Webinar Series in Applied Quantitative Analysis - Updated

Date	Topic
	Session One
February 29	Potential Outcomes and Omitted Variable Bias I (Theory)
March 7	Potential Outcomes and Omitted Variable Bias II (Application)
	Session Two
March 21	Difference-in-differences I (Theory)
March 28	Difference-in-differences II (Application)
	Session Three
April 25	Power analysis, clustering and sample size calculations I (Theory)
May 2	Power analysis, clustering and sample size calculations II (Application)
	Session Four
May 23	Propensity score matching (Theory)
May 30	Propensity score matching (Application)
	Session Five
June 20	Fixed-effects I (Theory)
June 27	Fixed-effects II (Application)
	Session Six
July 25	Instrumental variables I (Theory)
August 1	Instrumental Variables II (Application)
	Session Seven
August 22	Lagged dependent variables and the Arellano-Bond Estimator I (Theory)
August 29	Lagged dependent variables and the Arellano-Bond Estimator II (Application)

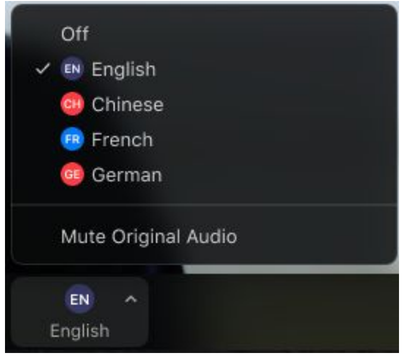
AIR®
Advancing Evidence.
Improving Lives.

https://www.airp3-africa.org/applied-quantitative-analysis-webinar-series

AERC
CREA

Écoute de l'interprétation d'une langue Windows | macOS

1. Dans les contrôles de votre réunion/webinaire, cliquez sur Interprétation .
2. Cliquez sur la langue que vous souhaitez entendre. (Nous aurons le français) Pas besoin de choisir l'anglais, c'est la langue de la salle Zoom principale



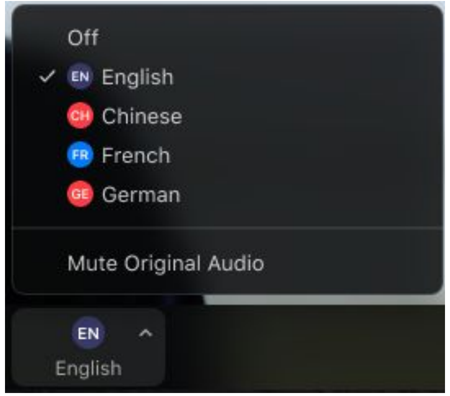
3. (Facultatif) Pour entendre uniquement la langue interprétée, cliquez sur Couper le son original.

Remarques:

- Vous devez rejoindre l’audio de la réunion via l’audio/VoIP de votre ordinateur. Vous ne pouvez pas écouter l’interprétation linguistique si vous utilisez les fonctions audio de connexion ou d’appel téléphonique.
- En tant que participant rejoignant une chaîne linguistique, vous pouvez retransmettre sur le canal audio principal canal si vous réactivez votre audio et parlez.

Listening to language interpretation
Windows | macOS

1. In your meeting/webinar controls, click Interpretation .
2. Click the language that you would like to hear. (We will have French) No need to choose English, that is the language in the main Zoom room



3. (Optional) To hear the interpreted language only, click Mute Original Audio.

Notes:

- You must join the meeting audio through your computer audio/VoIP. You cannot listen to language interpretation if you use the dial-in or call me phone audio features.
- As a participant joining a language channel, you can broadcast back into the main audio channel if you unmute your audio and speak.

Fixed Effects Theory

Professor Jeremy Moulton
Department of Public Policy
University of North Carolina at Chapel Hill

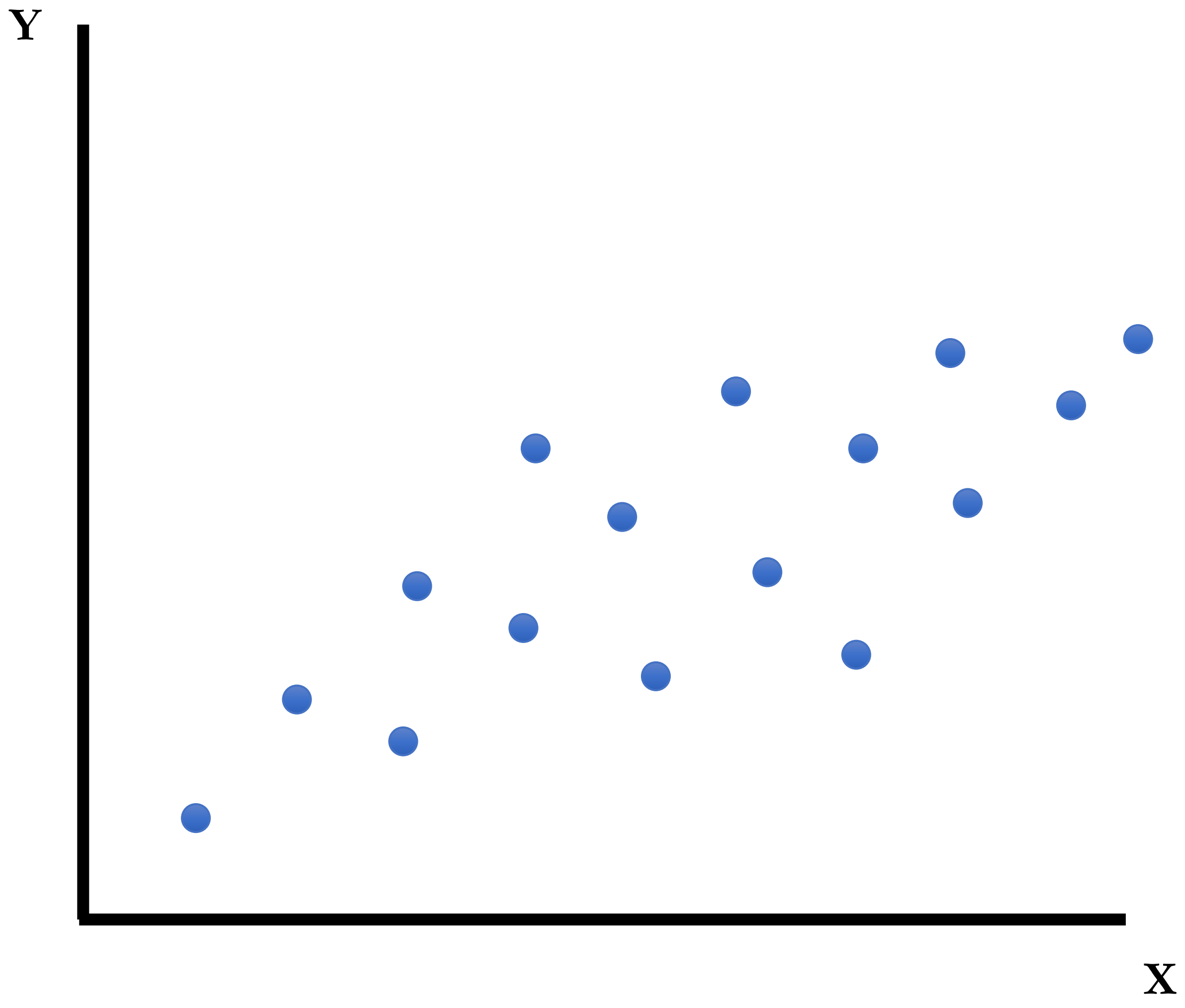
June 20, 2024

Advancing Evidence.
Improving Lives.

