

## Problem Set 5 (SOLUTIONS)

This problem set will revise some of the material covered in Handout 5 on panel data models. This will require you to familiarize yourself with Stata's panel-data commands.

```
help xtset
help xttab
help xtreg
```

You will be using a dataset that comes with Stata: `psidextract.dta`. The data is a correct version of the PSID sample in Cornwell and Rupert (1988), found in Baltagi and Khanti-Akom (1990). It includes a sample of 595 individuals observed for the years 1976-82.

### Preamble

<IPython.core.display.HTML object>

Create a do-file for this problem set and include a preamble that sets the directory and opens the data. For example,

```
clear
//or, to remove all stored values (including macros, matrices, scalars, etc.)
*clear all

* Replace $rootdir with the relevant path to on your local harddrive.
cd "$rootdir/problem-sets/ps-5"

cap log close
log using problem-set-5-log.txt, replace

use problem-set-5-data.dta, clear
```

```
C:\Users\neil_\OneDrive - University of Warwick\Documents\EC910\website\warwick
> -ec910\problem-sets\ps-5
```

```
-----
      name:  <unnamed>
      log:   C:\Users\neil_\OneDrive - University of Warwick\Documents\EC910\we
> bsite\warwick-ec910\problem-sets\ps-5\problem-set-5-log.txt
      log type:  smcl
      opened on:  11 Nov 2024, 14:07:07
(PSID wage data 1976-82 from Baltagi and Khanti-Akom (1990))
```

## Questions

1. Set the unit identifier and time variable using `xtset`. Note, you can also use `tsset` for this task. This will allow you to use `xt` package commands.

```
xtset id t
```

Panel variable: id (strongly balanced)

Time variable: t, 1 to 7

Delta: 1 unit

2. Describe and summarise the variables in the dataset using the normal `describe` and `summarize` commands.

```
des
sum id t lwage ed exper weeks south
```

Contains data from problem-set-5-data.dta

Observations: 4,165

PSID wage data 1976-82 from  
Baltagi and Khanti-Akom (1990)  
11 Nov 2024 11:19  
(\_dta has notes)

Variable name	Storage type	Display format	Value label	Variable label
exper	float	%9.0g		years of full-time work experience
weeks	float	%9.0g		weeks worked
occup	float	%9.0g		occupation; occ==1 if in a blue-collar occupation
industry	float	%9.0g		industry; ind==1 if working in a manufacturing industry
south	float	%9.0g		residence; south==1 if in the South area
smsa	float	%9.0g		smsa==1 if in the Standard metropolitan statistical area
ms	float	%9.0g		marital status
female	float	%9.0g		female or male

```

union      float    %9.0g          if wage set be a union contract
educ       float    %9.0g          years of education
black      float    %9.0g          black
lwage      float    %9.0g          log wage
id         float    %9.0g
t         float    %9.0g

```

---

Sorted by: id t

Variable	Obs	Mean	Std. dev.	Min	Max
id	4,165	298	171.7821	1	595
t	4,165	4	2.00024	1	7
lwage	4,165	6.676346	.4615122	4.60517	8.537
educ	4,165	12.84538	2.787995	4	17
exper	4,165	19.85378	10.96637	1	51
weeks	4,165	46.81152	5.129098	5	52
south	4,165	.2902761	.4539442	0	1

3. Describe and summarise the variables in the dataset using the panel commands: `xtdescribe` and `xtsummarize`. Comment on the information provided.

```

xtdescribe
xtsum id t lwage ed exper weeks south

```

```

id: 1, 2, ..., 595          n =      595
t:  1, 2, ..., 7           T =       7
Delta(t) = 1 unit
Span(t)  = 7 periods
(id*t uniquely identifies each observation)

```

```

Distribution of T_i:  min      5%      25%      50%      75%      95%      max
                    7         7         7         7         7         7

