True Panel versus Pooled Cross Section

• A true Panel data-set tracks specific individuals or firms over a period of time.

• A Pooled Cross-section on the other hand is a collection of cross-section datasets observed at different points in time. For example, if we collect data from randomly selected households in each year, but do not observe the same households in different years then we have pooled cross-section rather than a panel data set.

• Often loosely use the term panel data to refer to any data set that has both a cross-sectional dimension and a time-series dimension.
• More precisely it’s only data following the same cross-section units over time

• Otherwise it’s a pooled cross-section
Pooled Cross Sections

- We may want to pool cross sections just to get bigger sample sizes
- We may want to pool cross sections to investigate the effect of time
- We may want to pool cross sections to investigate whether relationships have changed over time
An Example of a Pooled Cross-Section: The Current Population Survey

- The Current Population Survey (CPS) is a monthly survey of about 50,000 households conducted by the Bureau of the Census for the Bureau of Labor Statistics. The survey has been conducted for more than 50 years.

- The CPS is the primary source of information on the labor force characteristics of the U.S. population. The sample is scientifically selected to represent the civilian noninstitutional population. Respondents are interviewed to obtain information about the employment status of each member of the household 15 years of age and older.

- However, published data focus on those ages 16 and over. The sample provides estimates for the nation as a whole and serves as part of model-based estimates for individual states and other geographic areas.
Estimates obtained from the CPS include employment, unemployment, earnings, hours of work, and other indicators. They are available by a variety of demographic characteristics including age, sex, race, marital status, and educational attainment.

They are also available by occupation, industry, and class of worker. Supplemental questions to produce estimates on a variety of topics including school enrollment, income, previous work experience, health, employee benefits, and work schedules are also often added to the regular CPS questionnaire.

CPS data are used by government policymakers and legislators as important indicators of our nation's economic situation and for planning and evaluating many government programs. They are also used by the press, students, academics, and the general public.
Example: Changes in the Return to Education and the Gender Gap

- This is a model that explains log(wage) observed in two years, 1978 and 1985 with sociodemographic and education variables.

\[
\log (\text{wage}) = \beta_0 + \delta_0 y_{85} + \beta_1 \text{educ} + \delta_1 y_{85} \cdot \text{educ} + \beta_2 \text{exper} \\
+ \beta_3 \text{exper}^2 + \beta_4 \text{union} + \beta_5 \text{female} + \delta_5 y_{85} \cdot \text{female} + u
\]

where \(y_{85}\) is a dummy variable set to 1 if log(wage) is observed in 1985, union is a dummy variable if the person belongs to a union, educ is the number of years of schooling, experience is defined as age-exper-6.

- There are 550 people in the sample in 1978 and a different set of 534 people in 1985.

- The intercept for 1987 is \(\beta_0\) and for 1985 it is \(\beta_0 + \delta_0\).
• The return to education in 1987 is $\beta_1$ and for 1985 it is $\beta_1 + \delta_1$. This means that $\delta_1$ measures the change to education between 1978 and 1985.

• The wage differential between men and women in 1978 is measured by $\beta_5$ and the wage differential in 1985 is measured by $\beta_5 + \delta_5$. The null hypothesis that the wage differential has not changed is $\delta_5 = 0$ against the alternative that the gap has narrowed $\delta_5 > 0$.

• Note that hourly wages are measured in nominal wages. Since wages are increasing solely due to inflation between 1978 we need to correct for that by deflating 1985 wages with a price deflator. We should use

$$\log(wage/\text{P85}) = \log(wage) - \log(\text{P85})$$

However, since P85 is the same for all individuals we can use the nominal log wage. The difference will simply be absorbed in the constant $\delta_0$. Note
that this only works because we are assuming that the price index is the same for everybody in the sample and because we are using log(wage) rather than the nominal wage directly.

- Note that if we interacted all explanatory variables with y85 then the estimation results would be identical to estimating two separate regressions, one for 1978 and one for 1985.
The underlying idea behind Differences-in-Differences is random assignment to treatment and control groups, like in a medical experiment. Random assignment guarantees that the treatment and control groups do not systematically differ except for the fact that one group receives the treatment while the other doesn’t. Compare this with a non-random assignment. For example, if in a medical trial medication were only given to the sickest patients then measuring the effects of the drugs by comparing the health of treated versus non-treated patients would likely be biased because the treated were less healthy than the non-treated to begin with.

One can then simply compare the change in outcomes across the treatment and control groups to estimate the treatment effect.
• For time 1,2, groups A, B

• Assume that treatment occurs for individuals in group B at time $t = 2$ while individuals in group A are not treated.

