

Problem Set 5
ANSWERS

1, 2 and 3 Suppose that Vermont has passed a law requiring employers to provide 6 months of paid maternity leave. You are concerned that women's wages will drop in order to pay for this new benefit. You find a data set that samples men and women in Vermont and in New Hampshire and has information on wages. You pool 2 cross-sections, one from the year before the law took effect and one from the year after and find that the mean wage for various groups is as follows:

	New Hampshire		Vermont	
	Before	After	Before	After
Women	\$9	\$12	\$8	\$10
Men	\$12	\$14	\$10	\$12

1. Suppose you estimate the following model using only data from Vermont:

$$wage = \beta_0 + \beta_1 after + \beta_2 women + \beta_3 after * women + u,$$

where *after* and *women* are dummy variables for the second period and being a woman respectively.

Your estimate of β_3 will be:

b) 0

This is just the difference-in-differences estimator. Thus β_3 will be 0, as seen below:

	Vermont		Differences
	Before	After	(After-Before)
Women	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_1 + \beta_3$
Men	β_0	$\beta_0 + \beta_1$	β_1
Difference (W-M)	β_2	$\beta_2 + \beta_3$	β_3

2. Suppose instead you estimate the following model on all of the data:

$$wage = \beta_0 + \beta_1 after + \beta_2 women + \beta_3 Vermont + \beta_4 after * women + \beta_5 after * Vermont + \beta_6 Vermont * women + \beta_7 after * women * Vermont + u,$$

where *after* and *women* are as before and *Vermont* is a dummy variable for Vermont. Your estimate of β_7 will be:

c) -1

Now we would be doing a difference-in-difference-in-differences! Thus, β_7 will be -1 as seen below:

	New Hampshire		Vermont	
	Before	After	Before	After
Women	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta_4$	$\beta_0 + \beta_2 + \beta_3 + \beta_6$	$\beta_0 + \beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5 + \beta_6 + \beta_7$
Men	β_0	$\beta_0 + \beta_1$	$\beta_0 + \beta_3$	$\beta_0 + \beta_1 + \beta_3 + \beta_5$
Difference	β_2	$\beta_2 + \beta_4$	$\beta_2 + \beta_6$	$\beta_2 + \beta_4 + \beta_6 + \beta_7$
Diff-in-Differences		β_4		$\beta_4 + \beta_7$
Diff-in-Diff-in-Diffs				β_7

3. Given the results of both models, the most reasonable conclusion is that

a) there was a small adverse effect of the law on women's wages

Obviously we would really want to run the regression and get standard errors, but our initial reaction to the diff-in-diffs – that the law had no effect, may be wrong. In that case, we found no difference in wage growth between men and women. However, in NH, without the law, women were actually gaining ground on men, implying that perhaps the VT law did have a negative effect on female wages.

4. and 5. Suppose a researcher is interested in whether having a lot of college students in a city affects the price of rental housing. Suppose that the true population model is $lrent_{it} = \beta_0 + \beta_1 lpop_{it} + \beta_2 lavginc_{it} + \beta_3 pctstu_{it} + \beta_4 y90_{it} + a_i + u_{it}$, where $lrent$ is the log of the rental price, $lpop$ is the log of the city's population, $lavginc$ is the log of per capita income, $pctstu$ is the student population as a percent of the city population (during the school year) and $y90=1$ if the year is 1990. The researcher uses the fixed effect estimator to obtain the following Stata output:

Regression with robust standard errors				Number of obs = 128			
				F(4, 60) = 691.38			
				Prob > F = 0.0000			
				R-squared = 0.9827			
				Adj R-squared = 0.9633			
				Root MSE = .06373			

	lrent	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
	lpop	.0722458	.0696803	1.04	0.304	-.0671357	.2116272
	lavginc	.3099605	.0893101	3.47	0.001	.1313138	.4886072
	pctstu	.0112033	.002936	3.82	0.000	.0053305	.0170761
	y90	.3855214	.0487188	7.91	0.000	.2880693	.4829735
	_cons	1.409384	1.162338	1.21	0.230	-.9156381	3.734405

city	absorbed	(64 categories)
------	----------	-----------------

4. Based on this, we can conclude that

d) all of the above

Using a fixed effect estimator removes the unobserved effect. If the researcher didn't think it was correlated with the other x's, it would only cause serial correlation, which could be dealt with by correcting the standard errors. Since robust standard errors are calculated, there is also a concern about heteroskedasticity. Thus, these estimates deal with all of the potential problems listed.