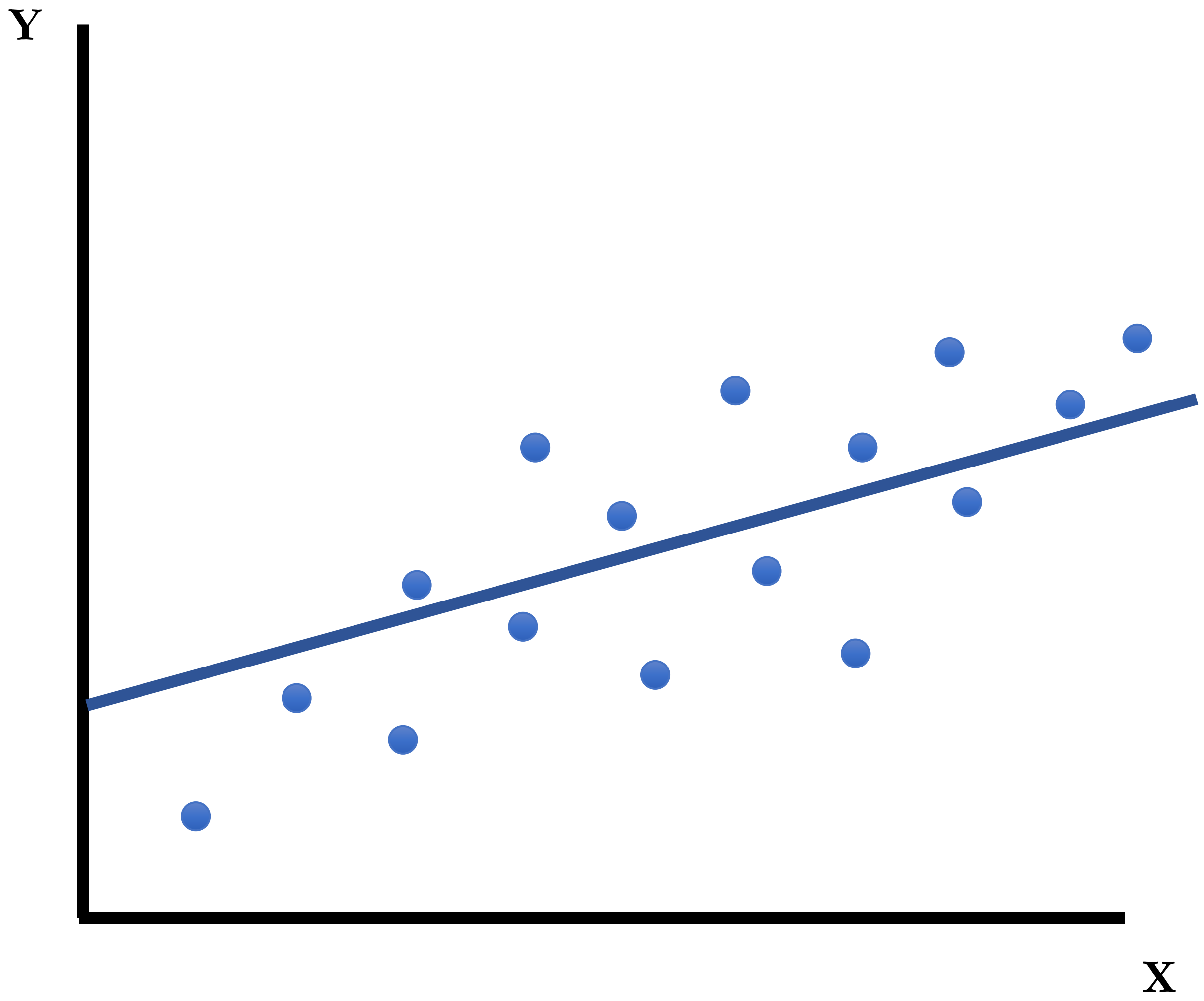
AERC
CREA

Panel Data

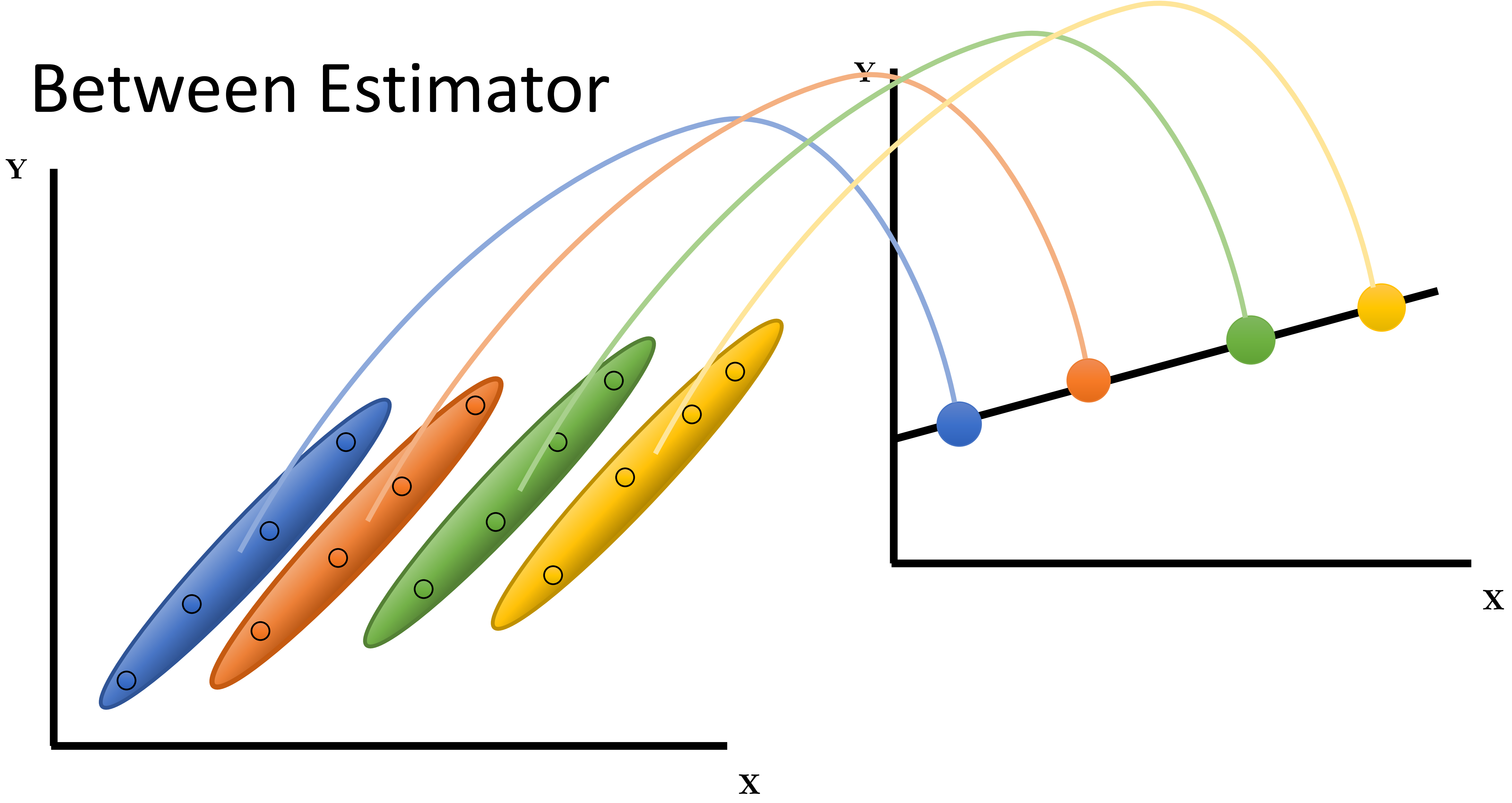
id	time		
1	1	2	0
1	2	2	0
1	3	5	1
2	1	2	0
2	2	4	1
2	3	4	1
3	1	5	1
3	2	5	1
3	3	5	1