• Assume that we are interested in the effect the treatment has on a certain outcome variable $y$. We use the notation $y_{t,g}$ for the outcome of individual $g = \{A, B\}$ at time $t = 1, 2$.

• The difference in differences of the outcomes is computed as

$$\left(y_{2,B} - y_{2,A}\right) - \left(y_{1,B} - y_{1,A}\right),$$

or equivalently

$$\left(y_{2,B} - y_{1,B}\right) - \left(y_{2,A} - y_{1,A}\right),$$

is the difference-in-differences
If we had a sample of observations with outcomes $y_{it}$ for each individual $i$ and for the before and after treatment times $t = 1, 2$ then we could estimate the so called treatment effect by computing the sample averages

$$(\bar{y}_{2,B} - \bar{y}_{1,B}) - (\bar{y}_{2,A} - \bar{y}_{1,A})$$

where $\bar{y}_{2,B}$ is the sample average over all outcome variables for individuals in period 2 and group $B$. The variables $\bar{y}_{1,B}$, $\bar{y}_{2,A}$ and $\bar{y}_{1,A}$ are defined correspondingly.
• A regression framework using time and treatment dummy variables can calculate this difference-in-difference as well

• Consider the model:

\[ y_{it} = \beta_0 + \beta_1 \text{treatment}_{it} + \beta_2 \text{after}_{it} + \beta_3 \text{treatment}_{it} \times \text{after}_{it} + u_{it} \]

where treatment\(_{it}\) is a dummy variable which is set =1 if individual \(i\) is in the treatment group, after\(_{it}\) is a dummy variable for time periods after the treatment is administered (this indicator is set=1 for all individuals after treatment time, even if they are not in the treatment group).

• The estimated \(\beta_3\) will be the difference-in-differences in the group means

• To see this consider the case where \(t = 1, 2\) and treatment occurs for group \(B\) at time 2. Then the predicted change in \(y_{it}\) for members of
group $B$ between time $t = 1$ and 2 is

$$y_{i2} - y_{i1} = \beta_0 + \beta_1 + \beta_2 + \beta_3 - (\beta_0 + \beta_1) = \beta_2 + \beta_3$$

- For the individuals in the control group the effect is

$$y_{i2} - y_{i1} = \beta_0 + \beta_2 - (\beta_0) = \beta_2$$

such that the difference between the two groups is

$$\beta_2 + \beta_3 - \beta_2 = \beta_3$$

- Thus, $\beta_3$ measures the difference in difference in the regression model.

- When we don’t truly have random assignment, the regression form becomes very useful.
• Additional $x$’s can be added to the regression to control for differences across the treatment and control groups

• Sometimes referred to as a “natural experiment” especially when a policy change is being analyzed

• The terminology natural experiment is justified if the policy change occurred for reasons that are exogenous to what determines the outcome variable and does not affect all the individuals in the sample. The policy change then acts as a random assignment mechanism that subjects certain randomly selected individuals in the sample to a changed environment.
Policy Analysis with Pooled Data

- Effect of Worker Compensation Laws on Weeks out of Work.

- Study the length of time in weeks that an injured worker receives workers compensation.

- On July 15, 198 Kentucky raised the cap on weekly earnings that were covered by workers compensation.

- An increase in the cap has no effect on the benefit for low income workers but it makes it less costly for a high income workers to stay on workers compensation.
Therefore the control group is low income workers and the treatment group is high income workers.

High income workers are defined as those who are subject to the per-policy change cap.

Using random samples both before and after the policy change we test whether more generous workers compensation causes people to stay out of work longer.

The estimated equation is

\[
\log(durat) = \beta_0 + \beta_1 \text{afchnge} + \beta_2 \text{highearn} \\
+ \beta_3 \text{afchnge} \cdot \text{highearn}
\]

where afchange is a dummy variable for observations after the policy change, highearn is a dummy variable for high earners.
The individuals in the sample are not necessarily the same pre and post the policy change. Thus we need to work with a pooled regression design.

The estimated coefficient $\beta_3$ is .191 which means that the time on workers compensation for high income workers increased by about 19% after the lifting of the cap.

We can include additional controls for gender, marital status, age, industry and type of injury to ensure that the differences between the pre and post policy sample are not due to systematic differences in the types of people and injuries covered in the sample.
Two-Period Panel Data

- It’s possible to use a panel just like pooled cross-sections, but can do more than that

- Panel data can be used to address some kinds of omitted variable bias

- If can think of the omitted variables as being fixed over time, then can model as having a composite error
Unobserved Fixed Effects

- Suppose the population model is defined for $t = 1, 2$ and $i = 1, 2, \ldots, n$ where $n$ is the number of cross-sectional units in the sample. Then,

$$y_{it} = \beta_0 + \delta_0 d_{2t} + \beta_1 x_{it,1} + \ldots + \beta_k x_{it,k} + a_i + u_{it}$$

where $d_{2t}$ is a dummy variable that equals zero when $t = 1$ and one when $t = 2$.

