Gov 2000: 13. Panel Data and Clustering

Matthew Blackwell
Harvard University
mblackwell@gov.harvard.edu

November 28, 2016

Where are we? Where are we going?

- Up until now: the linear regression model, its assumptions, and violations of those assumptions
- This week: what can we do with panel data?

PANEL DATA

Motivation

Is there a relationship between democracy and, say, infant mortality? We could run a big cross-national regression, but would that be convincing? Perhaps democratic countries are different from non-democracies in ways that we can't measure—they are richer, provide benefits more efficiently, developed longer ago, or posses some cultural trait that tends to make their health outcomes better. One idea is to look at countries over time to see if we can get a better estimate of the effect of democracy on infant mortality. It turns out that under certain assumptions, we can allow for violations of zero conditional mean error if we have panel data (repeated observations over time), such as the following:

```r
ross <- foreign::read.dta(”../data/ross-democracy.dta”)
head(ross[,c(”cty_name”, ”year”, ”democracy”, ”life”, ”infmort_unicef”)])
```

```
##      cty_name year democracy  life infmort_unicef
## 1 Afghanistan 1965         0 36.82            230
## 2 Afghanistan 1966         0    NA             NA
## 3 Afghanistan 1967         0 37.80             NA
## 4 Afghanistan 1968         0    NA             NA
## 5 Afghanistan 1969         0    NA             NA
## 6 Afghanistan 1970         0 38.40            215
```

Notation

Let i continue to denote the unit, but now let t denote the time period. There are still n units, each measured at T periods which we call a **balanced panel**. Of course, the labeling here is somewhat arbitrary and we could rewrite this in terms of general groups. Time is a typical application, but this could be other groupings such as counties within states, states within countries, people within coutries, etc.

We assume the following linear model for the outcome of unit i at time t:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}$$

Here, $\mathbf{x}_{it}$ is a vector of covariate which might be time-varying, a_i is an **unobserved** time-constant unit effect (**fixed effect** or **unit effect**), u_{it} are the unobserved time-varying "idiosyncratic" errors.

Pooled OLS, as its name implies, pools all observations into one regression and treats all unit-periods (each it) as iid units. If the true model has the unit effect, then we can think about how pooled OLS will work by writing the overall error as $v_{it} = a_i + u_{it}$:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + v_{it}$$

This has two potential problems: first, heteroskedasticity, and second possible violation of zero conditional mean errors. Both problems arise out of ignoring the **unmeasured heterogeneity** inherent in a_i.

For instance, imagine that some of the covariates are correlated with a_i, so that $\text{Cov}[x_{itj}, a_i] \neq 0$, where x_{itj} is one covariate in the $\mathbf{x}_{it}$ vector. Perhaps having democratic institutions is correlated with some unmeasured aspects of health outcomes, like quality of health system or a lack of ethnic conflict. If this is true, then the zero conditional mean error assumption will be violated in the pooled model because we have

$$\text{Cov}[x_{itj}, v_{it}] = \text{Cov}[x_{itj}, a_i] + \text{Cov}[x_{itj}, u_{it}] \neq 0$$

Note that this is a violation even if the idiosyncratic errors (u_{it}) are uncorrelated with the covariates. One way to think about this is that there is a time-constant omitted variable or a set of time-constant omitted variables that are determining both y_{it} and

$\mathbf{x}_{it}$. The unobserved effect, a_i, captures the total effect of all of these time-constant omitted variables. For instance, if i represents individuals, then this might contain their early childhood political socialization and other features that are fixed over time.

To move forward, we will make some assumptions on this model. First, we'll assume that if we could somehow measure a_i and include it in the regression, then zero conditional mean error will hold. That is, we'll assume that $\mathbb{E}[u_{it}|\mathbf{x}_{it}, a_i] = 0$ or, equivalently, that the CEF in each time period is:

$$\mathbb{E}[y_{it}|\mathbf{x}_{it}, a_i] = \mathbf{x}'_{it}\boldsymbol{\beta} + a_i.$$

We'll see in a second that this assumption will not be enough to get around the omitted variable problems, but this is a first step.

Finally, we'll assume that we do have an i.i.d. sample across units. Letting $\mathbf{X}_i$ be the $T \times k$ vector of covariates for unit i across all time periods (where each row of $\mathbf{X}_i$ is $\mathbf{x}'_{it}$) and let $\mathbf{y}_i$ be the $T \times 1$ vector of outcomes across time. Then, we'll assume that $\{(\mathbf{y}_i, \mathbf{X}_i) : i = 1, 2, \ldots, n\}$ are i.i.d. draws from a population distribution.

