y $X , be
est store be

xtreg $Y $X , mle
est store re_ML

outreg2 [GLS fe re_GLS be re_ML] using "tex/reg.tex"
   ///
, nose replace tex(frag) title("Regression Resutlts")
```

REGRESSION RESULTS

	(1)	(2)	(3)	(4)	(5)
	GLS	fe	re_GLS	be	re_ML
VARIABLES	LWAGE	LWAGE	LWAGE	LWAGE	LWAGE
EXP	0.0401***	0.113***	0.0821***	0.0319***	0.107***
EXP2	-0.000673***	-0.000418***	-0.000808***	-0.000566***	-0.000515***
WKS	0.00422***	0.000836	0.00103	0.00919**	0.000840
OCC	-0.140***	-0.0215	-0.0501***	-0.168***	-0.0251*
IND	0.0468***	0.0192	0.00374	0.0579**	0.0138
SOUTH	-0.0556***	-0.00186	-0.0166	-0.0571**	0.00577
SMSA	0.152***	-0.0425**	-0.0138	0.176***	-0.0475**
MS	0.0484**	-0.0297	-0.0746***	0.115**	-0.0414**
FEM	-0.368***		-0.339***	-0.317***	-0.176
UNION	0.0926***	0.0328**	0.0632***	0.109***	0.0387***
ED	0.0567***		0.0997***	0.0514***	0.136***
BLK	-0.167***		-0.210***	-0.158***	-0.261*
o.FEM		-			
o.ED		-			
o.BLK		-			
Constant	5.251***	4.649***	4.264***	5.121***	3.126***
Constant	5.251***	4.649***	4.264***	5.121***	0.839***
Constant	5.251***	4.649***	4.264***	5.121***	0.153***
Observations	4,165	4,165	4,165	4,165	4,165
R-squared		0.658		0.544	
Number of id	595	595	595	595	595

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Do the signs of the parameters agree with your intuition?

- EXP: Yes. as EXP increase, his wages will increase (coef. of EXP is positive), however this marginal effects decrease (coef. of EXP2 is negative).
- WKS: Yes. the coef. is positive which means more working hour, more wages.
- OCC: Yes. bule-collar works earn less than white-collar, it consist with intuition.
- IND: the regression result tells me manufacture industry earns more, maybe it's true.
- SOUTH, SMSA, MS: the 4 regression give different signs of coefficient.
- FEM: Yes, Even in the U.S, female earn less than male, it's also the case in China.
- UNION: Yes, being in the Union will have more wage, since they have more bargining power.
- EDU: Yes, Education attainment and wages is positively-correlated, Substaintial number of literatures in Labor Economic have studied this issue.
- BLK: Yes, people who are black earns less, whether it's due to race discrimination or ability which is unobervable, howerver, is not clear.

What variables are significant at 5% level?

- fe: EXP SMSA UNION
- re-GLS: EXP OCC MS FEM UNION ED BLK
- be: All varibles
- re-ML: EXP OCC SMSA MS UNION ED BLK

Is the estimation procedure appropriate for the context?Is the estimator consistent? Compare the results. Do you accept or reject the presence of individual effects? Why or why not?

The estimation of panel data models boils down to the choice between three estimators:
1. The pooled model should be used when there is no individual heterogeneity in the model.
2. When there is individual heterogeneity and it is not correlated with the independent variables of the model, the random effects model should be preferred. The Hausman test helps us decide whether this is the case or not.

3. If the individual heterogeneity is correlated with the independent variables, the fixed effects model should be used.
In this context, the dependent variable is WAGE, by intuition, we can not ignore heterogeneity, since different people have different abiliy, characterictics, pereferences and so on, these factors will affect WAGE and it varies accross individuals. The poold regression model is not reasonable.
So the pooled model in collumn 1 is not consistent, the fixed effect model in column 2,4 is always consistent, at the expense of efficiency. Whether random effect model is consistent depends on whether the assumption of random effect model are satisfied, which will disscuss later, if true, the random effect model is alse consistent as well as efficient(than fixed effect model).

2.5 Compute the estimates of specific effects (fixed and random). Comment. Would you prefer the fixed effects or the random effects results? Why? (you can use the Hausman test).

Panel data models acknowledge that different units behave differently by adding an individual heterogeneity term.There are two kinds of panel data models that account for individual heterogeneity in two different ways: fixed effects model and random effects model. Fixed effects model assumes that the heterogeneity term and the independent variables x are correlated, while random effects model assumes that the heterogeneity term and the independent variables x are not correlated.

In this wage equaiotn context, I think fixed effects model is better. Because here the heterogeneity term is correlated with the X's. for example, individual's ability or characteristic or working-leisure pereferance affacts his wage, and it's self-evident that people's ability is correlated with his education attainment. This reasoning is based on economic intuition. Also we should apply Statistic method to see if the assumption of random effect model holds, that is, Hausman Test:

