I. Longer Panels: First Differences and Fixed Effects

In Section 12a our introduction to panel data focused 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. See Section 12a notes for more examples and R code for working with panel data, including longer panels.

A. Taking Differences

With two-period panels, we talked about taking first differences as a way of eliminating the unobserved time-constant characteristics of the units of analysis, which we referred to as α_i. This is helpful, as it deals with a potentially important source of omitted variable bias.

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 MLR4 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.

B. Fixed Effects

In Section 12a we introduced fixed effects. We discussed both unit (what you analyze, e.g., people, cities, businesses, etc.) and time fixed effects. We can apply these with multiple period panel data as well—in fact, you need more than 2 periods in order to include time fixed effects.

Time fixed effects are $T-1$ dummy variables for each time period observed (minus one reference period). We often represent the vector of time period dummies (fixed effects) by δ_t or d_t. 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.

A **unit fixed effect** is a vector of $n-1$ dummy variables, where n is the number of unique units (e.g., people, cities, businesses, etc.) in your data. We often label these dummies as a_j or α_j. A given dummy variable α_j is coded as 1 for unit j 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). Effectively, we are aware that our data look like

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \alpha_i + \delta_t + u_{it}$$
$$y_{jt} = \beta_0 + \beta_1 x_{1jt} + \cdots + \beta_k x_{kjt} + \alpha_j + \delta_t + u_{jt}$$

and include a unit fixed effect as an alternative way besides first differences to account for the α_i term in the model. For notational simplicity when we write our regression models, we often just write α_i to represent the vector of all the individual dummy variables for each i (except for the reference unit).

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.

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

$$y_{it} - \bar{y}_i = \delta_t - \bar{\delta}_t + \beta_1(x_{1it} - \bar{x}_{1i}) + \cdots + \beta_k(x_{kit} - \bar{x}_{kit}) + u_{it} \quad (1)$$
$$y_{it} = \beta_0 + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \sum_i c_i \alpha_i + \delta_t + u_{it} \quad (2)$$

and these models are equivalent, but the fixed effects model (equation 2, 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.

C. Fixed Effects vs. First Differences

Both first (or other) differences and unit fixed effects approaches control for time-constant characteristics of the units of analysis. This is extremely useful because it reduces the number of potential omitted variables we could be concerned about, without having to know anything about what those omitted variables could be. We still will be concerned about time-varying omitted variables, but we have at least reduced the set of possible sources of OVB.

First differences and fixed effects both also require the same assumption of strict exogenity: a generalization of MLR4 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 MLR 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

$$cov(u_{i3} - u_{i2}, u_{i2} - u_{i1}) \neq 0 \tag{3}$$

$$cov(u_{i3} - \frac{1}{T}\sum_s u_{is}, u_{i2} - \frac{1}{T}\sum_s u_{is}) \neq 0 \tag{4}$$

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 be lower variance. If u_{it} are close to random within individuals over time, then fixed effects will be lower variance.

D. Other Considerations With Panel Data

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.

In R, you can compute clustered standard errors using 'vcovCL' or using the cluster functionality in 'felm'. 'vcovCL' outputs the variance-covariance matrix given some regression output, and allows you to specify the level(s) of clustering. The diagonal elements of this matrix are the variances of the beta coefficients. If we take the square root of these variances we get our standard errors. Let's see this using code from Lecture 23 (see that Jupyter notebook for how the output looks).

```
# Approach 1, using vcovCL
reg3 <- felm(lcrmrte~ lprbarr + lprbconv + lprbpris + lavgsen + lpolpc + d83
+ d84 + d85 + d86 + d87 | county, data = crimedata)
library(sandwich) # needed for vcovCL function
reg3$clus_se <- sqrt(diag(vcovCL(reg3, cluster = crimedata$county)))
# view the clustered standard errors alongside the coefficients
cbind(reg2$coefficients, reg3$clus_se)

# Approach 2, using felm's clustering option
reg4 <- felm(lcrmrte~ lprbarr + lprbconv + lprbpris + lavgsen + lpolpc + d83
+ d84 + d85 + d86 + d87 | county | 0 | county, data = crimedata)
```

Note the order of arguments in 'felm', separated by brackets. We first have the regression formula, then we specify the fixed effects (*county*), then we specify any instrumental variables we want to use (in this case we aren't using any so we put 0), then we specify how we want to cluster the standard errors (*county*, the cross-sectional unit of analysis).