First differencing

To see what the assumptions we'll need, it's useful to focus on a particular setup with two time periods and to think about using differencing to get rid of the unit effect. That is, we'll try to exploit panel data by changing our analysis from comparing levels to comparing **changes over time**. Intuitively, the levels include the unobserved heterogeneity, but changes over time should be free of this heterogeneity. It's easiest to see the logic when we only have two time periods:

$$\begin{aligned} y_{i1} &= \mathbf{x}'_{i1}\boldsymbol{\beta} + a_i + u_{i1} \\ y_{i2} &= \mathbf{x}'_{i2}\boldsymbol{\beta} + a_i + u_{i2} \end{aligned}$$

Let's create a new variable, the change in y over time and see what the model looks like for that variable:

$$\begin{aligned} \Delta y_i &= y_{i2} - y_{i1} \\ &= (\mathbf{x}'_{i2}\boldsymbol{\beta} + a_i + u_{i2}) - (\mathbf{x}'_{i1}\boldsymbol{\beta} - a_i - u_{i1}) \\ &= (\mathbf{x}'_{i2} - \mathbf{x}'_{i1})\boldsymbol{\beta} + (a_i - a_i) + (u_{i2} - u_{i1}) \\ &= \Delta\mathbf{x}'_i\boldsymbol{\beta} + \Delta u_i \end{aligned}$$

Note what changes in this first differenced model and what does not. For instance, the coefficient on the levels $\mathbf{x}_{it}$ is the same as the coefficient on the changes $\Delta\mathbf{x}_i$. Also note that the fixed effects/unobserved heterogeneity drops out since it is constant over time.

Can we apply regular OLS to the differences? What assumptions will we need? Linearity holds by our assumptions about linearity of the levels. The differences Δy_i and $\Delta \mathbf{x}_i$ will be i.i.d. across units. There are two more assumptions to consider for unbiasedness/consistency: no perfect collinearity and zero conditional mean error.

First, no perfect collinearity will also hold so long as $\mathbf{x}_{it}$ has to change over time for some units. If there is a variable that is constant over time, then it will be constantly equal to 0 when differenced and there will be perfect collinearity between it and the constant term. This shouldn't be too surprising since the whole point of this differencing is to remove the effects of time-constant omitted variables. We won't be able to separately net those effects out and estimate the effect of a time-constant variable.

Second, zero conditional mean error in the FD context would mean that $\mathbb{E}[\Delta u_i|\Delta \mathbf{x}_i] = 0$. Is this implied by the zero conditional mean error assumption on levels above, $\mathbb{E}[u_{it}|\mathbf{x}_{it} m a_i] = 0$? Not quite. The FD zero conditional mean error assumption would imply that the changes in the idiosyncratic error should be uncorrelated with the changes in the covariates. Let's see if that holds with a single covariate:

$$\begin{aligned}\text{Cov}[\Delta x_i, \Delta u_i] &= \mathbb{E}[(x_{i2} - x_{i1})(u_{i2} - u_{i1})] - \mathbb{E}[(x_{i2} - x_{i1})]\mathbb{E}[(u_{i2} - u_{i1})] \\ &= \mathbb{E}[(x_{i2} - x_{i1})(u_{i2} - u_{i1})] - 0 \\ &= \mathbb{E}[x_{i2}u_{i2}] + \mathbb{E}[x_{i1}u_{i1}] - \mathbb{E}[x_{i2}u_{i1}] - \mathbb{E}[x_{i1}u_{i2}]\end{aligned}$$

From above, we know that the idiosyncratic errors in time t will be uncorrelated with the covariates in time t, so that the first two terms will be equal to 0. But the other two terms measure the correlation between the idiosyncratic error in time 1 and the covariate in time 2, which the assumptions above did not address.

Thus, in order to use regular OLS to estimate the FD equation, we will need to make a stronger zero-conditional mean error assumption. There are two options. First, we can assume exactly what is needed for the FD, $\mathbb{E}[\Delta u_i|\Delta \mathbf{x}_i] = 0$, which says that changes in the idiosyncratic error term are uncorrelated with changes in the covariates. But just be sure to note that this is a stronger assumption that we made above. Another option is to assume a form of so-called **strict exogeneity** on the levels so that the idiosyncratic errors are uncorrelated with the covariates at **any** point in time:

$$\mathbb{E}[u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, a_i] = \mathbb{E}[u_{it}|\mathbf{x}_{it}, a_i] = 0$$

This assumption is stronger than either of the the zero-conditional mean error assumptions above (for levels or for differences). It implies that the correlation between the idiosyncratic errors and the covariates are uncorrelated at any time point, future or past:

$$\text{Cov}[\mathbf{x}_{is}, u_{it}] = 0 \qquad \forall t.$$

One obvious violation of this assumption is if there is a lagged dependent variable in the set of covariates.

With either of these two assumptions in hand, OLS on the differences will produce consistent estimates of the β vector.

