Tutorials Week 5

Pdf file on Blackboard	Dataset on Blackboard	Papers	Description
C 14.5	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	Fixed effects vs. first diff, pooled OLS, consequences of omitting dummy variables.
C 14.9	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	Estimation of models with pooled OLS, use of random effects, comparison of results from random effects vs. fixed effects.
C 14.8 (part (vi)	mathpnl.dta	Papke, Leslie (2005): “The Effects of Spending on Test Pass Rates: Evidence from Michigan” (2005), Journal of Public Economics 89, 821-839.	Estimation of models with pooled OLS, estimate the model with a fixed effect estimator, and verify serial correlation. For the extra questions: application of the Hausman test.

C 14.5

C5 (i) In the wage equation in Example 14.4, explain why dummy variables for occupation might be important omitted variables for estimating the union wage premium.

(ii) If every man in the sample stayed in the same occupation from 1981 through 1987, would you need to include the occupation dummies in a fixed effects estimation? Explain.

(iii) Using the data in WAGEPAN.RAW, include eight of the occupation dummy variables in the equation and estimate the equation using fixed effects. Does the coefficient on *union* change by much? What about its statistical significance?

Utrecht University

EXAMPLE 14.4 A WAGE EQUATION USING PANEL DATA

We again use the data in WAGEPAN.RAW to estimate a wage equation for men. We use three methods: pooled OLS, random effects, and fixed effects. In the first two methods, we can include *educ* and race dummies (*black* and *hispan*), but these drop out of the fixed effects analysis. The time-varying variables are *exper*, $exper^2$, *union*, and *married*. As we discussed in Section 14.1, *exper* is dropped in the FE analysis (although $exper^2$ remains). Each regression also contains a full set of year dummies. The estimation results are in Table 14.2.

TABLE 14.2 Three Different Estimators of a Wage Equation

Dependent Variable: log(*wage*)

Independent Variables	Pooled OLS	Random Effects	Fixed Effects
educ	.091 (.005)	.092 (.011)	——
black	−.139 (.024)	−.139 (.048)	——
hispan	.016 (.021)	.022 (.043)	——
exper	.067 (.014)	.106 (.015)	——
$exper^2$	−.0024 (.0008)	−.0047 (.0007)	−.0052 (.0007)
married	.108 (.016)	.064 (.017)	.047 (.018)
union	.182 (.017)	.106 (.018)	.080 (.019)

The coefficients on *educ*, *black*, and *hispan* are similar for the pooled OLS and random effects estimations. The pooled OLS standard errors are the usual OLS standard errors, and these underestimate the true standard errors because they ignore the positive serial correlation; we report them here for comparison only. The experience profile is somewhat different, and both the marriage and union premiums fall notably in the random effects estimation. When we eliminate the unobserved effect entirely by using fixed effects, the marriage premium falls to about 4.7%, although it is still statistically significant. The drop in the marriage premium is consistent with the idea that men who are more able—as captured by a higher unobserved effect, a_i—are more likely to be married. Therefore, in the pooled OLS estimation, a large part of the marriage premium reflects the fact that men who are married would earn more even if they were not married. The remaining 4.7% has at least two possible explanations: (1) marriage really makes men more productive or (2) employers pay married men a premium because marriage is a signal of stability. We cannot distinguish between these two hypotheses.

The estimate of θ for the random effects estimation is $\hat{\theta} = .643$, which helps explain why, on the time-varying variables, the RE estimates lie closer to the FE estimates than to the pooled OLS estimates.

EXPLORING FURTHER 14.3

The union premium estimated by fixed effects is about 10 percentage points lower than the OLS estimate. What does this strongly suggest about the correlation between *union* and the unobserved effect?

Utrecht University

TABLE 14.2

Three Different Estimators of a Wage Equation

Dependent Variable: log(*wage*)			
Independent Variables	**Pooled OLS**	**Random Effects**	**Fixed Effects**
educ	.091 (.005)	.092 (.011)	——
black	−.139 (.024)	−.139 (.048)	——
hispan	.016 (.021)	.022 (.043)	——
exper	.067 (.014)	.106 (.015)	——
exper^2	−.0024 (.0008)	−.0047 (.0007)	−.0052 (.0007)
married	.108 (.016)	.064 (.017)	.047 (.018)
union	.182 (.017)	.106 (.018)	.080 (.019)

i) The exercise refers to Example 14.4. in the book

- The question is why omitting dummy variables for occupations can be a problem. First, you should realize this would be a problem with the fixed-effects specifications only if the observed individuals changed their occupations in the sample period.
- If occupations are fixed over time, then their effect is removed by the fixed-effect transformation.
- In this sample, **individuals change their occupation**; hence, an omitted variables bias affects all our results.
- Omitting the **occupation** dummy causes a bias since (apart from the fact wages are likely to differ across occupations) the degree of unionization may vary by occupation; hence, when you exclude it from the model, your residual will be correlated with the variable union.

