1 Data Types

Data can come to us in various formats. Let's start by visualizing what these kinds of data look like. We will refer to "unit" as the individual people, cities, firms, etc. in a given dataset.

Cross-sectional data is a snapshot of a group of individuals/units at *one point in time*. When a sample of units is randomly selected, we can think of it as a representative "cross section" of the population. Since we use i to index units (people, firms, cities, etc.), the notation for cross sectional data is what you've seen before. The following example is for a cross-section of individuals.

$$wage_i = \beta_0 + \beta_1 edu_i + \beta_2 exper_i + \beta_3 female_i + u_i$$

indiv	wage	edu	exper	female
1	3.10	11	2	1
2	3.24	12	22	1
.	.	.	.	.
100	5.30	12	7	0

Pooled cross-sectional data is multiple snapshots of multiple groups of (randomly selected) units at *many points in time*. Suppose we have two cross-sectional datasets from two different years; *pooling* the data means to treat them as one larger sample and control for the fact that some observations are from a different year. We can use the same notation here as in cross-section, indexing each person, firm, city, etc. by i. In the example below, the variable $y2010_i$ captures whether house i from the pooled sample was observed in 2010 (as opposed to 2000).

$$hprice_i = \beta_0 + \beta_1 bdrms_i + \beta_2 bthrms_i + \beta_3 sqrft_i + \delta y2010_i + u_i$$

house	year	hprice	bdrms	bthrms	sqrft
1	2000	85,500	3	2.0	1600
2	2000	67,300	3	2.5	1400
.	.	.	.	.	.
100	2000	134,000	4	2.5	2000
101	2010	243,000	4	3.0	2600
102	2010	65,000	2	1.0	1250
.	.	.	.	.	.
200	2010	144,000	3	2	2000

Finally, **panel data** is more like a movie than a snapshot because it tracks particular people, firms, cities, etc. over time. We observe the *same cross-section* in *multiple time periods*. With panel data, we start indexing observations by t as well as i to distinguish between our observations of unit i at various points in time. The following example is panel data from cities on murder rates, population density and police arrests.

city	year	murder rate	pop density	police
1	2000	9.3	2.24	440
1	2001	11.6	2.38	471
2	2000	7.6	1.61	75
2	2001	10.3	1.73	75
.	.	.	.	.
100	2000	11.1	3.12	520
100	2001	17.2	3.34	493

$$murders_{it} = \beta_0 + \beta_1 pop_{it} + \beta_2 police_{it} + a_i + d_t + u_{it}$$

a_i is a vector of dummy variables for each unit i (except one reference unit) and d_t is a vector of dummy variables for each time period t (except one reference period). We call these "fixed effects" (more on this in the next section of the notes).

2 Two-Period Panel

2.1 Fixed Effects

In panel data where we observe the same units in multiple periods, we can include controls for specific units. We do this with what we call a **unit fixed effect**, which we can denote as a_i or α_i (for notational simplicity). A unit fixed effect is a vector of $n-1$ dummy variables, where n is the number of units in our data. Thus, a given dummy variable a_i is coded as 1 for unit i and 0 for all other units. The unit fixed effect captures all time-constant factors within the unit that affect Y_{it} (the fact that this term is not indexed by a time subscript t reminds us that it does not change over time). In the example above, this would be all city characteristics that don't change over time.

Unit fixed effects are extremely useful because they reduce the number of potential omitted variables we could be concerned about. Many time-constant variables could affect Y_{it} and cause bias in estimates of a particular coefficient of interest if we are using cross-sectional data. With panel data, we can include the unit fixed effects and control for all of those variables at once. We still will be concerned about time-varying omitted variables, but we have at least reduced the set of possible omitted variables. Note that we can only include unit fixed effects with panel data where the outcome variable changes over time. Otherwise, including the fixed effect perfectly predicts the outcome.

What does including a unit fixed effect do? We know that the unit fixed effect accounts for all characteristics of the unit that don't change over time, but what about characteristics that *do* change over time? The unit fixed effect essentially controls for the *mean* values of those variables within the given unit across the time periods observed. The mean of a variable within a unit is constant over time. Including unit fixed effects therefore means that the **source of variation** we use to identify our β_k for a given X_k is variation in those variables within units over time, relative to the means for those variables within units. As a consequence, including unit fixed effects means that we cannot include any additional X variables that don't change over time within units because they will already be subsumed by the fixed effect for our unit of interest—they do not vary relative to their within-unit mean. In addition, if any X_k only changes over time for certain units but not others, the estimated coefficient β_k will be identified just off the units that have variation in X_k, which may be particular subset of the whole sample.