First differences in R

```
pooled.mod <- lm(log(kidmort_unicef) ~ democracy + log(GDPcur), data = ross)
summary(pooled.mod)
```

```
##
## Call:
## lm(formula = log(kidmort_unicef) ~ democracy + log(GDPcur), data = ross)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4590 -0.5476  0.0945  0.5013  2.2643
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.76405    0.34491   28.31   <2e-16 ***
## democracy   -0.95525    0.06978  -13.69   <2e-16 ***
## log(GDPcur) -0.22828    0.01548  -14.75   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.7948 on 646 degrees of freedom
##   (5773 observations deleted due to missingness)
## Multiple R-squared:  0.5044, Adjusted R-squared:  0.5029
## F-statistic: 328.7 on 2 and 646 DF,  p-value: < 2.2e-16
```

```
library(plm)
fd.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur), data = ross,
              index = c(”id”, ”year”), model = ”fd”)
summary(fd.mod)
```

```
## Oneway (individual) effect First-Difference Model
##
## Call:
## plm(formula = log(kidmort_unicef) ~ democracy + log(GDPcur),
##     data = ross, model = ”fd”, index = c(”id”, ”year”))
```

```
##
## Unbalanced Panel: n=166, T=1-7, N=649
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -0.9060 -0.0956  0.0468  0.1410  0.3950
##
## Coefficients :
##              Estimate Std. Error  t-value Pr(>|t|)
## (intercept) -0.149469   0.011275 -13.2567  < 2e-16 ***
## democracy   -0.044887   0.024206  -1.8544  0.06429 .
## log(GDPcur) -0.171796   0.013756 -12.4886  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    23.545
## Residual Sum of Squares: 17.762
## R-Squared      :  0.24561
##       Adj. R-Squared :  0.24408
## F-statistic: 78.1367 on 2 and 480 DF, p-value: < 2.22e-16
```

Differences-in-differences

One place where this framework is very easy to understand is when trying to assess the impact of some treatment that was not randomly assigned, but there is some variation over the timing of its implementation. For instance, let x_{it} be an indicator of a unit being "treated" at time t. The most basic differences-in-differences estimation strategy usually has two time periods, where no one is treated in the first period ($x_{i1} = 0$ for all i) and some of the units are treated in the second. The goal is to know the effect of being treated. Here is the basic model:

$$y_{it} = \beta_0 + \delta_0 d_t + \beta_1 x_{it} + a_i + u_{it}$$

Here, d_t is a dummy variable for the second time period and β_1 is the quantity of interest: it's the effect of being treated.

To remove the fixed effect, let's take differences:

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1(x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

Here, we can interpret these quantities by remembering that $(x_{i2} - x_{i1}) = 1$ only for the treated group in this DD setup and that $(x_{i2} - x_{i1}) = 0$ only for the control group in this DD setup. Thus, δ_0 represents the difference in the average outcome from period 1 to period 2 in the untreated group, and β_1 represents the **additional** change in y over time (on top of δ_0) associated with being in the treated group. The key here is that we are comparing the change over time in the control group to the change over time in the treated group. The differences between these differences is our estimate of the causal effect:

$$\beta_1 = \overline{\Delta y}_{\text{treated}} - \overline{\Delta y}_{\text{control}}$$

Why is this more credible than simply looking at the treatment/control differences in period 2? Let's look at that equation:

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2}$$

Here there might be unmeasured reasons why the treated group has higher or lower outcomes than the control group. This is picked up a_i. If a_i is correlated with the treatment status, then zero conditional mean error fails and our estimates are no good. With differences-in-differences, we leverage the changes over time to help eliminate these time-constant problems. Thus, the power comes from the over time variation.

One example of this type of technique is Lyall (2009) which was interested in answering the following question: does Russian (that is, government) shelling of villages cause insurgent (that is, non-governmental) attacks? We might think that artillery shelling by Russians is targeted to places where the insurgency is the strongest. That is, part of the village fixed effect, a_i might be correlated with whether or not shelling occurs, x_{it}. This would cause any pooled OLS estimates to be biased. Instead Lyall takes a diff-in-diff approach: compare attacks over time (Δy_i) for shelled ($x_{i2} = 1$) and non-shelled ($x_{i2} = 0$) villages.

Let's cover the assumptions needed to get a consistent/unbiased estimate of β_1 in this context. From above, we need zero conditional mean error for the changes: $\mathbb{E}[\Delta u_i | \Delta x_i] = 0$. Notice here that because no one is treated in the first period, we can write $\Delta x_i = x_{i2}$, so that the key assumption is that treatment needs to be independent of the idiosyncratic shocks:

$$\mathbb{E}[(u_{i2} - u_{i1}) | x_{i2}] = 0$$

Implies that without treatment, the trend in the outcome over time is the same for the treated and control groups. This assumption is sometimes called the **parallel trends assumption**.

Of course, parallel trends might not be plausible. One way it might be violated is called Ashenfelter's dip, which is a empirical finding that people who enroll in job training programs see their earnings decline prior to that training. In the Lyall paper, it might be the case that insurgent attacks might be falling in places where there is shelling because rebels attacked in those areas and have moved on.

We could generalize the framework to handle covariates, which might make the parallel trends assumption more plausible. For instance, the independence of the treatment and idiosyncratic shocks might only hold conditional on covariates:

$$y_{i2} - y_{i1} = \delta_0 + \mathbf{z}_i'\tau + \beta(x_{i2} - A_{i1}) + (u_{i2} - u_{i1})$$

Here, $\mathbf{z}_i'$ is a vector of covariates and β is still the causal effect of the treatment. This is sometimes called "regression diff-in-diff."