Utrecht University

(ii) If every man in the sample stayed in the same occupation from 1981 through 1987, would you need to include the occupation dummies in a fixed effects estimation? Explain.

- As we saw under point (i), if all individuals remained in the same occupation, then the fixed effect (within-group) transformation would have removed the effect of occupations from the model.
- Not only were occupation dummies unnecessary to include, but they would be dropped from the regression because of the lack of variance over time.

(iii) Using the data in WAGEPAN.RAW, include eight of the occupation dummy variables in the equation and estimate the equation using fixed effects. Does the coefficient on *union* change by much? What about its statistical significance?

- Because the nine occupational categories (*occ1* through *occ9*) are exhaustive, we must choose one as the base group.
- Of course the group we choose does not affect the estimated union wage differential.
- The fixed effect estimate on *union*, to four decimal places, is .0804 with standard error = .0194.
- There is practically no difference between this estimate and standard error and the estimate and standard error without the occupational controls ($\hat{\beta}_{union}$ = .0800, se =0193).

NOTE: In Stata 11 you do not need "xi:" when adding dummy variables

```
. xi: xtreg lwage educ black hisp exper expersq married union i.year, fe
i.year            _Iyear_1980-1987    (naturally coded; _Iyear_1980 omitted)
note: educ omitted because of collinearity
note: black omitted because of collinearity
note: hisp omitted because of collinearity
note: _Iyear_1987 omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =      4360
Group variable: nr                              Number of groups   =       545

R-sq:  within  = 0.1806                         Obs per group: min =         8
       between = 0.0005                                        avg =       8.0
       overall = 0.0635                                        max =         8

                                                F(10,3805)         =     83.85
corr(u_i, Xb)  = -0.1212                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |          0  (omitted)
       black |          0  (omitted)
        hisp |          0  (omitted)
       exper |   .1321464   .0098247    13.45   0.000     .1128842    .1514087
     expersq |  -.0051855   .0007044    -7.36   0.000    -.0065666   -.0038044
     married |   .0466804   .0183104     2.55   0.011     .0107811    .0825796
       union |   .0800019   .0193103     4.14   0.000     .0421423    .1178614
 _Iyear_1981 |   .0190448   .0203626     0.94   0.350    -.0208779    .0589674
 _Iyear_1982 |   -.011322   .0202275    -0.56   0.576    -.0509798    .0283359
 _Iyear_1983 |  -.0419955   .0203205    -2.07   0.039    -.0818357   -.0021553
 _Iyear_1984 |  -.0384709   .0203144    -1.89   0.058    -.0782991    .0013573
 _Iyear_1985 |  -.0432498   .0202458    -2.14   0.033    -.0829434   -.0035562
 _Iyear_1986 |  -.0273819   .0203863    -1.34   0.179    -.0673511    .0125872
 _Iyear_1987 |          0  (omitted)
       _cons |    1.02764   .0299499    34.31   0.000     .9689201    1.086359
-------------+----------------------------------------------------------------
     sigma_u |   .4009279
     sigma_e |  .35099001
         rho |  .56612236   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(544, 3805) =     9.14             Prob > F = 0.0000
```