- Here we have added a time-constant component to the error,

$$v_{it} = a_i + u_{it}$$

- If $a_i$ is correlated with the $x$’s, OLS will be biased, since we $a_i$ is part of the error term
• With panel data, we can difference-out the unobserved fixed effect.

• This is called taking 'First-differences'.

• We can subtract one period from the other, to obtain

\[ y_{i2} - y_{i1} = \Delta y_i = \delta_0 + \beta_1 \Delta x_{i,1} + \ldots + \beta_k \Delta x_{i,k} + \Delta u_i \]

where

\[ \Delta y_i = y_{it} - y_{it-1} \]
\[ \Delta x_{i1} = x_{i2,1} - x_{i1,1} \]

• This model has no correlation between the \(x\)’s and the error term, so no bias.
Because the model now only has one time period (when represented in first difference form) it can be estimated with the usual regression or instrumental variable methods that we are already familiar with.

Need to be careful about organization of the data to be sure compute correct change.

Stata distinguishes two formats in which it stores panel data. The wide and the long form. In the wide form, each entry in the data matrix contains observations for the variables in both years. In the long form, all the observations in different years are stored for one individual after the other.
• Wide Form:

<table>
<thead>
<tr>
<th>State</th>
<th>( y_{90} )</th>
<th>( y_{95} )</th>
<th>( x_{90} )</th>
<th>( x_{95} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>AK</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>AL</td>
<td>3</td>
<td>4</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>AR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

To compute differences in wide form we simply use

\[
\text{gen } dy = y_{95} - y_{90}
\]

• Long Form:

<table>
<thead>
<tr>
<th>State</th>
<th>year</th>
<th>( y )</th>
<th>( x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>AK</td>
<td>90</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>AK</td>
<td>95</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>AL</td>
<td>90</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Many specialized commands for panel estimators require the data to be in long form. It is possible to convert panel data sets from long form to wide form or vice versa using the stata command \texttt{reshape}.

- A second issue that is important for the analysis of panel data are the assumptions about \( a_i \) and \( u_{it} \).

- It is usually assumed that both \( a_i \) and \( u_{it} \) are unobservable random factors. However, sometimes \( a_i \) is also assumed to be fixed value.

- If \( a_i \) is fixed then the mean of \( y_{it} \) is not

\[
E \left( y_{it} | x_{it,1}, \ldots x_{it,k} \right) = \beta_0 + \delta_0 d_{2t} + \beta_1 x_{it,1} + \ldots + \beta_k x_{it,k}
\]

but rather

\[
E \left( y_{it} | x_{it,1}, \ldots x_{it,k} \right) = \beta_0 + \delta_0 d_{2t} + \beta_1 x_{it,1} + \ldots + \beta_k x_{it,k} + a_i.
\]
This means that estimating the coefficients $\beta_0, \ldots, \beta_k$ and $\delta_0$ without taking $a_i$ into account is incorrect even if $a_i$ is fixed.

- When $a_i$ is random, the important distinction is whether $a_i$ is correlated with $x_{it,1}, \ldots, x_{it,k}$ or not.

- A second issue are the assumptions about $u_{it}$. In the simplest case we assume that $u_{it}$ is iid both across $i$ and $t$. In other words, $u_{it}$ is independent of $u_{js}$ whenever $i \neq j$ or $t \neq s$.

- This is often a very strong assumption. Especially the lack of correlation across time is often not satisfied, such that

$$\text{Cov}(u_{it}, u_{is}) \neq 0$$
• Serial correlation of $u_{it}$ does not prevent us from using OLS estimators to estimate the parameters of the model as long as $u_{it}$ is not correlated with the regressors. However, the standard errors are no longer correct in the presence of serial correlation and need to be adjusted accordingly. Stata has several options for doing this.

• We also assume that

$$\text{Cov}(a_i, u_{it}) = 0$$

and that

$$\text{Cov}(u_{it}, x_{it,j}) = 0$$

• The last condition may fail for the same reasons that we discussed for the instrumental variables estimators. Panel models can also be estimated using instruments.
The effect of drunk driving laws on traffic fatalities:

Many states in the US have adopted different policies in an attempt to curb drunk driving. Two types of laws are open container laws (which make it illegal for passengers to have open containers of alcoholic beverages) and administrative per se laws (which allow courts to suspend licenses after a driver is arrested for drunk driving but before the driver is convicted).