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 you're 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 your 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 you 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 you 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 you have data on wages for a bunch of individuals over time and only observe those individuals in periods where they are working, you might think there are important factors changing their circumstances in periods when they are not working where you do not observe them and they are earning 0 wages. This could matter if you 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 x variables of interest.

Note: See the Section 12a notes for additional discussion of first differences and fixed effects, including R code and examples. Several of those already used panel data with multiple time periods.

II. Exercise (solutions at the end)

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

Your 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

You first estimate the following regression:

$$hours_{iht} = \beta_0 + \beta_1 daycare_{ht} + \beta_2 sex_{ih} + \beta_3 sector_{iht} + month_t + u_{iht} \quad (5)$$

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:

$$hours_{iht} = \beta_0 + \beta_1 daycare_{ht} + \beta_2 sector_{iht} + \alpha_{ih} + month_t + u_{iht} \quad (6)$$

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. Where does our identifying variation come from for estimating β_1? In other words, what households are providing the information we use to estimate this coefficient in a unit fixed effects model?

6. You are concerned about correlation of your errors. What level of clustering would be appropriate in this model - which variable would you cluster by?

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

III. Instrumental Variables

A. Introduction

Panel data methods are tools that help use 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 also don't always have a nice threshold to use for an RD design, and will 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

   $\Rightarrow \ cov(z, x) \neq 0$

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.

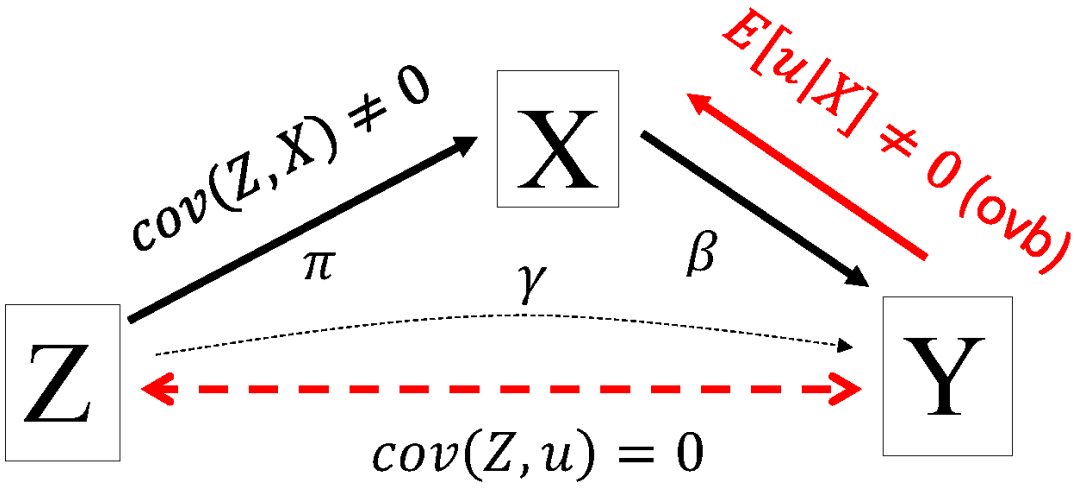
   $\Rightarrow \ cov(z, u) = 0$

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 MLR4, 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 $E[u|x = 0]$. This is shown by the solid red arrow going from y to x. But if we have an instrument z that is correlated with x (black 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 red arrow). This implies that the direct effect of z on y represented by the dashed black line is happening through x.

