1 Dynamic panel data

As we saw in the lecture notes #3, the strict exogeneity assumption \( E(\nu_{it}|x_{it1}, x_{it2}, \ldots, x_{iT}) = 0 \) is violated when one of the explanatory variable is a lagged dependent variable. This means that fixed effect estimates will be biased. To see this, consider the model:

\[
y_{it} = x_{it}\beta + y_{it-1}\delta + \theta_i + \nu_{it} \tag{1}
\]

First differencing yields:

\[
\Delta y_{it} = \Delta x_{it}\beta + \Delta y_{it-1}\delta + \Delta \nu_{it} \tag{2}
\]

The problem is that the explanatory variable \( \Delta y_{it-1} = y_{it-1} - y_{it-2} \) is correlated, by construction, with the error term \( \Delta \nu_{it} = \nu_{it} - \nu_{it-1} \), since \( \nu_{it-1} \) appears on the right hand side of the equation for \( y_{it-1} \):

\[
y_{it-1} = x_{it-1}\beta + y_{it-2}\delta + \theta_i + \nu_{it-1} \tag{3}
\]

A similar problem arises in the case of the within estimator, since the difference from means equation is:

\[
y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + (y_{it-1} - \bar{y}_{i-1})\delta + \nu_{it} - \bar{\nu}_i \tag{4}
\]

where
\[
\bar{y}_{i,-1} = (1/T) \sum_{t=0}^{T-1} y_{it} = ... + (1/T) \sum_{t=0}^{T-1} \nu_{it}
\]  

(5)

It follows that

\[
\text{cov} \left( [y_{it-1} - \bar{y}_{i,-1}], [\nu_{it} - \bar{\nu}_i] \right) = \text{cov} \left( \left[ \nu_{it-1} - (1/T) \sum_{t=0}^{T-1} \nu_{it} \right], \left[ \nu_{it} - (1/T) \sum_{t=1}^{T} \nu_{it} \right] \right)
\]

\[
= -(2/T)\sigma^2_\nu + [(T - 1)/T^2]\sigma^2_\nu
\]  

(6)

OLS estimates of \( \delta \) will be biased because of this non-zero correlation between the explanatory variable, \( y_{it-1} - \bar{y}_{i,-1} \), and the error term, \( \nu_{it} - \bar{\nu}_i \). Note, however, that the correlation goes to zero as \( T \) goes to infinity. This means that the bias in the within estimator will become quite small when the number of observations per individual, \( T \), becomes large.

Does this mean that the within estimator is consistent, i.e. that the bias goes to zero when the number of observations goes to infinity? The answer depends on whether we do the asymptotics with \( N \), the number of individuals, and/or \( T \) going to infinity. In the case where the strict exogeneity assumption holds, it follows that within estimates are consistent even if \( T \) is small and fixed (e.g. \( T = 2 \)) while \( N \) goes to infinity. In the case where there is a lagged dependent variable, however, \( T \) must also be going to infinity at some reasonably fast rate for the within estimator to be consistent.

In most applications based on micro data, \( T \) tends to be small while \( N \) is quite large. For example, the main individual-based panel data set that used to be collected by Statistics Canada (the Survey of Labour and Income Dynamics, or SLID) has about 30,000 individuals (\( N = 30,000 \)) followed over 6 years (\( T = 6 \)). In such a setting, it is unreasonable to think that \( T \) is “large enough” for the within estimator not to be biased when there is a lagged dependent variable. This assumption is much more reasonable in analyses of cross-country data where \( T \) and \( N \) are often of comparable sizes (e.g. 25 years of data for 25 OECD countries), or in some data sets on firms like COMPUSTAT where we have decades worth of information for thousands of firms.