Now with occupation dummies included:

```
. xi: xtreg lwage educ black hisp exper expersq married union i.year occ2-occ9, fe
i.year            _Iyear_1980-1987    (naturally coded; _Iyear_1980 omitted)
note: educ omitted because of collinearity
note: black omitted because of collinearity
note: hisp omitted because of collinearity
note: _Iyear_1987 omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =      4360
Group variable: nr                              Number of groups   =       545
```

```
R-sq:  within  = 0.1827                         Obs per group: min =         8
       between = 0.0021                                        avg =       8.0
       overall = 0.0696                                        max =         8

                                                F(18,3797)         =     47.14
corr(u_i, Xb)  = -0.1071                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |          0  (omitted)
       black |          0  (omitted)
        hisp |          0  (omitted)
       exper |   .1305982   .0099326    13.15   0.000     .1111245    .1500718
     expersq |  -.0050996    .000709    -7.19   0.000    -.0064898   -.0037095
     married |   .0459227   .0183429     2.50   0.012     .0099598    .0818855
       union |    .080381   .0194006     4.14   0.000     .0423445    .1184175
 _Iyear_1981 |   .0206195   .0203817     1.01   0.312    -.0193406    .0605797
 _Iyear_1982 |  -.0098058   .0202533    -0.48   0.628    -.0495141    .0299025
 _Iyear_1983 |  -.0395489   .0203629    -1.94   0.052    -.0794721    .0003744
 _Iyear_1984 |  -.0371896   .0203516    -1.83   0.068    -.0770906    .0027115
 _Iyear_1985 |  -.0424542   .0202617    -2.10   0.036    -.0821791   -.0027293
 _Iyear_1986 |  -.0287417   .0204052    -1.41   0.159    -.0687479    .0112644
 _Iyear_1987 |          0  (omitted)
        occ2 |  -.0136055   .0323164    -0.42   0.674    -.0769646    .0497536
        occ3 |  -.0621039   .0377516    -1.65   0.100    -.1361192    .0119114
        occ4 |   -.079205   .0307204    -2.58   0.010    -.1394351    -.018975
        occ5 |  -.0307397    .030348    -1.01   0.311    -.0902397    .0287603
        occ6 |  -.0280367   .0306803    -0.91   0.361     -.088188    .0321147
        occ7 |  -.0386446   .0338524    -1.14   0.254    -.1050153     .027726
        occ8 |  -.0611993    .066051    -0.93   0.354    -.1906982    .0682997
        occ9 |  -.0438229   .0342572    -1.28   0.201    -.1109872    .0233415
       _cons |   1.068066   .0394164    27.10   0.000     .9907871    1.145346
-------------+----------------------------------------------------------------
     sigma_u |   .3984261
     sigma_e |  .35091327
         rho |   .5631524   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(544, 3797) =     8.37           Prob > F = 0.0000
```

We find that after including the occupations in the fixed-effect specification the coefficient on the union did not change much, from 0.800 to 0.804, and the standard errors remained also very close.
Hence our fears about the effect of omitted occupations are not justified.

Utrecht University

C14.9 of third edition. Use the data in wagepan.dta for this exercise.

(i) Estimate the model

$$lwage_{it} = \beta 0 + \beta_1 educ_i + \beta_2 black_i + \beta_3 hisp_i + v_{it}$$

by pooled OLS en report the standard errors in the usual form (this means here: we do not yet control for heteroscedasticity or serial correlation).

(ii) Estimate the model from part (i) by random effects (thinking that $v_{it} = a_i + u_{it}$). How do the RE and pooled OLS point estimates of the β_j compare?

(iii) Are the RE and pooled OLS standard errors the same? Which ones are more reliable and why?

(iv) As example 14.4 (Table 14.2) shows, parameter estimates corresponding to time-varying explanatory variables such as *union*, may differ between pooled OLS and RE. What could be an explanation for this?

Utrecht University

C14.9

(i) We estimate the model as requested with pooled OLS.

```
. reg lwage educ black hisp

      Source |       SS       df       MS              Number of obs =    4360
-------------+------------------------------           F(  3,  4356) =  108.70
       Model |  86.1248357     3  28.7082786           Prob > F      =  0.0000
    Residual |  1150.40481  4356  .264096604           R-squared     =  0.0697
-------------+------------------------------           Adj R-squared =  0.0690
       Total |  1236.52964  4359  .283672779           Root MSE      =   .5139

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0770943   .0045613    16.90   0.000     .0681518    .0860367
       black |  -.1225637   .0247022    -4.96   0.000    -.1709926   -.0741348
        hisp |    .024623   .0222046     1.11   0.268    -.0189092    .0681553
       _cons |   .7523087   .0553537    13.59   0.000     .6437872    .8608301
------------------------------------------------------------------------------
```

Model Specification:

We estimated the model using pooled OLS with the following independent variables: education (*educ*), race (*black*), and hispanic status (*hisp*) to explain the logarithm of wages (*lwage*).

You can also interpret the results. For example:

Education (educ): Holding other factors constant, an additional year of education is associated with an increase of approximately 8.01% in wages. This result is statistically significant ($p < 0.001$).

