1 Panel Data Analysis (Continued)

1.1 Other Considerations With Panel Data

There are two questions we would like to talk about in detail. How do we estimate the standard errors of the estimators? What types of correlation can we allow in this form of the variance matrix?

$$\begin{array}{cccccc}
 & i=1 & i=2 & i=3 & \cdots & i=n \\
t=1 & u_{11} & u_{21} & u_{31} & \cdots & u_{n1} \\
\vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\
t=T & u_{1T} & u_{2T} & u_{3T} & \cdots & u_{nT}
\end{array}$$

The result of **correlated errors** within observations is that we have fewer actual observations for estimation purposes than the number in our sample. Essentially, if observations are correlated over time, we learn less from each individual observation.

We can correct for this in our estimation using what we call **clustering** of errors. This treats observations within a cluster as correlated rather than independent, and calculates standard errors assuming only different clusters are independent. Most often, we will cluster at the level of the unit of analysis, e.g., the individuals, cities, businesses that you observe over time. Clustering at the individual level recognizes that observations of a given individual over time are likely to be correlated with each other, but assumes that observations of different individuals are independent (valid under some circumstances). It is usually a good idea to cluster at the level of the cross-sectional unit with panel data. The clusters can be other groupings as well. You can also allow for correlation of u_{it} between individuals within a given group, as long as there is independence across groups. For example, i runs over individuals, the clusters can be families (correlation of u_{it} for i within each family, but not between families). Another term for this kind of asymptotic variance is heteroscedasticity-and-autocorrelation-consistent (HAC) asymptotic variance.

Another concern with panel data is covariates that violate the strict exogeneity assumption. One example of this would be **lagged dependent variables** (LDVs). A *lag* is a value of a variable from a prior period, e.g. Y_{it-1}. You can include a variety of different lags with panel data (of both the outcome variable and your independent variables).

LDVs can be attractive with panel data if we think the same unobserved variables influence our outcomes over time. If we control for a prior period's outcome variable using a lag, this effectively controls for all the unobserved factors that influenced the prior period's outcome and that might also influence the current period's outcome.

While LDVs are nice, they are essentially doing the same sort of thing as fixed effects or first differences, so we are basically controlling for the same thing twice. So in general, we would not want to do both at the same time, as that would violate the strict exogeneity assumption.

A final consideration with panel data is whether our panel is **balanced**. *Balanced* panels mean we have observations for each unit (e.g., each individual) in each time period. *Unbalanced* panels mean that for some (or all) units, there are some time periods where we do not observe them - we have missing data.

In unbalanced panels, taking first (or other) differences is tricky, because we can't take a difference across a period with missing data. Trying to take first differences in an unbalanced panel will thus result in dropping more observations than you would like. Fixed effects does not have this issue, as we can still include dummies for each unique unit, which controls implicitly for the mean variable values across time periods where each unit is observed. So fixed effects will not drop additional data.

Should we be concerned about unbalanced panels? Maybe, if we think there are factors determining whether a unit (e.g., an individual) is observed that are correlated with the outcome of interest. For example, if we have data on wages for a group of individuals over time and only observe those individuals in periods where they are working, we might think there are important factors changing their circumstances in periods when they are not working where we do not observe them while they are earning 0 wages. This could matter if we are looking for the impact of wages on some outcome, like health. Fixed effects will deal with this if factors affecting whether a unit is observed are constant over time, but coefficients will be biased if the factors determining whether a unit is observed are associated with changes in the Y variable of interest (in this case, their health outcome).

1.2 Exercise 1

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:

$$hours_{iht} = \beta_0 + \beta_1 daycare_{ht} + \beta_2 sex_{ih} + \beta_3 sector_{iht} + month_t + u_{iht} \tag{1}$$

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:

$$hours_{iht} = \beta_0 + \beta_1 daycare_{ht} + \beta_2 sector_{iht} + \alpha_{ih} + month_t + u_{iht} \tag{2}$$

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?

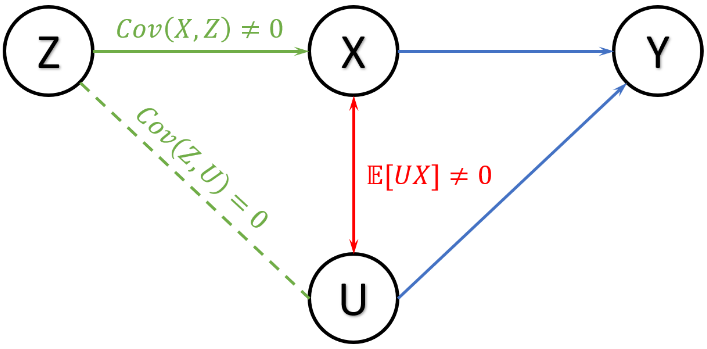
5. What households are providing the information we use to estimate β_1 in a unit fixed effects model?

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?

2 Instrumental Variables

2.1 Introduction

Panel data methods are tools that help deal with omitted variable bias (also referred to as endogeneity). However, we often don't have the luxury of panel data: carrying out surveys can often be a large logistical undertaking, also involving thousands of dollars. We also may suspect the omitted variables change over time, which limits the identification strategies that we can credibly employ. We will also not be able to randomize many treatments of interest.