In addition to unit fixed effects, with panel or repeated cross-sectional data, we can also include **time fixed effects**. These follow the same concept as unit fixed effects, but are dummy variables for each time period observed. We often represent the vector of time period dummies (fixed effects) by δ_t or d_t. These time fixed effects capture all variables that change over time in the same way across units. If a variable changes over time in a different way across units, it will not be included in the time fixed effects.

2.2 Example

Let's consider an example of panel data where you have data on crime and unemployment rates for 46 cities for 1982 and 1987. Therefore we have two time periods, and we can label them as $t = 1$ for 1982, and $t = 2$ for 1987.

Cross-sectional analysis: What happens if we use the 1987 cross section and run a simple regression of crime on unemployment?

$$\widehat{crmrte} = \underset{(20.76)}{128.38} - \underset{(3.42)}{4.16}unemp$$

If we were to interpret the coefficient on unemployment, we would infer that higher unemployment is associated with less crime. This seems backwards. The culprit? Well we might first think about omitted variable bias. The first solution that comes to mind is to control for more factors, such as land area, part of the country (West or East), police officers per square mile, law enforcement expenditure, and per capita income. We get the following result:

$$\widehat{crmrte} = \underset{(2.74)}{140.06} - \underset{(1.80)}{6.7}unem + \underset{(1.23)}{0.059}area - \underset{(1.79)}{21.963}west - \underset{(0.17)}{0.114}police + \underset{(1.15)}{0.021}lawexp - \underset{(0.53)}{0.002}pcinc$$

We still get this puzzling negative relationship between unemployment and crime. Is this the true relationship, or are there still omitted variables we are missing?

Pooled cross-sectional analysis: Since we have two time periods, we can take advantage of this by using both years of data and controlling for what time period an observation is in. This would account for factors that changed over time and are associated with both unemployment and crime. Doing this, we obtain

$$\widehat{crmrte} = 93.42 + 7.94d87 + 0.427unem$$

Here we recover the positive relationship we expected! But we are effectively treating the data as a pooled cross-section and not taking advantage of the fact that we observe a panel of the same cities multiple times. Doing so could help us address some remaining concerns about OVB even after controlling for time period.

Panel data analysis - first differences
One potential solution we can use with panel data is to take first differences. Because the set of cities in our dataset is constant over time, we can difference the data across the two years. Taking first differences (within cities) tells us how variables are changing within cities over time, and controls for all city characteristics that don't change over time. This is useful, since what we want to estimate is how a change in unemployment affects crime.

For an observation i measured in two time periods (where t is period 1 or period 2), we can think of separate regressions for each time period as follows:

$$Y_{i2} = (\beta_0 + \delta_0) + \beta_1 X_{i2} + \alpha_i + u_{i2}$$
$$Y_{i1} = \beta_0 + \beta_1 X_{i1} + \alpha_i + u_{i1}$$

In this specification, α_i is the subset of variables in $u_{i,t}$ that includes all time-constant characteristics of unit i that affect the outcome Y. We don't know what these are since we don't observe u_i, but we can assume that u_{it} includes some time-invariant variables, and for notational purposes we group them together in α_i. Note that we assume that the effect of X on Y is constant over time (β_1), but we allow there to be a different baseline level (intercept) of Y in the two time periods, where that difference is captured by δ_0.

Subtracting the second equation from the first gives:

$$\Delta Y_i = \delta_0 + \beta_1 \Delta X_i + \Delta u_i$$

The most important thing to note about this formula is that α_i has been "differenced away". We don't even have to know what those time-constant omitted variables are - we have accounted for them! We can analyze this expression using the same methods as before in the class, just defining the changes in variables as random variables (i.e., let $\tilde{Y} = \Delta Y$, etc.). You will recover causal estimates as long as the usual OLS assumptions hold, with some changes to account for the data structure. Most importantly, the zero conditional mean assumption now requires that Δu_i is uncorrelated with ΔX_i.

After taking first differences in our data, we obtain:

$$\widehat{\Delta crmrte} = 15.40 + 2.22\Delta unem$$

Here a one unit change in the unemployment rate over time within cities is associated with a 2.22 unit increase in the change in crime rate over time. By taking first differences, we have controlled for all city characteristics that do not change over time.