FIXED EFFECTS MODELS

When people talk about "fixed effects models" or that they "included fixed effects" what they usually mean is that they did a specific transformation of the data similar to, but distinct from the differences. In both, however, we transform the data to remove the unobserved effect, a_i. First note that taking the average of the y's over time for a given unit leaves us with a very similar model:

$$\begin{aligned}
\overline{y}_i &= \frac{1}{T}\sum_{t=1}^{T}\left[\mathbf{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}\right] \\
&= \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_{it}'\right)\boldsymbol{\beta} + \frac{1}{T}\sum_{t=1}^{T}a_i + \frac{1}{T}\sum_{t=1}^{T}u_{it} \\
&= \overline{\mathbf{x}}_i'\boldsymbol{\beta} + a_i + \overline{u}_i
\end{aligned}$$

They key fact here is that because the unobserved effect is constant over time, so the over-time mean of the a_i is just a_i itself. The **fixed effects**, **within**, or **time-demeaning** transformation is when we subtract off the over-time means from the original data:

$$(y_{it} - \overline{y}_i) = (\mathbf{x}_{it}' - \overline{\mathbf{x}}_i')\boldsymbol{\beta} + (u_{it} - \overline{u}_i)$$

If we write $\ddot{y}_{it} = y_{it} - \overline{y}_i$, then we can write this more compactly as:

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}'\boldsymbol{\beta} + \ddot{u}_{it}$$

Note that since a_i is time-constant, it drops out of this specification.

What assumptions do we need to apply OLS to this? Again, linearity and i.i.d. come from linearity and i.i.d. for the levels that we discussed above. No perfect collinearity will again be satisfied so long as do not include any covariates that are constant in time (since they will be a constant 0 column in the data). Finally, and again similar to first differencing, the zero conditional mean idiosyncratic error, $\mathbb{E}[u_{it}|\mathbf{x}_{it}, a_i] = 0$, is not sufficient for zero conditional mean error of the within transformation: $\mathbb{E}[\ddot{u}_{it}|\ddot{\mathbf{x}}_{it}] = 0$. This is for similar reasons to the FD model, since the levels version of zero conditional mean error only places restrictions on the correlation between u_{it} and $\mathbf{x}_{it}$, but the FE transformation is a function of all time periods.

Thus, in order to move forward, we will need to strengthen our zero conditional mean error assumption to a strict exogeneity assumption:

$$\mathbb{E}[u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, a_i] = \mathbb{E}[u_{it}|\mathbf{x}_{it}, a_i] = 0$$

This assumption will imply that the time-demeaned idiosyncratic errors ($\ddot{u}_{it}$) will be uncorrelated with the time-demeaned covariates $\ddot{\mathbf{x}}_{it}$. This is because u_{it} will be uncorrelated with all $\mathbf{x}_{is}$ and so will also be uncorrelated with $\overline{\mathbf{x}}_i$, which also implies that $\overline{u}_i$ will be uncorrelated with all $\mathbf{x}_{is}$ and $\overline{\mathbf{x}}_i$. Thus, under strict exogeneity, we can get unbiased and consistent estimates of the $\boldsymbol{\beta}$ vector by applying regular OLS to the time-demeaned data above. We do need to slightly modify the degrees of freedom we use for t-tests and variance estimation to account for the transformation to $nT - n - k - 1$. In particular, it is important to estimate the error variance as:

$$\widehat{\sigma}_u^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{nT - n - k - 1}$$

If you just demeaned the data yourself and ran the usual OLS, however, it would divide the sum of the square residuals by $nT - k - 1$, which would mean that the SEs would be understated. For this reason, it is useful to use functions dedicated to implementing fixed effects (like `plm()`) or to use the dummy variable approach below.

Fixed effects with Ross data

```
fe.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur), data = ross,
              index = c(”id”, ”year”), model = ”within”)
summary(fe.mod)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(kidmort_unicef) ~ democracy + log(GDPcur),
##     data = ross, model = ”within”, index = c(”id”, ”year”))
##
## Unbalanced Panel: n=166, T=1-7, N=649
```

```
##
## Residuals :
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -0.70500 -0.11700  0.00628  0.12200  0.75700
##
## Coefficients :
##             Estimate Std. Error  t-value  Pr(>|t|)
## democracy   -0.143233   0.033500  -4.2756 2.299e-05 ***
## log(GDPcur) -0.375203   0.011328 -33.1226 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Total Sum of Squares:    81.711
## Residual Sum of Squares: 23.012
## R-Squared      :  0.71838
##       Adj. R-Squared :  0.53242
## F-statistic: 613.481 on 2 and 481 DF, p-value: < 2.22e-16
```