1.1 IV estimation

A simple solution to the problem caused by lagged dependent variables in a fixed effect model is to estimate the first difference model using instrumental variables methods.
Remember from equation (2) that the problem with first differences is that the component \( y_{it-1} \) of \( \Delta y_{it-1} \) is correlated with the component \( \nu_{it-1} \) of the error term \( \Delta \nu_{it} \). But equation (3) shows that \( y_{it-1} \) depends on \( x_{it-1} \) and \( y_{it-2} \), neither of which appear in the equation for \( y_{it} \), or is correlated with \( \nu_{it-1} \). This means that both \( x_{it-1} \) and \( y_{it-2} \) can be used as instrumental variables for \( y_{it-1} \). So a simple solution is to estimate the first-differenced equation (2) by two-stage least squares using \( x_{it-1} \) and/or \( y_{it-2} \) as instruments for \( \Delta y_{it-1} \). Note, however, that \( y_{it-2} \) will not be a valid instrument in the highly plausible scenario where \( \nu_{it} \) exhibits some serial correlation, since \( \nu_{it-2} \) and thus \( y_{it-2} \) will then be correlated with the component \( \nu_{it-1} \) of the error term \( \Delta \nu_{it} \). But since

\[
y_{it-1} = x_{it-1} \beta + y_{it-2} \delta + \theta_i + \nu_{it-1} = x_{it-1} \beta + [x_{it-2} \beta + y_{it-3} \delta + \theta_i + \nu_{it-2}] \delta + \theta_i + \nu_{it-1}
\]

we can think of using \( x_{it-2} \) or \( y_{it-3} \) as alternative instruments for \( y_{it-1} \). In summary, we are faced with various choices of instrumental variables based on lagged values of \( x_{it} \) and \( y_{it} \), where the “best” choice of instrument depends on assumptions about the extent of serial correlation in \( \nu_{it} \) and on the predictive power of different lags in the first stage equation. The problem of how to best choose instruments in this context has been well studied in the literature, in particular in an influential 1991 Review of Economic Studies paper by Arellano and Bond. Their suggested estimator is programmed up in the “xtabond” procedure in Stata.

### 2 Repeated cross sections and difference-in-differences (DiD)

Collecting panel data is often difficult and costly. For example, in a panel of individuals like SLID, Statistics Canada has to track down people over time, which is often difficult to do since people move out, and may even leave the country. Unless this type of “attrition” is happening at random, it may result in systematic selection biases. In the case of very large surveys like the 20 percent sample of the Canadian Census (one household out of five gets a “long form” asking about income, education, etc.), it is simply impractical to follow millions of people over time. A potential solution to the problem is to use large longitudinal administrative data sets like the LAD in Canada (Longitudinal Administrative Databank), which is based on income tax data. These types of data sets are becoming increasingly available to researchers.
In some situations, it is also possible to pool together several years of cross-sectional data and create a “quasi-panel” that shares some (but not all) of the advantages of true panel data. Since it is generally easier to obtain series of cross-sections than true panels, quasi-panel data methods can be very useful in practice.

2.1 Quasi panels

Consider the case where we have cross-sections of individuals $i$ available for several years $t$ ranging from $t = 1$ to $T$. Individuals in each year are grouped together and $j$ indicates the group to whom individual $i$ belongs. As before, the group could be defined on a geographical basis (province, metropolitan area, etc.). Another interesting case is where $j$ represents a cohort of individuals, for instance those born in 1982. Consider the micro-level model

$$y_{ijt} = x_{ijt}\beta + \theta_{ij} + \delta_t + \nu_{ijt}$$

where $\delta_t$ is an unrestricted time effect (captured in practice by using year dummies in a regression). We can create a quasi-panel by averaging out all the variables within each group $j$ at time $t$. Using a notation similar to the one introduced for the within estimator, this leads to the group level model:

$$\overline{y}_{jt} = \overline{x}_j\beta + \overline{\theta}_j + \delta_t + \overline{\nu}_{jt}$$

This now looks like a standard panel data model where the group fixed (or random) effect $\overline{\theta}_j$ is now defined as the average of the individual fixed (or random) effects within the group $j$.