Log-level:

(ii) Estimate the model from part (i) by random effects (thinking that $v_{it} = a_i + u_{it}$). How do the RE and pooled OLS point estimates of the β_i compare?

```
. xtreg lwage educ black hisp , re

Random-effects GLS regression                   Number of obs      =      4360
Group variable: nr                              Number of groups   =       545

R-sq:  within  = 0.0000                         Obs per group: min =         8
       between = 0.1296                                        avg =       8.0
       overall = 0.0697                                        max =         8

                                                Wald chi2(3)       =     80.56
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0770943    .009177     8.40   0.000     .0591076    .0950809
       black |  -.1225637   .0496994    -2.47   0.014    -.2199728   -.0251546
        hisp |    .024623   .0446744     0.55   0.582    -.0629371    .1121831
       _cons |   .7523087   .1113686     6.76   0.000     .5340302    .9705872
-------------+----------------------------------------------------------------
     sigma_u |  .33894941
     sigma_e |  .38723169
         rho |   .4338047   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

•**Point Estimates**: The point estimates (coefficients) for all variables are the same between the pooled OLS and random effects model.

• **Implications**: This suggests that the random effects model did not alter the estimated relationships between the independent variables and *wage* compared to the pooled OLS model.

(iii) Are the RE and pooled OLS standard errors the same? Which ones are more reliable and why?

```
. reg lwage educ black hisp

      Source |       SS           df       MS      Number of obs   =     4360
-------------+----------------------------------   F(  3,  4356)   =   108.70
       Model |  86.1248357         3  28.7082786   Prob > F        =   0.0000
    Residual |  1150.40481     4356  .264096604   R-squared       =   0.0697
-------------+----------------------------------   Adj R-squared   =   0.0690
       Total |  1236.52964     4359  .283672779   Root MSE        =    .5139

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0770943   .0045613    16.90   0.000     .0681518    .0860367
       black |  -.1225637   .0247022    -4.96   0.000    -.1709926   -.0741348
        hisp |    .024623   .0222046     1.11   0.268    -.0189092    .0681553
       _cons |   .7523087   .0553537    13.59   0.000     .6437872    .8608301
```

```
. xtreg lwage educ black hisp , re

Random-effects GLS regression                   Number of obs      =      4360
Group variable: nr                              Number of groups   =       545

R-sq:  within  = 0.0000                         Obs per group: min =         8
       between = 0.1296                                        avg =       8.0
       overall = 0.0697                                        max =         8

                                                Wald chi2(3)       =     80.56
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0770943    .009177     8.40   0.000     .0591076    .0950809
       black |  -.1225637   .0496994    -2.47   0.014    -.2199728   -.0251546
        hisp |    .024623   .0446744     0.55   0.582    -.0629371    .1121831
       _cons |   .7523087   .1113686     6.76   0.000     .5340302    .9705872
-------------+----------------------------------------------------------------
     sigma_u |  .33894941
     sigma_e |  .38723169
         rho |   .4338047   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

- While the coefficients are the same, the standard errors are not. The standard errors for the RE estimates are higher for all coefficients.
- The **RE errors are more reliable** –for the usual unobserved effects model - because they account for the serial correlation in the composite error.
- One should **trust the results from the random-effect specifications**, since it corrects for the presence of first-order autocorrelation in the error term, which is caused by the presence of individual-specific effects. (See also, chapter 14.2 in the book)

Utrecht University

(iv)

TABLE 14.2 Three Different Estimators of a Wage Equation

Dependent Variable: log(*wage*)

Independent Variables	Pooled OLS	Random Effects	Fixed Effects
educ	.091 (.005)	.092 (.011)	——
black	−.139 (.024)	−.139 (.048)	——
hispan	.016 (.021)	.022 (.043)	——
exper	.067 (.014)	.106 (.015)	——
$exper^2$	−.0024 (.0008)	−.0047 (.0007)	−.0052 (.0007)
married	.108 (.016)	.064 (.017)	.047 (.018)
union	.182 (.017)	.106 (.018)	.080 (.019)

- The important conclusions are the same. The RE and OLS coefficient estimates are identical for all explanatory variables, but the RE standard errors differ.
- One difference is that the RE standard errors on the year dummies are smaller than the OLS standard errors on the year dummies.