Fixed effects with Ross data

- Pooled model with a time-constant variable, proportion Islamic:

```
p.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur) + islam,
             data = ross, index = c(”id”, ”year”), model = ”pooling”)
summary(p.mod)
```

```
## Oneway (individual) effect Pooling Model
##
## Call:
## plm(formula = log(kidmort_unicef) ~ democracy + log(GDPcur) +
##     islam, data = ross, model = ”pooling”, index = c(”id”, ”year”))
##
## Unbalanced Panel: n=136, T=1-7, N=583
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -2.3500 -0.4880  0.0807  0.4740  2.2200
##
## Coefficients :
##               Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept) 10.30607817  0.35951939  28.6663 < 2.2e-16 ***
## democracy   -0.80233845  0.07766814 -10.3303 < 2.2e-16 ***
## log(GDPcur) -0.25497406  0.01607061 -15.8659 < 2.2e-16 ***
## islam        0.00343325  0.00091045   3.7709 0.0001794 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    757.57
## Residual Sum of Squares: 331.63
## R-Squared      :  0.56224
##       Adj. R-Squared :  0.55839
## F-statistic: 247.884 on 3 and 579 DF, p-value: < 2.22e-16
```

- FE model, where the islam variable drops out:

```
fe.mod2 <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur) + islam,
               data = ross, index = c("id", "year"), model = "within")
summary(fe.mod2)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = log(kidmort_unicef) ~ democracy + log(GDPcur) +
##     islam, data = ross, model = "within", index = c("id", "year"))
##
## Unbalanced Panel: n=136, T=1-7, N=583
##
## Residuals :
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -0.6990 -0.1220  0.0109  0.1300  0.7490
##
## Coefficients :
##              Estimate Std. Error  t-value  Pr(>|t|)
## democracy   -0.129693   0.035865  -3.6162 0.0003332 ***
## log(GDPcur) -0.379997   0.011849 -32.0707 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    78.768
## Residual Sum of Squares: 21.855
## R-Squared      :  0.72254
##       Adj. R-Squared :  0.55151
## F-statistic: 579.423 on 2 and 445 DF, p-value: < 2.22e-16
```

Least squares dummy variable

As an alternative to the within transformation, we can also include a series of $n-1$ dummy variables for each unit. Gives the **exact** same point estimates as with time-demeaning and it automatically gets the right degrees of freedom so it can be implemented with standard functions like `lm()`. The main disadvantage of this approach is that it is computationally difficult with large n, since we have to run a regression with $n+k$ variables. It is often much faster to demean the data.

Example with Ross data

```
library(lmtest)
lsdv.mod <- lm(log(kidmort_unicef) ~ democracy + log(GDPcur) + as.factor(id),
               data = ross)
coeftest(lsdv.mod)[1:6,]
```

```
##                    Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)      13.7644887 0.26597312  51.751427 1.008329e-198
## democracy        -0.1432331 0.03349977  -4.275644  2.299393e-05
## log(GDPcur)      -0.3752030 0.01132772 -33.122568 3.494887e-126
## as.factor(id)AGO  0.2997206 0.16767730   1.787485  7.448861e-02
## as.factor(id)ALB -1.9309618 0.19013955 -10.155498  4.392512e-22
## as.factor(id)ARE -1.8762909 0.17020738 -11.023558  2.386557e-25
```

```
coeftest(fe.mod)
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error  t value  Pr(>|t|)
## democracy   -0.143233   0.033500  -4.2756 2.299e-05 ***
## log(GDPcur) -0.375203   0.011328 -33.1226 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Fixed effects versus first differences

Under strict exogeneity of the idiosyncratic error and time-constant unobserved error, both fixed effects and first differences are unbiased and consistent. In fact, with $T = 2$ the estimators produce identical estimates. So which one is better when $T > 2$? If they are both unbiased, we need to look to efficiency—which of them will have lower uncertainty in their estimates?

It turns out that when the idiosyncratic errors u_{it} are uncorrelated with each other, FE is more efficient and when the idiosyncratic errors u_{it} are serially correlated in a particular way (they follow a random walk), FD is more efficient. The truth is usually in between and so it isn't clear which will be more efficient in any particular example—depends on n, T, and the true model. One thing we can probably say is that large differences between the FE and FD estimates should make us worry about the validity of our assumptions.

RANDOM EFFECTS

Random effects model

With a random effects approach, we take the same basic model for the levels and impose a slightly stronger assumption on the heterogeneity. For instance, take the basic model:

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + a_i + u_{it}$$

The key **random effect assumption** is that we will assume that the covariates are uncorrelated with the unit effect, a_i. More precisely, we will assume that

$$\mathbb{E}[a_i|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}] = \mathbb{E}[a_i] = 0.$$

We also continue to assume strict exogeneity,

$$\mathbb{E}[u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}, a_i] = \mathbb{E}[u_{it}|\mathbf{x}_{it}, a_i] = 0,$$

which implies that a_i are uncorrelated with the u_{it}.

Obviously these assumptions are stronger than for FE or FD, since there we made no assumptions about the correlation between the unobserved unit effect and the covariates. Here, we assume that they are (mean) independent. Of course, this means that the a_i cannot be thought of as a confounder and so the random effects models assumes that a_i are not an omitted variable in the way that FE or FD models do. Specifically, we can treat $v_{it} = a_i + u_{it}$ as a combined error that satisfies zero conditional mean error:

$$\mathbb{E}[a_i + u_{it}|\mathbf{x}_{it}] = \mathbb{E}[a_i|\mathbf{x}_{it}] + \mathbb{E}[u_{it}|\mathbf{x}_{it}] = 0 + 0 = 0$$

Thus, under these random effects assumptions, pooled OLS is actually unbiased and consistent. It it not, however, since the Gauss-Markov assumptions are not satisfied. Furthermore, the standard errors from conventional OLS will be incorrect. This is what RE analyses are designed to fix.

Quasi-demeaned data

Random effects models usually transform the data via what is called **quasi-demeaning** or **partial pooling**:

$$y_{it} - \theta\overline{y}_i = (\mathbf{x}'_{it} - \theta\overline{\mathbf{x}}'_i) + (v_{it} - \theta\overline{v}_i)$$

Here θ is between zero and one, where $\theta = 0$ implies pooled OLS and $\theta = 1$ implies fixed effects. Doing some math shows that

$$\theta = 1 - \left[\frac{\sigma_u^2}{T\sigma_a^2 + \sigma_u^2}\right]^{1/2}.$$

Here, σ_u^2 is the variance of u_{it} and σ_a^2 is the variance of a_i. The **random effect estimator** runs pooled OLS on this model replacing θ with an estimate $\widehat{\theta}$, where the this based on estimates of σ_u^2 and σ_a^2.

You can do basic random effects with linear models using `plm()` and more general random effects models using `lmer()` from the `lme4` package.

Example with Ross data

