

Panel Data

Sometimes there are data sets in existence that follow the same economic agent (for example individual, firm, industry or country) over a period of time. These panel (or longitudinal) data can be useful if we are interested in the following sort of questions

Does a 10% unemployment rate mean that the same 10% of the labour force are unemployed all the time or that at any point in time a random 10% of the labour force will be unemployed?

Is firm growth caused by economies of scale (cross-section variation in input size) or technical change (time series variation given fixed inputs)

Only by following the same individuals over time can we deduce the answer

Panel data allow us to control for *unobserved individual effects* which may otherwise bias the estimation.

Since unobserved or omitted variables are often responsible for endogeneity and inconsistent OLS estimates, the availability of data sets that follow the same set of agents (individuals, firms) over time can be used to remove the influence of these *unobservables* from regressions and so produce consistent estimates.

Given

$$Y_{it} = a_t + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

The idea is to somehow purge the equation of the individual-specific component ϕ_i .

This would be impossible in a single cross section and so the presence of unobservables means that

$$\text{Cov}(X,u) = \text{Cov}(X, \phi_i + e_i) \neq 0$$

One simple approach is to assume that ϕ_i is constant over time (a *fixed effect*), so that by differencing the above the unobservable effect disappears and an OLS regression on the 1st difference gives consistent estimates of b_1 .

In period 1

$$Y_{i1} = a_1 + b_1 X_{i1} + \phi_i + e_{i1}$$

In period 2

$$Y_{i2} = a_2 + b_1 X_{i2} + \phi_i + e_{i2}$$

So

$$[Y_{i2} - Y_{i1}] = (a_2 + b_1 X_{i2} + \phi_i + e_{i2}) - (a_1 + b_1 X_{i1} + \phi_i + e_{i1})$$

or
$$\Delta Y = \delta + b_1 \Delta X + \Delta e \quad (\text{if } a_1 = a_2)$$

Hence the change in the dependent variable will not be influenced by the unobservables

IF

the unobservable effect stays fixed over time

(note that this technique also removes *any* variable that stays constant over time.

With more than two time periods, then subtract data on each unit at time 1 from data on the same unit at time 2 *and* subtract data on each unit at time 2 from data on the same unit at time 3 and then pool these differenced observations.

The easiest way to do this is to sort your data by individual unit and then time, (using a command like *sort idcode time* in Stata), so that the 1st observation in the data is the 1st unit at time 1, the 2nd observation is the 1st unit at time 2 etc.

Example.

Using the *train.dta* dataset which is a 3 year panel of firms with information on sales, employment and union recognition

You sort the data by firm and by year and generate the first differences in the variables using the following commands

```
. sort fcode year

. g dsales=sales-sales[_n-1] if fcode==fcode[_n-1]
(119 missing values generated)

/* Note the if command ensures that lagged values from other firms are not
assigned to the first observation of each new firm */

. g dunion=union-union[_n-1] if fcode==fcode[_n-1]
(119 missing values generated)

. list year fcode sales dsales union dunion in 91/111
```

| | year | fcode | sales | dsales | union | dunion |
|------|------|--------|----------|----------|-------|--------|
| 91. | 1987 | 410609 | 1650831 | . | 0 | . |
| 92. | 1988 | 410609 | 1817961 | 167130 | 0 | 0 |
| 93. | 1989 | 410609 | 1642441 | -175520 | 0 | 0 |
| 94. | 1987 | 410612 | 7000000 | . | 0 | . |
| 95. | 1988 | 410612 | 8500000 | 1500000 | 0 | 0 |
| 96. | 1989 | 410612 | 11000000 | 2500000 | 0 | 0 |
| 97. | 1987 | 410626 | 4600000 | . | 1 | . |
| 98. | 1988 | 410626 | 4900000 | 300000 | 1 | 0 |
| 99. | 1989 | 410626 | 5600000 | 700000 | 1 | 0 |
| 100. | 1987 | 410627 | 2900000 | . | 1 | . |
| 101. | 1988 | 410627 | 2800000 | -100000 | 1 | 0 |
| 102. | 1989 | 410627 | 2900000 | 100000 | 1 | 0 |
| 103. | 1987 | 410629 | 1100000 | . | 0 | . |
| 104. | 1988 | 410629 | 2050000 | 950000 | 0 | 0 |
| 105. | 1989 | 410629 | 2260000 | 210000 | 0 | 0 |
| 106. | 1987 | 410635 | 20000000 | . | 1 | . |
| 107. | 1988 | 410635 | 18000000 | -2000000 | 1 | 0 |
| 108. | 1989 | 410635 | 16000000 | -2000000 | 1 | 0 |
| 109. | 1987 | 410636 | 386807 | . | 0 | . |
| 110. | 1988 | 410636 | 734613 | 347806 | 0 | 0 |
| 111. | 1989 | 410636 | 518842 | -215771 | 0 | 0 |

Note that the value of the differenced dummy variable does not change but the differenced sales variable does.

It is also normal to include a constant in these differenced regressions, (even though differencing removes all constants).

The way to interpret the constant is that it represents the *change* in the value of the intercept over time, ie $a_1 \neq a_2$. (Remember also that the absence of a constant in a regression no longer restricts the R^2 coefficient to lie between 0 and 1).

Example using the data set train.dta we regress the log of sales on the log of employment in 1988

```
. reg lsales lempl union if year==1988
```

| Source | SS | df | MS | | | |
|----------|------------|-----------|------------|------------------------|----------------------|----------|
| Model | 92.2537446 | 2 | 46.1268723 | Number of obs = 115 | | |
| Residual | 50.224002 | 112 | .448428589 | F(2, 112) = 102.86 | | |
| | | | | Prob > F = 0.0000 | | |
| | | | | R-squared = 0.6475 | | |
| | | | | Adj R-squared = 0.6412 | | |
| | | | | Root MSE = .66965 | | |
| lsales | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
| lemploy | .8379075 | .0645985 | 12.97 | 0.000 | .7099139 | .9659011 |
| union | .2754602 | .1595039 | 1.73 | 0.087 | -.0405763 | .5914967 |
| _cons | 12.03388 | .2276874 | 52.85 | 0.000 | 11.58275 | 12.48502 |

and in 1989

```
. reg lsales lempl union if year==1989
```

| Source | SS | df | MS | | | |
|----------|------------|-----------|------------|------------------------|----------------------|----------|
| Model | 102.849229 | 2 | 51.4246145 | Number of obs = 115 | | |
| Residual | 37.7492552 | 112 | .337046922 | F(2, 112) = 152.57 | | |
| | | | | Prob > F = 0.0000 | | |
| | | | | R-squared = 0.7315 | | |
| | | | | Adj R-squared = 0.7267 | | |
| | | | | Root MSE = .58056 | | |
| lsales | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
| lemploy | .9069097 | .0555367 | 16.33 | 0.000 | .7968708 | 1.016949 |
| union | .1570459 | .1372818 | 1.14 | 0.255 | -.1149604 | .4290522 |
| _cons | 11.85956 | .1999808 | 59.30 | 0.000 | 11.46332 | 12.25579 |

The effect of union recognition is more pronounced in 1988 and the effect of firm size is larger in 1989. However we can't discount the probability that these regressions suffer from omitted variable bias.

So taking the difference of the variables gives

```
. reg dlsales dlemp dunion if year==1989
```

| Source | SS | df | MS | | | |
|----------|------------|-----|------------|-----------------|--------|--|
| Model | 3.72177543 | 1 | 3.72177543 | Number of obs = | 115 | |
| Residual | 14.7658904 | 113 | .130671597 | F(1, 113) = | 28.48 | |
| Total | 18.4876659 | 114 | .162172508 | Prob > F = | 0.0000 | |
| | | | | R-squared = | 0.2013 | |
| | | | | Adj R-squared = | 0.1942 | |
| | | | | Root MSE = | .36149 | |

| dlsales | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|---------|-----------|-----------|------|-------|----------------------|----------|
| dlemp | .6750276 | .1264844 | 5.34 | 0.000 | .4244392 | .9256161 |
| dunion | (dropped) | | | | | |
| _cons | .0580459 | .0347168 | 1.67 | 0.097 | -.0107345 | .1268262 |

It can't be that any fixed unobservables influence the change in sales so the change in sales must have been caused by other factors, of which employment appears to be significant.

Note that the regression is done using the differenced values for the later year in the sample. Differences for first year in sample do not exist.

Note that the coefficient on employment in the change regression is lower than in either single year regression suggesting the estimated employment effect in these earlier regressions was biased up by the presence of omitted variables.

Note also that you cannot estimate the union effect, because, like the fixed effect, it is constant over time and so is differenced away.

Panel data can be useful addition to the problem of **policy evaluation** outlined in earlier lectures. If the same agents appear in the data before and after an event then the difference in difference estimator will also net out any fixed effects that might otherwise influence the results. (Note that this also applies if the data are pooled and year dummy/policy interactions used instead, since the fixed effects drop out in this formulation).

Using the earlier example suppose there is now an individual fixed effect to add to the equations

$$\begin{aligned} \ln W_1 &= a_1 + b_1 \text{Treatment Dummy Variable}_1 + \phi_i && \text{Period Before} \\ \ln W_2 &= a_2 + b_2 \text{Treatment Dummy Variable}_2 + \phi_i && \text{Period After} \end{aligned}$$

The coefficients b_1 and b_2 give the differential impact of the treatment group on wages in each period.

The difference between these two coefficients gives the “difference in difference” estimator – the change in the treatment effect following an intervention.

The change in wages for the treatment group is

$$(a_2 + b_2 + \phi_i) - (a_1 + b_1 + \phi_i) = a_2 - a_1 + b_2 - b_1$$

and the change in wages for the control group is

$$(a_2 + \phi_i) - (a_1 + \phi_i) = a_2 - a_1$$

so the “difference in difference” estimator

$$\begin{aligned} &= \text{Change in wages for treatment} - \text{change in wages for control} \\ &= (a_2 - a_1 + b_2 - b_1) - (a_2 - a_1) \\ &= b_2 - b_1 \end{aligned}$$

ie the fixed effect drops out

One problem with the 1st differencing approach is that it can generate autocorrelation in the differenced error term.

$$\begin{aligned} \text{If } & \epsilon_t = e_t - e_{t-1} \quad \text{and} \\ \text{then } & \text{Cov}(\epsilon_t, \epsilon_{t-1}) = \text{Cov}(e_t - e_{t-1}, e_{t-1} - e_{t-2}) \neq 0 \end{aligned}$$

One solution is to include more time-varying variables that could account for the autocorrelation, (which often stems from missing variables in an equation).

There are at least two other ways of obtaining fixed effects estimates of b_1 .

The first is to pool the data across years and estimate (1) directly by including a dummy variable for each individual in the data to capture the fixed effect, (*least squares dummy variables*). This may be rather wasteful of degrees of freedom and will usually produce inconsistent estimates of the dummy variables (ie the fixed effects) if the time dimension of the panel is small, (which is usually the case).

$$Y_{it} = a_t + b_1 X_{it} + g_1 D_1 + g_2 D_2 + \dots g_{n-1} D_{N-1} + e_{it}$$

Where $D_i = 1$ for individual i
 $= 0$ for everybody else

For these reasons, the most-commonly used alternative method – which can be used to obtain consistent estimates of the fixed effects if desired - is to calculate the mean value for each observation for each individual and subtract the observation at time t from this mean.

If $Y_{it} = a_t + b_1 X_{it} + \phi_i + e_{it}$ (1)

then $\bar{Y}_i = a_1 + b_1 \bar{X}_i + \phi_i + \bar{e}_i$

(since the mean value of something that is constant over time is the value itself so $\bar{f}_i = f_i$)

$$Y_{it} - \bar{Y}_i = b_1 (X_{it} - \bar{X}_i) + (e_{it} - \bar{e}_i)$$

This *within-group estimator* approach also removes the fixed effect, (because the mean of the fixed effect is the same as the individual fixed effect value), and avoids the problem of introducing autocorrelation into the residuals, (since the mean value of the residual should be zero).

Problems with this approach arise if there is variation in the X variables across individuals, but less variation over time. Even for variables that do vary a little over time, inclusion of fixed effects will

produce estimates on the X variables that are close to zero. The fixed effect picks up the possibly true impact of variables that move only a little over time.

Within-Groups or First Difference?

Within-groups estimates can be quite sensitive, (though remain consistent), with large T and small N dimensions to the panel. First differencing means variables more likely to be stationary and not suffer from spurious regression. In practice it is probably better to do both to test the sensitivity of the results.

Note that the two methods will produce identical estimates when there are 2 time periods in the data.

Example

Consider the 1st difference regression in a 2 year panel, (*panel2.dta*) of the change in log sales on the change in log employment, (the `noconst` option in Stata removes the constant from the regression).

```
. reg clsales clemp if year==1988, noconst
```

| Source | SS | df | MS | Number of obs = | 115 |
|----------|------------|-----------|------------|-----------------|----------------------|
| Model | 6.61023185 | 1 | 6.61023185 | F(1, 114) = | 32.05 |
| Residual | 23.5097718 | 114 | .206226068 | Prob > F = | 0.0000 |
| ----- | | | | R-squared = | 0.2195 |
| Total | 30.1200036 | 115 | .261913075 | Adj R-squared = | 0.2126 |
| ----- | | | | Root MSE = | .45412 |
| clsales | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
| clemp | .8395683 | .1482926 | 5.66 | 0.000 | .5458018 1.133335 |

Now the within-group estimate gives identical slope estimate

```
. xtreg lsales lemp if year<1989, fe i(fcode)
```

```
Fixed-effects (within) regression      Number of obs   =      230
Group variable (i) : fcode             Number of groups =      115

R-sq:  within = 0.2195                  Obs per group:  min =      2
        between = 0.6957                  avg   =      2.0
        overall = 0.6704                  max   =      2

corr(u_i, Xb) = 0.0794                  F(1,114)        =      32.05
                                          Prob > F         =      0.0000
```

| lsales | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] |
|---------|-----------|-----------------------------------|-------|-------|----------------------|
| lemp | .8395692 | .1482924 | 5.66 | 0.000 | .5458032 1.133335 |
| _cons | 12.06415 | .5150927 | 23.42 | 0.000 | 11.04375 13.08454 |
| sigma_u | .60128965 | | | | |
| sigma_e | .32111175 | | | | |
| rho | .77809084 | (fraction of variance due to u_i) | | | |

```
F test that all u_i=0:      F(114, 114) =      6.97      Prob > F = 0.0000
```

However when the data are extended to 3 time periods, the results no longer co-incide.

```
. reg clsales clempl if year==1988 | year==1989, noc
```

| Source | SS | df | MS | | | |
|----------|------------|-----|------------|-----------------|--------|--|
| Model | 11.1708402 | 1 | 11.1708402 | Number of obs = | 235 | |
| Residual | 38.7381982 | 234 | .165547856 | F(1, 234) = | 67.48 | |
| | | | | Prob > F = | 0.0000 | |
| | | | | R-squared = | 0.2238 | |
| | | | | Adj R-squared = | 0.2205 | |
| Total | 49.9090384 | 235 | .212378887 | Root MSE = | .40688 | |

| clsales | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|---------|----------|-----------|------|-------|----------------------|----------|
| clempl | .7845282 | .0955053 | 8.21 | 0.000 | .5963681 | .9726882 |

```
. xtreg lsales lemp, fe i(fcode)
```

```
Fixed-effects (within) regression      Number of obs   =      345
Group variable (i) : fcode             Number of groups =      115
R-sq:  within = 0.3029                  Obs per group:  min =       3
      between = 0.7162                      avg =      3.0
      overall  = 0.6901                      max =       3
corr(u_i, Xb) = 0.1611                   F(1,229)        =     99.52
                                          Prob > F         =     0.0000
```

| lsales | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|--------|----------|-----------|-------|-------|----------------------|----------|
| lemp | .8092035 | .0811145 | 9.98 | 0.000 | .6493773 | .9690297 |
| _cons | 12.19437 | .2850389 | 42.78 | 0.000 | 11.63273 | 12.756 |


```
sigma_u | .58411543
sigma_e | .28530622
rho     | .80737938 (fraction of variance due to u_i)
```

```
F test that all u_i=0:      F(114, 229) =      12.25      Prob > F = 0.0000.
```

Random Effects Estimation

The alternative way of dealing with unobserved effects is to assume that they form part of the residual

$$Y_{it} = a_t + b_1 X_{it} + \phi_i + e_{it} \quad (1)$$

So that $u_{it} = \phi_i + e_{it}$

consists of an part that is individual specific but does not vary over time, ϕ_i and a part that varies across both individuals and time, e_{it}

This gives us an explicit functional form for the error term which can be modelled by a **Generalised Least Squares** estimator (rather like with autocorrelation and FGLS) that takes account of this correlation. This is the random effects estimator

Using the same data as above

```
xtreg lsales lemp, re
```

```
Random-effects GLS regression           Number of obs   =       345
Group variable (i): fcode              Number of groups =       115

R-sq:  within = 0.3029                 Obs per group:  min =        3
        between = 0.7162                    avg =       3.0
        overall = 0.6901                    max =        3

Random effects u_i ~ Gaussian          Wald chi2(1)    =     383.87
corr(u_i, X) = 0 (assumed)            Prob > chi2     =     0.0000
```

| lsales | Coef. | Std. Err. | z | P> z | [95% Conf. Interval] | |
|--------|----------|-----------|-------|-------|----------------------|----------|
| lemp | .8750523 | .0446621 | 19.59 | 0.000 | .7875161 | .9625885 |
| _cons | 11.96331 | .1657475 | 72.18 | 0.000 | 11.63845 | 12.28817 |

```
sigma_u | .55482492
sigma_e | .28530622
rho     | .79087026   (fraction of variance due to u_i)
```

Can see the employment variable is more significant than in the fixed effects estimation

(Note that there is no R^2 in this estimation)

Fixed or Random Effects?

The latter will allow estimation of the effects of time-invariant variables.

```
. xtreg lsale lemp union
```

```
Random-effects GLS regression           Number of obs   =       345
Group variable (i): fcode              Number of groups =       115

R-sq:  within = 0.3029                 Obs per group:  min =        3
        between = 0.7226                    avg =       3.0
        overall = 0.6962                    max =        3

Random effects u_i ~ Gaussian          Wald chi2(2)    =     391.59
corr(u_i, X) = 0 (assumed)            Prob > chi2    =     0.0000
```

| lsales | Coef. | Std. Err. | z | P> z | [95% Conf. Interval] | |
|---------|---|-----------|-------|-------|----------------------|----------|
| lemploy | .8529599 | .046184 | 18.47 | 0.000 | .762441 | .9434789 |
| union | .2383114 | .1350074 | 1.77 | 0.078 | -.0262983 | .5029211 |
| _cons | 11.98902 | .1654712 | 72.45 | 0.000 | 11.6647 | 12.31334 |
| sigma_u | .55061581 | | | | | |
| sigma_e | .28530622 | | | | | |
| rho | .78834004 (fraction of variance due to u_i) | | | | | |

```
. xtreg lsale lemp union, fe
```

```
Fixed-effects (within) regression      Number of obs   =       345
Group variable (i): fcode              Number of groups =       115

R-sq:  within = 0.3029                 Obs per group:  min =        3
        between = 0.7162                    avg =       3.0
        overall = 0.6901                    max =        3

corr(u_i, Xb) = 0.1611                 F(1,229)       =     99.52
                                         Prob > F       =     0.0000
```

| lsales | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|---------|---|-----------|-------|-------|----------------------|----------|
| lemploy | .8092035 | .0811145 | 9.98 | 0.000 | .6493773 | .9690297 |
| union | (dropped) | | | | | |
| _cons | 12.19437 | .2850389 | 42.78 | 0.000 | 11.63273 | 12.756 |
| sigma_u | .58411543 | | | | | |
| sigma_e | .28530622 | | | | | |
| rho | .80737938 (fraction of variance due to u_i) | | | | | |

```
F test that all u_i=0:      F(114, 229) =    11.97      Prob > F = 0.0000
```

The former allows for correlation with the X variables, the latter does not.

The answer depends on whether you believe the unobservables are likely to be correlated with the X variables. If you think they are

use the fixed effects estimator. If not use random effects. (If they are and you use random effects you may get biased estimates because $Cov(X,u) = Cov(X_{it}, \phi_i + e_{it}) \neq 0$ and this leads to endogeneity bias.

There is a test to help determine which method to use

Hausman test.

Under null that error are uncorrelated with x variables then both random and fixed effects estimators are both consistent.

If the null is false then only fixed effects is consistent.

Test is therefore based around a comparison of the estimates, allowing for sampling variation. If the estimates are sufficiently different, conclude that random effects assumption is untenable.

Stata will do this test automatically after the random effects command, just type.

```
. xthausman
Hausman specification test
----- Coefficients -----
      lsales |      Fixed      Random
             |      Effects      Effects      Difference
-----+-----
      lemploy |      .8092035      .8750523      -.0658488
Test:  Ho:  difference in coefficients not systematic

      chi2( 1) = (b-B)'[S^(-1)](b-B), S = (S_fe - S_re)
           =      0.95
      Prob>chi2 =      0.3308
```

In this case can't reject null that 2 estimation strategies produce different results. Conclude random effects is better.

Heteroskedasticity and Autocorrelation

It is possible that panel data suffer from both heteroskedasticity and autocorrelation (because of the combined cross-section time series nature of the data)

With 1st differenced data then testing for either of these is as before. Just apply the White robust estimator or the Breusch-Godfrey test to the 1st differenced residuals