Additional Material. Estimate with FE and compare with pooled OLS and RE

```
. xi: xtreg lwage educ black hisp i.year, fe
i.year            _Iyear_1980-1987    (naturally coded; _Iyear_1980 omitted)
note: educ omitted because of collinearity
note: black omitted because of collinearity
note: hisp omitted because of collinearity

Fixed-effects (within) regression               Number of obs      =      4360
Group variable: nr                              Number of groups   =       545

R-sq:  within  = 0.1625                         Obs per group: min =         8
       between =      .                                        avg =       8.0
       overall = 0.0752                                        max =         8

                                                F(7,3808)          =    105.56
corr(u_i, Xb)  = -0.0000                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |          0  (omitted)
       black |          0  (omitted)
        hisp |          0  (omitted)
 _Iyear_1981 |   .1193902    .021487     5.56   0.000     .0772631    .1615173
 _Iyear_1982 |   .1781901    .021487     8.29   0.000      .136063    .2203172
 _Iyear_1983 |   .2257865    .021487    10.51   0.000     .1836594    .2679135
 _Iyear_1984 |   .2968181    .021487    13.81   0.000      .254691    .3389452
 _Iyear_1985 |   .3459333    .021487    16.10   0.000     .3038063    .3880604
 _Iyear_1986 |   .4062418    .021487    18.91   0.000     .3641147    .4483688
 _Iyear_1987 |   .4730023    .021487    22.01   0.000     .4308753    .5151294
       _cons |   1.393477   .0151936    91.71   0.000     1.363689    1.423265
-------------+----------------------------------------------------------------
     sigma_u |  .39074676
     sigma_e |  .35469771
         rho |  .54824631   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(544, 3808) =     9.71             Prob > F = 0.0000
```

The coefficients and the standard errors of the year dummies are the same as in the RE specifications. Only the constant terms change.
We can hence observe that if we have only time-varying explanatory variables that are the same for all individuals within a year (year dummies), then there is no difference between the results from the RE or FE estimators.

Utrecht University

C8 Use the data in MATHPNL.RAW for this exercise. You will do a fixed effects version of the first differencing done in Computer Exercises 11 in Chapter 13. The model of interest is

$$math4_{it} = \delta_1 y94_t + \ldots + \delta_5 y98_t + \gamma_1 \log(rexpp_{it}) + \gamma_2 \log(rexpp_{i,t-1}) + \psi_1 \log(enrol_{it}) + \psi_2 lunch_{it} + a_i + u_{it},$$

where the first available year (the base year) is 1993 because of the lagged spending variable.

(i) Estimate the model by pooled OLS and report the usual standard errors. You should include an intercept along with the year dummies to allow a_i to have a nonzero expected value. What are the estimated effects of the spending variables? Obtain the OLS residuals, $\hat{v}_{it}$.

(ii) Is the sign of the $lunch_{it}$ coefficient what you expected? Interpret the magnitude of the coefficient. Would you say that the district poverty rate has a big effect on test pass rates?

(iii) Compute a test for AR(1) serial correlation using the regression $\hat{v}_{it}$ on $\hat{v}_{i,t-1}$. You should use the years 1994 through 1998 in the regression. Verify that there is strong positive serial correlation and discuss why.

(iv) Now, estimate the equation by fixed effects. Is the lagged spending variable still significant?

(v) Why do you think, in the fixed effects estimation, the enrollment and lunch program variables are jointly insignificant?

(vi) Define the total, or long-run, effect of spending as $\theta_1 = \gamma_1 + \gamma_2$. Use the substitution $\gamma_1 = \theta_1 - \gamma_2$ to obtain a standard error for $\hat{\theta}_1$. [*Hint:* Standard fixed effects estimation using $\log(rexpp_{it})$ and $z_{it} = \log(rexpp_{i,t-1}) - \log(rexpp_{it})$ as explanatory variables should do it.]

C.14.8

We will use the latest specification from C13.11.

(i) We need to estimate the model with pooled OLS (assuming homeskedasticity and no serial correlation of the error term).

No serial correlation: u_{it} is uncorrelated across time.

```
. reg math4 lenrol lrexpp lrexpp_1 lunch y94 y95 y96 y97 y98

      Source |       SS       df       MS              Number of obs =    3300
-------------+------------------------------           F(  9,  3290) =  373.34
       Model |  487253.951     9  54139.3279           Prob > F      =  0.0000
    Residual |  477099.485  3290  145.015041           R-squared     =  0.5053
-------------+------------------------------           Adj R-squared =  0.5039
       Total |  964353.436  3299  292.316895           Root MSE      =  12.042

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lenrol |   .5926719   .2050417     2.89   0.004     .1906496    .9946942
      lrexpp |   .5339314   2.428118     0.22   0.826    -4.226844    5.294706
    lrexpp_1 |   9.049175    2.30532     3.93   0.000     4.529168    13.56918
       lunch |  -.4067083   .0138353   -29.40   0.000     -.433835   -.3795817
         y94 |   6.377355   .7362674     8.66   0.000     4.933766    7.820944
         y95 |    18.6502   .7862743    23.72   0.000     17.10856    20.19183
         y96 |   18.03336   .7672536    23.50   0.000     16.52902    19.53771
         y97 |   15.34006   .7768368    19.75   0.000     13.81693    16.86319
         y98 |   30.39788   .7831877    38.81   0.000      28.8623    31.93347
       _cons |  -31.66156   10.30109    -3.07   0.002    -51.85876   -11.46436
------------------------------------------------------------------------------
```