```
re.mod <- plm(log(kidmort_unicef) ~ democracy + log(GDPcur), data = ross,
              index = c(”id”, ”year”), model = ”random”)
coeftest(re.mod)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 12.312868   0.255008  48.2842 < 2.2e-16 ***
## democracy   -0.191796   0.033957  -5.6482 2.431e-08 ***
## log(GDPcur) -0.360927   0.011009 -32.7839 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
coeftest(fe.mod)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value  Pr(>|t|)
## democracy   -0.143233   0.033500  -4.2756 2.299e-05 ***
## log(GDPcur) -0.375203   0.011328 -33.1226 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coeftest(fd.mod)
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error  t value Pr(>|t|)
## (intercept) -0.149469   0.011275 -13.2567  < 2e-16 ***
## democracy   -0.044887   0.024206  -1.8544  0.06429 .
## log(GDPcur) -0.171796   0.013756 -12.4886  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fixed effects versus random effects

Let's look at the errors in the random effects model:

$$v_{it} - \theta\overline{v}_i = (1-\theta)a_i + u_{it} - \theta\overline{u}_i.$$

If there is correlation between x_{it} and a_i, then there will also be correlation between $x_{it} - \theta\overline{x}_i$ and $v_{it} - \theta\overline{v}_i$. This correlation will cause zero conditional mean error to be violated. Again, this implies that it is crucial for RE models that the a_i be uncorrelated with the independent variables. Basically, RE models can help us fix up standard errors in a similar way the WLS helps us with SEs.

Only FE/FD can consistently estimate effects when there is unmeasured heterogeneity correlated with the independent variables. But RE does allow us to include time-constant covariates, while FE does not. This is an advantage, but the advantage is somewhat lost if we can't consistently estimate the effect of this time-constant variable. Wooldridge comes down hard on the side of FE: "FE is almost always much more convincing than RE for policy analysis using aggregated data."

A more general model is that of correlated random effects which allows for some structured dependence between x_{it} and a_i. See Wooldridge (2010) for more on these models.

CLUSTERING

We are often in situations where the data we have is not i.i.d., but that units are correlated within groups and independent across groups. We often call these groups *clusters* and they arise very often in applied work. For example, think back to the Gerber, Green, and Larimer (2008) social pressure mailer example. Their design randomly sample households and randomly assign them to different treatment conditions, but the measurement of turnout is at the individual level. From the point of view of the linear model we have been discussing, this could lead to a violation of the **iid/random sampling** assumption. This is because the errors of individuals within the same household are correlated. This will lead to problems with the development of our uncertainty estimates (standard errors, variances) since it relied on i.i.d.

More generally, we have **clustering** or **clustered dependence** when there are G clusters or groups, each with some number of units, which might be related. We will label the observations, y_{ig} where $g \in \{1, \ldots, m\}$ represents clusters, and the $i \in \{1, \ldots, n_g\}$ labels the units,where n_g is the number of units in cluster g and $n = \sum_{g=1}^{m} n_g$ is the total number of units. Here, units belong to a single cluster. For example, we might have: voters in households, individuals in states, students in classes, or rulings in judges.