One possible analysis is to use a single cross section of states to regress driving fatalities on dummy variable indicators for whether each law is present.
• This is unlikely to work well because states decide, through a legislative process, whether they need such laws. Therefore, the presence of laws is likely to be related to the average drunk driving fatalities in recent years.

• A more convincing analysis uses panel data over a time period where some states adopted new laws and some states may have repealed existing laws.

• Consider data for 1985 and 1990 for all 50 states and the District of Columbia.

• The dependent variable is the number of traffic deaths per 100 million miles driven.

• In 1985, 19 states had open container laws, while 22 states had such laws in 1990.
• In 1985, 21 states had per se laws with 29 states with such laws in 1990.

• The estimated equation is

\[ \Delta dthrt = \beta_0 + \beta_1 \Delta open + \beta_2 \Delta admn + u \]

• This is a true panel data set, because all states are observed in the sample for both time periods. It is important to include fixed effects because states may differ in their characteristics which may affect their traffic fatality rates.

• We might want to add other controls such as seat belt laws, motorcycle helmet laws, maximum speed limits, age and gender composition of drivers.
Differencing with Multiple Periods

- Can extend this method to more periods

- Simply difference adjacent periods

- So if 3 periods, then subtract period 1 from period 2, period 2 from period 3 and have 2 observations per individual

- Simply estimate by OLS, assuming the $\Delta u_{it}$ are uncorrelated over time.

- More specifically, consider the following simple model

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$$
for \( t = 1, \ldots, T \) and \( i = 1, \ldots, n \). After taking first differences we have

\[
\Delta y_{it} = \beta_1 \Delta x_{it} + \Delta u_{it}
\]

which is available for \( t = 2, \ldots, T \) (why?) and \( i = 1, \ldots n \).

- We can now use OLS to estimate the parameter \( \beta_1 \), treating all the data equally: This estimator is called the pooled OLS estimator because all data are in the same pool and used as if they were a single cross-section. We have

\[
\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \sum_{t=2}^{T} \Delta x_{it} \Delta y_{it}}{\sum_{i=1}^{n} \sum_{t=2}^{T} (\Delta x_{it})^2}
\]

and after substituting for \( y_{it} \)

\[
\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} \sum_{t=2}^{T} \Delta x_{it} \Delta u_{it}}{\sum_{i=1}^{n} \sum_{t=2}^{T} (\Delta x_{it})^2}
\]

From the usual results for OLS we know that the usual standard error formulas require that the innovations \( \Delta u_{it} \) are iid.
• However, this can not hold for this model and the standard errors need to be adjusted. Note that even if \( u_{it} \) are serially uncorrelated the same will not be the case for \( \Delta u_{it} \). To see this note that

\[
\text{Cov}(\Delta u_{it}, \Delta u_{it-1}) \\
= E( (u_{it} - u_{it-1})(u_{it-1} - u_{it-2}) ) \\
= E( u_{it}u_{it-1} - u_{it-1}^2 - u_{it}u_{it-2} + u_{it-1}u_{it-2} ) \\
= E( u_{it-1}^2 ) = \sigma^2
\]

and

\[
\text{Cov}(\Delta u_{it}, \Delta u_{it-j}) = 0
\]

for \( j > 1 \).

• As we saw earlier, the assumption that \( u_{it} \) is not serially correlated, is quite strong and likely not satisfied. In that case, the correlation pattern of \( \Delta u_{it} \) may be even more complicated.
• We can test for serial correlation as follows. We first estimate the differenced model by pooled OLS. Then we obtain the residuals $\hat{r}_{it}$ where $r_{it} = \Delta u_{it}$. If $r_{it}$ follows an autoregressive process then

$$r_{it} = \rho r_{it-1} + e_{it}$$

where $e_{it}$ is an independent error. In other words, if $r_{it}$ is serially correlated, then $\rho \neq 0$.

• We thus run the pooled regression of $\hat{r}_{it}$ on $\hat{r}_{it-1}$ for $t = 3, \ldots, T$ and $i = 1, \ldots, n$ and compute the $t$-statistic for the coefficient on $\hat{r}_{it-1}$.