5. Suppose another researcher had the same data and regressed $\Delta lrent_i$ on $\Delta lpop_i$, $\Delta lavginc_i$ and $\Delta pctstu_{it}$. We can say for sure that

b) the estimated constant term would be .3855214

There are exactly 2 periods for every city, meaning that first differences and fixed effects estimation will be identical. The only difference is that first differences does not estimate the original intercept. Rather, the intercept is now the coefficient on the year dummy, $y90$.

Part II. Stata Problems.

```
1. use mathpnl
```

```
. desc
```

```
Contains data from mathpnl.dta
```

```
obs:      3,850
vars:      52          25 May 2002 08:02
size:     662,200 (99.6% of memory free)
```

variable name	storage type	display format	value label	variable label
distid	float	%9.0g		district identifier
intid	byte	%9.0g		intermediate school district

```

lunch          float   %9.0g          % eligible for free lunch
enrol          float   %9.0g          school enrollment
ptr            float   %9.0g          pupil/teacher: 1995-98
found          int     %9.0g          foundation grant, $: 1995-98
expp           int     %9.0g          expenditure per pupil
revpp          int     %9.0g          revenue per pupil
avgsal         float   %9.0g          average teacher salary
drop           float   %9.0g          high school dropout rate, %
grad           float   %9.0g          high school grad. rate, %
math4          float   %9.0g          % satisfactory, 4th grade math
math7          float   %9.0g          % satisfactory, 7th grade math
choice         int     %9.0g          number choice students
psa            int     %9.0g          # public school academy studs.
year           int     %9.0g          1992-1998
staff          float   %9.0g          staff per 1000 students
avgben         int     %9.0g          avg teacher fringe benefits
y92            byte    %9.0g          =1 if year == 1992
y93            byte    %9.0g          =1 if year == 1993
y94            byte    %9.0g          =1 if year == 1994
y95            byte    %9.0g          =1 if year == 1995
y96            byte    %9.0g          =1 if year == 1996
y97            byte    %9.0g          =1 if year == 1997
y98            byte    %9.0g          =1 if year == 1998
lexpp          float   %9.0g          log(expp)
lfound         float   %9.0g          log(found)
lexpp_1        float   %9.0g          lexpp[_n-1]
lfnd_1         float   %9.0g          lfnd[_n-1]
lenrol         float   %9.0g          log(enrol)
lenrolsq       float   %9.0g          lenrol^2
lunchsq        float   %9.0g          lunch^2
lfndsq         float   %9.0g          lfnd^2
math4_1        float   %9.0g          math4[_n-1]
cmath4         float   %9.0g          math4 - math4_1
gexpp          float   %9.0g          lexpp - lexpp_1
gexpp_1        float   %9.0g          gexpp[_n-1]
gfound         float   %9.0g          lfound - lfnd_1
gfnd_1         float   %9.0g          gfound[_n-1]
clunch         float   %9.0g          lunch - lunch[_n-1]
clnchsq        float   %9.0g          lunchsq - lunchsq[_n-1]
genrol         float   %9.0g          lenrol - lenrol[_n-1]
genrolsq       float   %9.0g          genrol^2
expp92         int     %9.0g          expp in 1992
lexpp92        float   %9.0g          log(expp92)
math4_92       float   %9.0g          math4 in 1992
cpi            float   %9.0g          consumer price index
rexpp          float   %9.0g          real spending per pupil, 1997$
lrexpp         float   %9.0g          log(rexpp)
lrexpp_1       float   %9.0g          lrexpp[_n-1]
grexpp         float   %9.0g          lrexpp - lrexpp_1
grexpp_1       float   %9.0g          grexpp[_n-1]

