Lecture 13: Repeated Observations I

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

March 21, 2017

Identification with repeated observations

Data on “repeated observations” (panel data, time-series cross section data, spatially clustered data, etc.) provide opportunities beyond what you can do with simple cross sections:

Identification with repeated observations

Data on “repeated observations” (panel data, time-series cross section data, spatially clustered data, etc.) provide opportunities beyond what you can do with simple cross sections:

- More nuanced estimation of causal effects *under* randomization or CIA given observables.
  - Growth curves and trajectories.
  - Dynamic treatment regimes.

Dynamic treatment regimes

- Blackwell and Glynn (2013) discuss effects in longitudinal data.

Dynamic treatment regimes

- Blackwell and Glynn (2013) discuss effects in longitudinal data.
- Treatment sequences:
  - Past treatment history, $\underline{D}_{it} = (D_{i1}, ..., D_{it}, D_{it+1}, , ...)$.
  - Future treatment trajectory: $\underline{D}_{it,\tau}(d_{it}) = (D_{it+1}(d_{it}), ..., D_{it+\tau}(d_{it}))$

Dynamic treatment regimes

- Blackwell and Glynn (2013) discuss effects in longitudinal data.
- Treatment sequences:
  - Past treatment history, $\underline{D}_{it} = (D_{i1}, \ldots, D_{it}, D_{it+1}, , \ldots)$.
  - Future treatment trajectory: $\underline{D}_{it,\tau}(d_{it}) = (D_{it+1}(d_{it}), \ldots, D_{it+\tau}(d_{it}))$
- "Blip" effects:
  - Last period blip:
    $\mathrm{E}_{\underline{D}}\left[\mathrm{E}\left[Y_{it}(\underline{d}_{it-1}, 1) - Y_{it}(\underline{d}_{it-1}, 0) | \underline{D}_{it-1} = \underline{d}_{it-1}\right]\right]$.
  - First period blip: $\mathrm{E}\left[Y_{it+\tau}(1, \underline{D}_{it,\tau}(1)) - Y_{it+\tau}(0, \underline{D}_{it,\tau}(0))\right]$
  - Blip effects identified under usual assumptions.

Dynamic treatment regimes

- Blackwell and Glynn (2013) discuss effects in longitudinal data.
- Treatment sequences:
  - Past treatment history, $\underline{D}_{it} = (D_{i1}, ..., D_{it}, D_{it+1}, , ...)$.
  - Future treatment trajectory: $\underline{D}_{it,\tau}(d_{it}) = (D_{it+1}(d_{it}), ..., D_{it+\tau}(d_{it}))$
- "Blip" effects:
  - Last period blip:
    $\mathrm{E}_{\underline{D}}\left[\mathrm{E}\left[Y_{it}(\underline{d}_{it-1}, 1) - Y_{it}(\underline{d}_{it-1}, 0) | \underline{D}_{it-1} = \underline{d}_{it-1}\right]\right]$.
  - First period blip: $\mathrm{E}\left[Y_{it+\tau}(1, \underline{D}_{it,\tau}(1)) - Y_{it+\tau}(0, \underline{D}_{it,\tau}(0))\right]$
  - Blip effects identified under usual assumptions.
- Treatment regime effects:
  - Effects of sequences of treatments, $\underline{d}$.
  - Effects of simplified combinations of treatment sequences (e.g., total number of periods under treatment, treatment in last three periods, etc.)—"marginal structural models."
  - Sequence effects require "sequential ignorability":
    For every sequence $\underline{d}_t$, covariate history $\underline{X}_{it}$, and period t,
    $Y_{it}(\underline{d}_t) \perp\!\!\!\perp D_{it} | \underline{X}_{it}, \underline{D}_{it-1} = \underline{d}_{t-1}$.
- See Blackwell (2012) for more.

Identification with repeated observations

Data on "repeated observations" (panel data, time-series cross section data, spatially clustered data, etc.) provide opportunities beyond what you can do with simple cross sections:

- More nuanced estimation of causal effects *under* randomization or CIA given observables.
  - Growth curves and trajectories.
  - Dynamic treatment regimes.
- Possibility of identifying causal effects *when we do not have* randomization or CIA given observables.
  - Controlling for unobserved confounders.