Of course, if there are just clusters with no dependence within the cluster, then everything is iid and we can just proceed with OLS and the linear model as usual. Let's build a linear model that encodes the within-cluster dependence:

$$y_{ig} = \mathbf{x}'_{ig}\beta + v_{ig} = \mathbf{x}'_{ig}\beta + a_g + u_{ig}$$

We'll refer to the a_g as the cluster error component and u_{ig} as the unit error component, where a_g and u_{ig} are assumed to be independent of each other. Furthermore, we assume that the zero conditional mean error holds, $\mathbb{E}[v_{ig}|\mathbf{x}_{ig}] = 0$, so that OLS is unbiased and consistent. Thus, we are in a world more similar to the random effects model above rather than the fixed effects assumptions. Note that these assumptions are weaker than the RE ones since we don't have to assume strict exogeneity.

Let $\mathbf{X}_g$ be the $n_g \times k$ matrix of covariates for cluster g and let $\mathbf{y}_g$, $\mathbf{v}_g$, and $\mathbf{u}_g$ be the vectors of outcomes, combined errors, and unit errors for the cluster. With this, we can write the clustered or grouped linear model as:

$$\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta} + \mathbf{v}_g = \mathbf{X}_g\boldsymbol{\beta} + a_g + \mathbf{u}_g$$

Let $\mathbf{y}$ be the $n \times 1$ vector of outcomes across all clusters, with $\mathbf{X}$ and $\mathbf{v}$ defined similarly. Then, we can write the linear model as usual as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{v}$$

Thus, we can write the model now at the unit-level, cluster-level, or entire-sample level.

We'll continue to assume that zero conditional mean error holds so that $\mathbb{E}[\mathbf{v}_g|\mathbf{X}_g] = 0$. To investigate the properties of this model, let the variance of the two components be $\mathbb{V}[a_g|\mathbf{X}_g] = \rho\sigma^2$ and $\mathbb{V}[u_{ig}|\mathbf{X}_g] = (1-\rho)\sigma^2$, which implies that the variance of the overall error is $\mathbb{V}[v_{ig}|\mathbf{X}_g] = \sigma^2$:

$$\begin{aligned}\mathbb{V}[v_{ig}|\mathbf{X}_g] &= \mathbb{V}[a_g + u_{ig}|\mathbf{X}_g] \\ &= \mathbb{V}[a_g|\mathbf{X}_g] + \mathbb{V}[u_{ig}|\mathbf{X}_g] \\ &= \rho\sigma^2 + (1-\rho)\sigma^2 = \sigma^2\end{aligned}$$

We call $\rho \in (0,1)$ is called the within-cluster correlation. To see why, first note that we can calculate the covariance between two units i and s in the same cluster is $\rho\sigma^2$

$$\begin{aligned}\text{Cov}[v_{ig}, v_{sg}|\mathbf{X}_g] &= \text{Cov}[a_g + u_{ig}, a_g + u_{sg}|\mathbf{X}_g] \\ &= \text{Cov}[a_g, a_g|\mathbf{X}_g] + \text{Cov}[a_g, u_{sg}|\mathbf{X}_g] + \text{Cov}[u_{ig}, a_g|\mathbf{X}_g] + \text{Cov}[u_{ig}, u_{sg}|\mathbf{X}_g] \\ &= \mathbb{V}[a_g|\mathbf{X}_g] + 0 + 0 + 0 = \rho\sigma^2\end{aligned}$$

With this in hand, it is easy to see that correlation between units in the same group is gust ρ:

$$\text{Cor}[v_{ig}, v_{sg}|\mathbf{X}_g] = \frac{\text{Cov}[v_{ig}, v_{sg}|\mathbf{X}_g]}{\sqrt{\mathbb{V}[v_{ig}|\mathbf{X}_g]\mathbb{V}[v_{sg}|\mathbf{X}_g]}} = \frac{\rho\sigma^2}{\sqrt{\sigma^2\sigma^2}} = \rho$$

Finally, note that with this structure, the covariance of two units i and s in different clusters j and k:

$$\begin{aligned}\text{Cov}[v_{ig}, v_{sk}|\mathbf{X}] &= \text{Cov}[a_g + u_{ig}, v_k + u_{sk}|\mathbf{X}] \\ &= \text{Cov}[a_g, v_k|\mathbf{X}] + \text{Cov}[a_g, u_{sk}|\mathbf{X}] + \text{Cov}[u_{ig}, v_k|\mathbf{X}] + \text{Cov}[u_{ig}, u_{sk}|\mathbf{X}] \\ &= 0 + 0 + 0 + 0 = 0\end{aligned}$$

