UCI Paul Merage School of Business

Leadership for a Digitally Driven World™

MFIN 290: Financial Econometrics

Lecture 6-1

Last Time

- VAR Models
- Impulse Response Functions
- Lagged Dependent Variables with Time Series Effects: Bias and IV
- LASSO

Panel Data

- Panel data is where we have repeated observations for a group or individual i over time t.

- Typically, $i > t \Rightarrow$ panel datasets are wider than they are long.

- Example:
- Household spending
- Loan performance data
- Stock returns for a fixed set of firms

Panel Data

- Idea here is to provide an introduction to the structure and key issues regarding panel datasets.

- We will not be focused on proofs and deriving the properties shown in Hill and Lim (Chp 15) though they are useful to know once you have the intuition.

- Instead, we will focus on the sources of variation in the data, the common issues you are likely to run into when estimating relationships on a panel dataset, and some tricks that can give us multiple ways to identify the effects we care about.

Project STAR

1985-1989, TN

Followed a single cohort (grade) of students from K – 3rd. Children were randomly assigned into three types of classes: small (13-17 students), regular (22-25), and regular, with a full time teacher aide in addition to the teacher.

Achievement scores on tests and some basic information about the students, teachers, and schools. Random assignment helps deal with endogeneity.

Does this sample look random?

Want to know the effect of class size of test outcomes.

What would be some potential examples of endogeneity if we didn’t have this and instead families could choose which option they wanted?

Using star.dta (on the course website)

Project STAR

```
. summ totalscore small tchexper boy freelunch white_asian tchwhite tchmasters schurban
> schrural if regular ==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  totalscore |      2,005    918.0429    73.13799        635       1229
       small |      2,005           0           0          0          0
    tchexper |      2,005    9.068329    5.724446          0         24
         boy |      2,005     .513217      .49995          0          1
   freelunch |      2,005    .4738155    .4994385          0          1
-------------+---------------------------------------------------------
 white_asian |      2,005    .6812968    .4660899          0          1
    tchwhite |      2,005     .798005    .4015887          0          1
  tchmasters |      2,005    .3650873    .4815747          0          1
    schurban |      2,005    .3012469    .4589142          0          1
    schrural |      2,005    .4997506    .5001247          0          1

. summ totalscore small tchexper boy freelunch white_asian tchwhite tchmasters schurban
> schrural if small==1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
  totalscore |      1,738    931.9419    76.35863        747       1253
       small |      1,738           1           0          1          1
    tchexper |      1,738    8.995397    5.731568          0         27
         boy |      1,738    .5149597      .49992          0          1
   freelunch |      1,738    .4718067    .4993482          0          1
-------------+---------------------------------------------------------
 white_asian |      1,738    .6846951    .4647709          0          1
    tchwhite |      1,738    .8624856    .3444887          0          1
  tchmasters |      1,738    .3176064    .4656795          0          1
    schurban |      1,738     .306099     .461004          0          1
    schrural |      1,738    .4626007    .4987428          0          1
```

Consistent with random assignment, look fairly similar on controls.

Could make this formal with t-tests, but we are going to use dummies to control for differences in any event…

Project STAR

An intuitive way to check for random assignment is to regress SMALL (the class size assignment) on the available characteristics and check for any significant coefficients, or an overall significant relationship

If there is random assignment, we should not find any significant relationships…

SMALL is itself an indicator variable, when the y variable is a dummy, this is called a "linear probability model", but there are other methods (soon!)

Project STAR

```
. reg small boy white_asian tchexper freelunch

      Source |       SS           df       MS      Number of obs   =     5,766
-------------+----------------------------------   F(4, 5761)      =      2.32
       Model |  1.95484935         4  .488712338   Prob > F        =    0.0544
    Residual |  1212.17349     5,761  .210410257   R-squared       =    0.0016
-------------+----------------------------------   Adj R-squared   =    0.0009
       Total |  1214.12834     5,765  .210603354   Root MSE        =     .4587

------------------------------------------------------------------------------
       small |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         boy |  -.0002542   .0120979    -0.02   0.983    -.0239706    .0234621
 white_asian |   .0123603   .0144885     0.85   0.394    -.0160426    .0407632
    tchexper |  -.0029793   .0010545    -2.83   0.005    -.0050466    -.000912
   freelunch |  -.0087997   .0135262    -0.65   0.515    -.0353162    .0177167
       _cons |   .3251669   .0188346    17.26   0.000     .2882441    .3620898
------------------------------------------------------------------------------
```

Project STAR

Teacher experience is a bit of a concern.