Identification with repeated observations

Techniques we will consider:

- Fixed effects estimation.
- Difference in differences estimation and extensions.

Motivating Example

American Political Science Review Vol. 105, No. 3 August 2011

doi:10.1017/S0003055411000281

Do Democracies Select More Educated Leaders?

TIMOTHY BESLEY *London School of Economics and Political Science*
MARTA REYNAL-QUEROL *Universitat Pompeu Fabra*

If we pool data since 1848 on regime types and leader education, there is a clear pattern: leaders in democracies have more education. So, can we conclude that installing democratic institutions causes a country to select more educated leaders?

Motivating Example

American Political Science Review Vol. 105, No. 3 August 2011
doi:10.1017/S0003055411000281

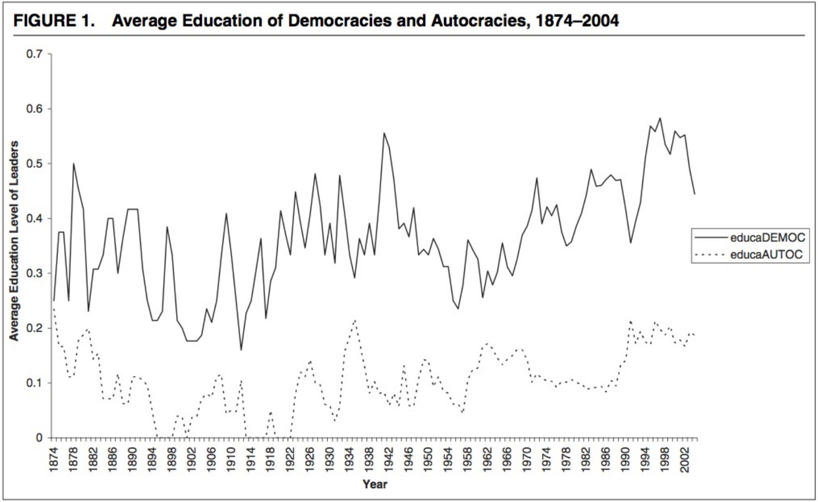
Do Democracies Select More Educated Leaders?

TIMOTHY BESLEY *London School of Economics and Political Science*
MARTA REYNAL-QUEROL *Universitat Pompeu Fabra*

If we pool data since 1848 on regime types and leader education, there is a clear pattern: leaders in democracies have more education. So, can we conclude that installing democratic institutions causes a country to select more educated leaders?

- What about global macro-trend of rising education as well as democracy?
- Could this mean that the relationship is spurious to this temporal coincidence?

Motivating Example



So the picture suggests that there is a persistent difference. Is this enough to conclude that installing democratic institutions *causes* a country to select more democratic leaders?

Fixed Effects

- Fixed effects (FE) methods allow us to use repeated observations to account for *fixed sources of confounding* in estimating causal effects.

Fixed Effects

- Fixed effects (FE) methods allow us to use repeated observations to account for *fixed sources of confounding* in estimating causal effects.
- Conventionally, these methods rely on strong functional form assumptions.
- We can loosen these assumptions somewhat.

Fixed Effects

The conventional FE setting, adapted for causal inference:

[1] Definition of these potential outcomes is tricky since they can depend on treatment histories. Assumptions below sidestep this issue with X specification.

Fixed Effects

The conventional FE setting, adapted for causal inference:

- We have a sample of units indexed by $i = 1, .., N$ observed over periods indexed by $t = 1, .., T_i$.
- $T_i \geq 2$ for all i, and number of units with same t value is at least 2 for all t. This is either panel or time-series cross-section data.

[1] Definition of these potential outcomes is tricky since they can depend on treatment histories. Assumptions below sidestep this issue with X specification.