```

Sorted by: distid year

. sum

Variable	Obs	Mean	Std. Dev.	Min	Max
distid	3850	46186.15	24159.92	1010	83070
intid	3850	45.52	24.02906	3	83

lunch		3850	27.20165	15.40591	0	91.27
enrol		3850	3043.734	8153.137	26	183151
ptr		2200	21.45305	2.895641	0	33.6
found		2200	5398.325	1190.792	0	10916
expp		3850	5237.257	1224.984	946	13982
revpp		2200	6388.37	1408.232	0	13049
avgsal		3850	38375.14	9044.325	0	76412
drop		3766	4.174198	5.592694	-45.59	100
grad		3639	86.25916	44.91319	1.8	2570.4
math4		3850	55.42753	18.19832	5.9	100
math7		3826	47.85781	17.47808	0	100
choice		1100	16.88364	66.43066	0	1589
psa		1259	29.07228	192.6558	0	5106
year		3850	1995	2.00026	1992	1998
staff		1650	102.1086	22.71165	0	384.1
avgben		1645	7124.01	2328.494	0	18754
y92		3850	.1428571	.3499726	0	1
y93		3850	.1428571	.3499726	0	1
y94		3850	.1428571	.3499726	0	1
y95		3850	.1428571	.3499726	0	1
y96		3850	.1428571	.3499726	0	1
y97		3850	.1428571	.3499726	0	1
y98		3850	.1428571	.3499726	0	1
lexpp		3850	8.536588	.2341271	6.852242	9.545527
lfound		2159	8.600313	.1515919	8.34284	9.297985
lexpp_1		3300	8.5073	.2326065	6.852242	9.266532
lfnd_1		1621	8.580548	.1540653	8.34284	9.283776
lenrol		3850	7.421071	1.066945	3.258096	12.11807
lenrolsq		3850	56.21037	15.59235	10.61519	146.8475
lunchsq		3850	977.2099	1047.249	0	8330.212
lfndsq		2159	73.98836	2.639172	69.60298	86.45253
math4_1		3300	52.22212	16.89562	5.9	100
cmath4		3300	6.255242	13.35616	-75	86
gexpp		3300	.0659901	.1011984	-.7317338	1.196007
gexpp_1		2750	.0743619	.1039131	-.7317338	1.196007
gfound		1602	.04035	.016493	.0142088	.1641006
gfnd_1		1068	.0416743	.0169462	.0145073	.1641006
clunch		3300	.6903151	3.792335	-39.13	44.7
clnchsq		3300	46.53831	278.3464	-2760.12	3611.532
genrol		3300	-.0084638	.2496736	-2.614188	2.614651
genrolsq		3300	-.1077835	3.493964	-39.04708	39.05521
expp92		3300	4181.165	932.9284	946	9041
lexpp92		3300	8.316376	.2064792	6.852242	9.109525
math4_92		3300	37.12855	13.17844	5.9	100
cpi		3850	1.522571	.077887	1.403	1.63
rexpp		3850	5495.538	1150.775	1082.202	13767.55
lrexpp		3850	8.590627	.2057544	6.986753	9.530069
lrexpp_1		3300	8.572922	.2082203	6.986753	9.33514
grexpp		3300	.0409956	.1004653	-.7570171	1.168062
grexpp_1		2750	.0474597	.1038768	-.7570171	1.168062

a) This implies the following regression:

```
. reg math4 lrexpp lenrol lunch y93-y98, cluster(distid)
```