So how do we estimate β? Let's call the total effect of z on y γ. We know from the exclusion restriction that the only path through which z impacts y is through x, which implies that $\gamma = \beta\pi$. We can't estimate β because of the OVB represented by the solid red arrow, but we can regress x on z to obtain π. This gives us enough information to solve for β – it's just $\frac{\gamma}{\pi}$. Of course, this is ill-defined if $\pi = 0$. This is why need the relevance assumption.



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 use intuition, theory, or clever reasoning to convience your audience that this holds.

Another possibility 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 due to the offer of 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.

B. Estimation Mechanics

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 $\hat{\beta}_1 \neq \beta_1$ from this regression. So, we use an instrument z. Transforming this equation, we can write:

$$cov(z, y) = \beta_1 cov(z, x) + cov(z, u)$$

If our assumptions are satisfied, we know $cov(z, x) \neq 0$ and $cov(z, u) = 0$, which allows us to solve for β_1:

$$\beta_1 = \frac{cov(z, y)}{cov(z, x)}$$

We can estimate this using our sample data:

$$\widehat{\beta_1^{IV}} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

which is our **instrumental variables estimator** of β_1.

With the previous assumptions, as well as $E[u^2|z] = \sigma^2 = \text{Var}(u)$, the variance of $\hat{\beta}_1^{IV}$ is

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$$

where σ_x^2 is the population variance of x, σ^2 is the population variance of u, and $\rho_{x,z}^2$ is the square of the population correlation between x and z. With a random sample, all of these components can be estimated. The asymptotic standard error of $\hat{\beta}_1^{IV}$ is the square root of the estimated asymptotic variance, which itself is

$$\frac{\hat{\sigma}^2}{SST_x \cdot R_{x,z}^2}$$

One can compare the variance of the IV estimator to that of the OLS estimator (in the SLR case):

$$\frac{\hat{\sigma}^2}{SST_x \cdot R_{x,z}^2} \text{ vs. } \frac{\hat{\sigma}^2}{SST_x}$$

Since $R^2 < 1$, the IV variance is always larger than the OLS variance. In particular, if $R_{x,z}^2$ is small, then the IV variance can be *much* larger than that of OLS. This makes sense: if your instrument z doesn't explain much of the variation in x, you will not obtain a very precise estimate for the effect of x on y when using z as an IV.

C. Weak Instruments

Setting aside assumptions about relevance and exogeneity, one can write the probability limit of the IV estimator $\widehat{\beta_1^{IV}}$ as

$$E[\widehat{\beta_1^{IV}}] = \frac{cov(z,y)}{cov(z,x)} = \beta_1 + \frac{cov(z,u)}{cov(z,x)} = \beta_1 + \frac{\text{Corr}(z,u)}{\text{Corr}(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$$

where σ_u and σ_x are the standard deviations of u and x in the population, respectively. If the exogeneity restriction holds, then $Corr(z,u) = 0$ and we recover β_1.

But observe that even if $\text{Corr}(z,u)$ is small, substantial bias can arise if $\text{Corr}(z,x)$ is small as well. Thus if we have a *weak instrument*, that is, one where $\text{Corr}(z,x)$ is small, IV estimation is vulnerable to bias if our assumption that $cov(z,u) = 0$ does not hold. Bias from failure of this assumption will be smaller if we have a very strong instrument.

D. Treatment Effects

Suppose we have

$$y = \beta_0 + \beta_1 x + u$$

where we are concerned that x is endogenous (MLR4 fails), and

$$x = \pi_0 + \pi_1 z + v$$

for some $\pi_1 \neq 0$ and $cov(u,z) = 0$, with all other MLR assumptions holding.