Y
X

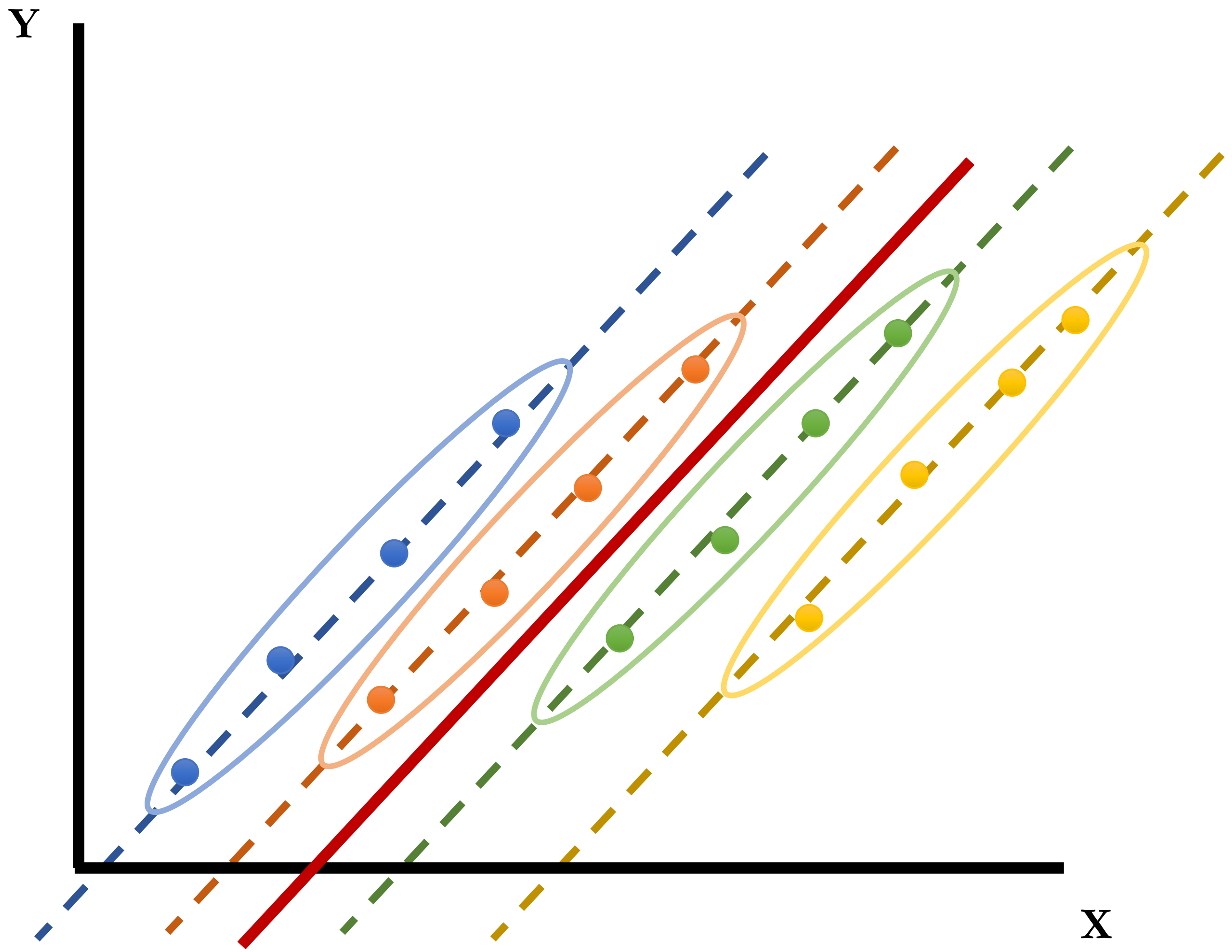
Pooled OLS (simple OLS regression)



Between Estimator



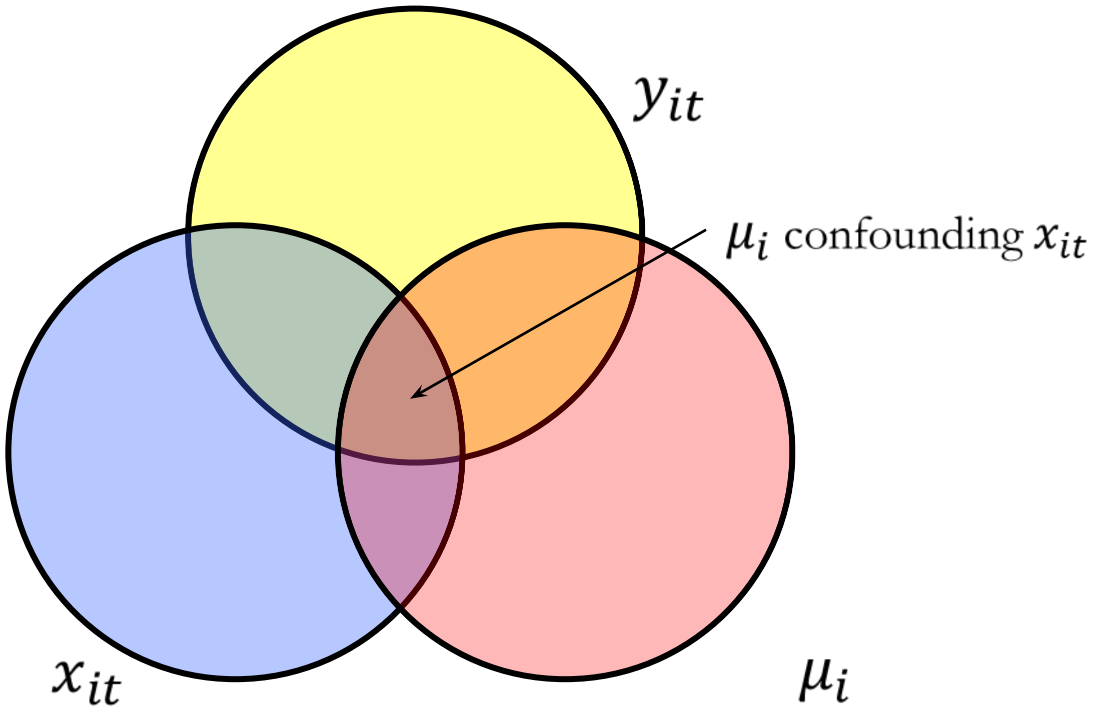
Within Estimator (Fixed Effects)



