Tutorials
Week 4

Pdf file on Blackboard	Dataset on Blackboard	Papers	Description
C 13.5	rental.dta		Pooled OLS, first different estimator, heteroskedasticity-robust standard errors, the consequences of omitted variables.
C 13.11	mathpnl.dta	Papke,Leslie (2005): The Effects of Spending on Test Pass Rates: Evidence from Michigan" (2005), Journal of Public Economics 89, 821-839.	Panel data analysis, first different estimator, heteroskedasticity-robust standard errors, and the consequences of omitted variables. (very big exercise - 45mins)
C 13.13	wagepan.dta	F. Vella and M. Verbeek (1998), "Whose Wages Do Unions Raise? A Dynamic Model of Unionism and Wage Rate Determination for Young Men," Journal of Applied Econometrics 13, 163-183	first differencing for estimate parameters on time-varying variables. Test hypothesis on fully robust specification, adding interaction terms.

C.13.5 Use the data in RENTAL.RAW for this exercise. The data for the years 1980 and 1990 include rental prices and other variables for college towns. The idea is to see whether a stronger presence of students affects rental rates. The unobserved effects model is

$$\log(rent_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 \log(pop_{it}) + \beta_2 \log(avginc_{it}) + \beta_3 pctstu_{it} + a_i + u_{it},$$

where *pop* is city population, *avginc* is average income, and *pctstu* is student population as a percentage of city population (during the school year).

(i) Estimate the equation by pooled OLS and report the results in standard form. What do you make of the estimate on the 1990 dummy variable? What do you get for $\hat{\beta}_{pctstu}$?

(ii) Are the standard errors you report in part (i) valid? Explain.

(iii) Now, difference the equation and estimate by OLS. Compare your estimate of β_{pctstu} with that from part (ii). Does the relative size of the student population appear to affect rental prices?

(iv) Obtain the heteroskedasticity-robust standard errors for the first-differenced equation in part (iii). Does this change your conclusions?

Utrecht University

i) Estimate the equation by pooled OLS and report the results in standard form. What do you make of the estimate on the 1990 dummy variable? What do you get for $\hat{\beta}_3 pctstu_{it}$?

$$\log(rent_{it}) = \beta_0 + \delta_0 y90_t + \beta_1 \log(pop_{it}) + \beta_2 \log(avginc_{it}) + \beta_3 pctstu_{it} + a_i + u_{it}$$

Note: for panel data the command '*xtset* city year, delta(10)' is usually used to let Stata know the data are panel data, instead of *tsset* which are usually used for time series data (but both are ok to use for panel data).

```
tsset city year, delta(10)
       panel variable:  city (strongly balanced)
        time variable:  year, 80 to 90
                delta:  10 units

. reg lrent y90 lpop lavginc pctstu

      Source |       SS           df       MS      Number of obs   =       128
-------------+----------------------------------   F(4, 123)       =    190.92
       Model |  12.1080112         4  3.02700281   Prob > F        =    0.0000
    Residual |   1.9501234       123  .015854662   R-squared       =    0.8613
-------------+----------------------------------   Adj R-squared   =    0.8568
       Total |  14.0581346       127  .110693974   Root MSE        =    .12592

------------------------------------------------------------------------------
       lrent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         y90 |   .2622267   .0347632     7.54   0.000     .1934151    .3310384
        lpop |   .0406863   .0225154     1.81   0.073    -.0038815    .0852541
     lavginc |   .5714461   .0530981    10.76   0.000     .4663417    .6765504
      pctstu |   .0050436   .0010192     4.95   0.000     .0030262     .007061
       _cons |  -.5688069   .5348808    -1.06   0.290    -1.627571    .4899568
```

- The positive and very significant coefficient on *y90* simply means that, other things in the equation fixed, nominal rents grew by over 29.98% over the 10-year period. (Log-level model)
- The coefficient on *pctstu* means that a one percentage point increase in *pctstu* increases *rent* by half a percent point (.5%). The variable ranges within [0-100], hence a unit increase is a 1 p.p. increase. **Log-Level.**
- The *t* statistic shows that, at least based on the usual analysis, *pctstu* is very statistically significant.

ii) Are the standard errors in i) valid? Explain.

- The standard errors from part (i) are not valid unless we think a_i does not really appear in the equation.
- If a_i is in the error term, the errors across the two time periods for each city are positively correlated, invalidating the usual OLS standard errors and t statistics.

iii) Now, difference the equation and estimated by OLS. Compare your estimate $\widehat{\beta}_3 pctstu_{it}$ with that from part ii). Does the relative size of the student population appear to affect rental prices?

```
. reg d.lrent d.lpop d.lavginc d.pctstu

      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(3, 60)        =      9.51
       Model |  .231738668         3  .077246223   Prob > F        =    0.0000
    Residual |  .487362198        60  .008122703   R-squared       =    0.3223
-------------+----------------------------------   Adj R-squared   =    0.2884
       Total |  .719100867        63  .011414299   Root MSE        =    .09013

------------------------------------------------------------------------------
     D.lrent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpop |
         D1. |   .0722456   .0883426     0.82   0.417     -.104466    .2489571
             |
     lavginc |
         D1. |   .3099605   .0664771     4.66   0.000     .1769865    .4429346
             |
      pctstu |
         D1. |   .0112033   .0041319     2.71   0.009     .0029382    .0194684
             |
       _cons |   .3855214   .0368245    10.47   0.000     .3118615    .4591813
------------------------------------------------------------------------------
```