Then we can write

$$\begin{aligned} y &= \beta_0 + \beta_1(\pi_0 + \pi_1 z + v) + u \\ &= (\beta_0 + \beta_1\pi_0) + \beta_1\pi_1 z + (\beta_1 v + u) \end{aligned}$$

Then we can run the reduced form regression $y = b_0 + b_1 z + e$. We will recover an unbiased estimator of b_1, such that $E[\hat{b}_1] = \beta_1\pi_1$. Here, we have estimated the effect of z on y, which is the effect of x on y weighted by the effect of z on y. We can think of this as an **ITT**, where π_1 is a measure of 'compliance': how z affects x.

With the data we have, we can also estimate π_1 by regressing $x = \pi_0 + \pi_1 z + v$. Using this, we can recover $E[\widehat{\beta_1^{IV}}] = \frac{E[\hat{b}_1]}{E[\hat{\pi}_1]}$. We divide the relationship between y and z by the relationship between x and z. Intuitively, we have $\frac{\Delta y}{\Delta z} / \frac{\Delta z}{\Delta x} = \frac{\Delta y}{\Delta x}$ which is what we care about. This gives us a **TOT** estimator, where we have assumed that all of the effect of z on y is through its effect on x.

E. Instrumental Variables and Multiple Regression

Suppose we have the following *structural equation* (so called because it reflects the causal relationships we want to estimate):

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + u$$

We assume $E[u] = 0$, and z_1 is exogenous (not correlated with u). In the equation, we suspect x of being correlated with u. Thus if we measure the above formula with OLS, all estimators will be biased and inconsistent, which prompts us to want to find an IV for x.

Can we use z_1 as an IV for x? No, as it is part of the true model, i.e., it has a effect on y outside of any effect through x. We will need another variable, z_2 that is uncorrelated with u but correlated with x, even after controlling for z_1.

How are our IV conditions modified? Write out the endogenous explanatory variable as linear function:

$$x = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v$$

where (by construction), $E[v] = 0$, $\text{Cov}(z_1, v) = 0$, $\text{Cov}(z_2, v) = 0$ and π_j are unknown. Note that you should include all exogenous explanatory variables in this model. In this set-up, the **key identification condition** (with those noted previously) is that $\pi_2 \neq 0$ (the relevance restriction). This can be tested via OLS through the above **reduced form equation**, which is what we call a regression of endogenous variable in terms of exogenous variables. We will usually report on the strength of the relationship between our IV (z_2) and the endogenous explanatory variable (x), so that we know whether we need to be worried about potential bias from using a weak instrument. Unfortunately, we still cannot test that z_1 and z_2 are uncorrelated with u.

F. Multiple IVs and Two Stage Least Squares (2SLS)

What if we have multiple candidate IVs for the same endogenous variable? We use a process called Two Stage Least Squares: **2SLS**. Essentially, we use multiple IVs to construct a single, stronger IV to use in our estimation.

Our structural model, as before:

$$y = \beta_0 + \beta_1 x + \beta_2 z_1 + u$$

but we have two exogenous variables z_2 and z_3 that 1) are both correlated with x, and 2) satisfy the exclusion restrictions (they do not appear in the structural model and are uncorrelated with the error u). So both are candidates as IVs for endogenous x. Thus

$$x = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v$$

Here the best IV for x will make use of all exogenous variables. The intuition for this is that including more valid instruments helps explain more of the exogenous variation in x (variation in x that is not correlated with u) and thus generate more precise estimates.

For IV to meet the relevance restriction, one needs at least $\pi_2 \neq 0$ or $\pi_3 \neq 0$, which one can test via an F-test. The larger our F-statistic, the stronger our instruments (and the less concern we have about potential bias from weak instruments).

Using these IVs, we can isolate the part of x that is not correlated with u by construction, since none of the z variables are correlated with u:

$$\hat{x} = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$

We can then use this exogenous part of x to identify the causal relationship between x and y.