2.2 Difference-in-differences

One special but important case of quasi-panel data is the so-called “difference-in-differences”, or DiD model. Consider two groups $j = 1$ and 2 that are observed over two time periods, $t = 1$ and 2. A relevant policy or other intervention is introduced for group $j = 2$ at time $t = 2$. Let’s capture this policy by a dummy variable $x_{ijt}$ where $x_{ijt} = 1$ in group $j = 2$ at time $t = 2$, and $x_{ijt} = 0$ otherwise. Since $x_{ijt}$ shares the exact same value for all individuals $i$ in group $j$ at time $t$, we can write $\overline{x}_{jt} = x_{jt} = 1$ for $j = 2$ at time $t = 2$, and $x_{jt} = 0$ otherwise. To simplify the notation, it is also convenient to write $\theta_j = \overline{\theta}_j$ and $\nu_{jt} = \overline{\nu}_{jt}$. The resulting model is
\[
\bar{y}_{jt} = x_{jt}\beta + \theta_j + \delta_t + \nu_{jt}
\] (10)

where \(\beta\) can now interpreted as a “treatment” or “policy” effect of the intervention captured by the dummy variable \(x_{jt}\). In the four relevant cases, we have:

\[
\begin{align*}
\bar{y}_{11} &= \theta_1 + \delta_1 + \nu_{11} \\
\bar{y}_{12} &= \theta_1 + \delta_2 + \nu_{12} \\
\bar{y}_{21} &= \theta_2 + \delta_1 + \nu_{21} \\
\bar{y}_{22} &= \beta + \theta_2 + \delta_2 + \nu_{22}
\end{align*}
\] (11-14)

The fixed group effects \(\theta_1\) and \(\theta_2\) can be eliminated by taking time differences:

\[
\begin{align*}
\bar{y}_{12} - \bar{y}_{11} &= (\delta_2 - \delta_1) + (\nu_{12} - \nu_{11}) \\
\bar{y}_{22} - \bar{y}_{21} &= \beta + (\delta_2 - \delta_1) + (\nu_{12} - \nu_{11})
\end{align*}
\] (15-16)

The time effects \(\delta_1\) and \(\delta_2\) are further removed by taking the “difference-in-differences”:

\[
DiD = (\bar{y}_{22} - \bar{y}_{21}) - (\bar{y}_{12} - \bar{y}_{11}) = \beta + (\nu_{22} - \nu_{21}) - (\nu_{12} - \nu_{11})
\] (17)

Since the expected value of all components of the error term is equal to zero, it follows that \(E(DD) = \beta\), i.e. that DD is a consistent estimate of \(\beta\).

### 2.3 Alternative ways of estimating difference-in-differences

There are two regression ways of computing the difference-in-differences estimates that yield exactly the same result. The first way consists of running a regression on the grouped data with a dummy variable for the group and a dummy variable for time, in addition to the “intervention” dummy \(x_{jt}\). For instance, let \(D_{jt}\) indicate the group \((D_{jt} = 0\) if \(j = 1\), \(D_{jt} = 1\) if \(j = 2\)) and \(T_{jt}\) indicate the time period \((T_{jt} = 0\) if \(t = 1\), \(T_{jt} = 1\) if \(t = 2\)) . We can then rewrite equation (10) as:

\[
\bar{y}_{jt} = c + x_{jt}\beta + D_{jt}\theta + T_{jt}\delta + \nu_{jt}
\] (18)

where \(c\) is the constant in the model \((c = \theta_1 + \delta_1)\), \(\theta\) is the group effect for group 2 relative to group 1 \((\theta = \theta_2 - \theta_1)\), and \(\delta\) is the time effect for time 2 relative to time 1 \((\delta = \delta_2 - \delta_1)\). Note that this is a very peculiar model since we have exactly four
parameters to estimate with four observations. As a result, the model will have a perfect fit and standard errors will be equal to zero. As it turns out, however, we would get the exact same result by running OLS on the micro-data that was originally used to construct the group means $\overline{y}_{jt}$. In other words, OLS estimates of $\beta$ (and the other parameters) in the equation:

$$y_{ijt} = c + x_{jt} \beta + D_{jt} \theta + T_{jt} \delta + \nu_{ijt}$$  \hspace{1cm} (19)

will be exactly the same as in equation (18). An important advantage of running the equation on the micro data is that we then get standard errors for the estimates and can test for the significance of $\beta$.