Interestingly, the effect of *pctstu* is over twice as large as we estimated in the pooled OLS equation. Now, a one percentage point increase in *pctstu* is estimated to increase rental rates by about 1.1%. Not surprisingly, we obtain a much less precise estimate when we difference (although the OLS standard errors from part (i) are likely to be much too small because of the positive serial correlation in the errors within each city). While we have differenced away *ai*, there may be other unobservables that change over time and are correlated with $\Delta pctstu$.

iv) Obtain the heteroskedasticity-robust standard errors for the first-differenced equation in part (iii). Does this change your conclusion?

Additional question: Is it actually necessary to include heteroskedasticity-robust standard errors?
Apply the Breusch-Pagan test for heteroscedasticity.

```
reg d.lrent d.lpop d.lavginc d.pctstu

      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(3, 60)        =      9.51
       Model |  .231738668         3  .077246223   Prob > F        =    0.0000
    Residual |  .487362198        60  .008122703   R-squared       =    0.3223
-------------+----------------------------------   Adj R-squared   =    0.2884
       Total |  .719100867        63  .011414299   Root MSE        =    .09013

------------------------------------------------------------------------------
     D.lrent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpop |
         D1. |   .0722456   .0883426     0.82   0.417     -.104466    .2489571
             |
     lavginc |
         D1. |   .3099605   .0664771     4.66   0.000     .1769865    .4429346
             |
      pctstu |
         D1. |   .0112033   .0041319     2.71   0.009     .0029382    .0194684
             |
       _cons |   .3855214   .0368245    10.47   0.000     .3118615    .4591813
------------------------------------------------------------------------------


. predict uhat, resid
(64 missing values generated)

. gen uhat_sq=uhat*uhat
(64 missing values generated)
```

Utrecht University

```
. reg uhat_sq d.lpop d.lavginc d.pctstu

      Source |       SS           df       MS      Number of obs   =        64
-------------+----------------------------------   F(3, 60)        =      1.29
       Model |  .000435522         3  .000145174   Prob > F        =    0.2855
    Residual |  .006743785        60  .000112396   R-squared       =    0.0607
-------------+----------------------------------   Adj R-squared   =    0.0137
       Total |  .007179307        63  .000113957   Root MSE        =     .0106

------------------------------------------------------------------------------
     uhat_sq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpop |
         D1. |  -.0109167   .0103919    -1.05   0.298    -.0317036    .0098703
             |
     lavginc |
         D1. |  -.0069714   .0078198    -0.89   0.376    -.0226135    .0086706
             |
      pctstu |
         D1. |   .0005589    .000486     1.15   0.255    -.0004133    .0015312
             |
       _cons |   .0118461   .0043317     2.73   0.008     .0031814    .0205109
------------------------------------------------------------------------------
. test d.lpop d.lavginc d.pctstu

 ( 1)  D.lpop = 0
 ( 2)  D.lavginc = 0
 ( 3)  D.pctstu = 0

       F(  3,    60) =    1.29
            Prob > F =    0.2855
```

Ho: beta1=beta2…=beta4=0   (homoskedasticity)
H1: Ho not true (heteroskedasticity)
If the Fvalue > Fcritical Value = reject Ho
1.29 < 2.60 at 5%, we fail to reject Ho
There is no heteroskedasticity.

If we perform a heteroskedasticity test, we find out that there is no evidence of heteroskedasticity (we fail to reject the null hypothesis of homoskedasticity). However, in this exercise, we proceed by using robust standard errors, as per question (iv). We do this by adding the option robust to our regression.

```
. reg d.lrent d.lpop d.lavginc d.pctstu, robust

Linear regression                               Number of obs     =         64
                                                F(3, 60)          =      11.30
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3223
                                                Root MSE          =     .09013

------------------------------------------------------------------------------
             |               Robust
     D.lrent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lpop |
         D1. |   .0722456   .0696796     1.04   0.304    -.0671344    .2116255
             |
     lavginc |
         D1. |   .3099605   .0893099     3.47   0.001     .1313141     .488607
             |
      pctstu |
         D1. |   .0112033    .002936     3.82   0.000     .0053305    .0170762
             |
       _cons |   .3855214   .0487186     7.91   0.000     .2880697    .4829731
------------------------------------------------------------------------------
```

- The heteroskedasticity-robust standard error on Δ*pctstu* is about .0029, which is actually much smaller than the usual OLS standard error.
- This only makes *pctstu* even more significant (robust *t* statistic ≈ 4). Note that serial correlation is no longer an issue because we have no time component in the first-differenced equation.