The procedure to carry out 2SLS is as follows:

- Regress x on z_1, z_2, and z_3 and obtain the fitted values $\hat{x}$: these will be uncorrelated with u. This is called the **first stage**.
- Verify that z_2 and z_3 are jointly significant in the **reduced form** (the regression of y on all exogenous variables) with an F-test.
- Then to estiamte the **structural equation** use $\hat{x}$ as IV for x, for the regression of y on $\hat{x}$ and z_1,

Note: you will want to use a command in R to do this automatically, since standard errors and test statistics obtained in this way are not valid. You can actually run a 2SLS model using the felm command, but you won't need to know how to do that for this course. In practice, we'll often use 2SLS to generate IV estimates even when we only have one IV.

G. Other notes for IVs

1. Multicollinearity can be a serious concern. In 2SLS asymptotic variance is described by $\sigma^2/[\widehat{SST_x}(1-\hat{R}_x^2)]$, where $\sigma^2 = \text{Var}(u)$, $\widehat{SST_x}$ is the total variation in $\hat{x}$ and $\hat{R}_x^2$ is the R^2 from a regression of $\hat{x}$ on all other exogenous variables appearing in the structural equation. The variance term is larger if the correlation between $\hat{x}$ and the exogenous variables is much higher than the correlation between x and these variables.
2. We are still worried about weak instruments. One commonly accepted rule of thumb is that we fulfill the relevance condition if the F-statistic previously mentioned (on the instruments in the reduced form) is larger than 10.
3. One can extend 2SLS to models with more than one endogenous explanatory variable. Note that you will need at least one valid instrument for each endogenous variable.
4. IVs are useful to deal with measurement error (15.4 in Wooldridge), as they can help recover the true variation in a variable of interest and get rid of the noise.
5. RCTs with imperfect compliance are ideal settings for IVs. You use the random assignment to treatment as an IV for actual treatment received to recover the TOT.
6. Can use IVs with pooled cross sections and panel data.

IV. Exercise (solutions at the end)

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? What treatment effect is this most similar to if we're interested in the effect of education—ITT or TOT? Can we recover the other type of treatment effect with the output we have?

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

Solutions

Section II

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

1. What does the month fixed effect control for?
   The time fixed effect controls for all factors that vary in the same way across time for all individuals in the zip code. This would include things like weather which changes in the same way for all individuals in a small geographic area. It also includes the number of daycare centers in the zip code and the prices charged at daycare centers in the zip code, which could be changing over time but do not vary across households. It does not control for something like distance to the closest daycare center, however, as this might vary at the household level over time as new daycare centers open.

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?
   This is unlikely to recover a causal impact of daycare. There are many potential omitted variable that could affect the decision to use daycare and with work hours. For example, households with one high-earning member might choose to have the other member not work and therefore not use daycare. Households with many young children might not be able to afford to send all of them to daycare, so might have one parent working less to care for children even while sending some children to daycare.

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

3. What does the individual fixed effect control for?
   α_{ih} controls for all characteristics of the individual (and their household) that do not change over time. This includes time-invariant characteristics like their sex as well as hard-to-measure factors like their work motivation, parenting preferences, attitudes towards daycare, etc. It also controls for means within individual across the sample period, such as their mean age, mean years of education, mean number of children, etc. These are all potentially important omitted variables. It does not control for individual variables that do change over time, such as their actual number of children (if they have more during the survey period) or their sector of employment (if they switch jobs).

4. Why do we no longer include *sex* in the model?
   We assume that an individual's sex does not vary over time, and therefore it is absorbed in the individual fixed effect. Effectively, if there are no individuals whose sex varies over time, after including an individual fixed effect we cannot estimate how changes in sex are associated with changes in work hours.

5. Where does our identifying variation come from for estimating β_1? In other words, what households are providing the information we use to estimate this coefficient in a unit fixed

effects model?
Recall that unit fixed effects control for the mean values of all variables, so the coefficient for a particular variable is estimated using within-individual variation across time around its mean. For *daycare*, the mean across the sample period will be the proportion of months where the household had a child enrolled in daycare. If some households never had a child in daycare or always had a child in daycare over these 12 months, they have no variation in *daycare* after including individual fixed effects. Therefore, the households that provide the information we use to estimate β_1 is the subset of households that vary their use of daycare in the survey period. We might worry that these households are different from households that always use daycare, and that the conditions determining whether they use daycare might also be associated with their work hours.