The Set Up

- $y_{it} = \alpha + \beta x_{it} + \gamma z_{it} + \delta z_i + \varepsilon_{it}$
  - $\varepsilon_{it} = \nu_{it} + \mu_i$

- x_{it} = independent variable of interest
- z_{it} = time-varying controls
- z_i = time-invariant controls
- μ_i = time-invariant individual specific error
  - Example: innate ability or health endowment



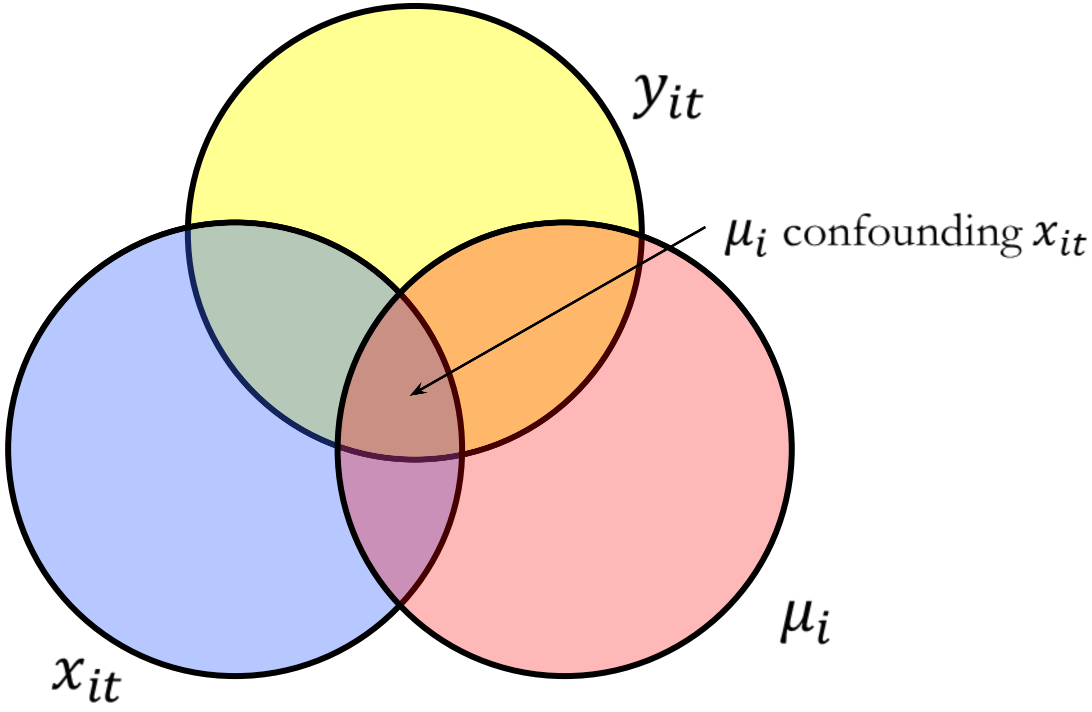
First Difference (T = 2)

- $$\begin{array}{rl} & y_{i2} = \alpha + \beta x_{i2} + \gamma z_{i2} + \delta z_i + \nu_{i2} + \mu_i \\ - & y_{i1} = \alpha + \beta x_{i1} + \gamma z_{i1} + \delta z_i + \nu_{i1} + \mu_i \\ \hline & \Delta y_{i2} = \qquad \beta \Delta x_{i2} + \gamma \Delta z_{i2} \qquad + \Delta \nu_{i2} \end{array}$$

What are you doing?

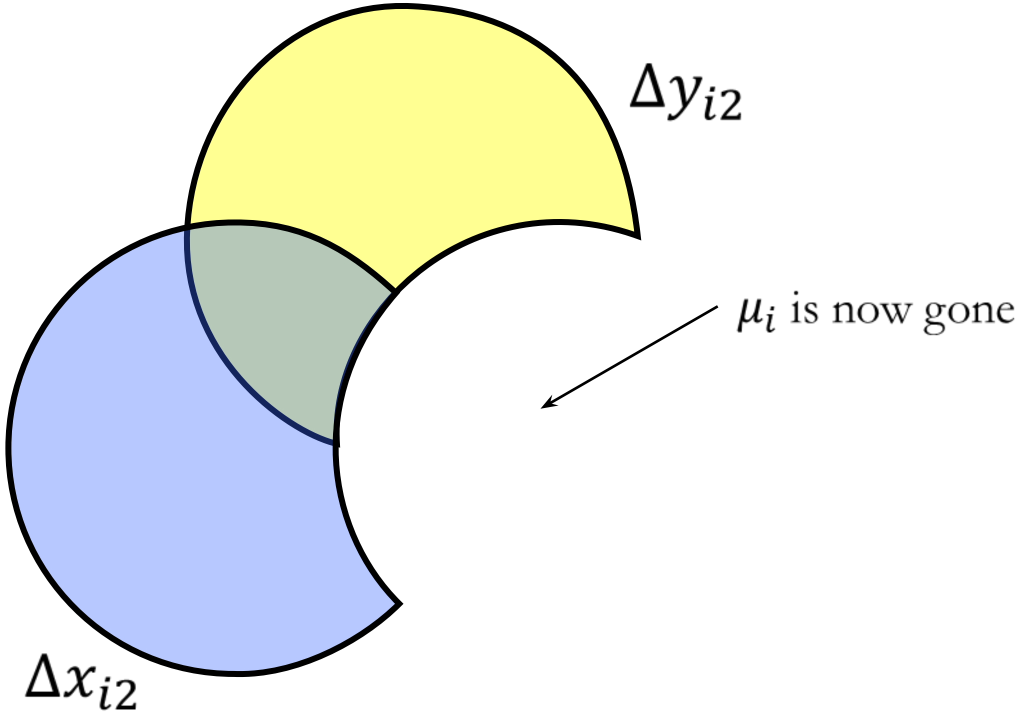
- $\Delta y_{i2} = \beta \Delta x_{i2} + \gamma \Delta z_{i2} + \Delta \nu_{i2}$

- Estimate a regression using the differenced variables
- The time-invariant variables drop out:
  - z_i: We cannot include time-invariant variables in the model
  - μ_i: But, the time-invariant unobservables “drop out” too.