Utrecht University

C11 The file MATHPNL.RAW contains panel data on school districts in Michigan for the years 1992 through 1998. It is the district-level analogue of the school-level data used by Papke (2005). The response variable of interest in this question is *math4*, the percentage of fourth graders in a district receiving a passing score on a standardized math test. The key explanatory variable is *rexpp*, which is real expenditures per pupil in the district. The amounts are in 1997 dollars. The spending variable will appear in logarithmic form.

(i) Consider the static unobserved effects model

$$math4_{it} = \delta_1 y93_t + \ldots + \delta_6 y98_t + \beta_1 \log(rexpp_{it}) + \beta_2 \log(enrol_{it}) + \beta_3 lunch_{it} + a_i + u_{it},$$

where $enrol_{it}$ is total district enrollment and $lunch_{it}$ is the percentage of students in the district eligible for the school lunch program. (So $lunch_{it}$ is a pretty good measure of the district-wide poverty rate.) Argue that $\beta_1/10$ is the percentage point change in $math4_{it}$ when real per-student spending increases by roughly 10%.

(ii) Use first differencing to estimate the model in part (i). The simplest approach is to allow an intercept in the first-differenced equation and to include dummy variables for the years 1994 through 1998. Interpret the coefficient on the spending variable.

(iii) Now, add one lag of the spending variable to the model and reestimate using first differencing. Note that you lose another year of data, so you are only using changes starting in 1994. Discuss the coefficients and significance on the current and lagged spending variables.

Utrecht University

(iv) Obtain heteroskedasticity-robust standard errors for the first-differenced regression in part (iii). How do these standard errors compare with those from part (iii) for the spending variables?

(v) Now, obtain standard errors robust to both heteroskedasticity and serial correlation. What does this do to the significance of the lagged spending variable?

(vi) Verify that the differenced errors $r_{it} = \Delta u_{it}$ have negative serial correlation by carrying out a test of AR(1) serial correlation.

(vii) Based on a fully robust joint test, does it appear necessary to include the enrollment and lunch variables in the model?

Utrecht University

Solution:

Question C13.11

(i) Argue that $\beta_1/10$ *is the percentage point change in math4it when real per-student spending increases by roughly 10%.*

This is a level-log specification. For the explanation, let's take a simple example:

$$Y = \beta_0 + \beta_1 \ln(X) + u$$
$$\frac{dY}{dX} = \frac{\beta_1}{X}$$
$$dY = \frac{\beta_1}{100}\left(\frac{dX}{X} * 100\right)$$

Hence, $\frac{\beta_1}{100}$ is the change in Y due to a one-percent change in X (i.e. when $\frac{dX}{X} * 100 = 1$). Hence, a 10% change in X relates to a $\frac{\beta_1}{100} * 10$ change in Y.

OR:

1. A one-percent change in X, increases ln(X) with 0.01 (Note: ln(aX)=ln(a)+ln(X) with a=1.01 => ln(a) about 0.01)
2. A 0.01 increase in ln(X) increases Y with $\beta_1 * 0.01$
3. In our case, since X grows by 10%, the effect on Y is about $\beta_1 * 0.01 * 10$ units.

Utrecht University

ii) Use first differencing to estimate the model in part i). The simplest approach is to allow an intercept in the first-differenced equation and to include dummy variables for the years 1994 through 1998. Interpret the coefficient on the spending variable.

Solution:

- The equation, estimated by pooled OLS in first differences (except for the year dummies), is

$$\Delta math4 = \underset{(.52)}{5.95} + \underset{(.73)}{.52}\, y94 + \underset{(.78)}{6.81}\, y95 - \underset{(.73)}{5.23}\, y96 - \underset{(.72)}{8.49}\, y97 + \underset{(.72)}{8.97}\, y98$$

$$- \underset{(2.76)}{3.45}\, \Delta \log(rexpp) + \underset{(1.029)}{.635}\, \Delta \log(enroll) + \underset{(.055)}{.025}\, \Delta lunch$$

$$n = 3{,}300, \quad R^2 = .208.$$

- Taken literally, the spending coefficient implies that a 10% increase in real spending per pupil decreases the *math4* pass rate by about 3.45/10 ≈0.35 percentage points.

```
. reg d.math4 d.lrexpp d.lenrol d.lunch y94 y95 y96 y97 y98

      Source |       SS           df       MS      Number of obs   =     3,300
-------------+----------------------------------   F(8, 3291)      =    108.03
       Model |  122398.669         8  15299.8336   Prob > F        =    0.0000
    Residual |  466100.028     3,291  141.628693   R-squared       =    0.2080
-------------+----------------------------------   Adj R-squared   =    0.2061
       Total |  588498.697     3,299  178.386995   Root MSE        =    11.901

------------------------------------------------------------------------------
     D.math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lrexpp |
         D1. |  -3.447268   2.760079    -1.25   0.212    -8.858913    1.964377
             |
      lenrol |
         D1. |   .6345335   1.028603     0.62   0.537    -1.382233      2.6513
             |
       lunch |
         D1. |    .025074   .0554734     0.45   0.651     -.083692    .1338399
             |
         y94 |   .5210521   .7283771     0.72   0.474    -.9070661     1.94917
         y95 |   6.812446   .7786636     8.75   0.000     5.285732    8.339161
         y96 |   -5.23489   .7271019    -7.20   0.000    -6.660508   -3.809272
         y97 |  -8.488463   .7222014   -11.75   0.000    -9.904472   -7.072453
         y98 |   8.967841   .7192335    12.47   0.000      7.55765    10.37803
       _cons |   5.954963   .5182347    11.49   0.000     4.938868    6.971058
------------------------------------------------------------------------------
```