```

Freq.	Percent	Cum.	Pattern
595	100.00	100.00	1111111
595	100.00		XXXXXXX

Variable		Mean	Std. dev.	Min	Max	Observations	
id	overall	298	171.7821	1	595	N =	4165
	between		171.906	1	595	n =	595
	within		0	298	298	T =	7
t	overall	4	2.00024	1	7	N =	4165
	between		0	4	4	n =	595
	within		2.00024	1	7	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
educ	overall	12.84538	2.787995	4	17	N =	4165
	between		2.790006	4	17	n =	595
	within		0	12.84538	12.84538	T =	7
exper	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
weeks	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
south	overall	.2902761	.4539442	0	1	N =	4165
	between		.4489462	0	1	n =	595
	within		.0693042	-.5668667	1.147419	T =	7

4. Use the command `xttab` and `xtrans`, `freq` to describe transitions over time in the variable `south`.

```
xttab south
xttrans south, freq
```

south	Overall		Between		Within
	Freq.	Percent	Freq.	Percent	Percent
0	2956	70.97	428	71.93	98.66
1	1209	29.03	182	30.59	94.90

-----+-----					
Total		4165	100.00	610	102.52
(n = 595)					
-----+-----					
residence;					
south==1		residence;	south==1		
if in the		if in the	South area		
South area		0	1		Total
-----+-----					
0		2,527	8		2,535
		99.68	0.32		100.00
-----+-----					
1		8	1,027		1,035
		0.77	99.23		100.00
-----+-----					
Total		2,535	1,035		3,570
		71.01	28.99		100.00

5. Create the variable: `expsq=exper^2/1000`. Why would you scale the variable in this way?

```
gen expsq=exp*exp/1000
```

6. Estimate the following model using pooled OLS, between-group, feasible GLS, within-group, LSDV, and first-difference. For the first-difference estimator, you can define a first-difference in Stata using the time-series operator: `D.variable`.

$$\ln(Wage_{it}) = \beta_1 + \beta_2 Exper_{it} + \beta_3 Exper_{it}^2 + \beta_4 Weeks_{it} + \beta_5 Eduyrs_{it} + \varepsilon_{it}$$

With each model, store the results using `estimates store`. For example,

```
* clear existing stored estimates
est clear

* Pooled OLS
regress lwage exper expsq weeks ed
est store OLS

* alternatively,
eststo OLS: regress lwage exper expsq weeks ed
```

```

est clear
* Pooled-OLS
eststo OLS: regress lwage exper expsq weeks ed

* Between-group
eststo BG: xtreg lwage exper expsq weeks ed, be

* Feasible-GLS
eststo FGLS: xtreg lwage exper expsq weeks ed, re theta

* Within-group
eststo WG: xtreg lwage exper expsq weeks ed, fe

* LSDV
eststo LSDV: areg lwage exper expsq weeks ed, absorb(id)

* First-difference
eststo FD: reg D.(lwage exper expsq weeks), noconst

```

Source		SS	df	MS	Number of obs	=	4,165
-----+-----					F(4, 4160)	=	411.62
Model		251.491445	4	62.8728612	Prob > F	=	0.0000
Residual		635.413457	4,160	.152743619	R-squared	=	0.2836
-----+-----					Adj R-squared	=	0.2829
Total		886.904902	4,164	.212993492	Root MSE	=	.39082

	lwage		Coefficient	Std. err.	t	P> t	[95% conf. interval]
-----+-----							
	exper		.044675	.0023929	18.67	0.000	.0399838 .0493663
	expsq		-.715631	.0527938	-13.56	0.000	-.8191351 -.6121268
	weeks		.005827	.0011827	4.93	0.000	.0035084 .0081456
	educ		.0760407	.0022266	34.15	0.000	.0716754 .080406
	_cons		4.907961	.0673297	72.89	0.000	4.775959 5.039963
-----+-----							

Between regression (regression on group means)	Number of obs	=	4,165
Group variable: id	Number of groups	=	595

R-squared:	Obs per group:
Within = 0.1357	min = 7

Between = 0.3264  
Overall = 0.2723

avg = 7.0  
max = 7

sd(u\_i + avg(e\_i.)) = .324656  
F(4,590) = 71.48  
Prob > F = 0.0000

lwage	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
exper	.038153	.0056967	6.70	0.000	.0269647	.0493412
expsq	-.631272	.1256812	-5.02	0.000	-.8781089	-.384435
weeks	.0130903	.0040659	3.22	0.001	.0051048	.0210757
educ	.0737838	.0048985	15.06	0.000	.0641632	.0834044
_cons	4.683039	.2100989	22.29	0.000	4.270407	5.095672