Regression with robust standard errors

Number of obs = 3850
F(9, 549) = 664.53
Prob > F = 0.0000
R-squared = 0.5672

Number of clusters (distid) = 550 Root MSE = 11.986

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
math4						
lrexpp	8.42089	2.076137	4.06	0.000	4.342745	12.49903
lenrol	.4763886	.4115497	1.16	0.248	-.3320162	1.284793
clunch	-.4141128	.0271626	-15.25	0.000	-.4674682	-.3607574
y93	6.470661	.4662991	13.88	0.000	5.554712	7.386609
y94	12.4951	.542871	23.02	0.000	11.42874	13.56146
y95	24.38428	.6979197	34.94	0.000	23.01336	25.7552
y96	24.689	.756721	32.63	0.000	23.20258	26.17543
y97	21.98691	.8091444	27.17	0.000	20.39752	23.57631
y98	37.19954	.7943905	46.83	0.000	35.63912	38.75996
_cons	-27.35903	17.06238	-1.60	0.109	-60.87456	6.156506

Based on this regression, it appears that spending more does result in higher passing rates on the 4th grade math test. Specifically, a 10% increase in spending would increase the passing rate by .84 percentage points. (Recall that the coefficient on a logged x variable implies the effect of a 100% change, so a 10% change is the coefficient/10).

b) This now implies the following regression. Note that with changes, we have one less year:

```
. reg cmath4 grexpp genrol clunch y94-y98, cluster(distid)
```

Regression with robust standard errors Number of obs = 3300
F(8, 549) = 98.85
Prob > F = 0.0000
R-squared = 0.2080
Number of clusters (distid) = 550 Root MSE = 11.901

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
cmath4						
grexpp	-3.447268	4.869028	-0.71	0.479	-13.01147	6.116937
genrol	.6345335	1.36618	0.46	0.643	-2.049047	3.318114
clunch	.025074	.1491805	0.17	0.867	-.2679604	.3181084
y94	.5210521	.8969629	0.58	0.562	-1.240847	2.282951
y95	6.812446	.8707643	7.82	0.000	5.102009	8.522884
y96	-5.23489	.7746066	-6.76	0.000	-6.756446	-3.713335
y97	-8.488463	.7102396	-11.95	0.000	-9.883582	-7.093343
y98	8.967841	.7517733	11.93	0.000	7.491136	10.44454
_cons	5.954963	.5374064	11.08	0.000	4.899338	7.010587

Now, the effect of spending is insignificant, and the point estimate implies a .34 percentage point decline in the passing rate for a 10% increase. It appears that there were unobserved effects in the original model that were correlated with spending, leading to a positive bias in the OLS estimates.

c) This now implies the following regression. Note that with a lag, we have even one less year:

```
. reg cmath4 grexpp grexpp_1 genrol clunch y95-y98, cluster(distid)
```

Regression with robust standard errors Number of obs = 2750
F(8, 549) = 95.27
Prob > F = 0.0000
R-squared = 0.2376

Number of clusters (distid) = 550 Root MSE = 12.087

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
cmath4						
grexpp	-1.410699	4.944102	-0.29	0.775	-11.12237	8.300972
grexpp_1	11.04026	5.131503	2.15	0.032	.9604803	21.12004
genrol	2.140017	1.64512	1.30	0.194	-1.091482	5.371517
clunch	.0728056	.1654564	0.44	0.660	-.2521995	.3978107
y95	5.704738	.9093617	6.27	0.000	3.918484	7.490992
y96	-6.795939	.8745341	-7.77	0.000	-8.513781	-5.078097
y97	-8.989378	.7718688	-11.65	0.000	-10.50556	-7.473201
y98	8.453018	.7838456	10.78	0.000	6.913315	9.992721
_cons	6.158613	.6572949	9.37	0.000	4.867492	7.449734

Now, it looks like lagged spending has a big positive effect on the passing rate. A 10% increase in spending will increase next year's passing rate by a full percentage point. This makes sense if much of the preparation for the 4th grade exam occurs during 3rd grade, since the exam is early in the year.

d) This now implies the following regression. Note that with lag, we have one less year than before:

```
. reg math4 lrexpp lrexpp_1 lenrol lunch y94-y98, cluster(distid)
```

Regression with robust standard errors Number of obs = 3300
F(9, 549) = 467.35
Prob > F = 0.0000
R-squared = 0.5053
Root MSE = 12.042