We save the residuals:

```
. predict vhat,res
(550 missing values generated)
```

The effect of spending on success rate is about 0.096 p.p. per 1% increase in spending.

```
. nlcom _b[lrexpp]+_b[lrexpp_1]

       _nl_1:  _b[lrexpp]+_b[lrexpp_1]

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   9.583107   1.245504     7.69   0.000     7.141964    12.02425
```

ii) The coefficient of the lunch variable is negative. This suggests that districts with more pupils eligible for free lunch (poor districts) have lower success rates on average. The *lunch* variable is the percentage of students in the district eligible for free or reduced-price lunches, which is determined by poverty status. Therefore, *lunch* is effectively a poverty rate.
We see that the district poverty rate has a large impact on the math pass rate: a one percentage point increase in *lunch* reduces the pass rate by about 0.41 percentage points.

Utrecht University

(iii) We should regress the residual on the lagged residual

```
. tsset distid year, delta(1)
       panel variable:  distid (strongly balanced)
        time variable:  year, 1992 to 1998
                delta:  1 unit

. reg vhat l.vhat

      Source |       SS           df       MS      Number of obs   =     2,750
-------------+----------------------------------   F(1, 2748)      =    892.82
       Model |  101287.401         1  101287.401   Prob > F        =    0.0000
    Residual |  311751.296     2,748  113.446614   R-squared       =    0.2452
-------------+----------------------------------   Adj R-squared   =    0.2450
       Total |  413038.697     2,749  150.250526   Root MSE        =    10.651

------------------------------------------------------------------------------
        vhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        vhat |
         L1. |   .5043244   .0168783    29.88   0.000      .471229    .5374198
             |
       _cons |   1.21e-09   .2031091     0.00   1.000     -.398262     .398262
------------------------------------------------------------------------------
```

```
. reg vhat l.vhat lrexpp l1.lrexpp lenrol lunch y94 y95 y96 y97 y98
note: y98 omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =     2,750
-------------+----------------------------------   F(9, 2740)      =     99.17
       Model |  101482.737         9  11275.8596   Prob > F        =    0.0000
    Residual |   311555.96     2,740  113.706555   R-squared       =    0.2457
-------------+----------------------------------   Adj R-squared   =    0.2432
       Total |  413038.697     2,749  150.250526   Root MSE        =    10.663

------------------------------------------------------------------------------
        vhat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        vhat |
         L1. |    .504918   .0169119    29.86   0.000     .4717566    .5380794
             |
      lrexpp |
         --. |   .7747519   2.395187     0.32   0.746    -3.921803    5.471307
         L1. |   -.396761   2.226345    -0.18   0.859    -4.762246    3.968724
             |
      lenrol |   .0928598   .1979226     0.47   0.639    -.2952328    .4809523
       lunch |   .0156079   .0133579     1.17   0.243    -.0105847    .0418006
         y94 |   .0721523   .6940439     0.10   0.917     -1.28875    1.433054
         y95 |  -.0108298   .6950408    -0.02   0.988    -1.373687    1.352027
         y96 |   .0263861   .6451815     0.04   0.967    -1.238705    1.291478
         y97 |   .0102643   .6442674     0.02   0.987    -1.253035    1.273563
         y98 |          0  (omitted)
       _cons |  -4.428925   10.71066    -0.41   0.679    -25.43072    16.57287
------------------------------------------------------------------------------
```

There is indeed a positive first order autocorrelation. P-value of vhat L1 is 0,000

There are many reasons for positive serial correlation. In the context of panel data, it indicates the presence of a time-constant unobserved effect, *ai*.

(iv) We estimate the model with a FE estimator.