```
     hausman fe re_GLS
```

```
---- Coefficients ----
(b)          (B)             (b-B)     sqrt(diag(V_b-V_B))
fe          re_GLS        Difference          S.E.

EXP             .1132083      .0820544         .0311539               .
EXP2     -.0004184     -.0008084         .0003901                .
WKS             .0008359      .0010347        -.0001987               .
OCC             -.0214765     -.0500664         .0285899               .
IND      .0192101      .0037441          .015466                 .
SOUTH    -.0018612     -.0166176         .0147564         .0217436
SMSA     -.0424692     -.0138231        -.0286461                .
MS              -.0297258     -.0746283         .0449025               .
UNION    .0327849      .0632232        -.0304384                .
```

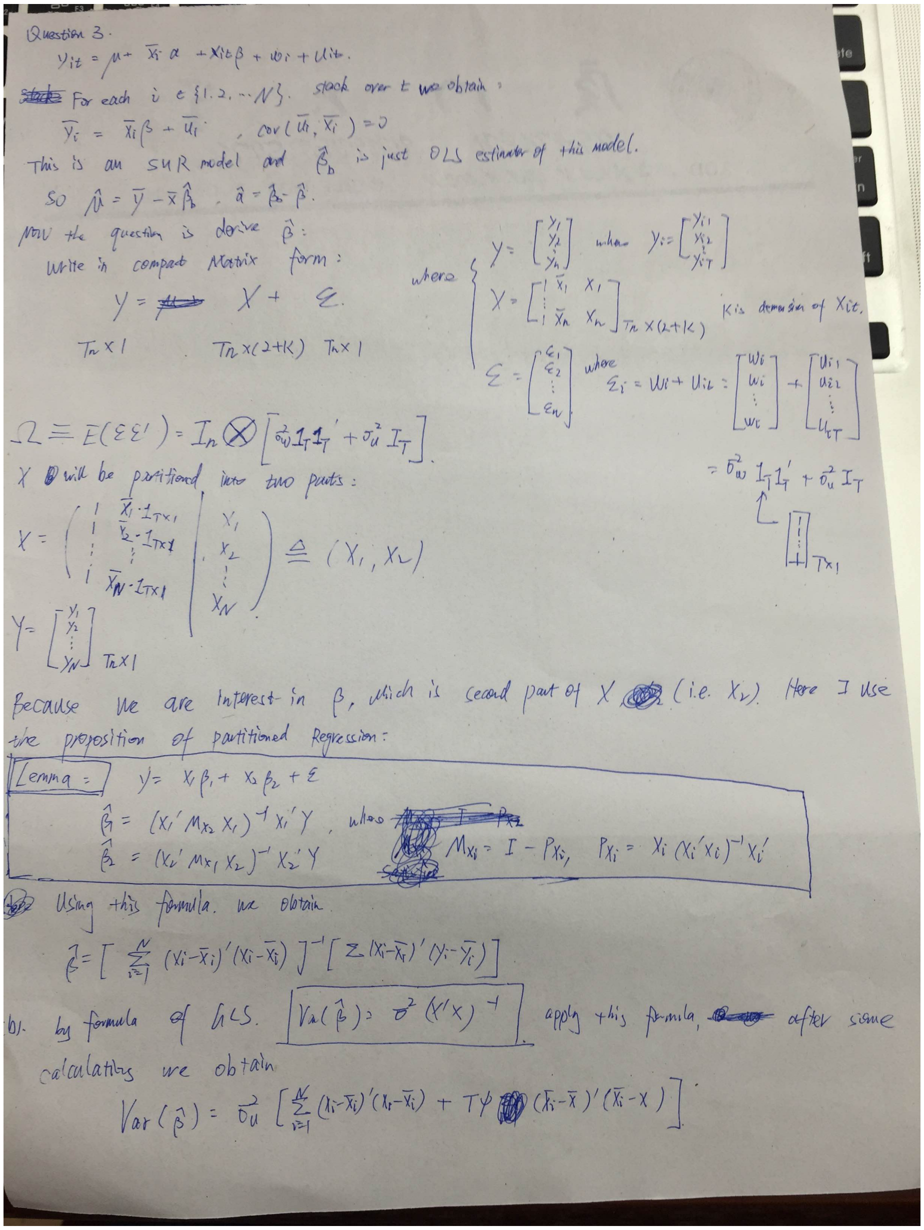
```
b = consistent under Ho and Ha; obtained from xtreg
B =      inconsistent under Ha, efficient under Ho; obtained from
    xtreg