Number of clusters (distid) = 550

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
lrexpp	.5339314	2.512543	0.21	0.832	-4.401443	5.469305
lrexpp_1	9.049175	2.7953	3.24	0.001	3.558383	14.53997
lenrol	.5926719	.4112799	1.44	0.150	-.2152029	1.400547
lunch	-.4067083	.0281132	-14.47	0.000	-.4619309	-.3514858
y94	6.377355	.5290088	12.06	0.000	5.338226	7.416484
y95	18.6502	.6065382	30.75	0.000	17.45878	19.84162
y96	18.03336	.7198465	25.05	0.000	16.61937	19.44735
y97	15.34006	.7567352	20.27	0.000	13.85361	16.82651
y98	30.39788	.7801259	38.97	0.000	28.86549	31.93028
_cons	-31.66156	18.43073	-1.72	0.086	-67.86494	4.541829

And the following fixed-effect regression:

```
. areg math4 lrexpp lrexpp_1 lenrol lunch y94-y98, absorb(distid) robust
```

Regression with robust standard errors Number of obs = 3300
F(9, 2741) = 459.61
Prob > F = 0.0000
R-squared = 0.7700
Adj R-squared = 0.7231
Root MSE = 8.9962

	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
math4						

lrexpp		-.4111804	3.873469	-0.11	0.915	-8.006393	7.184032
lrexpp_1		7.002988	3.480938	2.01	0.044	.1774612	13.82851
lenrol		.2450874	1.205097	0.20	0.839	-2.117903	2.608078
lunch		.061527	.0967727	0.64	0.525	-.1282279	.2512818
y94		6.177316	.6102614	10.12	0.000	4.980697	7.373934
y95		18.09267	.8616857	21.00	0.000	16.40305	19.78229
y96		17.9404	1.013277	17.71	0.000	15.95353	19.92726
y97		15.19184	1.092992	13.90	0.000	13.04867	17.33501
y98		29.88319	1.195098	25.00	0.000	27.53981	32.22657
_cons		-16.08091	35.73831	-0.45	0.653	-86.15765	53.99582

 distid | absorbed (550 categories)

With the lagged spending included, there is not as big a difference in the implications between the OLS and fixed-effect estimates as there was before, since both imply significant positive effects. However, the OLS estimates still appear positively biased – implying a .9 increase instead of a .7 increase for a 10% increase in spending.

e) I would advise the school district that it looks like increased spending can have a small affect on scores, but that they shouldn't expect immediate results.

```
2. . use murder, clear
```

```
. desc
```

```
Contains data from murder.dta
```

```
obs:      153
vars:      13          13 Sep 2000 15:34
size:      5,049 (100.0% of memory free)
```

```
-----
variable name  storage  display  value  variable label
               type   format   label
-----
id             byte    %9.0g
state          str2    %9s
year          byte    %9.0g
mrdrte        float   %9.0g
exec          byte    %9.0g
unem          float   %9.0g
d90           byte    %9.0g
d93           byte    %9.0g
cmrdrte       float   %9.0g
cexec         byte    %9.0g
cunem         float   %9.0g
cexec_1       byte    %9.0g
cunem_1       float   %9.0g
-----
state identifier
postal code
87, 90, or 93
murders per 100,000 population
total executions, past 3 years
annual unem. rate
=1 if year == 90
=1 if year == 93
mrdrte - mrdrte[t-1]
exec - exec[t-1]
unem - unem[t-1]
cexec[t-1]
cunem[t-1]
```

```
Sorted by:
```

```
. sum
```

```
-----
Variable | Obs      Mean      Std. Dev.  Min      Max
-----
id       | 153      26       14.76794   1        51
state    | 0
year     | 153      90       2.457534   87       93
mrdrte   | 153      8.070588 9.192867   .8       78.5
exec     | 153      1.228758 3.791432   0        34
```

unem		153	5.973203	1.680617	2.2	12
d90		153	.3333333	.4729527	0	1
d93		153	.3333333	.4729527	0	1
cmrdрте		102	.8421568	4.290271	-2.6	41.6
cexec		102	.1862745	2.950853	-11	23
cunem		102	.0058823	1.658272	-5.8	3.6
cexec_1		51	-.2745098	2.191606	-11	5
cunem_1		51	-.8862745	1.7339	-5.8	3.1