First Differences Model

$$\Delta y_{i2} = \beta \Delta x_{i2} + \gamma \Delta z_{i2} + \Delta v_{i2}$$



First Difference (T > 2)

- $$\begin{array}{l} y_{i2} = \alpha + \beta x_{i2} + \gamma z_{i2} + \delta z_i + \nu_{i2} + \mu_i \\ \underline{- y_{i1} = \alpha + \beta x_{i1} + \gamma z_{i1} + \delta z_i + \nu_{i1} + \mu_i} \\ \Delta y_{i2} = \quad \beta \Delta x_{i2} \quad + \gamma \Delta z_{i2} \qquad\qquad + \Delta \nu_{i2} \end{array}$$

$$\begin{array}{l} y_{i3} = \alpha + \beta x_{i3} + \gamma z_{i3} + \delta z_i + \nu_{i3} + \mu_i \\ \underline{- y_{i2} = \alpha + \beta x_{i2} + \gamma z_{i2} + \delta z_i + \nu_{i2} + \mu_i} \\ \Delta y_{i3} = \quad \beta \Delta x_{i3} \quad + \gamma \Delta z_{i3} \qquad\qquad + \Delta \nu_{i3} \end{array}$$

The First Difference Data

id	time				
1	1	2	0		
1	2	2	0	0	0
1	3	5	1	3	1
2	1	2	0		
2	2	4	1	2	1
2	3	4	1	0	0
3	1	5	1		
3	2	5	1	0	0
3	3	5	1	0	0

Fixed Effects Model

- $y_{it} = \alpha + \beta x_{it} + \gamma z_{it} + \delta z_i + \varepsilon_{it}$
  - $\varepsilon_{it} = \nu_{it} + \mu_i$
- Include an indicator (or fixed effect) for all individuals (or whatever you are following):
  - $y_{it} = \alpha + \beta x_{it} + \gamma z_{it} + \delta z_i + \theta Individual_i + \varepsilon_{it}$
  - Or, we often just simplify to using α_i
  - $y_{it} = \alpha_i + \beta x_{it} + \gamma z_{it} + \delta z_i + \varepsilon_{it}$
- α_i is collinear with μ_i. So if we include α_i, we have "controlled" for the time-invariant unobservables.

Between Effects Model

- The **Between Effects Model** is estimated using only the mean values for each individual:
  - $\bar{y}_i = \beta \bar{x}_i + \gamma \bar{z}_i + \bar{\varepsilon}_i$
  - Where: $\bar{y}_i = \frac{1}{T} \sum_{t=1}^{T} y_{it}$

Fixed Effect: De-Meaned version

- The **Fixed Effects De-Meaned Version** is estimated by first subtracting an individuals' mean value for each variable
- $(y_{it} - \bar{y}_i) = (\alpha - \bar{\alpha}) + \beta(x_{it} - \bar{x}_i) + \gamma(z_{it} - \bar{z}_i) + \delta(z_i - \bar{z}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$
  - $(\alpha - \bar{\alpha}) = 0$
  - $\delta(z_{it} - \bar{z}_i) = 0$: Time-invariant variables "drop out"
  - $(\varepsilon_{it} - \bar{\varepsilon}_i) = (\nu_{it} - \bar{\nu}_i) + (\mu_i - \bar{\mu}_i) = (\nu_{it} - \bar{\nu}_i)$: But, the time-invariant unobservables also drop out.

Let’s go back to an earlier class

How do we actually estimate $\widehat{\beta_1}$?

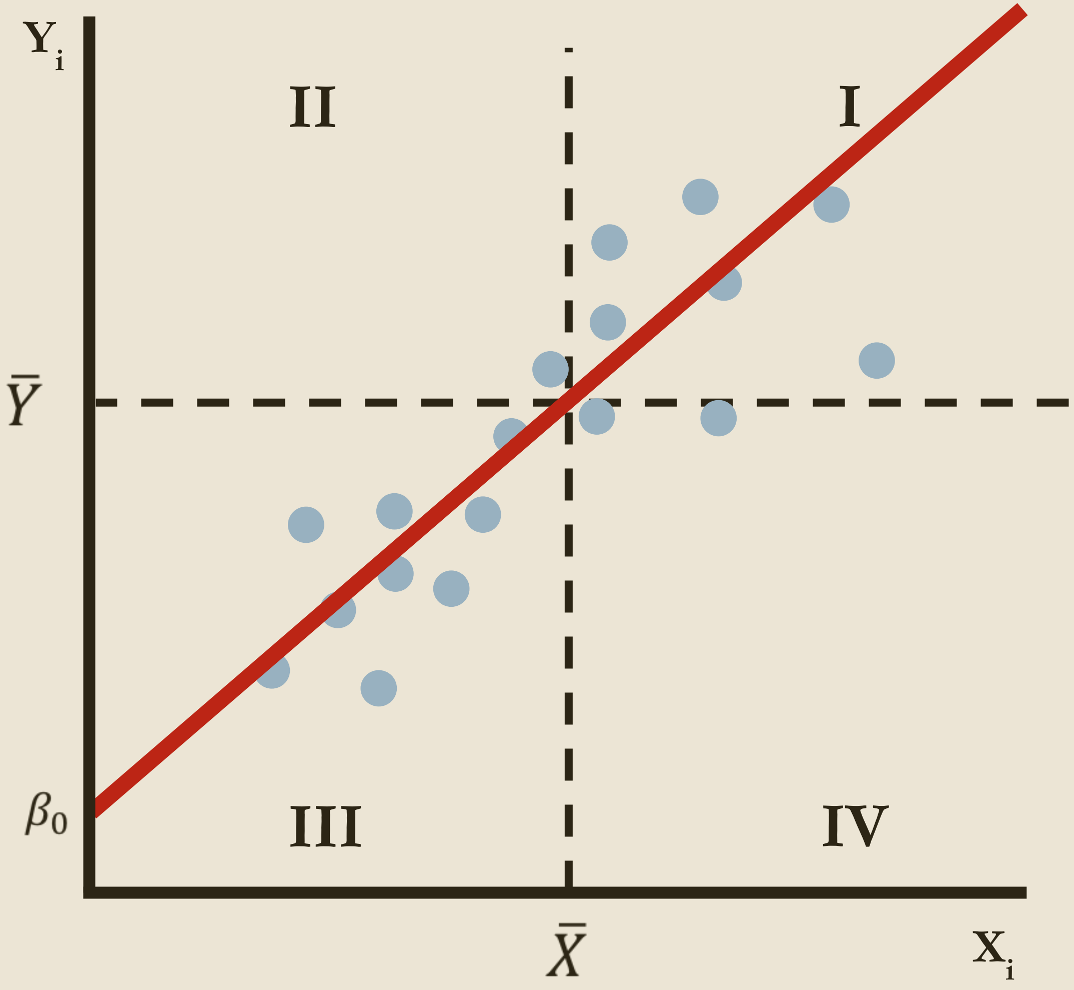
$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- While the denominator is always positive, the numerator is not squared, so it can be positive or negative.
- So the numerator that will determine whether the slope ($\widehat{\beta_1}$) is positive or negative.