Fixed Effects

The conventional FE setting, adapted for causal inference:

- We have a sample of units indexed by $i = 1, .., N$ observed over periods indexed by $t = 1, .., T_i$.
- $T_i \geq 2$ for all i, and number of units with same t value is at least 2 for all t. This is either panel or time-series cross-section data.
- $D_{it} = 0, 1$ is treatment assigned to i in period, t.
- X_{it} is a vector of covariates that vary for i over t.

[1] Definition of these potential outcomes is tricky since they can depend on treatment histories. Assumptions below sidestep this issue with X specification.

Fixed Effects

The conventional FE setting, adapted for causal inference:

- We have a sample of units indexed by $i = 1, .., N$ observed over periods indexed by $t = 1, .., T_i$.
- $T_i \geq 2$ for all i, and number of units with same t value is at least 2 for all t. This is either panel or time-series cross-section data.
- $D_{it} = 0, 1$ is treatment assigned to i in period, t.
- X_{it} is a vector of covariates that vary for i over t.
- We have Y_{1it} and Y_{0it}, *period-specific* potential outcomes under treatment or control, respectively.[1]
- We observe $Y_{it} = D_{it}Y_{1it} + (1 - D_{it})Y_{0it} = Y_{0it} + D_{it}(Y_{1it} - Y_{0it})$.

[1] Definition of these potential outcomes is tricky since they can depend on treatment histories. Assumptions below sidestep this issue with X specification.

Fixed Effects

The conventional FE setting, adapted for causal inference:

- We have a sample of units indexed by $i = 1, .., N$ observed over periods indexed by $t = 1, .., T_i$.
- $T_i \geq 2$ for all i, and number of units with same t value is at least 2 for all t. This is either panel or time-series cross-section data.
- $D_{it} = 0, 1$ is treatment assigned to i in period, t.
- X_{it} is a vector of covariates that vary for i over t.
- We have Y_{1it} and Y_{0it}, *period-specific* potential outcomes under treatment or control, respectively.[1]
- We observe $Y_{it} = D_{it}Y_{1it} + (1 - D_{it})Y_{0it} = Y_{0it} + D_{it}(Y_{1it} - Y_{0it})$.
- A_i is vector of "time-invariant" attributes of i.
- S_t is vector of "time-specific" conditions that apply to all i in time t.
- A_i and S_t may be unmeasured (unobserved).

[1] Definition of these potential outcomes is tricky since they can depend on treatment histories. Assumptions below sidestep this issue with X specification.

Fixed Effects

- Assumption 1: D_{it} is conditionally mean independent in any given period, with the conditioning set including the covariate as well as unit- and time-specific effects:

$$\mathrm{E}[Y_{0it}|A_i, S_t, X_{it}, D_{it}] = \mathrm{E}[Y_{0it}|A_i, S_t, X_{it}]$$

  This satisfied under CIA conditional on A_i and S_t, which may be unmeasured, as well as X_{it}.[2]

[2] Note that X_{it} may contain previous treatment assignment histories, lags, etc.

Fixed Effects

- Assumption 1: D_{it} is conditionally mean independent in any given period, with the conditioning set including the covariate as well as unit- and time-specific effects:

$$\mathrm{E}\left[Y_{0it}|A_i, S_t, X_{it}, D_{it}\right] = \mathrm{E}\left[Y_{0it}|A_i, S_t, X_{it}\right]$$

  This satisfied under CIA conditional on A_i and S_t, which may be unmeasured, as well as X_{it}.[2]

- Assumption 2: Y_{0it} can be characterized via the linear expression,

$$\mathrm{E}\left[Y_{0it}|A_i, S_t, X_{it}\right] = \mu + A_i'\gamma + S_t'\zeta + X_{it}'\beta,$$

  such that $A_i'\gamma$ is fixed over time, $S_t'\zeta$ is fixed over units, and μ is a global constant.

[2] Note that X_{it} may contain previous treatment assignment histories, lags, etc.