Interested in two models:

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + e$$

$$TOTALSCORE = \beta_1 + \beta_2 SMALL + \beta_3 TCHEXPER + e$$

Hope that we get similar results. If not, it means the correlation with teacher experience matters (remember omitted variables bias)

Probably would want to run these with and without school controls to check robustness/confirm random assignment across schools.

Project STAR

```
. codebook schid

schid                                                    school id

                  type:  numeric (long)

                 range:  [112038,264945]          units:  1
         unique values:  79                   missing .:  0/5,786

                  mean:    211002
             std. dev:   38381.9

           percentiles:        10%       25%       50%       75%       90%
                            161183    176329    215533    244764    252885
```

79 different schools.

Can control for school with 78 additional dummy variables δ_j equal to one if in school j, zero otherwise

A student in school j has:

$$E\left(TOTALSCORE_i\right) = \begin{cases} (\beta_1 + \delta_j) + \beta_3 TCHEXPER_i & \text{If student is in regular class} \\ (\beta_1 + \delta_j + \beta_2) + \beta_3 TCHEXPER_i & \text{If student is in small class} \end{cases}$$

Project STAR

	(1)	(2)	(3)	(4)
C	918.0429***	907.5643***	917.0684***	908.7865***
	(1.6672)	(2.5424)	(1.4948)	(2.5323)
SMALL	13.8990***	13.9833***	15.9978***	16.0656***
	(2.4466)	(2.4373)	(2.2228)	(2.2183)
TCHEXPER		1.1555***		0.9132***
		(0.2123)		(0.2256)
SCHOOL EFFECTS	No	No	Yes	Yes
N	3743	3743	3743	3743
adj. R^2	0.008	0.016	0.221	0.225
SSE	20847551	20683680	16028908	15957534

Standard errors in parentheses
Two-tail *p*-values: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Natural Experiments

Randomized controlled experiments like Project STAR are rare in economics because they are expensive and involve human subjects

Natural experiments, also called quasi-experiments, rely on observing real-world conditions that approximate what would happen in a randomized controlled experiment

In these cases, treatment appears as if it were randomly assigned

Validity of your regression / the experiment hinges on how “natural” it truly is!

Diff in Diff Example

Card, Krueger (1994)

Interested in effect of Minimum Wage on Employment

April 1, 1992. NJ increases their minimum wage, Pennsylvania remains fixed.

Data collected on 410 fast food restaurants in NJ (treatment) and Eastern PA (control) from Feb – Nov 1992.

Diff in Diff Example

Card, Krueger (1994)

Interested in effect of Minimum Wage on Employment

April 1, 1992. NJ increases their minimum wage, Pennsylvania remains fixed.

Data collected on 410 fast food restaurants in NJ (treatment) and Eastern PA (control) from Feb – Nov 1992.

Parallel Trend: employment in the two regions was following the same pattern. Are these regions comparable? Are these industries relevant? How can we check?

What does this assume about the relationship between government policy and preferences?

This caveats apply to current empirical work as well!

Diff in Diff Example

$$FTE_{it} = \beta_1 + \beta_2 NJ_i + \beta_3 D_t + \delta\,(NJ_i \times D_t) + e_{it}$$

FTE = full time employment

NJ = 1 if from NJ

D = 1 if after April, 1992

δ = Diff in Diff estimate.