Positive slope

- If most of the observations are in quadrants I and III, then the slope is positive
  - Quadrant I: X and Y are above their mean
  - Quadrant III: X and Y are below their mean

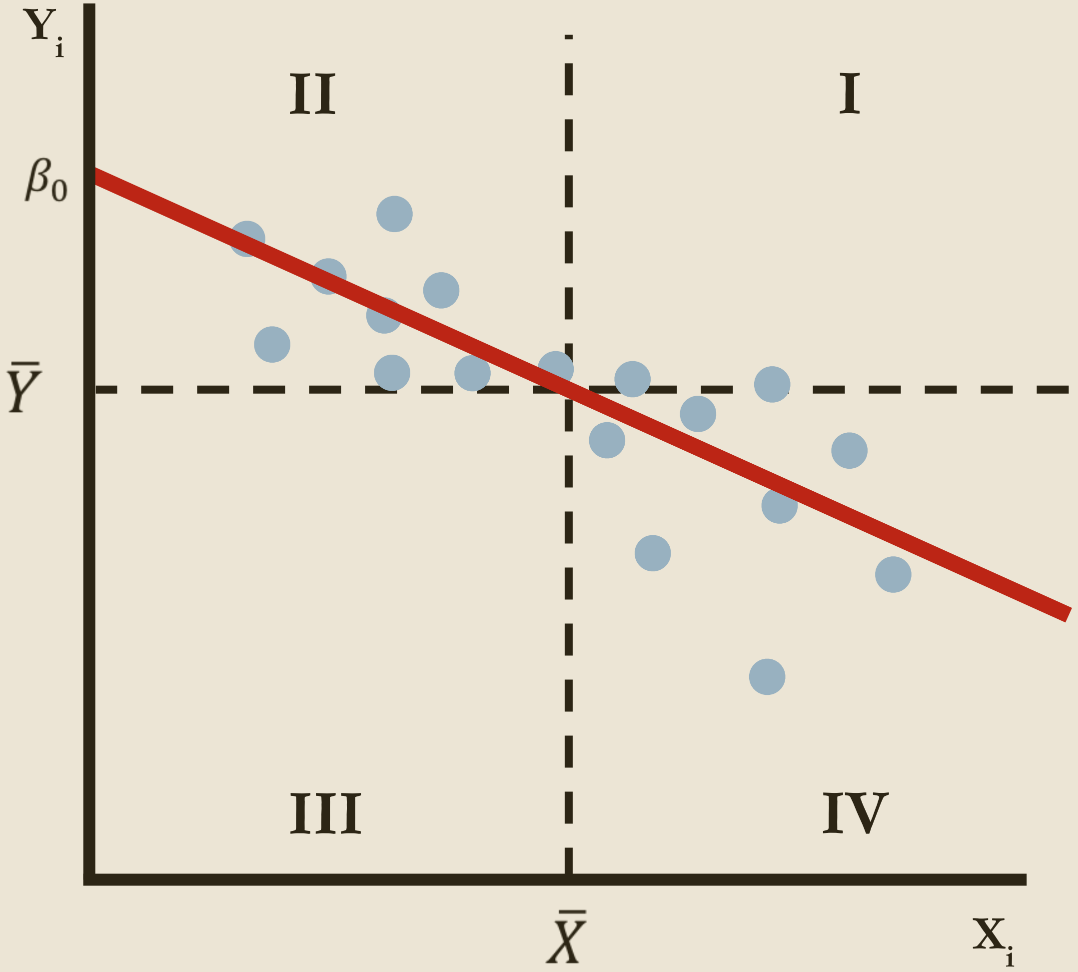
$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$



Negative slope

- If most of the observations are in quadrants II and IV, then the slope is negative
  - Quadrant II: X is below and Y above their mean
  - Quadrant IV: X is above and Y below their mean

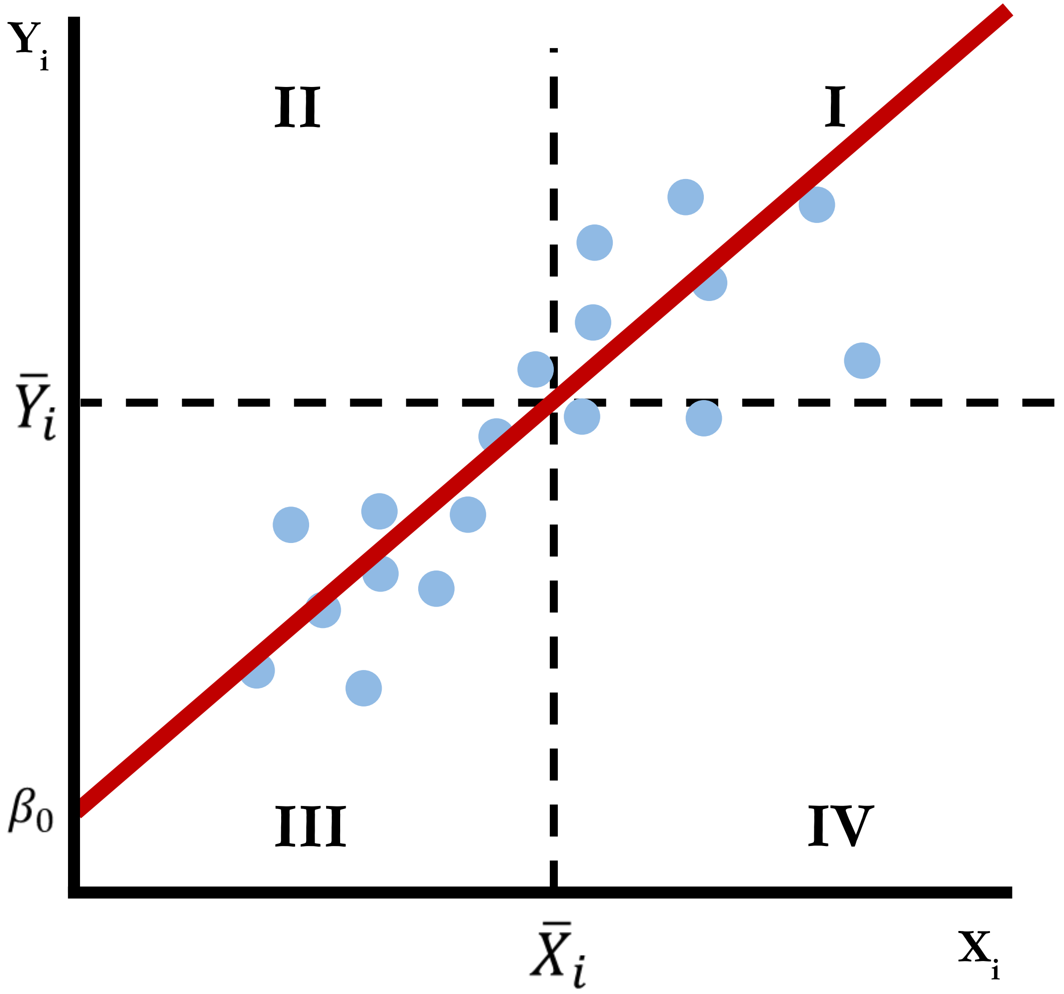
$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$



OK, now go back to the future

Do that again, but with Fixed Effects

- **Fixed Effects** changes the reference point from the overall mean to the individual mean.
  - How do you in one time period compare to your average – instead of how do you in one time period compare to everyone's average?