Fixed Effects

- Assumption 1: D_{it} is conditionally mean independent in any given period, with the conditioning set including the covariate as well as unit- and time-specific effects:

$$\mathrm{E}[Y_{0it}|A_i, S_t, X_{it}, D_{it}] = \mathrm{E}[Y_{0it}|A_i, S_t, X_{it}]$$

  This satisfied under CIA conditional on A_i and S_t, which may be unmeasured, as well as X_{it}.[2]

- Assumption 2: Y_{0it} can be characterized via the linear expression,

$$\mathrm{E}[Y_{0it}|A_i, S_t, X_{it}] = \mu + A_i'\gamma + S_t'\zeta + X_{it}'\beta,$$

  such that $A_i'\gamma$ is fixed over time, $S_t'\zeta$ is fixed over units, and μ is a global constant.

- Assumption 3: Causal effects are constant and additive over i and t:

$$\mathrm{E}[Y_{1it}|A_i, S_t, X_{it}] = \mathrm{E}[Y_{0it}|A_i, S_t, X_{it}] + \rho.$$

  Thus ρ defines a *constant per-period treatment effect*, the target causal parameter of interest.

[2] Note that X_{it} may contain previous treatment assignment histories, lags, etc.

Fixed Effects

- Define the abbreviated terms $\alpha_i = A_i'\gamma$ and $\lambda_t = S_t'\zeta$.

Fixed Effects

- Define the abbreviated terms $\alpha_i = A_i'\gamma$ and $\lambda_t = S_t'\zeta$.
- Define $\varepsilon_{it} = Y_{0it} - \mathrm{E}[Y_{0it}|A_i, X_{it}, t]$.

Fixed Effects

- Define the abbreviated terms $\alpha_i = A_i'\gamma$ and $\lambda_t = S_t'\zeta$.
- Define $\varepsilon_{it} = Y_{0it} - \mathrm{E}[Y_{0it}|A_i, X_{it}, t]$.
- Then, putting it together, observed outcomes are given by,

$$Y_{it} = \mu + \alpha_i + \lambda_t + \rho D_{it} + X_{it}'\beta + \varepsilon_{it}.$$

  This is a model with unit-specific (α_i) and time-specific (λ_t) "fixed effects."
- We could estimate this via OLS, using unit-specific and time-specific dummy variables to estimate α_i and λ_t.

Fixed Effects

- Define the abbreviated terms $\alpha_i = A_i'\gamma$ and $\lambda_t = S_t'\zeta$.
- Define $\varepsilon_{it} = Y_{0it} - \mathrm{E}[Y_{0it}|A_i, X_{it}, t]$.
- Then, putting it together, observed outcomes are given by,

$$Y_{it} = \mu + \alpha_i + \lambda_t + \rho D_{it} + X_{it}'\beta + \varepsilon_{it}.$$

  This is a model with unit-specific (α_i) and time-specific (λ_t) "fixed effects."
- We could estimate this via OLS, using unit-specific and time-specific dummy variables to estimate α_i and λ_t.
- Note what this implies: we don't have to *measure* the components of A_i and S_t in order to take advantage of Assumption 1. We only have to measure whatever X_{it} are needed for Assumption 1 to hold.

Fixed Effects

- This is why FE models are touted as allowing for the identification of causal effects despite *unobserved* unit or period-specific confounding.

Fixed Effects

- This is why FE models are touted as allowing for the identification of causal effects despite *unobserved* unit or period-specific confounding.
- By construction, the FE remove A_i or S_t from the analysis.

Fixed Effects

- This is why FE models are touted as allowing for the identification of causal effects despite *unobserved* unit or period-specific confounding.
- By construction, the FE remove A_i or S_t from the analysis.
- This is *not* a problem (contrary to what you might hear):
  - This research design presumes that what interests us is ρ.
  - If what interests us are effects of variables in A_i or S_t, then X_{it} and D_{it} are post-treatment!
  - Nothing in the above implies that causal effects of A_i or S_t are identified anyway.
  - To study effects of variables in A_i or S_t you need another identification strategy and another research design.
  - I will mention these points again later.

Fixed Effects

- We have expressed the FE model in terms of units over time. A similar logic applies when we have individuals nested within groups (e.g., people within states).

