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:

<table>
<thead>
<tr>
<th></th>
<th>New Hampshire</th>
<th>Vermont</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
<td>Before</td>
</tr>
<tr>
<td>Women</td>
<td>$9</td>
<td>$12</td>
<td>$8</td>
</tr>
<tr>
<td>Men</td>
<td>$12</td>
<td>$14</td>
<td>$10</td>
</tr>
</tbody>
</table>

1. Suppose you estimate the following model using only data from Vermont:
   \[ \text{wage} = \beta_0 + \beta_{\text{after}} + \beta_{\text{women}} + \beta_{\text{after}*\text{women}} + u, \]
   where \( \text{after} \) and \( \text{women} \) are dummy variables for the second period and being a woman respectively. Your estimate of \( \beta_3 \) will be:
   b) 0
   This is just the difference-in-differences estimator. Thus \( \beta_3 \) will be 0, as seen below:

<table>
<thead>
<tr>
<th></th>
<th>Vermont</th>
<th>Differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
</tr>
<tr>
<td>Women</td>
<td>( \beta_0 + \beta_2 )</td>
<td>( \beta_0 + \beta_1 + \beta_2 + \beta_3 )</td>
</tr>
<tr>
<td>Men</td>
<td>( \beta_0 )</td>
<td>( \beta_0 + \beta_1 )</td>
</tr>
<tr>
<td>Difference (W-M)</td>
<td>( \beta_2 )</td>
<td>( \beta_2 + \beta_3 )</td>
</tr>
</tbody>
</table>

2. Suppose instead you estimate the following model on all of the data:
   \[ \text{wage} = \beta_0 + \beta_{\text{after}} + \beta_{\text{women}} + \beta_{\text{Vermont}} + \beta_{\text{after}*\text{women}} + \beta_{\text{after}*\text{Vermont}} + \beta_{\text{Vermont}*\text{women}} + \beta_{\text{after}*\text{women}*\text{Vermont}} + u, \]
   where \( \text{after} \) and \( \text{women} \) are as before and \( \text{Vermont} \) is a dummy variable for Vermont. Your estimate of \( \beta_7 \) will be:
   c) -1
   Now we would be doing a difference-in-difference-in-differences! Thus, \( \beta_7 \) will be –1 as seen below:

<table>
<thead>
<tr>
<th></th>
<th>New Hampshire</th>
<th>Vermont</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Before</td>
<td>After</td>
</tr>
<tr>
<td>Women</td>
<td>( \beta_0 + \beta_2 )</td>
<td>( \beta_0 + \beta_1 + \beta_2 + \beta_4 )</td>
</tr>
<tr>
<td>Men</td>
<td>( \beta_0 )</td>
<td>( \beta_0 + \beta_1 )</td>
</tr>
<tr>
<td>Difference</td>
<td>( \beta_2 )</td>
<td>( \beta_2 + \beta_4 )</td>
</tr>
<tr>
<td>Diff-in-Differences</td>
<td>( \beta_4 )</td>
<td></td>
</tr>
<tr>
<td>Diff-in-Diff-in-Diffs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 \( lrent_{it} = \beta_0 + \beta_1 lpop_{it} + \beta_2 lavginc_{it} + \beta_3 pctstu_{it} + \beta_4 y90_t + a_i + u_{it} \), 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

|               | Coef.   | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|---------------|---------|-----------|-------|-------|----------------------|
| lrent         |         |           |       |       |                      |
| lpop          | .0722458| .0696803  | 1.04  | 0.304 | -.0671357 to .2116272|
| lavginc       | .3099605| .0893101  | 3.47  | 0.001 | .1313138 to .4886072 |
| pctstu        | .0112033| .002936   | 3.82  | 0.000 | .0053305 to .0170761 |
| y90           | .3855214| .0487188  | 7.91  | 0.000 | .2880693 to .4829735 |
| _cons         | 1.409384| 1.162338  | 1.21  | 0.230 | -.915638 to 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
   size: 662,200 (99.6% of memory free)

<table>
<thead>
<tr>
<th>variable name</th>
<th>type</th>
<th>format</th>
<th>display</th>
<th>value</th>
<th>variable label</th>
</tr>
</thead>
<tbody>
<tr>
<td>distid</td>
<td>float</td>
<td>%9.0g</td>
<td>district identifier</td>
<td></td>
<td></td>
</tr>
<tr>
<td>intid</td>
<td>byte</td>
<td>%9.0g</td>
<td>intermediate school district</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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, $: 1995-98
expp int %9.0g expenditure per pupil
revpp int %9.0g revenue per pupil
avgsal float %9.0g average teacher salary
drop float %9.0g high school dropout rate, %
grad float %9.0g high school grad. rate, %
math4 float %9.0g % satisfactory, 4th grade math
math7 float %9.0g % satisfactory, 7th grade math
choice int %9.0g number choice students
psa int %9.0g # public school academy studs.
year int %9.0g 1992-1998
staff float %9.0g staff per 1000 students
avgben int %9.0g avg teacher fringe benefits
y92 byte %9.0g =1 if year == 1992
y93 byte %9.0g =1 if year == 1993
y94 byte %9.0g =1 if year == 1994
y95 byte %9.0g =1 if year == 1995
y96 byte %9.0g =1 if year == 1996
y97 byte %9.0g =1 if year == 1997
y98 byte %9.0g =1 if year == 1998
lexpp float %9.0g log(expp)
lfound float %9.0g log(found)
lexpp_1 float %9.0g lexpp[_n-1]
lfnd_1 float %9.0g lfnd[_n-1]
lenrol float %9.0g log(enrol)
lenrolsq float %9.0g lenrol^2
lunchsq float %9.0g lunch^2
lfndsq float %9.0g lfnd^2
math4_1 float %9.0g math4[_n-1]
cmath4 float %9.0g math4 - math4_1
gexpp float %9.0g lexpp - lexpp_1
gexpp_1 float %9.0g gexpp[_n-1]
gfound float %9.0g lfound - lfnd_1
gfnd_1 float %9.0g gfnd[_n-1]
clunch float %9.0g lunch - lunch[_n-1]
clunchsq float %9.0g lunchsq - lunchsq[_n-1]
genrol float %9.0g genrol - genrol[_n-1]
genrolsq float %9.0g genrol^2
expp92 int %9.0g exp in 1992
lexpp92 float %9.0g log(expp92)
math4_92 float %9.0g math4 in 1992
cpi float %9.0g consumer price index
rexpp float %9.0g real spending per pupil, 1997$
1rexpp float %9.0g log(rexpp)
1rexpp_1 float %9.0g 1rexpp[_n-1]
grexpp float %9.0g grexpp - grexpp_1
grexpp_1 float %9.0g grexpp[_n-1]