Random-effects GLS regression  
Group variable: id

Number of obs = 4,165  
Number of groups = 595

R-squared:

Within = 0.6340  
Between = 0.1716  
Overall = 0.1830

Obs per group:

min = 7  
avg = 7.0  
max = 7

corr(u\_i, X) = 0 (assumed)  
theta = .82280511  
Wald chi2(4) = 3012.45  
Prob > chi2 = 0.0000

lwage	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
exper	.0888609	.0028178	31.54	0.000	.0833382	.0943837
expsq	-.772565	.0622619	-12.41	0.000	-.894596	-.6505339
weeks	.0009658	.0007433	1.30	0.194	-.000491	.0024226
educ	.1117099	.0060572	18.44	0.000	.0998381	.1235818
_cons	3.829366	.0936336	40.90	0.000	3.645848	4.012885
sigma_u	.31951859					
sigma_e	.15220316					
rho	.81505521	(fraction of variance due to u_i)				

note: educ omitted because of collinearity.

```

Fixed-effects (within) regression              Number of obs   =    4,165
Group variable: id                           Number of groups =     595

R-squared:                                   Obs per group:
    Within = 0.6566                               min =          7
    Between = 0.0276                              avg =         7.0
    Overall = 0.0476                              max =          7

corr(u_i, Xb) = -0.9107                      F(3, 3567)      =   2273.74
                                              Prob > F        =    0.0000

```

lwage	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
exper	.1137879	.0024689	46.09	0.000	.1089473	.1186284
expsq	-.4243693	.0546316	-7.77	0.000	-.5314816	-.317257
weeks	.0008359	.0005997	1.39	0.163	-.0003399	.0020116
educ	0	(omitted)				
_cons	4.596396	.0389061	118.14	0.000	4.520116	4.672677
sigma_u	1.0362039					
sigma_e	.15220316					
rho	.97888036	(fraction of variance due to u_i)				

```

F test that all u_i=0: F(594, 3567) = 53.12                Prob > F = 0.0000
note: educ omitted because of collinearity.

```

```

Linear regression, absorbing indicators          Number of obs   =    4,165
Absorbed variable: id                        No. of categories =     595
                                              F(3, 3567)      =   2273.74
                                              Prob > F        =    0.0000
                                              R-squared       =    0.9068
                                              Adj R-squared   =    0.8912
                                              Root MSE       =    0.1522

```

lwage	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
exper	.1137879	.0024689	46.09	0.000	.1089473	.1186284
expsq	-.4243693	.0546316	-7.77	0.000	-.5314816	-.317257
weeks	.0008359	.0005997	1.39	0.163	-.0003399	.0020116
educ	0	(omitted)				
_cons	4.596396	.0389061	118.14	0.000	4.520116	4.672677



F test of absorbed indicators: F(594, 3567) = 53.118 Prob > F = 0.000

Source	SS	df	MS	Number of obs	=	3,570
Model	33.3371458	3	11.1123819	F(3, 3567)	=	337.12
Residual	117.57812	3,567	.032962747	Prob > F	=	0.0000
				R-squared	=	0.2209
				Adj R-squared	=	0.2202
Total	150.915266	3,570	.042273184	Root MSE	=	.18156

D.lwage	Coefficient	Std. err.	t	P> t	[95% conf. interval]	
exper						
D1.	.1170654	.0063106	18.55	0.000	.1046927	.1294381
expsq						
D1.	-.5321208	.1392741	-3.82	0.000	-.8051857	-.259056
weeks						
D1.	-.0002683	.0005648	-0.47	0.635	-.0013757	.0008392

7. Using the formula from Handout 5, replicate the value of  $\theta$  reported above by the FGLS estimator. Note, you will need to use the stored values of  $\sigma_\varepsilon^2$  and  $\sigma_\alpha^2$ .

```
qui xtreg lwage exper expsq weeks ed, re theta
display "theta = " 1 - sqrt(e(sigma_e)^2 / (7*e(sigma_u)^2+e(sigma_e)^2))
```

theta = .82280511

8. Make a table of the computed estimates. You can either use `estimates table` or `esttab`. The latter is part of the `estout` package, which you may need to install: `ssc install estout`.

```
esttab OLS BG FGLS, se scalar(N r2 r2_o r2_b r2_w sigma_u sigma_e rho) mtitle("OLS" "BG" "FGLS")
esttab WG LSDV FD, se scalar(N r2 r2_o r2_b r2_w sigma_u sigma_e rho) rename(D.exper exper D
```