$\bar{Y}$ $\bar{X}$

Fixed Effects: De-Meaned version

id	time				
1	1	2	0	3	0.33
1	2	2	0	3	0.33
1	3	5	1	3	0.33
2	1	2	0	3.33	0.67
2	2	4	1	3.33	0.67
2	3	4	1	3.33	0.67
3	1	5	1	5	1
3	2	5	1	5	1
3	3	5	1	5	1

Fixed Effects: De-Meaned version

id	time						
1	1	2	0	3	0.33	-1	-.33
1	2	2	0	3	0.33	-1	-.33
1	3	5	1	3	0.33	2	.67
2	1	2	0	3.33	0.67	-1.33	-.67
2	2	4	1	3.33	0.67	.67	.33
2	3	4	1	3.33	0.67	.67	.33
3	1	5	1	5	1	0	0
3	2	5	1	5	1	0	0
3	3	5	1	5	1	0	0

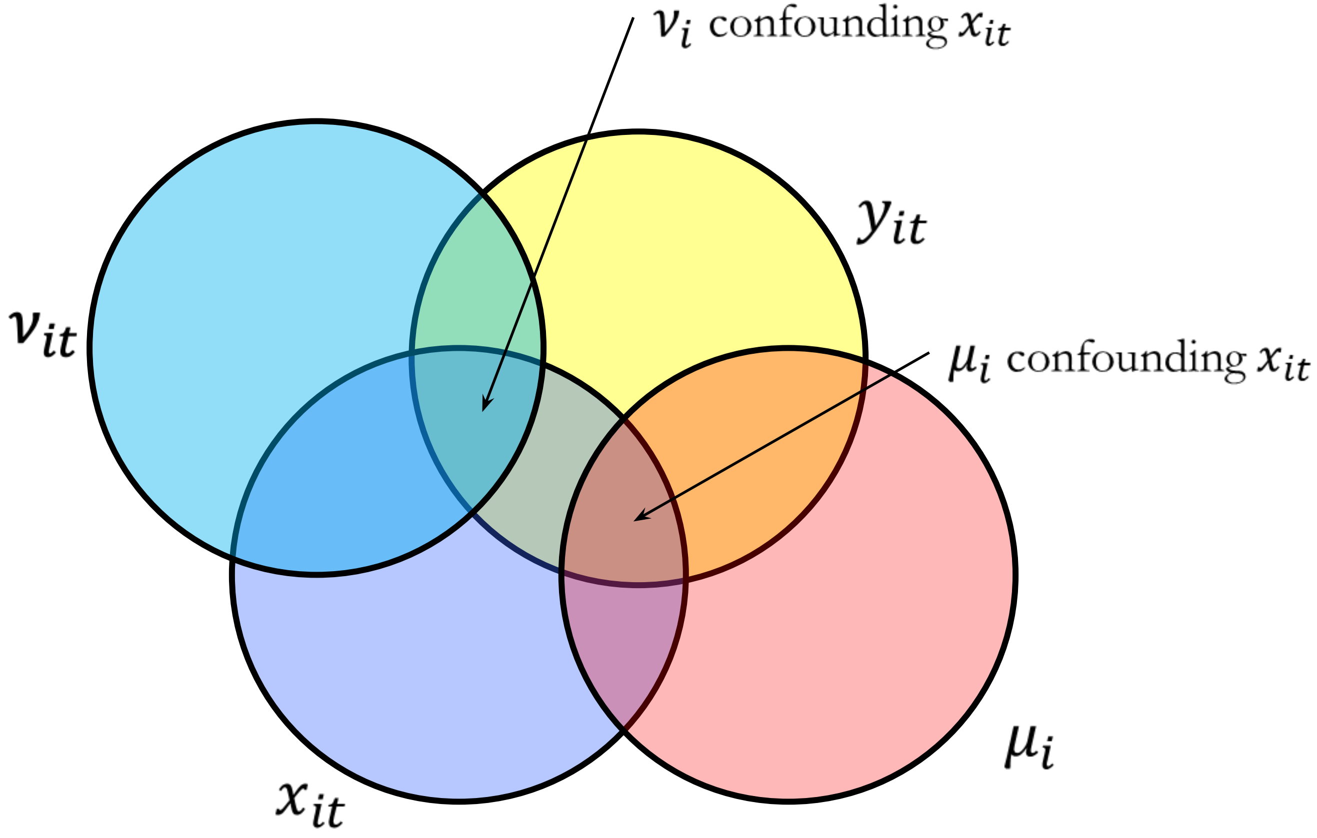
Ok, but how do I actually do it?

- **<u>First Difference:</u>**
  - `reg D.(y x)`
- **<u>Fixed Effects:</u>**
- You could create a fixed effect for every individual:
  - `tab id, gen(idFE)`
  - `reg y x idFE*`
- You could include a temporarily created fixed effect for ever individual:
  - `reg y x i.id`
- You could use the xt version:
  - `xtset id year`
  - `xtreg y x, fe`
- Or other options:
  - `areg y x, absorb(id)`
  - `reghdfe y x, absorb(id)`
    - *You need to install this ado file, but it is very helpful when using a large number of fixed effects*

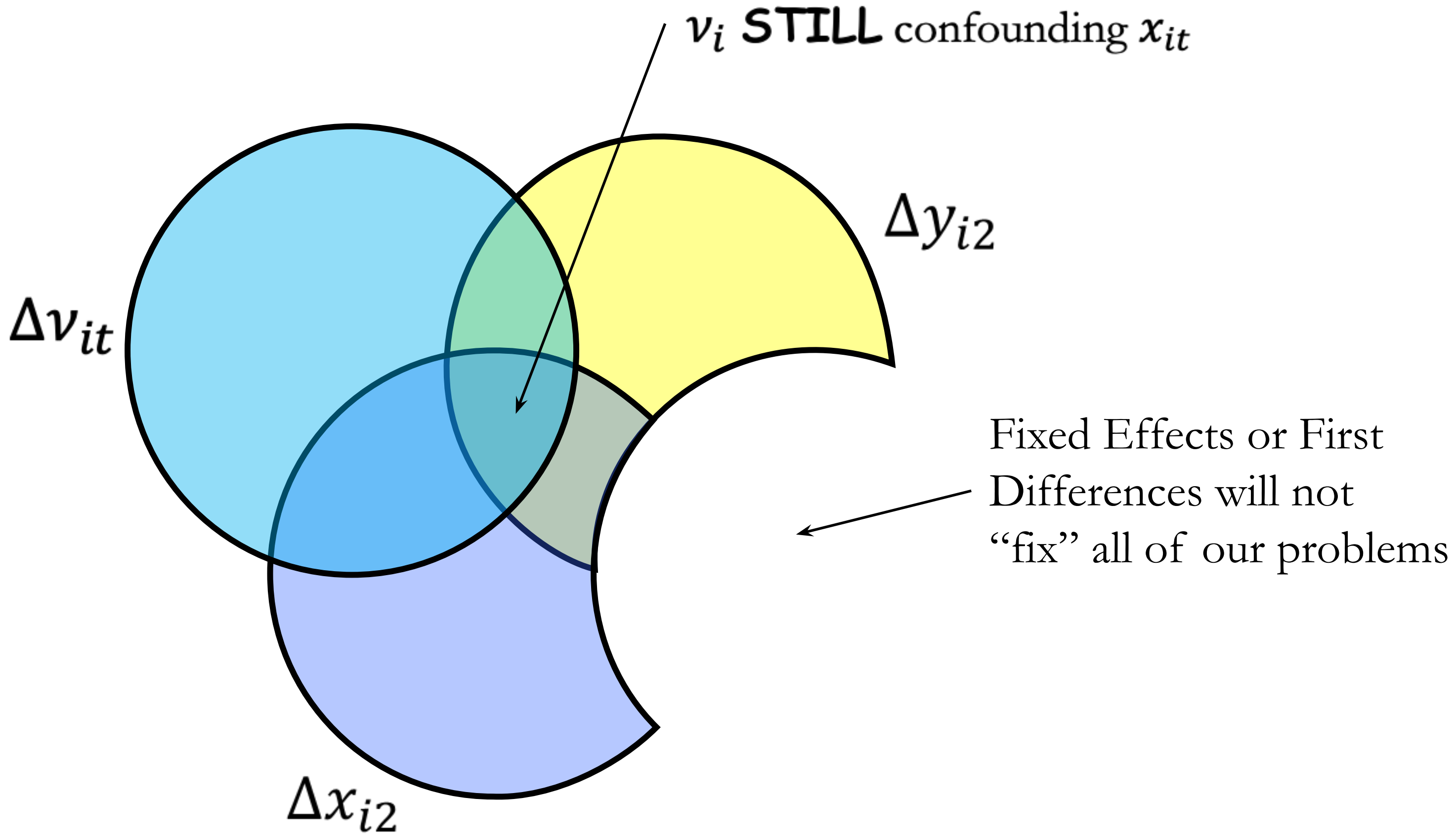
But, what about time-varying unobservables?