Panel data analysis - fixed effects
First differences allowed us to eliminate all unobserved, time-constant factors that affect crime rates in a given city i from our estimation. An alternative method to accomplish this is to directly control for those city-specific time-constant factors. We can do this by introducing *unit fixed effects*. In this case, with cities, we can create a dummy for each city. We get the following result:

$$\widehat{crmrte} = \underset{(1.95)}{91.618} + \underset{(1.80)}{2.932unem} + \underset{(1.03)}{1.838police} - \underset{(0.51)}{0.007lawexp} - \underset{(1.00)}{0.006pcinc} + \alpha_2 city2 + \cdots + \alpha_{46} city46 + \delta d87$$

As with the first differences approach, we now have something that makes much more sense: an increase in the unemployment rate is associated with an increase in the crime rate. In fact, with two time periods, fixed effects and first differences give the same results for the independent variables of interest: the difference with the previous regression using first differences is due to added time-varying controls in this fixed effects specification.

Comparing this regression to those above, you can see that we include a control for year (*d87*) to capture any factors broadly affecting crime rate over time, as we could with repeated cross-sections. The panel nature of the data allows us to include unit fixed effects, which accomplish the same thing as taking first differences. Because unit fixed effects capture time-constant variables, we can no longer include *area* or *west* (included in our initial cross-sectional attempt to deal with OVB) because these do not change over time so are absorbed in the fixed effect.

3 Assumptions for FE model

Consider the following model:

$$Y_{it} = \beta_1 X_{it1} + \beta_2 X_{it2} + \cdots + \beta_k X_{itk} + \alpha_i + u_{it}$$

1. Assumption 1: $\mathbb{E}[u_{it}|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}, \alpha_i] = 0$, or equivalently $E(\Delta u_i|\Delta X_i) = 0$ (since the former case rules out the time-invariant parts of u and X and the common trends over time in these variables). This assumption says that we don't want changes in the u's to be correlated with changes in the $\mathbf{X}$'s. There are no omitted lagged effects and there is not feedback from u to future $\mathbf{X}$.
2. Assumption 2: $\{\mathbf{W}_i\}_{i=1}^n$ are i.i.d., where $\mathbf{W}_i = \{Y_{i1}, \ldots, Y_{iT}, \mathbf{X}'_{i1}, \ldots, \mathbf{X}'_{iT}\}$ is the data for unit i. This is satisfied if units are randomly sampled from their population by simple random sampling, then data for those units are collected over time. This does not require observations to be i.i.d. over time for the same unit.
3. Assumption 3: $\mathbb{E}[\sum_{t=1}^T \tilde{\mathbf{X}}_{it}\tilde{\mathbf{X}}'_{it}]$ has rank k. This is the no multicollinearity assumption for the fixed effect case. Each explanatory variable changes over time (for at least some i), and no perfect linear relationships exist among the explanatory variables. This is important: we can't use first differences or fixed effects to analyze impacts of independent variables that don't change over time.
4. Assumption 4: $(\mathbf{X}'_{it}, u_{it})$ have nonzero finite fourth moments. Large outliers are unlikely.

As before, from Assumption $A1 \rightarrow A4$, we get that $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ is unbiased for β and converges to β. From the above assumption, we also get an expression we can estimate for $var(\hat{\beta})$.

Exercise

Consider the two panel data regressions below, where i indexes individuals and t indexes time in months:

$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + ... + \beta_k X_{k,it} + u_{it} \quad (1)$$
$$Y_{it} = \beta_0 + \beta_1 X_{1,it} + ... + \beta_k X_{k,it} + a_i + u_{it} \quad (2)$$

1. What are the zero conditional mean assumptions for each model?

2. We've talked about how OVB is a violation of the zero conditional mean assumption. What kind of omitted variable bias is mitigated by using model (2) instead of model (1)? [Why is model (2) *better* than model (1)?]

4 Longer Panels

The previous section focuses on the case with two time periods. But panels can be of any length - as many time periods (which can be seconds, hours, days, months, quarters, etc.) as you can collect data for. With longer panels, we will generally use the same approaches we talked about previously but with some small tweaks.