a) To do a Hausman test, we need to estimate the random effects model and use xthaus afterward:

```
. xtreg mrdрте unem exec, i(id) re
```

```
Random-effects GLS regression           Number of obs   =       153
Group variable (i) : id                 Number of groups =        51

R-sq:  within = 0.0015                   Obs per group:  min =         3
        between = 0.0732                               avg =        3.0
        overall = 0.0433                               max =         3

Random effects u_i ~ Gaussian           Wald chi2(2)    =         0.90
corr(u_i, X) = 0 (assumed)              Prob > chi2     =        0.6369
```

	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
mrdрте						
unem		.2560543	.2708762	0.95	0.345	-.2748532 .7869619
exec		-.0351956	.1619968	-0.22	0.828	-.3527036 .2823124
_cons		6.584371	2.001338	3.29	0.001	2.661819 10.50692
sigma_u		8.1923983				
sigma_e		3.612922				
rho		.83717807	(fraction of variance due to u_i)			

```
. xthaus
```

```
Hausman specification test
```

---- Coefficients ----				
	Fixed	Random		
mrdрте	Effects	Effects	Difference	
unem		.095914	.2560543	-.1601403
exec		-.1140743	-.0351956	-.0788787

```
Test: Ho: difference in coefficients not systematic
```

```
chi2( 2) = (b-B)'[S^(-1)](b-B), S = (S_fe - S_re)
        = 6.79
Prob>chi2 = 0.0336
```

The p-value implies we can reject the null at the .034 level. The null in this case is that any unobserved effect is uncorrelated with *unem* and *exec*. Technically, the test is for whether the difference in coefficients is systematic, but this is the same thing. Why? Because the idea behind a Hausman test is that under the null (in this case the unobserved effect is uncorrelated with x's) 2 different estimators are both consistent (in this case random effects and fixed effects). If they are both

consistent, we would expect only random variation in the estimates. If there is a systematic difference, we must reject the null and conclude that the unobserved effect is correlated with the x's.

b) A first difference model is another way of removing the correlated unobserved effect:

```
. reg cmrdrtc cunem cexec, robust
```

```
Regression with robust standard errors
```

	Number of obs =	102
	F(2, 99) =	8.93
	Prob > F =	0.0003
	R-squared =	0.0090
	Root MSE =	4.3139

		Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
cmrdrtc							
cunem		-.0447929	.1346571	-0.33	0.740	-.3119818	.2223961
cexec		-.1313545	.0311427	-4.22	0.000	-.1931484	-.0695606
_cons		.8668883	.429753	2.02	0.046	.014165	1.719611

To do a serial correlation test, we need to get the residual and its lag. We need to be careful in computing the lag to make sure the data is sorted and to use by id: to avoid using the previous state's residual.

```
. predict uhat, resid
(51 missing values generated)
```

```
. sort id year
```

```
. by id: gen uhat_1=uhat[_n-1]
(102 missing values generated)
```

```
. reg uhat uhat_1, robust
```

```
Regression with robust standard errors
```

	Number of obs =	51
	F(1, 49) =	1.19
	Prob > F =	0.2798
	R-squared =	0.0028
	Root MSE =	1.071

		Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
uhat							
uhat_1		.0094241	.0086225	1.09	0.280	-.0079035	.0267518
_cons		-.4596857	.1506809	-3.05	0.004	-.7624903	-.1568811

Since the coefficient on the lagged residual is not significantly different from zero, we cannot reject the null of no serial correlation. Note that I did a heteroskedasticity robust test, but the result is the same without that correction.

e) While the random effects model showed no significant deterrent effect, the first difference model did. Since the Hausman test implied that unobserved effects were a problem, the first difference model should be preferred. Thus, it appears that every execution in a state reduces its murder rate by .13 percentage points. This is very small, but is a significant effect.