Sorted by: distid year

. sum

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>distid</td>
<td>3850</td>
<td>46186.15</td>
<td>24159.92</td>
<td>1010</td>
<td>83070</td>
</tr>
<tr>
<td>intid</td>
<td>3850</td>
<td>45.52</td>
<td>24.02906</td>
<td>3</td>
<td>83</td>
</tr>
</tbody>
</table>
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
```
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:

```
.b. 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
------------------------------------------------------------------------------
|               Robust
|       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
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   .7571733     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:

```
c. 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      =  11.901
------------------------------------------------------------------------------
|               Robust
|       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
grexpp |  -8.488463   .7102396   -11.95   0.000    -9.883582   -7.093343
genrol |   .6345335    1.36618     0.46   0.643    -2.049047    3.318114
clunch |   .025074   .1491805     0.17   0.867    -.2679604    .3181084
y95 |   8.967841   .7571733     11.93   0.000     7.491136    10.44454
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   .7571733     11.93   0.000     7.491136    10.44454
_cons |   5.954963   .5374064    11.08   0.000     4.899338    7.010587
------------------------------------------------------------------------------
```
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

|                  | Coef. | 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           | -.4067083 | .0281132 | -14.47 | 0.000 | [-3514858  1984162] |
| lunch            | 6.377355 | .5290088 | 12.06 | 0.000 | [5.338226  7.461648] |
| y94              | 18.6502  | .6065382 | 30.75 | 0.000 | [17.45878  19.84162] |
| y95              | 18.03336 | .7198465 | 25.05 | 0.000 | [16.61937  19.44735] |
| y96              | 15.34006 | .7567352 | 20.27 | 0.000 | [13.85361 16.82651]  |
| y97              | 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

|                  | Coef. | 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           | -.4067083 | .0281132 | -14.47 | 0.000 | [-3514858  1984162] |
| lunch            | 6.377355 | .5290088 | 12.06 | 0.000 | [5.338226  7.461648] |
| y94              | 18.6502  | .6065382 | 30.75 | 0.000 | [17.45878  19.84162] |
| y95              | 18.03336 | .7198465 | 25.05 | 0.000 | [16.61937  19.44735] |
| y96              | 15.34006 | .7567352 | 20.27 | 0.000 | [13.85361 16.82651]  |
| y97              | 30.39788 | .7801259 | 38.97 | 0.000 | [28.86549 31.93028]  |
| _cons            | -31.66156 | 18.43073 | -1.72 | 0.086 | [-67.86494  4.541829] |
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)

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

Sorted by:
.

sum

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>id</td>
<td>153</td>
<td>26</td>
<td>14.76794</td>
<td>1</td>
<td>51</td>
</tr>
<tr>
<td>state</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>year</td>
<td>153</td>
<td>90</td>
<td>2.457534</td>
<td>87</td>
<td>93</td>
</tr>
<tr>
<td>mrdrte</td>
<td>153</td>
<td>8.070588</td>
<td>9.192867</td>
<td>.8</td>
<td>78.5</td>
</tr>
<tr>
<td>exec</td>
<td>153</td>
<td>1.228758</td>
<td>3.791432</td>
<td>0</td>
<td>34</td>
</tr>
</tbody>
</table>
unem | 153  5.973203   1.680617    2.2     12
d90 | 153  .333333   .4729527     0       1
d93 | 153  .333333   .4729527     0       1
cmrdt | 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 xhaus afterward:

```
. xtreg mrdrte 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
```

```
-------------+----------------------------------------------------------------
 mrdrte |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    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)
-------------+----------------------------------------------------------------
```

```.
exhaus
```

Hausman specification test

```
----- Coefficients -----
          | Fixed     Random
-------------+----------------
 mdrdrte | Effects  Effects
-------------+----------------
    unem |  .095914  .2560543  -.1601403
    exec | -1.140743 -.0351956  -.0788787

Test: Ho: difference in coefficients not systematic

ch2(  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
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 cmrdrte 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.    | Std. Err. | t     | P>|t|     | [95% Conf. Interval] |
|----------|----------|-----------|-------|---------|--------------------|
| cunem    | -.044793 | .134657   | -.33  | .740    | -.3119818 to .2223961 |
| cexec    | -.131354 | .031143   | -4.22 | .000    | -.1931484 to -.0695606 |
| _cons    | .866888  | .429753   | 2.02  | .046    | .014165 to 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.    | Std. Err. | t     | P>|t|     | [95% Conf. Interval] |
|----------|----------|-----------|-------|---------|--------------------|
| uhat_1   | .009424  | .008623   | 1.09  | .280    | -.0079035 to .0267518 |
| _cons    | -.459686 | .150681   | -3.05 | .004    | -.7624903 to -.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.