$H_0 = \delta \geq 0$, H_1: "policy reduced FTE in NJ

Diff in Diff Example

	(1)	(2)	(3)
C	23.3312***	25.9512***	25.3205***
	(1.072)	(1.038)	(1.211)
NJ	−2.8918*	−2.3766*	−0.9080
	(1.194)	(1.079)	(1.272)
D	−2.1656	−2.2236	−2.2119
	(1.516)	(1.368)	(1.349)
D_NJ	2.7536	2.8451	2.8149
	(1.688)	(1.523)	(1.502)
KFC		−10.4534***	−10.0580***
		(0.849)	(0.845)
ROYS		−1.6250	−1.6934*
		(0.860)	(0.859)
WENDYS		−1.0637	−1.0650
		(0.929)	(0.921)
CO_OWNED		−1.1685	−0.7163
		(0.716)	(0.719)
SOUTHJ			−3.7018***
			(0.780)
CENTRALJ			0.0079
			(0.897)
PA1			0.9239
			(1.385)
N	794	794	794
R^2	0.007	0.196	0.221
adj. R^2	0.004	0.189	0.211

Standard errors in parentheses
Two-tail *p*-values: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Looks like we cannot reject the null that D_NJ is positive at typical levels of confidence, and this result is robust to a variety of controls.

Data does not provide evidence that increasing the minimum wage lowers full time employment.

Remember, rejection region here would be negative! This is almost evidence that increasing minimum wages increased employment!

Diff in Diff Example

In this DD analysis, we did not exploit one very important feature of the data, namely the fact that the same fast food restaurants were observed on two occasions

We have “before” and “after” data

These are called paired data observations, or repeat data observations, or panel data observations.

Using panel data we can control for unobserved individual-specific characteristics by using individual indicators for each restaurant (called fixed effects)

Diff in Diff Example

$$FTE_{it} = \beta_1 + \beta_2 NJ_i + \beta_3 D_t + \delta\left(NJ_i \times D_t\right) + \boxed{c_i} + e_{it}$$

Similar point estimate (2.75), but now significant and positive (SE = 1.154).

Data still does not provide evidence that increasing the minimum wage lowers full time employment (on the contrary!)

Panel Data

- Panel data is where we have repeated observations for a group or individual i over time t. Also called “longitudinal data”.
- If all groups have the same number of observations, we call it a “balanced panel”.

Panel Data

- Panel data is where we have repeated observations for a group or individual i over time t.

- For each i, we have a time series of data.

- For each t, we have a cross-section of data.

Panel Data

- Panel data is where we have repeated observations for a group or individual i over time t.

- Suggests that there are two sources of *identifying variation* within a panel:

- Variation within a group over time (called “within <groups>”), and variation between groups at a point in time (called “between <groups>”). It will be useful to think about how effects are identified as we work through this lecture and you run into these models in your career.

Panel Data

- Panel data is where we have repeated observations for a group or individual i over time t.

- Can think about splitting up the data into variables that change over time and those that do not.

- $y_{it} = x_{it}\beta + z_i\Gamma + u_{it}$

- x_{it} = variables that change through time
- z_i = variables that do not change through time (inc. constant term)

Panel Data

- $y_{it} = x_{it}\beta + z_i\Gamma + u_{it}$

- x_{it} = variables that change through time
- z_i = variables that do not change through time (inc. constant term)

- z_i reflects the individual level heterogeneity/individual effects that are relevant to the model. Race, sex, location, industry, etc. In the limit, if everything relevant was observed and controlled, and everyone shared the same β, every individual would have their own specific effect…

Panel Data

- $y_{it} = x_{it}\beta + c_i + u_{it}$

- x_{it} = variables that change through time
- c_i = individual effects

- However, c_i also in principle includes a great deal of unobserved information (such as individual ability, or unknown firm decisions/probabilities). Note that if we did have all of the controls, we could just estimate this via OLS and be done!

Panel Data

- To make this concrete, recall the example of regressing earnings (y_{it}) on years of schooling (x_{it}) and some controls (z_i).

- In this setup, it seems plausible that there may be some individual effect that's unobserved but correlated with our x_{it}: if someone's ability is high– if they are smart/talented, school is marginally easier to attend -- they will attend more and earn more due to that rather than due to their extra schooling per se.

- Hard to imagine answering this question without a panel data setup…

Panel Data

- $y_{it} = x_{it}\beta + c_i + u_{it}$

- x_{it} = variables that change through time
- c_i= individual effects

- If $E(u_{it}|x_i, c_i) = 0$ => we have the needed OLS assumption for unbiasedness.