Instrumental Variables (IVs) are another method we can apply to deal with omitted variable bias (OVB). This is a powerful method, as it can generate unbiased β estimates even in the presence of OVB. Essentially, IVs are special X variables (usually denoted Z) that satisfy *two specific conditions*, allowing the researcher to overcome OVB and estimate the causal effect of and endogenous X (correlated with the error) on Y:

1. **Relevance:** Z must be related to our endogenous variable of interest X

   $\Rightarrow$ $Cov(X,Z) \neq 0$

2. **Exogeneity / Exclusion:** Z should be unrelated to Y except indirectly via its effect on X. In other words, Z should be uncorrelated with all omitted variables.

   $\Rightarrow$ $Cov(Z,U) = 0$

Intuitively, we recognize that our variable of interest X is not as good as random. But we find something that *is* as good as random, Z which impacts X, and we use Z to learn about the effect of X on Y using only the variation in X that is as good as random (explained by Z).

The exogeneity/exclusion restriction is analogous to the zero conditional mean assumption, except now we don't need X to be unrelated to U, just Z.

Graphically, the intuition for IV is as follows. We are concerned about omitted variables and reverse causality, meaning we won't get a causal estimate of β from Y on X as $E[UX] \neq 0$. This is shown by the solid red arrow between X and U. But if we have an instrument Z that is correlated with X (green arrow on the left) and only affects Y through its effect on X (none of the effect goes through the path highlighted by the dashed green line). This implies that the effect of Z on Y is only happening through X.

So how do we estimate β_1? The equation we wanted to estimate is

$$Y = \beta_0 + \beta_1 X + U$$

But X is correlated with omitted variables, therefore we know that $\hat{\beta}_1$ does not converge in probability to β_1 from this regression. If both IV restrictions are satisfied, then we have

$$\begin{aligned}
Cov\left(Y,Z\right) &= Cov(\beta_0 + \beta_1 X + U, Z) \\
&= Cov(\beta_0, Z) + Cov(\beta_1 X, Z) + Cov(U,Z) \\
&= \beta_1 Cov\left(X,Z\right) && \text{since } Cov(Z,U) = 0. \\
\beta_1 &= \frac{Cov\left(Y,Z\right)}{Cov\left(X,Z\right)} && \text{since } Cov(X,Z) \neq 0.
\end{aligned}$$

By analogy principle, the IV estimator replaces these population covariances with the sample covariances

$$\hat{\beta}_1^{IV} = \frac{s_{yz}}{s_{zx}} = \frac{\sum_i (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_i (Z_i - \bar{Z})(X_i - \bar{X})}$$

where s_{ab} is the sample covariance between variables A and B.

Satisfying both of the IV restrictions at the same time can be surprisingly difficult, with the latter particularly so. Imagine that we want to estimate the effect of school attendance on standardized test scores. However, we know that attending school is correlated with many other omitted variables (e.g., family income). One might think that a family's distance to the school can be a valid instrument for attendance. We will need to consider whether this is actually the case, by thinking through the two conditions above:

1. Relevance: Distance likely satisfies the relevance condition - children living close to school attend more. This is easily testable in the data.
2. Exclusion: Distance to school is likely correlated with omitted variables that matter for standardized test scores - e.g., income, environmental quality, etc. This means that distance is not a good candidate IV. This assumption is *fundamentally untestable* - you will have to use intuition, theory, or clever reasoning to convince your audience that this holds.

Another potential instrument for school attendance is a merit scholarship that some schools offered to provide free college tuition if a student attended more than 95% of school days. Do you think that this is a good instrument?

1. Relevance: This likely satisfies the relevance conditions - more students will attend upon receiving the scholarship.
2. Exclusion: Whether this satisfies the exclusion restriction depends on which schools choose to adopt this policy. If only schools in certain areas (e.g., in low-income areas) offer this program, then this will fail, as the scholarship offer will be correlated with other factors which could affect test scores. However, if the program was randomly assigned across schools (perhaps due to budget constraints that meant not all schools could benefit), then this is a good instrument.

2.2 Exercise 2

We want to estimate the return to education in the simple regression model

$$\log(wage) = \beta_0 + \beta_1 educ + u$$

Worrying about omitted variable bias, we use a plausible instrument: $fatheduc$ (father's years of education). Testing for relevance, we obtain:

$$\widehat{educ} = 10.24 + 0.269 fatheduc$$

$$(0.28) \quad (0.029)$$

Using $fatheduc$ as an IV for $educ$ gives:

$$\widehat{log(wage)} = 0.441 + 0.059 educ$$

$$(0.446) \quad (0.035)$$

1. Does the relevance condition hold? Does father's education seem like a strong instrument for an individual's education?

2. Do you think the exclusion restriction is likely to hold? If not, why not?

3. We have $\widehat{\beta_1^{IV}} = 0.059$. Is it statistically significant at the 5% level?

4. Can you think of a variable that would be a better IV that you could use in this scenario?