6. You are concerned about correlation of your errors. What level of clustering would be appropriate in this model - which variable would you cluster by?
Errors will almost definitely be correlated within individuals, so we could cluster our standard errors at the individual level using a combination of *hhid* and *individ* (to create a unique individual ID across households). But we might also be concerned that our errors will be correlated within households as well - the unobserved factors determining daycare use and work hours are likely to be similar for both adults in a household. Conservatively then, we would cluster our standard errors at the household level using *hhid*. This treats our data as if the households are independent observations, but the individuals within households are not.

7. Does this regression recover the causal impact of daycare on work hours? If not, what could be a possible source of omitted variable bias?
This regression likely comes closer to recovering a causal impact of daycare, but could still be affected by OVB. We no longer have to worry about a scenario like the one we described for the regression with no individual fixed effects, where households with one high-earning member might choose to have the other member not work and therefore not use daycare, since the fixed effects will control for the presence of a high-earning household member. But there could be variables that are changing over time that are correlated with daycare use. For example, a household with a child in daycare could have a new baby and one parent might stop working and remove their child from daycare to care for both children at home. Or one parent might lose their job, leading the household to remove their child from daycare to save money while the other parent increases their work hours to help make up for lost income.

Section IV

Relevance regression:

$$\widehat{educ} = 10.24 + 0.269 fatheduc$$
$$(0.28) \quad (0.029)$$

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?
   Yes, the relevance condition holds. We have $t = 0.269/0.029 = 9.28$ which is quite larger, so we can reject the null that there is no relationship between father's education and own education with a high level of confidence. This seems like a relatively strong instrument, though we would need to do an F test to check whether the rule of thumb that $F > 10$ holds.

2. Do you think the exclusion restriction is likely to hold? If not, give an example of a possible omitted variable that would be correlated with father's education?
   It seems unlikely that father's education would only affect an individual's wages through the individual's own education. For example, individuals with highly-educated fathers might benefit from their fathers' business connections to find high-paying jobs. Individuals with highly-educated fathers might also be more likely to live in high-income areas, where jobs also pay more than in other areas.

3. We have $\widehat{\beta_1^{IV}} = 0.059$. Is it statistically significant? What treatment effect is this most similar to if we're interested in the effect of education—ITT or TOT? Can we recover the other type of treatment effect with the output we have?
   We have $t = 0.059/0.035 = 1.686$. If our sample has at least 40 individuals then this is significant at a 10% significance level, but not at a 5% level.
   $\widehat{\beta_1^{IV}}$ is a TOT estimator. It divides the estimated relationship between $fatheduc$ and $wage$ by the estimated relationship between $fatheduc$ and $educ$. This gives us the effect of education on wages, assuming all of the effect of father's education on wages is through its effect on own education.
   We can recover the ITT equivalent—the reduced form effect of $fatheduc$ on $wage$, allowing for non-compliance in how $fatheduc$ affected $educ$—by multiplying $\widehat{\beta_1^{IV}}$ by the coefficient on $fatheduc$ from the relevance equation. We obtain $\beta_{ITT} = 0.059 * 0.269 = 0.016$. Thus an additional year of father's education increases wages by 1.6 percent.

4. Can you think of a variable that would be a better IV that you could use in this scenario?
   We need a variable that is correlated with individual education and uncorrelated with anything else that affects individual wages. Ideally, we would look for something that randomly created variation in education. For example, a college scholarship that was randomly offered to students in the study area, or a policy to increase funding to help dropouts return to school and get a high school diploma that was randomly implemented across school districts. Both of these should meet the relevance restriction, and because of the randomness of their design they should be uncorrelated with other factors which might affect wages.