If we write the overall model errors across all groups as $\mathbf{v}$, then we can figure out what the covariance matrix of this vector of errors is. Remember that in the linear model, the covariance matrix of the error is diagonal, with variances along the diagonal and 0s off the diagonal. With clustered dependence, this won't be the structure. To see what the structure will be, let's focus on an example with two clusters (though we will need far more to actually perform inference):

$$\mathbf{v} = \left[\begin{array}{cccccc} v_{1,1} & v_{2,1} & v_{3,1} & v_{4,2} & v_{5,2} & v_{6,2} \end{array}\right]'$$

$$\mathbb{V}[\mathbf{v}|\mathbf{X}] = \mathbf{\Sigma} = \left[\begin{array}{cccccc} \sigma^2 & \sigma^2\cdot\rho & \sigma^2\cdot\rho & 0 & 0 & 0 \\ \sigma^2\cdot\rho & \sigma^2 & \sigma^2\cdot\rho & 0 & 0 & 0 \\ \sigma^2\cdot\rho & \sigma^2\cdot\rho & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & \sigma^2\cdot\rho & \sigma^2\cdot\rho \\ 0 & 0 & 0 & \sigma^2\cdot\rho & \sigma^2 & \sigma^2\cdot\rho \\ 0 & 0 & 0 & \sigma^2\cdot\rho & \sigma^2\cdot\rho & \sigma^2 \end{array}\right]$$

In general, we can write the covariance matrix as a **block diagonal**, which means that there are matrices along the diagonal and 0s elsewhere. By independence, the errors are uncorrelated across clusters:

$$\mathbb{V}[\mathbf{v}|\mathbf{X}] = \mathbf{\Sigma} = \left[\begin{array}{c|c|c|c} \mathbf{\Sigma}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{\Sigma}_2 & \dots & \mathbf{0} \\ \hline & & \ddots & \\ \hline \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Sigma}_m \end{array}\right]$$

Here, $\mathbf{\Sigma}_g = \mathbb{V}[\mathbf{v}_g|\mathbf{X}_g]$ is the $n_g \times n_g$ covariance matrix of the errors for cluster g. Again, this says that errors can be correlated within clusters (on the block diagonal), but errors are uncorrelated across clusters.

Correcting for clustering

1. Including a dummy variable for each cluster (fixed effects, next week)
2. Random effects models
3. Cluster-robust ("clustered") standard errors
4. Aggregate data to the cluster-level and use OLS $\overline{y}_g = \frac{1}{n_g}\sum_i y_{ig}$
    - If n_g varies by cluster, then cluster-level errors will have heteroskedasticity
    - Can use WLS with cluster size as the weights

Cluster-robust standard errors

- Leads to this matrix:

$$\mathbb{V}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{j=1}^{m}\mathbf{X}_j'\mathbf{\Sigma}_j\mathbf{X}_j\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- Way to estimate this matrix: replace $\boldsymbol{\Sigma}_j$ with an estimate based on the within-cluster residuals, $\widehat{\mathbf{v}}_j$:
$$\widehat{\boldsymbol{\Sigma}}_j = \widehat{\mathbf{v}}_j \widehat{\mathbf{v}}_j'$$
- Final expression for our cluster-robust covariance matrix estimate:
$$\widehat{\mathbb{V}}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{j=1}^{m} \mathbf{X}_j' \widehat{\mathbf{v}}_j \widehat{\mathbf{v}}_j' \mathbf{X}_j \right) (\mathbf{X}'\mathbf{X})^{-1}$$
- With small-sample adjustment (which is what most software packages report):
$$\widehat{\mathbb{V}}_a[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \frac{m}{m-1} \frac{n-1}{n-k-1} (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{j=1}^{m} \mathbf{X}_j' \widehat{\mathbf{v}}_j \widehat{\mathbf{v}}_j' \mathbf{X}_j \right) (\mathbf{X}'\mathbf{X})^{-1}$$

Example

```
load(”../data/gerber_green_larimer.RData”)
social$voted <- 1 * (social$voted == ”Yes”)
social$treatment <- factor(social$treatment, levels = c(”Control”, ”Hawthorne”, ”Civic Duty”, ”Neighbors”, ”Self
mod1 <- lm(voted ~ treatment, data = social)
```

```
source(”vcovCluster.R”)
library(lmtest)
coeftest(mod1, vcov = vcovCluster(mod1, ”hh_id”))
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)         0.2966383  0.0013096 226.5172 < 2.2e-16 ***
## treatmentHawthorne  0.0257363  0.0032579   7.8997 2.804e-15 ***
## treatmentCivic Duty 0.0178993  0.0032366   5.5302 3.200e-08 ***
## treatmentNeighbors  0.0813099  0.0033696  24.1308 < 2.2e-16 ***
## treatmentSelf       0.0485132  0.0033000  14.7009 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Just a few things to keep in mind about CRSEs. First, CRSEs do not change our estimates $\widehat{\boldsymbol{\beta}}$ and so they cannot fix bias. Instead, CRSE is consistent estimator of $\mathbb{V}[\widehat{\boldsymbol{\beta}}]$

given clustered dependence that relies on independence between clusters but doesn't depend on the correct specification of the correlational structure within clusters. For instance, the RE models above require this structure to be correctly specified. CRSEs are usually bigger than conventional SEs so it is typically safe to use them as a "conservative" approach. Finally, remember that consistency of the CRSE are in the number of groups, not the number of individuals within each group. Because of this, CRSEs can be incorrect with a small (< 50 maybe) number of clusters.