Utrecht University

```
. xtreg math4 lenrol lrexpp lrexpp_1 lunch y94 y95 y96 y97 y98, fe

Fixed-effects (within) regression               Number of obs      =      3300
Group variable: distid                          Number of groups   =       550

R-sq:  within  = 0.6027                         Obs per group: min =         6
       between = 0.0399                                        avg =       6.0
       overall = 0.3202                                        max =         6

                                                F(9,2741)          =    462.02
corr(u_i, Xb)  = -0.0464                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lenrol |   .2450874   1.100381     0.22   0.824    -1.912572    2.402747
      lrexpp |  -.4111804   2.457658    -0.17   0.867     -5.23023    4.407869
    lrexpp_1 |   7.002988   2.369184     2.96   0.003     2.357421    11.64856
       lunch |    .061527   .0514661     1.20   0.232    -.0393892    .1624432
         y94 |   6.177316   .5601833    11.03   0.000     5.078892     7.27574
         y95 |   18.09267   .6905329    26.20   0.000     16.73865    19.44669
         y96 |    17.9404     .75684    23.70   0.000     16.45636    19.42443
         y97 |   15.19184   .7993116    19.01   0.000     13.62452    16.75915
         y98 |   29.88319   .8374619    35.68   0.000     28.24107    31.52531
       _cons |  -16.08091   23.80746    -0.68   0.499    -62.76329    30.60147
-------------+----------------------------------------------------------------
     sigma_u |  11.487395
     sigma_e |  8.9961899
         rho |  .61984689   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(549, 2741) =     5.75             Prob > F = 0.0000
```

- The lagged expenditure is still statistically significant.
- The coefficient on the lagged spending variable has gotten somewhat smaller, but its *t* statistic is still almost three. Therefore, there is still evidence of a lagged spending effect after controlling for unobserved district effects.

(v) They are jointly insignificant:

```
. test lunch lenrol

 ( 1)  lunch = 0

 ( 2)  lenrol = 0

       F(  2,  2741) =    0.73
            Prob > F =    0.4818
```

- The change in the coefficient and significance on the *lunch* variable is most dramatic.
- Both *enrol* and *lunch* are slow to change over time, which means that their effects are largely captured by the unobserved effect, *ai*. Plus, because of the time demeaning, their coefficients are hard to estimate.
- The spending coefficients can be estimated more precisely because of a policy change during this period, where spending shifted markedly in 1994 after the passage of Proposal A in Michigan, which changed the way schools were funded.

(vi)The Long-term effect is:

```
. nlcom _b[lrexpp]+_b[lrexpp_1]

       _nl_1:  _b[lrexpp]+_b[lrexpp_1]

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _nl_1 |   6.591808   2.637934     2.50   0.012     1.421552    11.76206
------------------------------------------------------------------------------
```

The estimated long-run spending effect is $\hat{\theta}_1 = 6.59$, $se(\hat{\theta}_1) = 2.64$.

Specifically, for each 1% increase in school expenditures, we expect the math4 scores to increase by approximately 6.59 percentage points, holding all other factors constant.

Estimating this yields:

```
. xtreg math4 lenrol lrexpp d.lrexpp lunch y94 y95 y96 y97 y98, fe

Fixed-effects (within) regression               Number of obs      =      3300
Group variable: distid                          Number of groups   =       550

R-sq:  within  = 0.6027                         Obs per group: min =         6
       between = 0.0399                                        avg =       6.0
       overall = 0.3202                                        max =         6

                                                F(9,2741)          =    462.02
corr(u_i, Xb)  = -0.0464                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lenrol |   .2450874   1.100381     0.22   0.824    -1.912572    2.402747
             |
      lrexpp |
         --. |   6.591808   2.637934     2.50   0.013     1.419268    11.76435
         D1. |  -7.002988   2.369184    -2.96   0.003    -11.64856   -2.357421
             |
       lunch |    .061527   .0514661     1.20   0.232    -.0393892    .1624432
         y94 |   6.177316   .5601833    11.03   0.000     5.078892     7.27574
         y95 |   18.09267   .6905329    26.20   0.000     16.73865    19.44669
         y96 |    17.9404     .75684    23.70   0.000     16.45636    19.42443
         y97 |   15.19184   .7993116    19.01   0.000     13.62452    16.75915
         y98 |   29.88319   .8374619    35.68   0.000     28.24107    31.52531
       _cons |  -16.08091   23.80746    -0.68   0.499    -62.76329    30.60147
-------------+----------------------------------------------------------------
     sigma_u |  11.487395
     sigma_e |  8.9961899
         rho |  .61984689   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(549, 2741) =     5.75           Prob > F = 0.0000
```

Which yields exactly the same what we obtained with nlcom.

Aditional Material

Extra questions to C14.8:
(vii)Do you think that the real expenditure per pupil is a strictly exogenous variable? Explain.