Utrecht University

(iii) Now, add one lag of the spending variable to the model and reestimate using first differencing. Note that you lose another year of data, so you are only using changes starting in 1994. Discuss the coefficients and significance on the current and lagged spending variables.

Solution: When we add the lagged spending change, and drop another year, we get

```
. reg d.math4 d.lenrol d.lrexpp d.l.lrexpp d.lunch y95 y96 y97 y98

      Source |       SS           df       MS      Number of obs   =     2,750
-------------+----------------------------------   F(8, 2741)      =    106.75
       Model |  124773.729         8  15596.7161   Prob > F        =    0.0000
    Residual |   400464.32     2,741  146.101539   R-squared       =    0.2376
-------------+----------------------------------   Adj R-squared   =    0.2353
       Total |  525238.048     2,749  191.065132   Root MSE        =    12.087

------------------------------------------------------------------------------
     D.math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lenrol |
         D1. |   2.140017   1.176886     1.82   0.069    -.1676554    4.447689
             |
      lrexpp |
         D1. |  -1.410699   3.037452    -0.46   0.642    -7.366626    4.545228
         LD. |   11.04026   2.785833     3.96   0.000     5.577717    16.50281
             |
       lunch |
         D1. |   .0728056   .0614869     1.18   0.236    -.0477598     .193371
             |
         y95 |   5.704738   .7743696     7.37   0.000     4.186331    7.223145
         y96 |  -6.795939   .7896773    -8.61   0.000    -8.344362   -5.247516
         y97 |  -8.989378   .7376818   -12.19   0.000    -10.43585    -7.54291
         y98 |   8.453018   .7435231    11.37   0.000     6.995096     9.91094
       _cons |   6.158613    .551317    11.17   0.000     5.077574    7.239652
------------------------------------------------------------------------------
```

$$\Delta math4 = \underset{(.55)}{6.16} + \underset{(.77)}{5.70}\, y95 - \underset{(.79)}{6.80}\, y96 - \underset{(.74)}{8.99}\, y97 + \underset{(.74)}{8.45}\, y98$$

$$- \underset{(3.04)}{1.41}\, \Delta\log(rexpp) + \underset{(2.79)}{11.04}\, \Delta\log(rexpp_{-1}) + \underset{(1.18)}{2.14}\, \Delta\log(enroll)$$

$$+ \underset{(.061)}{.073}\, \Delta lunch$$

$$n = 2{,}750, \quad R^2 = .238.$$

- The contemporaneous spending variable, while still having a negative coefficient, is not at all statistically significant.
- The coefficient on the lagged spending variable is very statistically significant and implies that a 10% increase in spending last year increases the *math4* pass rate by about 1.1 percentage points. Given the timing of the tests, a lagged effect is not surprising.
- In Michigan, the fourth grade math test is given in January, and so if preparation for the test begins a full year in advance, spending when the students are in third grade would at least partly matter.

Utrecht University

(iv) Obtain heteroskedasticity-robust standard errors for the first-differenced regression in part (iii). How do these standard errors compare with those from part (iii) for the spending variables?

```
. reg d.math4 d.lenrol d.lrexpp d.l.lrexpp d.lunch y95 y96 y97 y98, robust

Linear regression                               Number of obs     =      2,750
                                                F(8, 2741)        =     107.96
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2376
                                                Root MSE          =     12.087

------------------------------------------------------------------------------
             |               Robust
     D.math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lenrol |
         D1. |   2.140017   1.392986     1.54   0.125    -.5913921    4.871426
             |
      lrexpp |
         D1. |  -1.410699   4.282755    -0.33   0.742    -9.808452    6.987054
         LD. |   11.04026   4.379687     2.52   0.012     2.452441    19.62808
             |
       lunch |
         D1. |   .0728056   .1412903     0.52   0.606    -.2042407    .3498519
             |
         y95 |   5.704738   .7941769     7.18   0.000     4.147492    7.261983
         y96 |  -6.795939   .8399903    -8.09   0.000    -8.443017   -5.148861
         y97 |  -8.989378   .7516618   -11.96   0.000    -10.46326   -7.515497
         y98 |   8.453018   .7713763    10.96   0.000      6.94048    9.965556
       _cons |   6.158613      .5833    10.56   0.000     5.014861    7.302365
------------------------------------------------------------------------------
```

The heteroskedasticity-robust standard error for $\hat{\beta}_{\Delta\log(rexpp)}$ is about 4.28, which reduces the significance of $\Delta\log(rexpp)$ even further. The heteroskedasticity-robust standard error of $\hat{\beta}_{\Delta\log(rexpp-1)}$ is about 4.38, which substantially lowers the t statistic. Still, Δlog (*rexpp*-1) is statistically significant at just over the 1% significance level against a two-sided alternative

Utrecht University

(v) Now, obtain standard errors robust to both heteroskedasticity and serial correlation. What does this do to the significance of the lagged spending variable?