The regression approach based on either group means or the micro data becomes particularly useful in cases where we have more than two groups and more than two time periods. In such cases, the intervention variable $x_{jt}$ may switch to one at different times for different groups, and we can no longer compute the simplest DiD estimator. It is still possible, however, to run either a group means regression:

$$\overline{y}_{jt} = c + x_{jt} \beta + \sum_{k=2}^{J} D_{jt}^{(k)} \theta_j + \sum_{l=2}^{T} T_{jt}^{(l)} \delta_t + \nu_{jt}$$  \hspace{1cm} (20)

or a regression on the micro data:

$$y_{ijt} = c + x_{jt} \beta + \sum_{k=2}^{J} D_{jt}^{(k)} \theta_j + \sum_{l=2}^{T} T_{jt}^{(l)} \delta_t + \nu_{ijt}$$  \hspace{1cm} (21)

where the $D_{jt}^{(k)}$ are dummies for groups and $T_{jt}^{(l)}$ are dummies for time periods. The other advantage of running regressions is the case where we have additional explanatory variables $z_{ijt}$ that we also want to control for. The average value of $z_{ijt}$, $\overline{z}_{jt}$, can be included in the group level regression:

$$\overline{y}_{jt} = c + x_{jt} \beta + \overline{z}_{jt} \gamma + \sum_{k=2}^{J} D_{jt}^{(k)} \theta_j + \sum_{l=2}^{T} T_{jt}^{(l)} \delta_t + \nu_{jt}$$  \hspace{1cm} (22)

while in the micro-level regression we have:

$$y_{ijt} = c + x_{jt} \beta + z_{ijt} \gamma + \sum_{k=2}^{J} D_{jt}^{(k)} \theta_j + \sum_{l=2}^{T} T_{jt}^{(l)} \delta_t + \nu_{ijt}$$  \hspace{1cm} (23)

Note that it is good practice to estimate DiD models with and without covariates to
assess the robustness of the estimate of $\beta$. Including group-level linear trends is also a useful robustness check that can be performed.

### 2.4 Event study approach

In the difference-in-differences design we just discussed, the treatment effect $\beta$ is assumed to be constant over time. But because of adjustment costs, initial low program take-up, etc., it is reasonable to think that the treatment effect may be growing over time. This is can be easily captured in the regression model by replacing the treatment dummy $x_{jt}$ with a set of treatment dummies keeping track of how long ago the treatment was introduced. For instance, consider a case where we allow the treatment effect to grow from $\beta^{(1)}$ in the first year after the treatment is introduced, to $\beta^{(2)}$ in the second year, and then to stabilize to a long term value of $\beta^{(3)}$ after two years. This more general model could be estimated as:

$$y_{ijt} = c + x_{jt}^{(1)} \beta^{(1)} + x_{jt}^{(2)} \beta^{(2)} + x_{jt}^{(3)} \beta^{(3)} + z_{ijt} \gamma + \sum_{k=2}^{J} D_{jt}^{(k)} \theta_{j} + \sum_{l=2}^{T} T_{jt}^{(l)} \delta_{l} + \nu_{ijt}$$  \hspace{1cm} (24)$$

where $x_{jt}^{(1)}$ is a dummy variable equal to 1 in the first year after the treatment takes place, $x_{jt}^{(2)}$ is a dummy variable equal to 1 in the second year after the treatment takes place, and $x_{jt}^{(3)}$ is a dummy variable equal to 1 in the third (or more) year after the treatment takes place.