- Called "Strict Exogeneity"

Panel Data

- $y_{it} = x_{it}\beta + c_i + u_{it}$

- x_{it} = variables that change through time
- c_i = individual effects

- If $E(u_{it}|x_i, c_i) = 0$ => we have the needed OLS assumption for unbiasedness.

- This is relatively unlikely... remember the schooling/ability example, and how it is violated informs what kinds of approaches are best.

Pooled Regression

- $y_{it} = x_{it}\beta + z_i\Gamma + u_{it}$

- Pooled Regression:

- If z_i contains only a constant term, the OLS is a consistent (and efficient) estimate of both the constant and slope term (β)

Fixed Effects

- $y_{it} = x_{it}\beta + z_i\Gamma + u_{it} \Rightarrow \; y_{it} = x_{it}\beta + \alpha_i + u_{it}$

- Fixed Effects:

- If some relevant z_i is unobserved, but correlated with x_{it}, then an OLS estimate of β will be biased due to omitted variables bias/endogeneity, etc. However, if we observed it, we would combine the $z_i\Gamma$ for each observation into a single number, α_i for each individual.

- We can still run this regression!

Fixed Effects

- $y_{it} = x_{it}\beta + z_i\Gamma + u_{it}$

- Fixed Effects:
- $y_{it} = x_{it}\beta + \alpha_i + u_{it}$

- Includes a group-specific constant term into the regression. Identifies the slope using the variation *within groups only.* Sometimes called a "within" estimator for this reason.
- Essentially greatly increasing the degrees of freedom in the model but allows everyone to have their own intercept.

Random Effects

- $y_{it} = x_{it}\beta + z_i\Gamma + u_{it}$

- Random Effects:

- If we have unobserved individual heterogeneity that is NOT correlated with x_{it}, then we can split apart the model as follows:

- $y_{it} = x_{it}\beta + E(z_i\Gamma) + (z_i\Gamma - E(z_i\Gamma)) + u_{it}$
- $y_{it} = x_{it}\beta + \alpha + v_i + u_{it}$

Random Effects

- $y_{it} = x_{it}\beta + z_i\Gamma + u_{it}$

- Random Effects:

- $y_{it} = x_{it}\beta + \alpha + v_i + u_{it}$

- This is a regression model with a compound disturbance term $(v_i + u_{it})$ that is, by assumption, uncorrelated with x_{it}. There is a unique form of heteroskedasticity in the errors or this model. Still everyone can have their own intercept, but we don't have to estimate it, and if this structure is correct, we know the form of heteroskedasticity.

Fixed and Random Effects

- Fixed v Random Effects:

- Fixed effects are poorly named in the sense that they are still stochastic estimates. The key difference in assumptions is whether or not the unobserved individual effects covary with the other regressors in the model.

- If they do, then fixed effects is required. If they do not, then random effects is fine (and is more efficient than either fixed effects or pooled OLS).

Fixed and Random Effects

- Example:

- Imagine we are interested in predicting the effect of education on earnings over the lifecycle and have a panel dataset of different individuals.

- “Ability” is not directly observable, and likely affects wages.

- IMPORTANTLY for our purposes, it is very likely correlated with education as well => Need fixed effects!

Random Coefficients/Parameters

- Random effects is essentially a regression with a random constant term. We could just as easily have a specification with similar randomness in the slope:

- $y_t = x_{it}(\beta + w_i) + (\alpha + v_i) + u_{it}$

Random Coefficients/Parameters

- Let's ignore the intercept (or embed it in X with a column of 1's):

- $y_{it} = x_{it}(\beta + w_i) + u_{it}$
- $y_{it} = x_{it}\beta + x_{it}w_i + u_{it}$

- If we have $E(w_i|x_i) = E(u_i|x_i = 0)$, that w_i and u_i are uncorrelated, and $E(u_i u_i'|x_i) = \sigma_u^2 I_T$ as usual and let $E(w_i w_i'|x_i) = \Gamma$, then for each group/individual we have a regression as follows:

- $y_i = x_i\beta + (x_i w_i + u_i) = x_i\beta + v_i$

Random Coefficients/Parameters

- $y_i = x_i\beta + (x_i w_i + u_i) = x_i\beta + v_i$

- $E(v_i v_i' | x_i) = E((x_i w_i + u_i)\ (x_i w_i + u_i)' | x_i))$
- $= \sigma_u^2 I_T + x_i \Gamma x_i{}'$

- This gives us the form of the variance covariance matrix for each group => we can stack these groups on one another to run GLS on the whole system to get unbiased (and efficient!) estimates of the coefficients where every group gets their own sensitivity!
- Again though, these differences in sensitivity need to be uncorrelated with the other regressors. Otherwise, have to control for with fixed effects (with interactions here).

Pooled Regression Model

- $y_{it} = \alpha + x'_{it}\beta + e_{it}$
- $i = 1, \dots, n; t = 1, \dots T$
- $E(e_{it}|x_{i1}, x_{i2}, \dots, x_{iT}) = 0$
- $E\left(e_{it}e_{js}|x_{i1}, x_{i2}, \dots, x_{iT}\right) = \sigma_e^2$ if $i = j$ and $t = s$ and zero otherwise

- Here, OLS is efficient and nothing more is needed. This is a nice way of seeing what would need to be true for you to be satisfied with pooled OLS in a panel context… it is rarely plausible (particularly the last one).

Pooled Regression Model

- $y_{it} = \alpha + x'_{it}\beta + e_{it}$
- $i = 1, \ldots, n; t = 1, \ldots T$
- $E(e_{it} | x_{i1}, x_{i2}, \ldots, x_{iT}) = 0$
- $E\left(e_{it}e_{js} | x_{i1}, x_{i2}, \ldots, x_{iT}\right) = \sigma_e^2$ if $i = j$ and $t = s$ and zero otherwise

- The problem is that these assumptions (particularly the last one) are unlikely to be met.

- We can see how by writing out other estimators…

Random Effects

- $y_{it} = \alpha_i + x'_{it}\beta + e_{it}$
- $i = 1, \dots, n; t = 1, \dots T$

- Let $E(\alpha_i|x_i) = \alpha$, and we can plug in:
- $y_{it} = \alpha + x'_{it}\beta + e_{it} + (\alpha_i - E(\alpha_i|x_i))$
- $y_{it} = \alpha + x'_{it}\beta + e_{it} + u_i$
- $y_{it} = \alpha + x'_{it}\beta + w_{it} \Rightarrow$
- $E(w_{it}w_{is}|x_{i1}, x_{i2}, \dots, x_{iT}) = \sigma_u^2$ if t $\neq s$ and zero otherwise. This is different than the pooled assumption!

Random Effects

- $y_{it} = \alpha_i + \; x'_{it}\beta + e_{it}$
- $i = 1, \dots, n; t = 1, \dots T$
- $E(w_{it}w_{is}|x_{i1}, x_{i2}, \dots, x_{iT}) = \sigma_u^2$ if t $\neq s$ and zero otherwise.

- The errors within a group will be autocorrelated over time. This makes OLS inefficient, like any other issue with non-spherical errors.

- The fix is as before! GLS (Feasible GLS) is asymptotically efficient when we have a parameterized variance covariance matrix of the residuals.

Clustering

- Clustering represents a situation where we have errors correlated across groups that have some other characteristic in common.
- For example, surveys may sample everyone in a particular block, or school test scores data may include a number of observations from a single class, a single teacher, a single school, or a single district. Each lever here can plausibly introduce correlated variation in the same way as the group level random effects model.
- This extra correlation reduces the power in the data from what you would think. Many packages are available to incorporate clustered standard errors.
- This is always worth thinking about, and if it materially changes inference, it’s likely to be an issue.

Bertrand, Duflo, Mullainathan (2003)

“How much should we trust Difference in Difference Estimates”?

Shows (by simulating placebo laws in the Current Population Survey and estimating effects in DD framework) that DD standard errors are usually incorrect.

Similar to the Spurious Regression finding and exposition!

45% of placebo interventions are statistically significant at the 5% level.