Solution:

```
. reg d.math4 d.lenrol d.lrexpp d.l.lrexpp d.lunch y95 y96 y97 y98, cluster(distid)

Linear regression                               Number of obs     =      2,750
                                                F(8, 549)         =      95.27
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2376
                                                Root MSE          =     12.087

                               (Std. Err. adjusted for 550 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
     D.math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lenrol |
         D1. |   2.140017     1.64512     1.30   0.194    -1.091483    5.371517
             |
      lrexpp |
         D1. |  -1.410699    4.944102    -0.29   0.775    -11.12237    8.300972
         LD. |   11.04026    5.131503     2.15   0.032     .9604799    21.12004
             |
       lunch |
         D1. |   .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   -5.078096
         y97 |  -8.989378    .7718688   -11.65   0.000    -10.50556   -7.473201
         y98 |   8.453018    .7838456    10.78   0.000     6.913315    9.992722
       _cons |   6.158613    .6572949     9.37   0.000     4.867492    7.449734
------------------------------------------------------------------------------
```

Cluster at the district level

The fully robust standard error for for $\hat{\beta}_{\Delta\log(rexpp)}$ is about 4.94, which even further reduces the *t* statistic for Δlog(*rexpp*). The fully robust standard error for $\hat{\beta}_{\Delta\log(rexpp-1)}$ is about 5.13,which gives Δlog($rexpp_{-1}$) a *t* statistic of about 2.15. The two-sided *p*-value is about .032.

Utrecht University

(vi) Verify that the differenced errors $r_{it} = \Delta u_{it}$ have negative serial correlation by carrying out a test of AR(1) serial correlation.

We can conduct a Breusch-Godfrey test assuming strict exogeneity or not (in this exercise, strict exogeneity is assumed). In this case, for the test, we regress $\Delta(u_{it})$ only on its lag, whereas when we don't assume strict exogeneity, we include the rest of the explanatory variables in the regression. Here, we show the B-G test with strict exogeneity.

First,we predict rhat.

```
. predict rhat, resid
(1,100 missing values generated)
```

Then we carry out the test:

```
. reg rhat l.rhat

      Source |       SS           df       MS      Number of obs   =     2,200
-------------+----------------------------------   F(1, 2198)      =    588.06
       Model |  67536.1844         1  67536.1844   Prob > F        =    0.0000
    Residual |  252432.528     2,198  114.846464   R-squared       =    0.2111
-------------+----------------------------------   Adj R-squared   =    0.2107
       Total |  319968.712     2,199  145.506463   Root MSE        =    10.717

------------------------------------------------------------------------------
        rhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        rhat |
         L1. |  -.4628881   .0190883   -24.25   0.000     -.500321   -.4254552
             |
       _cons |   2.41e-09   .2284796     0.00   1.000    -.4480586    .4480586
------------------------------------------------------------------------------
```

We can use four years of data for this test. Doing a pooled OLS regression of $\widehat{r_{it}}$on $\widehat{r_{i,t-1}}$ using years 1995, 1996, 1997, and 1998 gives $\widehat{\rho = 0}$.423 (se = .019), which is strong negative serial correlation.

Utrecht University

(vii) **Based on a fully robust joint test, does it appear necessary to include the enrollment and lunch variables in the model?**

```
. reg d.math4 d.lenrol d.lrexpp d.lrexpp_1 d.lunch y95 y96 y97 y98, cluster(distid)

Linear regression                               Number of obs     =      2,750
                                                F(8, 549)         =      95.27
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2376
                                                Root MSE          =     12.087

                                (Std. Err. adjusted for 550 clusters in distid)
------------------------------------------------------------------------------
             |               Robust
     D.math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lenrol |
         D1. |   2.140017    1.64512     1.30   0.194    -1.091483    5.371517
             |
      lrexpp |
         D1. |  -1.410699   4.944102    -0.29   0.775    -11.12237    8.300972
             |
    lrexpp_1 |
         D1. |   11.04026   5.131503     2.15   0.032     .9604799    21.12004
             |
       lunch |
         D1. |   .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   -5.078096
         y97 |  -8.989378   .7718688   -11.65   0.000    -10.50556   -7.473201
         y98 |   8.453018   .7838456    10.78   0.000     6.913315    9.992722
       _cons |   6.158613   .6572949     9.37   0.000     4.867492    7.449734
------------------------------------------------------------------------------
test d.lenrol d.lunch

 ( 1)  D.lenrol = 0
 ( 2)  D.lunch = 0

       F(  2,   549) =    0.93
            Prob > F =    0.3951
```

The fully robust "*F*" test for $\Delta\log(enroll)$ and $\Delta lunch$, reported by Stata 7.0, is .93. With 2 and 549 *df*, this translates into *p*-value = .40. So we would be justified in dropping these variables.