Fixed Effects

- We have expressed the FE model in terms of units over time. A similar logic applies when we have individuals nested within groups (e.g., people within states).
- If the groups are indexed by i and individuals by t, then group-specific one-way FE model is,

$$Y_{it} = \mu + \alpha_i + \rho D_{it} + X_{it}'\beta + \varepsilon_{it}.$$

Fixed Effects

- We have expressed the FE model in terms of units over time. A similar logic applies when we have individuals nested within groups (e.g., people within states).
- If the groups are indexed by i and individuals by t, then group-specific one-way FE model is,

$$Y_{it} = \mu + \alpha_i + \rho D_{it} + X_{it}'\beta + \varepsilon_{it}.$$

- If individuals are partitioned by strata that cross-cut groups (e.g., occupational strata across states), we can write a two-way FE,

$$Y_{ist} = \mu + \alpha_i + \lambda_s + \rho D_{ist} + X_{ist}'\beta + \varepsilon_{ist},$$

  where s indexes the cross-cutting strata.

Fixed Effects

- We have expressed the FE model in terms of units over time. A similar logic applies when we have individuals nested within groups (e.g., people within states).
- If the groups are indexed by i and individuals by t, then group-specific one-way FE model is,

$$Y_{it} = \mu + \alpha_i + \rho D_{it} + X_{it}'\beta + \varepsilon_{it}.$$

- If individuals are partitioned by strata that cross-cut groups (e.g., occupational strata across states), we can write a two-way FE,

$$Y_{ist} = \mu + \alpha_i + \lambda_s + \rho D_{ist} + X_{ist}'\beta + \varepsilon_{ist},$$

  where s indexes the cross-cutting strata.
- Thus, FE models are ways to characterize arbitrary “unmeasured heterogeneity” across strata.

Mechanics of FE

- FE via OLS using dummy variables is equivalent to other procedures that do not require estimating dummy variable coefficients for each i and t value.

Mechanics of FE

- FE via OLS using dummy variables is equivalent to other procedures that do not require estimating dummy variable coefficients for each i and t value.
- Consider again our unit- and time-specific FE model:

$$\begin{aligned} Y_{it} &= \mu + \alpha_i + \lambda_t + \rho D_{it} + X_{it}'\beta + \varepsilon_{it} \\ &= \mu + \sum_{j=1}^{N} \alpha_j 1(i=j) + \sum_{s=1}^{T} \lambda_s 1(t=s) + \rho D_{it} + X_{it}'\beta + \varepsilon_{it} \end{aligned}$$

Mechanics of FE

- FE via OLS using dummy variables is equivalent to other procedures that do not require estimating dummy variable coefficients for each i and t value.
- Consider again our unit- and time-specific FE model:

$$\begin{aligned} Y_{it} &= \mu + \alpha_i + \lambda_t + \rho D_{it} + X_{it}'\beta + \varepsilon_{it} \\ &= \mu + \sum_{j=1}^{N} \alpha_j 1(i=j) + \sum_{s=1}^{T} \lambda_s 1(t=s) + \rho D_{it} + X_{it}'\beta + \varepsilon_{it} \end{aligned}$$

- Let's look at account just for α_i first.
- By FWL, residualizing with respect to $1(i=j)$ implies subtracting off mean values for unit j and leaving other units untouched. Going through all $j = 1, ..., N$, this yields:

$$(Y_{it} - \bar{Y}_i) = (\lambda_t - \frac{1}{T}) + \rho(D_{it} - \bar{D}_i) + (X_{it} - \bar{X}_i)'\beta + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

Mechanics of FE

- We could thus account for α_i by demeaning the data directly.
- Let $\mathbf{W}_i$ be the matrix containing all of the stacked regressors for unit i (including the constant and FEs) and let θ be the vector of all of the coefficients. Then,

$$Y_i = \mathbf{W}_i \theta + \varepsilon_i.$$

- We can define an idempotent "sweep" matrix for each unit,

$$\mathbf{Q}_T := \mathbf{I}_T - \bar{\mathbf{J}}_T, \text{ where } \bar{\mathbf{J}}_T := \frac{1}{T} \iota_T \iota_T'$$

  where ι_T is a T-vector of ones.
- Pre-multiplication of each unit's data by $\mathbf{Q}_T$ yields deviations from unit means, which in turn "sweeps" away the α_i's.