Bertrand, Duflo, Mullainathan (2003)

“How much should we trust Difference in Difference Estimates”?

Several corrections attempted, finds that the ones that work the best are the ones that treat the data as coming from two periods: pre and post law change, and allow for the errors within each state to be correlated (they actually average all of the pre and post data into one observation each!).

Clustering within states in the pre and post periods is one way to address this problem.

Estimation using Group Means

- Define as $\bar{y}_{i.}$ and $\bar{x}_{i.}$ as the sample means of x_{it} and y_{it} over time.
- $\bar{y}_{i.} = \frac{1}{T}\sum_{t=1}^{T} y_{it} = \frac{1}{T} i' y_i$
- where i' is a row vector of ones of length T
- In this model, we recover
- $\bar{y}_{i.} = \bar{x}_{i.}\beta + \bar{w}_{i.}$
- Where $\bar{w}_{i.}$ has heteroscedastic variances given by $\frac{1}{T^2} i' \Omega_i i$ for some unknown Ω_i

Estimation using Group Means

- Note that this will not solve any issues with clustering or random effects, though it will diminish them through averaging.
- Note that if we actually estimated this model, there would only be N observations, one for each individual/group i
- It is however, a useful way to think of one source of variation in the data (we will return to this). This is sometimes referred to as the "between" estimator, as all of the identifying variation comes from differences in the overall averages between groups.

Estimation using Differenced Data

- Imagine we have a random effects model given by

- $y_{it} = c_i + x_{it}\beta + e_{it}$

- If we difference the model, we recover
- $y_{it} - y_{it-1} = (c_i - c_i) + (x_{it} - x_{it-1})\beta + (e_{it} - e_{it-1})$
- $y_{it} - y_{it-1} = (x_{it} - x_{it-1})\beta + u_{it}$

- This removes any individual level heterogeneity from the model! This is great… but…

Estimation using Differenced Data

- $y_{it} - y_{it-1} = (x_{it} - x_{it-1})\beta + u_{it}$

- This removes any individual level heterogeneity from the model!
- Unfortunately, it also removes our ability to estimate any time-invariant effects, like the effect of gender, education status, race, or greatly reduces the identifying variation (number of children may change over time, but relatively rarely).

- The error terms in this equation are now a moving average, which means that the estimator can be made more efficient (or inference can be corrected) by accommodating this as well.

Within Groups estimation

- Now we can put together these various components to talk about identifying variation.

- The pooled regression model is
- $y_{it} = \alpha + x_{it}\beta + e_{it}$
- And we can calculate the regression in terms of the group means:
- $\bar{y}_{i.} = \alpha + \bar{x}_{i.}\beta + \bar{e}_{i.}$
- Which leaves us the deviations from the group means:

- $y_{it} - \bar{y}_{i.} = (x_{it} - \bar{x}_{i.})\beta + (e_{it} - \bar{e}_{i.})$

Within Groups estimation

- Which leaves us the deviations from the group means:
- $y_{it} - \bar{y}_{i.} = (x_{it} - \bar{x}_{i.})\beta + (e_{it} - \bar{e}_{i.})$
- This is called the "within groups" (or within) estimator and it is identified solely based on changes in the x_{it} within groups over time.
- You can construct the pooled OLS estimate as a weighted combination of the within and between estimates. We will not prove this in this course, though it should be intuitive that the combined estimation combines these two sources of variation.

Fixed Effects Model

- $y_{it} = \alpha_i + \ x'_{it}\beta + e_{it}$
- $i = 1, \dots, n; t = 1, \dots T$
- $E(e_{it}|x_{i1}, x_{i2}, \dots, x_{iT}) = 0$
- $E\left(e_{it}e_{js}|x_{i1}, x_{i2}, \dots, x_{iT}\right) = \sigma_e^2$ if $i = j$ and $t = s$ and zero otherwise
- α_i, α_j independent

But now the individual effect term (that we do not observe) can be correlated with the other variables:

$$E(\alpha_i|\ x_{i1}, x_{i2}, \dots, x_{iT}) = h(X_i)$$

Fixed Effects Model

- $y_{it} = \alpha_i + x'_{it}\beta + e_{it}$

$E(\alpha_i | x_{i1}, x_{i2}, \dots, x_{iT}) = h(X_i)$

Since we can't observe α_i, we can't estimate $y_{it} = \alpha_i + x_{it}\beta + e_{it} = x_{it}\beta + w_{it}$ since the $x's$ are now correlated with the error terms.