4.1 Example: General Period Panel Data Analysis

Let's consider an example of panel data, where the unit of observation is a city-year, and suppose we have data for 3 cities for 3 years—so 9 total observations in our dataset. So in contrast to the previous example, we now have multiple years of data. The data look as follows (note the unit and time period dummy variables):

i	t	murder rate	pop density	City1	City2	City3	Yr00	Yr01	Yr02
1	2000	9.3	2.24	1	0	0	1	0	0
1	2001	11.6	2.38	1	0	0	0	1	0
1	2002	11.8	2.42	1	0	0	0	0	1
2	2000	7.6	1.61	0	1	0	1	0	0
2	2001	10.3	1.73	0	1	0	0	1	0
2	2002	11.9	1.81	0	1	0	0	0	1
3	2000	11.1	6.00	0	0	1	1	0	0
3	2001	17.2	6.33	0	0	1	0	1	0
3	2002	20.3	6.42	0	0	1	0	0	1

Since we have multiple observations for each city, we can run the following regression:

$$murders_{it} = \beta_0 + \beta_1 popden_{it} + \alpha_2 City2 + \alpha_3 City3 + \delta_2 Yr01 + \delta_3 Yr02 + u_{it}$$

1. How do we interpret β_1, α_3 or δ_3 here?

2. How would we get the predicted murder rate for city 3 in the year 2002?

For fixed effect regressions, we usually save time by writing an α_i instead of writing out each dummy variable for the unit fixed effects. You can imagine that if we had 40 cities instead of 3, writing out each

dummy variable would get super tedious. Similarly, we usually write δ_t instead of writing out each dummy variable for the time fixed effects.

$$murder_{it} = \beta_0 + \beta_1 popden_{it} + \alpha_i + \delta_t + u_{it}$$

Note the subscripts on these variables: for a given city, its city dummy variable isn't going to vary by year, and for a given year, its year dummy variable isn't going to vary by city. Now that we have both time and city fixed effects, we can only include additional X variables that vary across both time t and across units i. X variables that vary across time in the same way for all cities are captured by the time fixed effects. X variables that do not vary across time within cities are captured by the city fixed effects.

4.2 Taking Differences

In the context of multiple-period panels, we can still think of **first differences** as subtracting off the value of the outcome variable y for the prior time period. Suppose our model for each time period looks like the following:

$$\begin{aligned} Y_{i1} &= \beta_0 + \beta_1 X_{1i1} + \cdots + \beta_k X_{ki1} + \alpha_i + u_{i1} \\ Y_{i2} &= \beta_0 + \beta_1 X_{1i2} + \cdots + \beta_k X_{ki2} + \alpha_i + \delta_2 + u_{i2} \\ Y_{i3} &= \beta_0 + \beta_1 X_{1i3} + \cdots + \beta_k X_{ki3} + \alpha_i + \delta_3 + u_{i3} \end{aligned}$$

Then taking first differences gives us

$$\begin{aligned} Y_{i2} - Y_{i1} &= \delta_2 + \beta_1 (X_{1i2} - X_{1i1}) + \cdots + \beta_k (X_{ki2} - X_{ki1}) + (u_{i2} - u_{i1}) \\ Y_{i3} - Y_{i3} &= (\delta_3 - \delta_2) + \beta_1 (X_{1i3} - X_{1i2}) + \cdots + \beta_k (X_{ki3} - X_{ki2}) + (u_{i3} - u_{i2}) \end{aligned}$$

Note that you can't take first differences for observations in the first time period. Essentially, we drop that set of observations in order to be able to calculate the first differences in all the other time periods.

This looks similar to what we saw before:

$$\Delta Y_{it} = \Delta \delta_t + \beta_1 \Delta X_{1it} + \cdots + \beta_k \Delta X_{kit} + \Delta u_{it}$$

though we now have a different intercept in each year. In practice, we'll include time period dummies for each period in the data to deal with this.

One thing to note is that now we have the same error term appearing in multiple differenced observations: for example, u_{i2} appears in both first differenced models above. To recover causal estimates we then need to slightly modify zero conditional mean assumption to require **strict exogeneity**:

$$cov(X_{jit}, u_{is}) = 0 \; \forall t, s, j$$

This means that the unobserved term in every period s is uncorrelated with all of your X_j variables in every period t, not just when $s = t$.

With multiple periods of panel data, we are not restricted to taking the *first* difference. Taking *any* difference will achieve the same objective of differencing out the α_i term. But in general, taking first differences is the most straightforward.