- Just because we removed time-invariant bias (μ_i), does not mean everything is fixed.
- There could still be bias from time **varying** unobservables:
  - For example: Insurance and Healthcare Expenditures, but assume you do not observe Current Health Status or Health Endowment
    - Health Endowment is time-invariant, so the fixed effects will fix this issue
    - BUT, if sicker individuals find it harder to obtain health insurance and have higher health expenditures, then we are in trouble.
    - Or someone that is getting in shape may find health insurance is cheaper, but uses less of it

Time Varying Unobservables



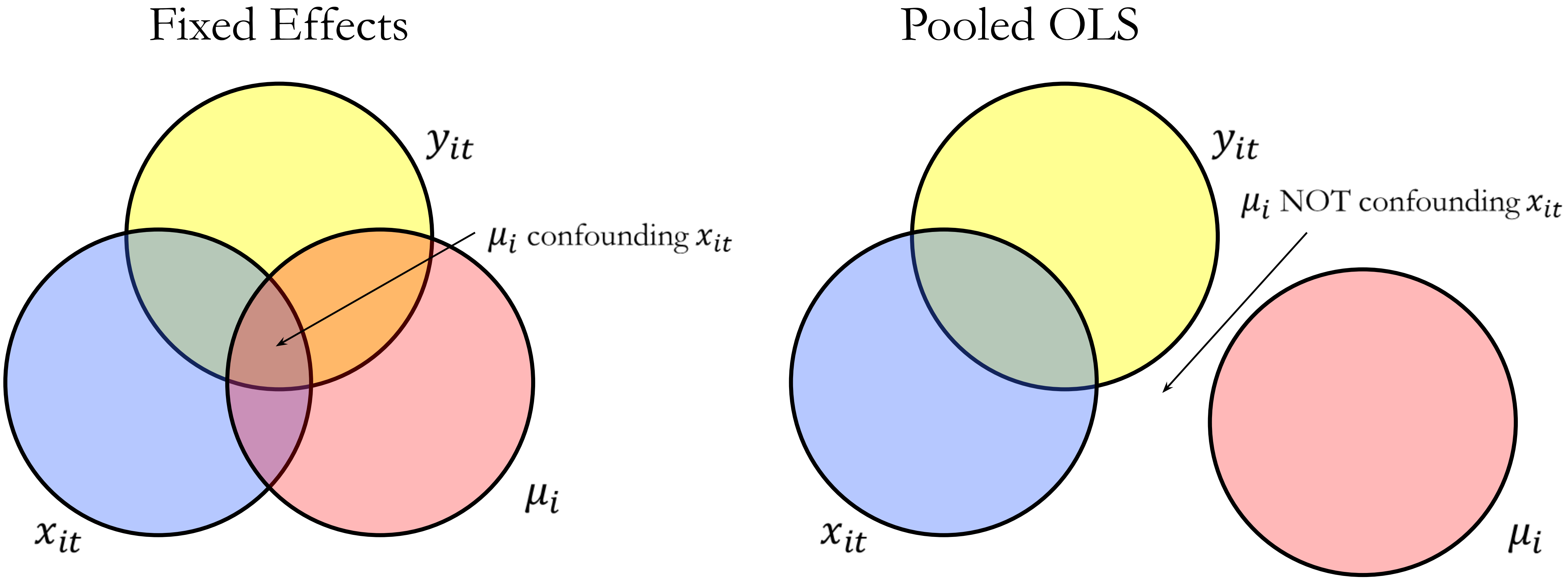
Time Varying Unobservables



Time invariant independent variables

- $(y_{it} - \bar{y}_i) = (\alpha - \bar{\alpha}) + \beta(x_{it} - x_i) + \gamma(z_{it} - \bar{z}_i) + \delta(z_i - \bar{z}_i) + (\varepsilon_{it} - \bar{\varepsilon}_i)$
- $\Delta y_{i2} = \beta \Delta x_{i2} + \gamma \Delta z_{i2} + \Delta \nu_{i2}$
- But what if my variable of interest was z_i?
  - Can estimate heterogenous treatment effects for different groups
  - Or can predict α_i from Fixed Effects model
    - Then use α_i as outcome in regression to determine how much of α_i can be explained by z_i
  - Any other ideas?

Fixed Effects vs. Pooled OLS



How do I know if OLS or FE is better?

- We can test if the individual specific intercepts $\alpha_{-i} = 0$, meaning that they don't explain any of the variation in y_{it}
  - If $\alpha_{-i} = 0$, then Pooled OLS is probably the right model
  - If $\alpha_{-i} \neq 0$, then Fixed Effects might be the right model
  - We can conduct this test in Stata using an F-test of the individual fixed effect coefficients:
    - `reg y x i.id`
    - `testparm i*.id`
  - Or it is included at the bottom of the `xtreg y x, fe` regression output

Test difference in coefficients

- We can also test if the coefficients from OLS and Fixed Effects are different using a Hausman test.
  - If the same, then you want to use the efficient estimator (Pooled OLS)
  - If different, then you want to use the consistent estimator (Fixed Effects)
- Consistent Estimator (Fixed Effects)
  - Converges in probability to population parameter as N grows without bound
- Efficient Estimator (Pooled OLS)
  - Has a smaller variance
- $(\beta^{FE} - \beta^{OLS})[V(\beta^{FE}) - V(\beta^{OLS})]^{-1}(\beta^{FE} - \beta^{OLS})$
- `reg y x`
- `estimates store my_ols`
- `xtreg y x, fe`
- `estimates store my_fe`
- `hausman my_fe my_ols, sigmamore`
  - Be sure to put the consistent estimator before the efficient estimator. sigmamore specifies that the covariance matrices be based on the estimated disturbance variance be based on the estimated disturbance variance from the efficient estimator. This forces the difference between the covariance matrices to be positive definite.