Test:  Ho:      difference in coefficients not systematic

chi2(9) = (b-B)'[(V_b-V_B)^(-1)](b-B)
=      5075.25
Prob>chi2 =      0.0000
(V_b-V_B is not positive definite)
```

Because the p-value is very small(Prob>chi2 =0.0000), which means that the we should reject H0, that is, use the fixed effect model.

3 Question 3



Question 3.

$y_{it} = \mu + \bar{x}_i \alpha + x_{it}\beta + w_i + u_{it}$.

~~stack~~ For each $i \in \{1, 2, \cdots N\}$. stack over t we obtain:

$\bar{y}_i = \bar{x}_i \beta + \bar{u}_i$, $cov(\bar{u}_i, \bar{x}_i) = 0$

This is an SUR model and $\hat{\beta}_b$ is just OLS estimator of this model.

So $\hat{\mu} = \bar{y} - \bar{x}\hat{\beta}_b$, $\hat{a} = \hat{\beta}_b - \hat{\beta}$.

Now the question is derive $\hat{\beta}$:

Write in compact Matrix form:

$$\underset{Tn \times 1}{y} = \underset{Tn \times (2+K)}{X} + \underset{Tn \times 1}{\varepsilon}.$$

where
$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ where $y_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}$

$X = \begin{bmatrix} 1 & \bar{x}_1 & x_1 \\ \vdots & & \\ 1 & \bar{x}_n & x_n \end{bmatrix}_{Tn \times (2+K)}$ K is demension of x_{it}.

$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$ where $\varepsilon_i = w_i + u_{it} = \begin{bmatrix} w_i \\ w_i \\ \vdots \\ w_i \end{bmatrix} + \begin{bmatrix} u_{i1} \\ u_{i2} \\ \vdots \\ u_{iT} \end{bmatrix}$

$$\Omega \equiv E(\varepsilon\varepsilon') = I_n \otimes \left[\sigma_w^2 1_T 1_T' + \sigma_u^2 I_T\right].$$

$= \sigma_w^2 1_T 1_T' + \sigma_u^2 I_T$

$1_T = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}_{T \times 1}$

X will be partitiond into two parts:

$$X = \left(\begin{array}{cc|c} 1 & \bar{x}_1 \cdot 1_{T\times 1} & x_1 \\ 1 & \bar{x}_2 \cdot 1_{T\times 1} & x_2 \\ \vdots & \vdots & \vdots \\ 1 & \bar{x}_N \cdot 1_{T\times 1} & x_N \end{array}\right) \triangleq (X_1, X_2)$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{Tn \times 1}$$

Because we are interest in β, which is second part of X (i.e. X_2). Here I use the proposition of partitioned Regression:

$\boxed{\text{Lemma}:}$ $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$

$\hat{\beta}_1 = (X_1' M_{X_2} X_1)^{-1} X_1' Y$, where $M_{X_i} = I - P_{X_i}$, $P_{X_i} = X_i(X_i'X_i)^{-1}X_i'$

$\hat{\beta}_2 = (X_2' M_{X_1} X_2)^{-1} X_2' Y$

Using this formula. we obtain

$$\hat{\beta} = \left[\sum_{i=1}^{N} (x_i - \bar{x}_i)'(x_i - \bar{x}_i)\right]^{-1} \left[\sum (x_i - \bar{x}_i)'(y_i - \bar{y}_i)\right]$$

b). by formula of GLS. $\boxed{Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}}$. apply this formula, after some calculating we obtain

$$Var(\hat{\beta}) = \sigma_u^2 \left[\sum_{i=1}^{N} (x_i - \bar{x}_i)'(x_i - \bar{x}_i) + T\psi (\bar{x}_i - \bar{x})'(\bar{x}_i - x)\right].$$

Figure 1: