Notes Grouped Panel Data

While the term panel data refers to a data structure in which we observe individuals across time, it is sometimes very useful to use panel data methods with other data structures. The main two features of the data must be:

- Observations can be divided in groups that share some features.
- We observe a large number of groups.

Examples might include: firms that are grouped by sector or investors grouped by geographical location. To illustrate the methods, consider we are interested to know how sensitive to performance is CEOs' compensation, and we only observe a cross-section of US CEOs:

$$y_i = \alpha + \gamma x_i + \varepsilon_i \tag{1}$$

where:

- y_i is the increase in compensation of manager i.
- x_i is the return on equity of CEO i's firm.

The first possibility is to run OLS and estimate γ:

```
. reg changecomp roe
```

Source	SS	df	MS		
Model	232149.856	1	232149.856	Number of obs	= 4,982
Residual	6859717.35	4,980	1377.45328	F(1, 4980)	= 168.54
				Prob > F	= 0.0000
Total	7091867.2	4,981	1423.78382	R-squared	= 0.0327
				Adj R-squared	= 0.0325
				Root MSE	= 37.114

changecomp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
roe	20.49241	1.57851	12.98	0.000	17.39784	23.58699
_cons	2.567188	.5260385	4.88	0.000	1.535921	3.598455

We might be concerned, however, that shareholders use an industry benchmark to pay CEOs, that is:

$$y_i = \beta(x_i - r_g) + u_i = \eta_g + \beta x_i + u_i$$

where r_g is the return on the industry g portfolio (that we do not observe). If this is the case, the inference based on OLS is wrong as $cov(\varepsilon_i, \varepsilon_j) \neq 0$ if i and j belong to the same industry. Instead, we can apply a proper panel data method. In particular, if we believe $cov(x_i, \eta_g) = 0$ we can use random effects; otherwise we must use the FE estimator. The FD estimator is not implemented as observations inside a panel (industry) do not usually have a natural ordering. These are the results with both types of estimators:

```
. xtreg changecomp roe,fe

Fixed-effects (within) regression               Number of obs      =       4,982
Group variable: naics                           Number of groups   =          81

R-sq:                                           Obs per group:
     within  = 0.0181                                         min =           2
     between = 0.2339                                         avg =        61.5
     overall = 0.0327                                         max =         707

                                                F(1,4900)          =       90.51
corr(u_i, Xb)  = 0.1462                         Prob > F           =      0.0000

------------------------------------------------------------------------------
  changecomp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         roe |   16.57201   1.741899     9.51   0.000     13.15711    19.98691
       _cons |   2.529485    .515107     4.91   0.000     1.519644    3.539325
-------------+----------------------------------------------------------------
     sigma_u |  16.157275
     sigma_e |   36.33871
         rho |  .16506338   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(80, 4900) = 3.68                     Prob > F = 0.0000
```

```
. xtreg changecomp roe,re

Random-effects GLS regression                   Number of obs      =       4,982
Group variable: naics                           Number of groups   =          81

R-sq:                                           Obs per group:
     within  = 0.0181                                         min =           2
     between = 0.2339                                         avg =        61.5
     overall = 0.0327                                         max =         707

                                                Wald chi2(1)       =      105.53
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =      0.0000
```

changecomp	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
roe	17.5122	1.704701	10.27	0.000	14.17105	20.85335
_cons	2.24985	1.611876	1.40	0.163	-.9093687	5.409069
sigma_u	12.388633					
sigma_e	36.33871					
rho	.10412503	(fraction of variance due to u_i)				

In reality shareholders might not only use returns of the industry portfolio as benchmark, but also returns on direct competitors or a narrower definition of industry:

$$y_i = \beta(x_i - r_g) + \theta(x_i + r_c) + v_i = \eta_g + \iota_c + \lambda x_i + v_i$$

where r_c is the return of the most direct competitor or the performance of a sub-industry. In contrast to the previous case we do not know exactly the level of the benchmark (sub-industry, sub-sub-industry, competitor...); therefore we cannot eliminate ι_c as in the FE estimator, neither we know the actual correlation matrix of u_i to use random effects. Nonetheless, we know that our previous estimation does not lead to the correct inference since it assumes $cov(u_i, u_j) = 0$ if $i \neq j$ and it is not true if i and j belong to the same cluster c.

In the case that $cov(\iota_c, x_i|\eta_g) \neq 0$, x_i is endogenous and we would need an instrument to perform inference. However, under the assumption that $cov(\iota_c, x_i|\eta_g) = 0$, the estimate is consistent, therefore we just need to correct the standard errors to account for the existence of correlation inside an industry. This can be easily done in Stata using `vce(cluster industry)` where industry is a variable that takes the values of g.

```
. xtreg changecomp roe,fe vce(cluster naics)

Fixed-effects (within) regression               Number of obs      =       4,982
Group variable: naics                           Number of groups   =          81

R-sq:                                           Obs per group:
     within  = 0.0181                                         min =           2
     between = 0.2339                                         avg =        61.5
     overall = 0.0327                                         max =         707

                                                F(1,80)            =       42.96
corr(u_i, Xb)  = 0.1462                         Prob > F           =      0.0000

                                    (Std. Err. adjusted for 81 clusters in naics)
------------------------------------------------------------------------------
             |               Robust
  changecomp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         roe |    16.57201   2.528311     6.55   0.000     11.54051    21.60351
       _cons |    2.529485   .0243154   104.03   0.000     2.481096    2.577874
-------------+----------------------------------------------------------------
     sigma_u |   16.157275
     sigma_e |    36.33871
         rho |    .16506338   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The resulting standard error is much higher which indicates that, indeed there is some within group correlation.

Cluster Variance

Consider the model:

$$y_i = \mu + \beta x_i + w_i \quad i = 1, ..., N \tag{2}$$

Moreover, $cov(w_i, w_j)$ might be different from 0 if i and j belong to the same group g. If we follow the previous example, using the FE transformation, y_i is the CEO's compensation minus the industry mean compensation and x_i is the difference between the firm and the industry returns. The OLS estimator corresponds to

$$\hat{\beta} = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2} = \beta + \frac{\frac{1}{N}\sum_{i=1}^{N} w_i(x_i - \bar{x})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

If we assume $\mathbb{E}(w_i|X)$, we obtain the usual result that $\hat{\beta}$ is unbiased:

$$\mathbb{E}(\hat{\beta}) = \beta$$

In this case, we are interested in the variance of the estimator. For that we need an extra assumption: $cov(w_i, w_j) = 0$ if i and j belong to different groups. Since β is deterministic ($Var(\beta|x_1, ..., x_N) = 0$):

$$Var(\hat{\beta}|x_1, ..., x_N) = Var\left(\frac{\frac{1}{N}\sum_{i=1}^{N} w_i(x_i - \bar{x})}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}|x_1, ..., x_N\right) = \frac{Var\left(\frac{1}{N}\sum_{i=1}^{N} w_i(x_i - \bar{x})|x_1, ..., x_N\right)}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^2}$$

The denominator is straightforward to compute as we observe x_i. The numerator, however, is more convoluted as it involves the variance of a sum; thus variances and covariances. Nonetheless, we have assumed that two observations in different groups are uncorrelated which simplifies the expression:

$$\begin{aligned} Var\left(\frac{1}{N}\sum_{i=1}^{N} w_i(x_i - \bar{x})|x_1, ..., x_N\right) =& Var\left(\frac{1}{G}\sum_{g=1}^{G}\frac{1}{N_g}\sum_{i\in g} w_i(x_i - \bar{x})|x_1, ..., x_N\right) \\ =& \frac{1}{G^2}\sum_{g=1}^{G} Var\left(\frac{1}{N_g}\sum_{i\in g} w_i(x_i - \bar{x})|x_1, ..., x_N\right) \end{aligned}$$

where $\sum_{i\in g}$ indicates that we sum every observation in group g, and N_g is the number of observations in group g. If G is large we can estimate the standard error of β as:

$$\hat{se}(\hat{\beta}|x_1, ..., x_N) = \frac{\sqrt{\frac{1}{G^2}\sum_{g=1}^{G}\left(\frac{1}{N_g}\sum_{i\in g} w_i(x_i - \bar{x})\right)^2}}{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{3}$$

Using this formula we allow any type of dependence between observations inside the same group.

Special case: common correlation

While (3) is the usual formula due to its flexibility, it is very hard to interpret. It simplifies significantly, however, if we assume $corr(w_i, w_j|x_1, ..., x_N) = \rho$ if i and j belong to the same group and zero otherwise. That is, every observation inside a cluster has the same

correlation. In this case, the formula reduces to:

$$\hat{se}(\hat{\beta}|x_1,...,x_N) = \sqrt{\left[1 + \left(\frac{Var(N_g)}{\bar{N}} + \bar{N} - 1\right)\rho_x\rho\right]} \sqrt{\left[\frac{\sum_{i=1}^{N} w_i^2(x_i - \bar{x})^2}{\left(\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^2}\right]} \quad (4)$$

where ρ_x is the intragroup correlation of x and $\bar{N}$ is the mean size of the groups. Note that the second part of the equation (blue) is the usual OLS standard error robust to heteroscedasticity. The first part is the "inflation" factor due to group intracorrelation, also known as Moulton factor.

We observe that if residuals are not correlated ($\rho = 0$) the formula reduces to the standard one. This is also the case if the explanatory variables inside a group are uncorrelated ($\rho_x = 0$) even if the errors are not. The problem arises when **both**, explanatory variables and errors present intragroup correlation, and it is actually worse the higher are these correlations. Furthermore, in the very likely scenario that both correlations have the same sign ($\rho_x\rho > 0$), the usual OLS standard error formula is biased downwards; as a consequence the t-statistic is biased upwards which lead us to "too much" rejection of the null hypothesis.

One relevant case is when the explanatory variable does not change within a group ($\rho_x = 1$). For instance, we might want to know the effect of industry R&D investment on firm performance, or the effect on wages of state market conditions. The left panel in the following table summarizes the Moulton factor in this case (assuming $V(N_g) = 0$, or equivalently that every group has the same number of members in the sample):

	Moulton Factor			Type-I Error		
N/ρ	0.05	0.50	0.95	0.05	0.50	0.95
2	1.05	1.5	1.95	6%	19%	31%
5	1.20	3.00	4.80	10%	51%	68%
10	1.45	5.50	9.55	18%	72%	84%
50	3.45	25.50	47.55	57%	94%	97%
100	5.95	50.50	95.05	74%	97%	98%

We observe that the factor is high even if the correlation between the errors is small and groups are not too big. In fact, it is not uncommon to have more than 10 individuals inside a group (consider how many firms can be in a given industry). The right side of

the table states the probability of type-I error if we do not correct for the intragroup correlation when testing $\beta = 0$ at the 5% confidence level. Recall that we expect this probability to be 5% but, indeed it already doubles with 5 individuals per group and a small correlation between the errors. Since the distortion is huge even for low level of intragroup correlation, it is highly recommended to use the cluster-robust standard error formula if at least one of our covariates does not vary within a cluster.

Back to Panel Data

The same correction for the variance can be applied in the usual panel data set-up in which we observe individuals across time periods. Note that this correction takes into account correlation within a cluster (individual); therefore we are indeed correcting for time series correlation. Consider we extend the sample, and we actually observe a panel of CEOs and their compensation. We want to estimate the following model:

$$y_{i,t} = \delta_i + \eta_g + \iota_c + \beta x_{i,t} + \varepsilon$$

where $y_{i,t}$ is the increase in compensation at time t and $x_{i,t}$ is the return on equity during that year. Moreover, the model takes into account that CEOs' skills might influence the level of wage increments (δ_i). We also assume that $cov(\varepsilon_{i,t}, \varepsilon_{j,s}) = 0$ for all s and t, and all $i \neq j$; furthermore we assume $\mathbb{E}(\varepsilon|X) = 0$ and $Var(\varepsilon|X) = \sigma^2$ where X indicates the set of all observations.

If we apply the FE transformation we get:

$$y_{i,t} - \frac{1}{T}\sum_{s=1}^{T} y_{i,s} = \beta(x_{i,t} - \frac{1}{T}\sum_{s=1}^{T} x_{i,s}) + \varepsilon - \frac{1}{T}\sum_{s=1}^{T} \varepsilon_{i,s}$$

that can be relabeled as:

$$\ddot{y}_{i,t} = \beta \ddot{x}_{i,t} + \ddot{\varepsilon}_{i,t} \tag{5}$$

Under the reasonable assumption that CEOs do not move across industries or clusters, the FE transformation eliminates the unobserved component due to industry and cluster even if we do not observe the actual cluster definition. One possible estimation strategy is to use OLS on equation (5):

```
. reg DMchangecomp DMroe,noconstant
```

Source	SS	df	MS		
Model	2379316.06	1	2379316.06	Number of obs	= 119,421
Residual	125338290	119,420	1049.55861	F(1, 119420)	= 2266.97
				Prob > F	= 0.0000
				R-squared	= 0.0186
				Adj R-squared	= 0.0186
Total	127717606	119,421	1069.47359	Root MSE	= 32.397

DMchangecomp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
DMroe	14.74568	.3097006	47.61	0.000	14.13867	15.35269

or we can use Stata FE estimation package (`xtreg x,fe`):

```
. xtreg changecomp roe,fe

Fixed-effects (within) regression               Number of obs      =       119,421
Group variable: execnum                         Number of groups   =        28,219

R-sq:                                           Obs per group:
     within  = 0.0186                                         min =             1
     between = 0.0370                                         avg =           4.2
     overall = 0.0250                                         max =            24

                                                F(1,91201)         =       1731.28
corr(u_i, Xb)  = 0.0398                         Prob > F           =        0.0000
```

changecomp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
roe	14.74568	.3543894	41.61	0.000	14.05108	15.44028
_cons	1.980083	.1089532	18.17	0.000	1.766536	2.19363

While the estimate is the same, standard errors are different. The reason is that the OLS estimation we performed assumes that errors are uncorrelated, and they are not by construction.[1] One possible solution is to correct the standard errors using the cluster formula:

[1] Even if $cov(\varepsilon_{i,t}, \varepsilon_{i,s}) = 0$ for all $t \neq s$, $cov(\ddot{\varepsilon}_{i,t}, \ddot{\varepsilon}_{i,s}) = -\frac{\sigma^2}{T}$.

```
. reg DMchangecomp DMroe,noconstant vce(cluster execnum)

Linear regression                               Number of obs     =    119,421
                                                F(1, 28218)       =    1476.75
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0186
                                                Root MSE          =     32.397

                                   (Std. Err. adjusted for 28,219 clusters in execnum)
------------------------------------------------------------------------------
             |               Robust
DMchangecomp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       DMroe |    14.74568   .3837167    38.43   0.000     13.99357    15.49778
------------------------------------------------------------------------------
```

The result is still different but now OLS standard errors are higher. The difference between OLS with cluster-robust standard errors and the FE estimation is that while the former allows any type of time series dependence, the later assumes $cov(\varepsilon_{i,t}, \varepsilon_{i,s}) = 0$ (or equivalently $cov(\ddot{\varepsilon}_{i,t}, \ddot{\varepsilon}_{i,s}) = -\frac{\sigma^2}{T}$). We can provide more flexibility to the inference based on the FE estimator by also using the cluster-robust standard error:

```
. xtreg changecomp roe,fe vce(cluster execnum)

Fixed-effects (within) regression               Number of obs     =    119,421
Group variable: execnum                         Number of groups  =     28,219

R-sq:                                           Obs per group:
     within  = 0.0186                                         min =          1
     between = 0.0370                                         avg =        4.2
     overall = 0.0250                                         max =         24

                                                F(1,28218)        =    1476.74
corr(u_i, Xb)  = 0.0398                         Prob > F          =     0.0000

                                   (Std. Err. adjusted for 28,219 clusters in execnum)
------------------------------------------------------------------------------
             |               Robust
  changecomp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         roe |    14.74568   .3837183    38.43   0.000     13.99357    15.49778
       _cons |    1.980083    .020621    96.02   0.000     1.939665    2.020501
------------------------------------------------------------------------------
```

In this case the error of both models are subject to the same assumptions; hence we get the same results.

Supplementary reading:

"A Practitioner's Guide to Cluster-Robust Inference" A. Colin Cameron and Dougleas L. Miller, Journal of Human Resources (Chapters I-IV,VIII except II.F)

"An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units" Brent R. Moulton, The Review of Economics and Statistics Vol 72 No 2 (May 1990)