• The model above is not a very good specification for most empirical work because it does not have a constant term and does not allow for year specific factors that affect all individuals equally. It is thus more common
to consider for $t = 1, \ldots, T$,

$$y_{it} = \tilde{\alpha}_0 + \tilde{\alpha}_1 d_{1t} + \tilde{\alpha}_2 d_{2t} + \ldots + \tilde{\alpha}_T dt_t + \beta_1 x_{it,1} + \ldots + \beta_k x_{it,k} + u_{it}$$

After taking first differences we obtain for $t = 2, \ldots, T$

$$\Delta y_{it} = \tilde{\alpha}_2 \Delta d_{2t} + \ldots + \tilde{\alpha}_T \Delta dt_t + \beta_1 \Delta x_{it,1} + \ldots + \beta_k \Delta x_{it,k} + \Delta u_{it}$$

where now in period $t = 2$, $\Delta d_{2t} = 1$ and in $t = 3$, $\Delta d_{2t} = -1$ and so on. This model does not contain an intercept which is inconvenient for example for the calculation of an $R^2$ measure. It is thus better to specify the model

$$\Delta y_{it} = \alpha_0 + \alpha_3 d_{3t} + \ldots + \alpha_T dt_t + \beta_1 \Delta x_{it,1} + \ldots + \beta_k \Delta x_{it,k} + \Delta u_{it}$$

It can be shown that this model is equivalent to the specification with differenced year dummies, as far as the parameters $\beta_1, \ldots, \beta_k$ are concerned
(not as far as $\alpha$ is concerned). Thus, if the interest is in these parameters and not in the $\alpha$, then the last specification is preferred.

- Example: Cornwell and Trumbull (1994) used data on 90 counties in north Carolina, for the years 1981 through 1987 to estimate an unobserved effects model of crime; they explain the crime rate with factors such as the probability of arrest, the probability of conviction, the probability of serving time in jail, and the length of the sentences.

- A first difference specification is used to eliminate the unobserved effect $a_i$. Various factors including geographical location, attitudes toward crime, historical records and reporting conventions might be contained in $a_i$.

- The crime rate is the number of crimes per person, prbarr is the estimated probability of arrest, prbconv is the estimated probability of conviction
(given arrest), prbpris is the probability of serving time in prison (given a conviction), avgsen is the average sentence length served, and polpc is the number of police officers per capita.

- Logs of all variables are used to estimate elasticities.

- Also include a full set of year dummies to control for state trends in crime rates.

- Can use the years 1982 through 1987 to estimate the differenced equation

\[
\Delta \log (\text{crmrte}_t) = \alpha_0 + \alpha_1 d83_t + \alpha_2 d84_t + \ldots + \alpha_5 d87_t \\
+ \beta_1 \Delta \log (\text{prbarr}) + \beta_2 \Delta \log (\text{prbconv}) \\
+ \beta_3 \Delta \log (\text{prbpris}) + \beta_4 \Delta \log (\text{avgsen}) \\
+ \beta_5 \Delta \log (\text{polpc}) + u_{it}
\]

where \( dx_t \) are year dummy variables.
There is quite strong evidence of serial correlation and heteroskedasticity in the error terms of this equation.
Example: Cellular Telephone regulation (Hausman and Kuersteiner, 2008, Section 5)  
In the U.S. for the first 12 years of operation, 1983-1995, cellular telephone operated as a duopoly. However, the two facilities-based carriers were required to sell cellular airtime to resellers who also sold cellular service to consumers. In the U.S. each of 51 state regulatory commissions decided on whether to regulate cellular prices or to use market outcomes. In an interesting natural experiment 26 states regulated cellular prices, while the other 25 did not.

A “natural experiment” occurred that allowed a further test of the regulatory hypothesis. In 1993 U.S. Congress instructed the Federal Communications Commission (FCC) to deregulate cellular prices unless a given state that was regulating cellular prices could show price regulation was
“necessary”. Eight states petitioned the FCC to continue price regulation, and the FCC turned them down in late 1994. One state appealed, but regulation completely ended in 1995. Thus, Congress and the FCC provided a natural experiment that permitted an analysis of how cellular prices changed in the regulated and unregulated states, after price regulation was prohibited.

- A complicating factor arose because cellular prices decreased significantly in 1995-96 both because of new PCS entry and because of deregulation. Thus, the econometric specification, a time effect for each year, which allows for the effect of new entry. A single indicator variable allows for the effect of price regulation. The econometric specification was estimated over 11 years of data with 7 years prior to the end of price regulation and 4 years after the end of regulation. Given the 30 MSAs we have a total of 330 observations.
We estimate the model

\[ y_{it} = \beta_t + T_{it}\gamma + \alpha_i + u_{it} \]

where \( \beta_t \) is the time fixed effect for period \( t \), \( T_{it} \) is an indicator variable which is one for the MSA \( i \) that deregulated in 1994 and for \( t > 1994 \).