Remember, when that happens, we're not (only) inefficient, we're biased!

Fixed Effects Model

- $y_{it} = \alpha_i + \ x'_{it}\beta + e_{it}$

$E(\alpha_i |\ x_{i1}, x_{i2}, \ldots, x_{iT}) = h(X_i)$

But we could estimate a model where each individual observation has their own intercept by including a series of dummy variables, one for each group:

$$y_{it} = \delta_i + \ x'_{it}\beta + e_{it}$$

Our interpretation of β in this context would be based on the deviation from X and Y from their time series mean within groups (remember the Frisch Waugh Theorem!).

This is exactly the within groups estimator!!!

Fixed Effects Model

$$y_{it} = \delta_i + x'_{it}\beta + e_{it}$$

Unfortunately, this means that it has the same shortcomings… namely, that we can’t identify any time invariant effects, and that we are losing a great deal of degrees of freedom.

Which to use?

$$y_{it} = \alpha_i + x'_{it}\beta + e_{it}$$

The key is whether or not there are individual specific effects

And whether or not the individual specific effects are correlated with other observables in the model.

If they are correlated, then we have to use fixed effects/within estimation to avoid bias.

If they are not correlated, we can use random effects (with a lot more degrees of freedom/power) without the fear of bias.

Hausman Specification Test

$$y_{it} = \alpha_i + x'_{it}\beta + e_{it}$$

Another way of stating this is that if FE is necessary, it is unbiased, and RE is not, while

If FE is not necessary, it is inefficient, and RE is the efficient estimator.

If one estimator is biased (inconsistent) under the null that FE is necessary, but they both have the same probability limit when FE is unnecessary, we can compare the coefficients to determine if they are the same!

Hausman Specification Test

In this case, we will make an exception to derive and discuss the test statistic.

If the coefficients are the same, we know they are asymptotically normal and thus

$(\hat{\beta}_{FE} - \hat{\beta}_{RE})' V(\hat{\beta}_{FE} - \hat{\beta}_{RE})^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE})$ will be a Chi-squared distribution with K -1 degrees of freedom (where K is the number of regressors in X).

What is $V(\hat{\beta}_{FE} - \hat{\beta}_{RE})^{-1}$?

Hausman Specification Test

$$V\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \quad = var\left(\hat{\beta}_{FE}\right) + var\left(\hat{\beta}_{RE}\right) - 2cov\left(\hat{\beta}_{RE}, \hat{\beta}_{FE}\right)$$

But if RE is truly efficient, and they have the same plim, then its covariance with any other unbiased estimator has to be zero (or a more efficient estimator could be constructed!)

That is, $cov\left(\hat{\beta}_{RE}, \hat{\beta}_{FE} - \hat{\beta}_{RE}\right) = cov\left(\hat{\beta}_{RE}, \hat{\beta}_{FE}\right) - var\left(\hat{\beta}_{RE}\right) = 0 \Rightarrow$

$$cov\left(\hat{\beta}_{RE}, \hat{\beta}_{FE}\right) = var\left(\hat{\beta}_{RE}\right)$$

$$\begin{aligned} V\left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \quad &= var\left(\hat{\beta}_{FE}\right) + var\left(\hat{\beta}_{RE}\right) - 2cov\left(\hat{\beta}_{RE}, \hat{\beta}_{FE}\right) \\ &= var\left(\hat{\beta}_{FE}\right) - var\left(\hat{\beta}_{RE}\right) \end{aligned}$$

Hausman Specification Test

Asymptotic test, can tell us something about exclusion restriction in our specification

Finite sample performance can be questionable, and may not get a positive definite matrix as our estimate of the variance (which is obviously a problem). There are equivalent alternatives, though this test is one of the few that targets the key identifying assumption of least squares models.