(1)

(2)

(3)

	OLS	BG	FGLS
exper	0.0447*** (0.00239)	0.0382*** (0.00570)	0.0889*** (0.00282)
expsq	-0.716*** (0.0528)	-0.631*** (0.126)	-0.773*** (0.0623)
weeks	0.00583*** (0.00118)	0.0131** (0.00407)	0.000966 (0.000743)
educ	0.0760*** (0.00223)	0.0738*** (0.00490)	0.112*** (0.00606)
_cons	4.908*** (0.0673)	4.683*** (0.210)	3.829*** (0.0936)
N	4165	4165	4165
r2	0.284	0.326	
r2_o		0.272	0.183
r2_b		0.326	0.172
r2_w		0.136	0.634
sigma_u			0.320
sigma_e			0.152
rho			0.815

Standard errors in parentheses

\* p<0.05, \*\* p<0.01, \*\*\* p<0.001

	(1) WG	(2) LSDV	(3) FD
exper	0.114*** (0.00247)	0.114*** (0.00247)	0.117*** (0.00631)
expsq	-0.424*** (0.0546)	-0.424*** (0.0546)	-0.532*** (0.139)
weeks	0.000836 (0.000600)	0.000836 (0.000600)	-0.000268 (0.000565)
educ	0	0	

	(.)	(.)	
_cons	4.596*** (0.0389)	4.596*** (0.0389)	
-----			
N	4165	4165	3570
r2	0.657	0.907	0.221
r2_o	0.0476		
r2_b	0.0276		
r2_w	0.657		
sigma_u	1.036		
sigma_e	0.152		
rho	0.979		

-----  
Standard errors in parentheses

\* p<0.05, \*\* p<0.01, \*\*\* p<0.001

9. Perform a Hausman test comparing the results of the FLGS and WG estimators. You should use the `hausman` command, with the option `sigmamore`. Be sure to get the order of the estimates correct. What do you learn from the test?

```
hausman WG FGLS, sigmamore
```

---- Coefficients ----				
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	WG	FGLS	Difference	Std. err.
-----				
exper	.1137879	.0888609	.0249269	.0012778
expsq	-.4243693	-.772565	.3481957	.0284727
weeks	.0008359	.0009658	-.0001299	.0001108

b = Consistent under H0 and Ha; obtained from xtreg.

B = Inconsistent under Ha, efficient under H0; obtained from xtreg.

Test of H0: Difference in coefficients not systematic

$$\begin{aligned}\chi^2(3) &= (b-B)'[(V_b-V_B)^{-1}](b-B) \\ &= 1513.02\end{aligned}$$

Prob >  $\chi^2$  = 0.0000

10. Estimate FGLS for the model below:

$$\ln(Wage_{it}) = \beta_1 + \beta_2 Exper_{it} + \beta_3 Exper_{it}^2 + \beta_4 Weeks_{it} + \beta_5 Eduyrs_{it} \\ + \gamma_2 \overline{Exper}_i + \gamma_3 \overline{Exper}_{it}^2 + \gamma_4 \overline{Weeks}_i + \varepsilon_{it}$$

You will need to manually create the variables:  $\{\overline{Exper}_i, \overline{Exper}_{it}^2, \overline{Weeks}_i\}$  - the individual-level averages of each variable. This is referred to as the Mundlack correction. Once you have estimated the model, repeat the Hausman test comparing these results with those of the WG estimator. What is the significance of the Mundlack correction?

```
foreach var in exper expsq weeks{
    bys id: egen av`var' = mean(`var')
}

eststo MUN: xtreg lwage exper expsq weeks ed avexper avexpsq avweeks, re theta

esttab WG LSDV FD MUN, se scalar(N r2 r2_o r2_b r2_w sigma_u sigma_e rho) rename(D.exper exp

hausman MUN FGLS, sigmamore
```