Utrecht University

C13 Use the data in WAGEPAN.RAW for this exercise.

(i) Consider the unobserved effects model

$$lwage_{it} = \beta_0 + \delta_1 d81_t + \ldots + \delta_7 d87_t + \beta_1 educ_i$$
$$+ \gamma_1 d81_t educ_i + \ldots + \delta_7 d87_t educ_i + \beta_2 union_{it} + a_i + u_{it},$$

where a_i is allowed to be correlated with $educ_i$ and $union_{it}$. Which parameters can you estimate using first differencing?

(ii) Estimate the equation from part (i) by FD, and test the null hypothesis that the return to education has not changed over time.

(iii) Test the hypothesis from part (ii) using a fully robust test, that is, one that allows arbitrary heteroskedasticity and serial correlation in the FD errors, Δu_{it}. Does your conclusion change?

(iv) Now allow the union differential to change over time (along with education) and estimate the equation by FD. What is the estimated union differential in 1980? What about 1987? Is the difference statistically significant?

(v) Test the null hypothesis that the union differential has not changed over time, and discuss your results in light of your answer to part (iv).

Utrecht University

Solution:

(i)Which parameters can you estimate using first differencing?

Using first differencing, we can estimate all parameters on time-varying variables. We cannot estimate the intercept β_0 because it is the intercept of the base year. We cannot estimate β_1 because education is not a time-varying variable. Its within variation is equal to zero (within standard deviation is 0).

```
. xtsum educ

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
educ     overall |  11.76697   1.746181          3         16 |     N =    4360
         between |             1.747585          3         16 |     n =     545
         within  |                    0   11.76697   11.76697 |     T =       8
```

Formalize a bit:

$$\ln(w_{it}) = \beta_0 + \gamma_t + \beta_1 educ_i + \beta_2 union_{it} + \alpha_i + u_{it}$$

$$\Delta \ln(w_{it}) = \Delta\gamma_t + \beta_2 \Delta union_{it} + \Delta u_{it}$$

Note: $\Delta\beta_0 = 0$; $\Delta\alpha_i = 0$; $\Delta educ_i = 0$.

Interaction terms with time.

$$\ln(w_{it}) = \beta_0 + \gamma_t + \beta_1 educ_i + \sum_{s=1981}^{1987} \delta_s(educ_i \times I(s = t)) + \beta_2 union_{it} + \alpha_i + u_{it}$$

Simplify, suppose on 3 years of data (1980-1982):

$$\ln(w_{it}) = \beta_0 + \gamma_t + \beta_1 educ_i + \delta_{1981}(educ_i \times I(t = 1981)) + \delta_{1982}(educ_i \times I(t = 1982)) + \beta_2 union_{it} + \alpha_i + u_{it}$$

$$\Delta \ln(w_{it}) = \Delta\gamma_t + \delta_{1981}\Delta(educ_i \times I(t = 1981)) + \delta_{1982}\Delta(educ_i \times I(t = 1982)) + \beta_2 \Delta union_{it} + \Delta u_{it}$$

$$\Delta(educ_i \times I(t = 1981)) = (educ_i \times I(t = 1981)) - (educ_i \times I(t - 1 = 1981))$$

$$\Delta(educ_i \times I(t = 1981)) = educ_i \times (I(t = 1981) - I(t - 1 = 1981))$$

The important thing to note is that this latter interaction term varies over time, hence its effect is identified when using a fist-difference estimator.

(ii) Estimate equation from (i) in FD and test if return to education changed over time.

To test if return to education changed over time, we need to include interaction terms between education and year dummies in our specification and test their joint significance.

First, we need to generate our interaction terms:

```
. foreach x of numlist 81/87 {
  2. gen edu`x'=educ*d`x'
  3. }
```