4.3 Fixed Effects

With this in mind, we can think of unit fixed effects regression as a specific type of difference regression: one where we *difference off the mean* within units. This is called the *within transformation*. We can write

$$Y_{it} - \bar{Y}_i = \delta_t - \bar{\delta}_t + \beta_1 (X_{1it} - \bar{X}_{1i}) + \cdots + \beta_k (X_{kit} - \bar{X}_{kit}) + u_{it} \tag{3}$$

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \cdots + \beta_k X_{kit} + \sum_i c_i \alpha_i + \delta_t + u_{it} \tag{4}$$

and these models are equivalent, but the fixed effects model (equation (4), where the α terms are unit dummies and the c terms are coefficients on those dummies) is much easier to estimate. The interpretation for c_i is the mean difference across time in outcome Y for unit i relative to the reference unit, after controlling for other observed characteristics X. Note that when we actually run fixed effects regressions, we don't usually care about these coefficients, so we use specific fixed effects regression functions that don't output all of those unit dummy coefficients.

Fixed Effects vs. First Differences

First differences and fixed effects both also require the same assumption of strict exogeneity: a generalization of the zero conditional mean assumption that holds across both variables and time periods: $cov(X_{jit}, u_{is}) = 0$ for all t, s, j. We saw above why this is for first differences. For unit fixed effects, it is because including unit fixed effects implicitly controls for the within-unit mean for each variable, meaning X variables for different time periods are included in the same model (think about how a mean is calculated). So in general, in both cases we will still have a concern that changes in u could be correlated with changes in some X variables.

So what is different between these two approaches? With two time periods, regression using unit fixed effects and first differences will be exactly identical. In both cases you end up estimating effects of within-unit variation relative to one left-out time period. With multiple time periods, first differences and unit fixed effects give slightly different results, because you end up estimating effects of different variation (relative to a prior period vs. a within-unit mean).

In the case of multiple time periods, if strict exogeneity (and the other regression assumptions) hold then the estimators will be unbiased under both approaches. What might distinguish them is the standard errors—whether fixed effects or first differences has lower standard errors depends on u.

The issue with the u_{it} terms is that both first differences and unit fixed effects result in error terms that are correlated over time. With first differences, the error terms for a given unit in two adjacent periods will include one of the same terms (the error in the first of the two periods), while with fixed effects the error term in every period will include the mean of the error term. Since those terms appear in the composite error terms, the covariance of error terms within units over time is not 0. In math for example with observations in periods 2 and 3, we would have

$$cov(u_{i3} - u_{i2}, u_{i2} - u_{i1}) \neq 0 \tag{5}$$

$$cov(u_{i3} - \frac{1}{T}\sum_s u_{is}, u_{i2} - \frac{1}{T}\sum_s u_{is}) \neq 0 \tag{6}$$

for first differences and fixed effects, respectively. It is not possible to say which correlation is bigger (and thus which method will have lower variance of the residuals). If u_{it} are similar to u_{it-1} then first differences may have a lower variance. If u_{it} are close to random within individuals over time, then fixed effects will have a lower variance.

Exercise

Suppose we want to analyze the impact of daycare for young children on parents' hours of work. We have data from both parents in 100 households with children aged 5 and under in the same zip code over 12 months in a calendar year.

Our data include the following variables:

- *hours*: Hours of work in the last 7 days, coded as 0 if the adult is not working
- *month*: An indicator of what month t it is
- *daycare*: A dummy variable taking a value of 1 if the household has a child in daycare and 0 otherwise
- *sex*: The sex of the adult

- *sector*: The ISIC code for the sector of employment for the adult, coded as 0 if the adult is not working
- *hhid*: A unique ID for the household h an adult is in
- *individ*: An ID number for each adult i within the household

We first estimate the following regression:

$$hours_{iht} = \beta_0 + \beta_1 daycare_{ht} + \beta_2 sex_{ih} + \beta_3 sector_{iht} + month_t + u_{iht} \tag{7}$$

where $month_t$ is a month fixed effect. Note the subscripts, which indicate which variables vary at the individual i, household h, and time t levels.

1. What does the month fixed effect control for?

2. Does this regression recover the causal impact of daycare on work hours? If not, what could be a possible source of omitted variable bias?

You are concerned that there might be omitted variable bias from household or individual characteristics that might be associated with the decision to put a child in daycare and with work hours. You therefore estimate the following fixed effects regression:

$$hours_{iht} = \beta_0 + \beta_1 daycare_{ht} + \beta_2 sector_{iht} + \alpha_{ih} + month_t + u_{iht} \tag{8}$$

where $month_t$ is a month fixed effect and α_{ih} is an individual fixed effect for person i in household h.

3. What does the individual fixed effect control for?

4. Why do we no longer include *sex* in the model?

5. What households are providing the information we use to estimate β_1 in a unit fixed effects model?

6. Does this regression recover the causal impact of daycare on work hours? If not, what could be a possible source of omitted variable bias?