- The answer is no.
- Strict exogeneity requires that the error term is independent of a variable in all periods.
- This can not be the case if there is a feedback mechanism from success rate towards expenditure. For example, if districts with lower pass rate will get more expenditure in the future, then the present values of the error-term will be correlated with future values of real expenditure.

(vii) Estimate the model mentioned in part (i) using random effects and compare the random effects results with the fixed effects results. Perform a Hausman test on the equality of the FE and RE coefficients. What do you conclude?

(viii) The Hausman test:

```
. . xtreg math4 lenrol lrexpp lrexpp_1 lunch   y94 y95 y96 y97 y98, fe

Fixed-effects (within) regression               Number of obs      =      3300
Group variable: distid                          Number of groups   =       550

R-sq:  within  = 0.6027                         Obs per group: min =         6
       between = 0.0399                                        avg =       6.0
       overall = 0.3202                                        max =         6

                                                F(9,2741)          =    462.02
corr(u_i, Xb)  = -0.0464                        Prob > F           =    0.0000

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lenrol |   .2450874   1.100381     0.22   0.824    -1.912572    2.402747
      lrexpp |  -.4111804   2.457658    -0.17   0.867     -5.23023    4.407869
    lrexpp_1 |   7.002988   2.369184     2.96   0.003     2.357421    11.64856

       lunch |    .061527   .0514661     1.20   0.232    -.0393892     .1624432
         y94 |   6.177316   .5601833    11.03   0.000     5.078892      7.27574
         y95 |   18.09267   .6905329    26.20   0.000     16.73865     19.44669
         y96 |    17.9404     .75684    23.70   0.000     16.45636     19.42443
         y97 |   15.19184   .7993116    19.01   0.000     13.62452     16.75915
         y98 |   29.88319   .8374619    35.68   0.000     28.24107     31.52531
       _cons |  -16.08091   23.80746    -0.68   0.499    -62.76329     30.60147
-------------+----------------------------------------------------------------
     sigma_u |  11.487395
     sigma_e |  8.9961899
         rho |  .61984689   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(549, 2741) =     5.75          Prob > F = 0.0000
```

```
. estimate store fe

xtreg math4 lenrol lrexpp lrexpp_1 lunch   y94 y95 y96 y97 y98, re

Random-effects GLS regression                   Number of obs      =      3300
Group variable: distid                          Number of groups   =       550

R-sq:  within  = 0.5941                         Obs per group: min =         6
       between = 0.3873                                        avg =       6.0
       overall = 0.5018                                        max =         6

                                                Wald chi2(9)       =   4335.94
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
       math4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      lenrol |   .7867612   .3467126     2.27   0.023     .1072169    1.466306
      lrexpp |   .3821852   2.060344     0.19   0.853    -3.656014    4.420384
    lrexpp_1 |   7.805654   1.925119     4.05   0.000      4.03249    11.57882
       lunch |  -.3337942   .0226692   -14.72   0.000     -.378225   -.2893635
         y94 |   6.356631   .5595344    11.36   0.000     5.259963    7.453298
         y95 |   18.64244   .6295381    29.61   0.000     17.40856    19.87631
         y96 |   18.20443   .6514805    27.94   0.000     16.92755    19.48131
         y97 |   15.51763   .6733494    23.05   0.000     14.19789    16.83737
         y98 |   30.54315   .6905471    44.23   0.000      29.1897     31.8966
       _cons |  -23.22425   14.75159    -1.57   0.115    -52.13683    5.688332
-------------+----------------------------------------------------------------
     sigma_u |  7.8856386
     sigma_e |  8.9961899
         rho |  .43449962   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Utrecht University

```
. estimate store re

. hausman fe re

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       fe           re         Difference          S.E.
-------------+----------------------------------------------------------------
      lenrol |    .2450874     .7867612       -.5416739        1.044331
      lrexpp |   -.4111804     .3821852       -.7933657        1.339801
    lrexpp_1 |    7.002988     7.805654       -.8026657        1.380924
       lunch |     .061527    -.3337942        .3953212        .0462046
         y94 |    6.177316     6.356631       -.1793149        .0269548
         y95 |    18.09267     18.64244       -.5497643         .283756
         y96 |     17.9404     18.20443       -.2640343        .3852012

         y98 |    29.88319     30.54315       -.6599615        .4738007
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(9) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =       74.27
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```

We find that the two estimators lead to statistically different results. Since FE is consistent under less stringent assumptions (it allows for a correlation between the explanatory variables and the district-specific effect), we should prefer FE over RE.