Then we run a regression in first differences, including first differences of the interaction terms we just generated and suppressing the intercept.

```
. reg d.lwage d.union d82 d83 d84 d85 d86 d87 d.edu81 d.edu82 d.edu83 d.edu84
d.edu85 d.edu86 d.edu87

      Source |       SS           df       MS      Number of obs   =     3,815
-------------+----------------------------------   F(14, 3800)     =      1.18
       Model |  3.25847076        14  .232747911   Prob > F        =    0.2812
    Residual |  747.935458     3,800  .196825121   R-squared       =    0.0043
-------------+----------------------------------   Adj R-squared   =    0.0007
       Total |  751.193929     3,814  .196956982   Root MSE        =    .44365

------------------------------------------------------------------------------
     D.lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |
         D1. |   .0413336   .0197394     2.09   0.036     .0026328    .0800344
             |
         d82 |   .0355376   .1831518     0.19   0.846    -.3235478     .394623
         d83 |   .0438366   .1831453     0.24   0.811    -.3152361    .4029092
         d84 |   .0998409   .1831754     0.55   0.586    -.2592906    .4589724
         d85 |   -.023556   .1831096    -0.13   0.898    -.3825586    .3354466
         d86 |   .0380151   .1831159     0.21   0.836    -.3209998      .39703
         d87 |     .05583   .1832165     0.30   0.761    -.3033821    .4150421
             |
       edu81 |
         D1. |   .0120416   .0108871     1.11   0.269    -.0093036    .0333868
             |
       edu82 |
         D1. |   .0158817   .0153938     1.03   0.302    -.0142992    .0460626
             |
       edu83 |
         D1. |   .0181288   .0188525     0.96   0.336    -.0188332    .0550908
             |
       edu84 |
         D1. |   .0175501   .0217688     0.81   0.420    -.0251295    .0602298
             |
       edu85 |
         D1. |   .0257116   .0243386     1.06   0.291    -.0220063    .0734296
             |
       edu86 |
         D1. |   .0295403   .0266623     1.11   0.268    -.0227334     .081814
             |
       edu87 |
         D1. |   .0321777   .0287973     1.12   0.264     -.024282    .0886374
             |
       _cons |  -.0222273   .1295114    -0.17   0.864    -.2761458    .2316913
------------------------------------------------------------------------------
```

```
. test d.edu81 d.edu82 d.edu83 d.edu84 d.edu85 d.edu86 d.edu87

 ( 1)  D.edu81 = 0
 ( 2)  D.edu82 = 0
 ( 3)  D.edu83 = 0
 ( 4)  D.edu84 = 0
 ( 5)  D.edu85 = 0
 ( 6)  D.edu86 = 0
 ( 7)  D.edu87 = 0

       F(  7,  3800) =    0.31
            Prob > F =    0.9518
```

To test if return to education changed over time we jointly test the interaction terms between education and year variables. F-stat is 0.31 and p-val=0.95 -> they are not jointly significant. Return to education does not change over time.

Utrecht University

(iii) Test hypothesis from (ii) on fully-robust specification.

```
. reg d.lwage d.union d82 d83 d84 d85 d86 d87 d.edu81 d.edu82 d.edu83 d.edu84
d.edu85 d.edu86 d.edu87, cluster (nr)

Linear regression                               Number of obs     =      3,815
                                                F(14, 544)        =       1.26
                                                Prob > F          =     0.2252
                                                R-squared         =     0.0043
                                                Root MSE          =     .44365

                                  (Std. Err. adjusted for 545 clusters in nr)
------------------------------------------------------------------------------
             |               Robust
     D.lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |
         D1. |   .0413336   .0220219     1.88   0.061    -.0019248     .084592
             |
         d82 |   .0355376   .2087356     0.17   0.865     -.374489    .4455642
         d83 |   .0438366   .1683955     0.26   0.795    -.2869485    .3746216
         d84 |   .0998409   .1950307     0.51   0.609    -.2832646    .4829464
         d85 |   -.023556   .1725279    -0.14   0.891    -.3624584    .3153464
         d86 |   .0380151   .1892503     0.20   0.841    -.3337358     .409766
         d87 |     .05583   .1877267     0.30   0.766    -.3129279     .424588
             |
       edu81 |
         D1. |   .0120416   .0121436     0.99   0.322    -.0118125    .0358957
             |
       edu82 |
         D1. |   .0158817   .0117999     1.35   0.179    -.0072972    .0390606
             |
       edu83 |
         D1. |   .0181288   .0130959     1.38   0.167     -.007596    .0438537
             |
       edu84 |
         D1. |   .0175501   .0138452     1.27   0.205    -.0096464    .0447467
             |
       edu85 |
         D1. |   .0257116   .0135124     1.90   0.058    -.0008312    .0522545
             |
       edu86 |
         D1. |   .0295403   .0147353     2.00   0.045     .0005953    .0584853
             |
       edu87 |
         D1. |   .0321777   .0135337     2.38   0.018      .005593    .0587623
             |
       _cons |  -.0222273   .1435815    -0.15   0.877    -.3042694    .2598149
------------------------------------------------------------------------------
test d.edu81 d.edu82 d.edu83 d.edu84 d.edu85 d.edu86 d.edu87

 ( 1)  D.edu81 = 0
 ( 2)  D.edu82 = 0
 ( 3)  D.edu83 = 0
 ( 4)  D.edu84 = 0
 ( 5)  D.edu85 = 0
 ( 6)  D.edu86 = 0
 ( 7)  D.edu87 = 0

       F(  7,   544) =    1.00
            Prob > F =    0.4315
```

Employee number level

The fully robust F statistic (obtained with option cluster at the employee id number level) is about 1.00, with p-value = .432. So the conclusion really does not change. The gammas are jointly insignificant.