Mechanics of FE

- We can apply this to the whole dataset at once using,

$$\mathbf{Q} = \mathbf{I}_N \otimes \mathbf{Q}_T = \mathbf{I}_{NT} - (\mathbf{I}_N \otimes \bar{\mathbf{J}}_T) \qquad \text{(also idempotent)}$$

Mechanics of FE

- We can apply this to the whole dataset at once using,

$$\mathbf{Q} = \mathbf{I}_N \otimes \mathbf{Q}_T = \mathbf{I}_{NT} - (\mathbf{I}_N \otimes \bar{\mathbf{J}}_T) \qquad \text{(also idempotent)}$$

- Let $\mathbf{W}^{tv}$ refer to the matrix of regressors excluding the unit FEs and constant, and define θ^{tv} as the vector of coefficients that exclude the same (tv = time-varying).
- Then, by the above, we can obtain the same OLS estimates of the time-dummies, ρ and β using,

$$\begin{pmatrix} \lambda \\ \rho \\ \beta \end{pmatrix} = \left(\mathbf{W}^{tv\prime}\mathbf{Q}\mathbf{W}^{tv}\right)^{-1}\mathbf{W}^{tv\prime}\mathbf{Q}Y. \tag{1}$$

- This is how panel regression functions like Stata's `areg` and `xtreg` and R's `plm` actually carry out one-way FE.

Mechanics of FE

- We can apply this to the whole dataset at once using,

$$\mathbf{Q} = \mathbf{I}_N \otimes \mathbf{Q}_T = \mathbf{I}_{NT} - (\mathbf{I}_N \otimes \bar{\mathbf{J}}_T) \qquad \text{(also idempotent)}$$

- Let $\mathbf{W}^{tv}$ refer to the matrix of regressors excluding the unit FEs and constant, and define θ^{tv} as the vector of coefficients that exclude the same (tv = time-varying).
- Then, by the above, we can obtain the same OLS estimates of the time-dummies, ρ and β using,

$$\begin{pmatrix} \lambda \\ \rho \\ \beta \end{pmatrix} = \left(\mathbf{W}^{tv\prime}\mathbf{Q}\mathbf{W}^{tv}\right)^{-1}\mathbf{W}^{tv\prime}\mathbf{Q}Y. \tag{1}$$

- This is how panel regression functions like Stata's `areg` and `xtreg` and R's `plm` actually carry out one-way FE.
- Algebraically equivalent to the dummy variable regression.
- Calculate standard errors from (1) in usual way (accounting for residual clustering if need be–e.g., for serial dependence).

Mechanics of FE

FE estimation is also called "within" estimation. To see why, consider again the following algorithm for one-way FE estimator (for just α_i):

Mechanics of FE

FE estimation is also called "within" estimation. To see why, consider again the following algorithm for one-way FE estimator (for just α_i):

1. For each of the FE strata (e.g., units), do a stratum-specific regression and get the stratum-specific coefficients.

Mechanics of FE

FE estimation is also called "within" estimation. To see why, consider again the following algorithm for one-way FE estimator (for just α_i):

1. For each of the FE strata (e.g., units), do a stratum-specific regression and get the stratum-specific coefficients.
2. Compute the weighted averages of each those stratum-specific coefficients:
   - Weights for coefficient β_k equals the stratum-specific variances of the associated residualized regressor, $\tilde{X}_{itk}$.

Mechanics of FE

- You already know this!

$$\delta_R = \frac{\sum_x \delta_X \text{Var}\left[D_{it}|X_{it}=x\right]\Pr[X_{it}=x]}{\sum_x \text{Var}\left[D_{it}|X_{it}=x\right]\Pr[X_{it}=x]}$$

where in this case the x's refer to the FE strata and X_{it} is unit i's stratum identifier.