Random-effects GLS regression	Number of obs	=	4,165
Group variable: id	Number of groups	=	595
R-squared:	Obs per group:		
Within = 0.6566	min =		7
Between = 0.3264	avg =		7.0
Overall = 0.4160	max =		7
	Wald chi2(7)	=	7107.12
corr(u_i, X) = 0 (assumed)	Prob > chi2	=	0.0000
theta = .82280511			

lwage	Coefficient	Std. err.	z	P> z	[95% conf. interval]	
exper	.1137879	.0024689	46.09	0.000	.1089489	.1186268
expsq	-.4243693	.0546316	-7.77	0.000	-.5314452	-.3172934
weeks	.0008359	.0005997	1.39	0.163	-.0003395	.0020112
educ	.0737838	.0048985	15.06	0.000	.0641829	.0833846
avexper	-.0756349	.0062087	-12.18	0.000	-.0878036	-.0634662
avexpsq	-.2069027	.1370415	-1.51	0.131	-.4754991	.0616937
avweeks	.0122544	.0041099	2.98	0.003	.0041991	.0203097

_cons		4.683039	.2100989	22.29	0.000	4.271253	5.094826
-----							
sigma_u		.31951859					
sigma_e		.15220316					
rho		.81505521	(fraction of variance due to u_i)				
-----							

	(1)	(2)	(3)	(4)
	WG	LSDV	FD	Mundlack
-----				
exper	0.114*** (0.00247)	0.114*** (0.00247)	0.117*** (0.00631)	0.114*** (0.00247)
expsq	-0.424*** (0.0546)	-0.424*** (0.0546)	-0.532*** (0.139)	-0.424*** (0.0546)
weeks	0.000836 (0.000600)	0.000836 (0.000600)	-0.000268 (0.000565)	0.000836 (0.000600)
educ	0 (.)	0 (.)		0.0738*** (0.00490)
avexper				-0.0756*** (0.00621)
avexpsq				-0.207 (0.137)
avweeks				0.0123** (0.00411)
_cons	4.596*** (0.0389)	4.596*** (0.0389)		4.683*** (0.210)
-----				
N	4165	4165	3570	4165
r2	0.657	0.907	0.221	
r2_o	0.0476			0.416
r2_b	0.0276			0.326
r2_w	0.657			0.657
sigma_u	1.036			0.320
sigma_e	0.152			0.152
rho	0.979			0.815

-----  
 Standard errors in parentheses  
 \* p<0.05, \*\* p<0.01, \*\*\* p<0.001

Note: the rank of the differenced variance matrix (3) does not equal the number of coefficients being tested (4); be sure this is what you expect, or there may be problems computing the test. Examine the output of your estimators for anything unexpected and possibly consider scaling your variables so that the coefficients are on a similar scale.

	---- Coefficients ----			
	(b)	(B)	(b-B)	sqrt(diag(V_b-V_B))
	MUN	FGLS	Difference	Std. err.
exper	.1137879	.0888609	.0249269	.0012778
expsq	-.4243693	-.772565	.3481957	.0284727
weeks	.0008359	.0009658	-.0001299	.0001108
educ	.0737838	.1117099	-.0379262	.0009972

b = Consistent under H0 and Ha; obtained from xtreg.  
 B = Inconsistent under Ha, efficient under H0; obtained from xtreg.

Test of H0: Difference in coefficients not systematic

chi2(3) = (b-B)'[(V\_b-V\_B)<sup>-1</sup>](b-B)  
 = 1513.02  
 Prob > chi2 = 0.0000  
 (V\_b-V\_B is not positive definite)

11. Export the results as a single CSV/Excel file. You can use `esttab` for .csv or `outreg2` for .xlsx.

```
esttab using "problem-set-5-results.csv", replace se scalar(N r2 r2_o r2_b r2_w sigma_u sigma_v)
```

(output written to problem-set-5-results.csv)

## Postamble

```
log close
```

```
name: <unnamed>
log: C:\Users\neil_\OneDrive - University of Warwick\Documents\EC910\we
> bsite\warwick-ec910\problem-sets\ps-5\problem-set-5-log.txt
log type: smcl
closed on: 11 Nov 2024, 14:07:10
-----
```