(iv) Allow union differential to change over time. What is it in 1980? In 1987? Are they significantly different?

To allow union differential to change over time, we need to include interaction terms between union and years. We first generate the union*year interaction terms.

```
. foreach x of numlist 81/87 {
  2. gen union`x'=union*d`x'
  3. }
```

After that, we run the regression in first differences, with cluster.

```
. reg d.lwage d.union d82 d83 d84 d85 d86 d87 d.union81 d.union82 d.union83
d.union84 d.union85 d.union86 d.union87 d.edu81 d.ed
> u82 d.edu83 d.edu84 d.edu85 d.edu86 d.edu87, cluster(nr)

Linear regression                               Number of obs     =      3,815
                                                F(21, 544)        =       1.22
                                                Prob > F          =     0.2306
                                                R-squared         =     0.0062
                                                Root MSE          =     .44364

                                 (Std. Err. adjusted for 545 clusters in nr)
------------------------------------------------------------------------------
             |               Robust
     D.lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       union |
         D1. |   .1056242   .0507495     2.08   0.038     .0059353    .2053132
             |
         d82 |   .0318354   .2131808     0.15   0.881    -.3869231    .4505938
         d83 |   .0359202   .1714342     0.21   0.834    -.3008339    .3726743
         d84 |   .0779319    .196595     0.40   0.692    -.3082465    .4641102
         d85 |  -.0381174   .1730352    -0.22   0.826    -.3780164    .3017817
         d86 |   .0346367   .1923653     0.18   0.857    -.3432331    .4125065
         d87 |   .0623818   .1891906     0.33   0.742    -.3092517    .4340153
             |
     union81 |
         D1. |   -.019609   .0572557    -0.34   0.732    -.1320784    .0928604
             |
     union82 |
         D1. |  -.0691526    .056471    -1.22   0.221    -.1800805    .0417754
             |
     union83 |
         D1. |  -.0881435   .0565481    -1.56   0.120    -.1992229    .0229359
             |
     union84 |
         D1. |  -.0585253    .059866    -0.98   0.329    -.1761221    .0590715
             |
     union85 |
         D1. |  -.0486768   .0596433    -0.82   0.415    -.1658361    .0684825
             |
     union86 |
         D1. |  -.1075604    .064063    -1.68   0.094    -.2334016    .0182807
             |
     union87 |
         D1. |  -.1471487   .0684309    -2.15   0.032    -.2815699   -.0127274
             |
      edu81 |
         D1. |   .0113406   .0122286     0.93   0.354    -.0126804    .0353616
```

edu82 D1.	.0154221	.0118275	1.30	0.193	-.0078111	.0386554
edu83 D1.	.0176074	.0131507	1.34	0.181	-.008225	.0434397
edu84 D1.	.0171421	.0138039	1.24	0.215	-.0099733	.0442575
edu85 D1.	.0252403	.013563	1.86	0.063	-.0014018	.0518825
edu86 D1.	.0293022	.0147808	1.98	0.048	.0002677	.0583366
edu87 D1.	.0313275	.0135471	2.31	0.021	.0047165	.0579386
_cons	-.0089671	.1475466	-0.06	0.952	-.298798	.2808638

The estimated union differential in 1980 is the parameter on Δ(union)= 0.105 (significant at 5% level). In 1987, this is equal to the sum of Δ(union) and Δ(union*y87):

```
. nlcom _b[d.union87]+ _b[d.union]

       _nl_1:  _b[d.union87]+ _b[d.union]
```

D.lwage	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	-.0415244	.0444042	-0.94	0.350	-.1285551	.0455062

Thus, the estimated union differential in 1987 is equal -0.042

The difference between the union differential in 1980 and 1987 is equal the parameter on the interaction term Δ(union*y87). It is equal to -.147 (-14.7%) and t-stat= 2.15. It is significant at the 5% level (p-val=0.032).

(v) Test null hypothesis that union differential has not changed over time.

We can do this by jointly testing the coefficients on the Δ (union*year) interaction terms from the regression above.

```
. test d.union81 d.union82 d.union83 d.union84 d.union85 d.union86
d.union87

 ( 1)  D.union81 = 0
 ( 2)  D.union82 = 0
 ( 3)  D.union83 = 0
 ( 4)  D.union84 = 0
 ( 5)  D.union85 = 0
 ( 6)  D.union86 = 0
 ( 7)  D.union87 = 0

       F(  7,   544) =    1.15
            Prob > F =    0.3310
```

The fully robust, joint test has an F-stat=1.15 and p-val=0.33. We fail to reject H0 of no significant change over time. Thus, we conclude that the union differential has not changed significantly over time. Some of the differences are individually significant but not jointly so. This could be because lumping several insignificant coefficients together with a couple of significant ones drives the F-statistic down. Another problem could be with the strict exogeneity assumption: perhaps union membership next year depends on unexpected wage changes this year.