Note that if the difference-in-differences design is valid, adding “placebo” treatment dummies that go back in time should not be significant. Otherwise it would indicate that the treatment and control groups were on different trajectories prior to the treatment taking place, a violation of the “parallel trend” assumption. This can be tested in the following example of an extended model:

$$y_{ijt} = c + x_{jt}^{(-1)} \beta^{(-1)} + x_{jt}^{(-2)} \beta^{(-2)} + x_{jt}^{(1)} \beta^{(1)} + x_{jt}^{(2)} \beta^{(2)} + x_{jt}^{(3)} \beta^{(3)} + \ldots$$  \hspace{1cm} (25)$$

where $x_{jt}^{(-1)}$ is a dummy variable equal to 1 in the first year before the treatment took place, and $x_{jt}^{(-2)}$ is a dummy variable equal to 1 in the second year before the treatment took place. Note that the omitted category in the regression is three or more years before the treatment took place, meaning that all the $\beta$’s are defined relative to that base. Finding that $\beta^{(-1)}$ and $\beta^{(-2)}$ are statistically significant would represent a violation
of the parallel trend assumption, and put into question the validity of the difference-in-differences design.

### 2.5 Standard errors in difference-in-differences models

In the last version of the micro data model above, notice that the key variable of interest, $x_{jt}$, is the same for all individuals $i$ in group $j$ at time $t$. This is thus a classic case where standard errors may be biased if we don’t control for group effects by clustering. From the equations above, we can rewrite the micro level error term $\nu_{ijt}$ as

$$\nu_{ijt} = \nu_{jt} + \xi_{ijt}$$  \hspace{1cm} (26)

where $\nu_{jt}$ is the group level error while $\xi_{ijt}$ can be thought as a purely random term. Since the group level error term $\nu_{jt}$ varies both over $j$ and $t$, at first pass it sounds reason to cluster standard errors at the “jt” level. For example, if $j = 1, 2, ..., 10$ and $t = 1, 2, ..., 5$, we should create a new variable $k$, for example $k = j + (t - 1) \times 10$, that would go from $k = 1$ to $k = 50$, and cluster on $k$ (or use the Stata “egen k=group(j t)” command discussed below). A very important paper on this topic is the QJE piece by Bertrand, Duflo and Mullainathan where they indeed point out the importance of clustering.\(^1\) They also point out that if $\nu_{jt}$ is correlated over time, standard errors can be further biased for the usual reason (autocorrelation). One of their proposed solution where the number of groups $j$ (states in their paper) is large enough is to do block bootstrap.

### 3 Stata commands

The command of choice for creating a quasi-panel (group means) from repeated cross sections is the “collapse” command. Say we have data by state and year and three variables: $y$ (outcome), $x$ (intervention), and $z$ (covariate). The following command can be used to create a new data set where we have the mean of $y$, $x$, and $z$ by state and year:

```
collapse (mean) y x z, by(state year)
```

Say we have 50 states and 10 years of data. We can now create state and year dummies using

\[
\text{tab state, gen(st)}
\]
\[
\text{tab year, gen(yr)}
\]

We can then run the regression at the group level:

\[
\text{reg y x z st2-st50 yr2-yr10}
\]

or use the “i.” option:

\[
\text{reg y x z i.state i.year}
\]

A minor advantage of the first option is that you can choose the base year and region to be excluded, while Stata does it automatically in the second case. Yet another option is to do:

\[
\text{xtreg y x z yr2-yr10, fe i(state)}
\]

The alternative is to go at the micro level and cluster at the state-year or state level. To create state-year dummies we can do:

\[
\text{egen state_yr=group(state year)}
\]

We can then run the micro-level regression and cluster on state_yr:

\[
\text{reg y x z i.state i.year, vce(cluster state_yr)}
\]

To do block bootstrap as suggested by Bertrand et al. just do:

\[
\text{reg y x z i.state i.year, vce(bootstrap, reps(200) cluster(state))}
\]