Mechanics of FE

- You already know this!

$$\delta_R = \frac{\sum_x \delta_X \mathrm{Var}\left[D_{it}|X_{it}=x\right]\Pr[X_{it}=x]}{\sum_x \mathrm{Var}\left[D_{it}|X_{it}=x\right]\Pr[X_{it}=x]}$$

  where in this case the x's refer to the FE strata and X_{it} is unit i's stratum identifier.
- This provides a nice way to visualize FE estimation:
  - You do separate regressions in each of the FE strata, and then taking the weighted average of the results.

Mechanics of FE

- You already know this!

$$\delta_R = \frac{\sum_x \delta_X \text{Var}\left[D_{it}|X_{it}=x\right]\Pr[X_{it}=x]}{\sum_x \text{Var}\left[D_{it}|X_{it}=x\right]\Pr[X_{it}=x]}$$

  where in this case the x's refer to the FE strata and X_{it} is unit i's stratum identifier.
- This provides a nice way to visualize FE estimation:
  - You do separate regressions in each of the FE strata, and then taking the weighted average of the results.

Mechanics of FE

- We can also define a sweep transformation to account for both α_i and λ_t, although the math is more complicated and so is the interpretation (Baltagi, 2005, *Ec. An. Panel Data*, pp. 35-6):

$$\mathbf{Q}_{TW} = \mathbf{I}_N \otimes \mathbf{I}_T - \mathbf{I}_N \otimes \bar{\mathbf{J}}_T - \bar{\mathbf{J}}_N \otimes \mathbf{I}_T + \bar{\mathbf{J}}_N \otimes \bar{\mathbf{J}}_T,$$

  where all terms are defined analogously to **Q**.
- Then each element is of the form,

$$\tilde{y}_{it} = y_{it} - \bar{y}_i - \bar{y}_t + \bar{y}$$

- Here, the "within" interpretation is not so clean.
- Also, with respect to causal effects, there are some complications (Imai and Kim, 2012)– we will return to this when we discuss difference-in-differences.

Sources of Confusion

- Aggregation bias as distinct from confounding bias.
- Regressors that do not vary within strata or units and FE.
- Clustering standard errors by FE strata.
- Lags with FE.

Aggregation bias as distinct from confounding bias

Motivation for FE:

- Confounding due to correlation between D_{it} and A_i or between D_{it} and S_t
- Aspects of A_i or S_t that generate the confounding are unmeasured.

Aggregation bias as distinct from confounding bias

The FE estimator computes,

$$\delta_R = \frac{\sum_x \delta_X \mathrm{Var}\left[D_{it}|X_{it}=x\right]\Pr[X_{it}=x]}{\sum_x \mathrm{Var}\left[D_{it}|X_{it}=x\right]\Pr[X_{it}=x]}$$

Even though we have accounted for the confounding, this estimator is still biased (and inconsistent) if what we really want is

$$\rho = \sum_x \delta_X \Pr[X_{it}=x].$$

The nature of this bias is "aggregation bias."

Aggregation bias as distinct from confounding bias

To recover ρ, we can either

- Compute stratified estimator directly (sample analogue of ρ),
- Weight the FE regression by $1/\text{Var}\left[D_i|X_i = x\right]$, or
- Compute the centered-interaction FE model (cf. Imbens & Wooldridge 2009, p. 28).

See R simulation...

Regressors that are constant within strata

- If a regressor is constant within an FE stratum, then it is perfectly collinear with that FE stratum dummy.
  - E.g., a time-invariant regressor in the panel/TSCS context.
- When you fit FE, these within-stratum-invariant (or time-invariant) regressors must be dropped.
- (Recall that with multi-way FE, what matters is whether the "swept" variables are time-invariant or not.)

Regressors that are constant within strata

- This has led some to conclude that FE "throws the baby out with the bath water" (cf. Green et al. vs. Beck & Katz), and that other approaches (e.g., RE or OLS with adequate controls) are "better."

Regressors that are constant within strata

- This has led some to conclude that FE "throws the baby out with the bath water" (cf. Green et al. vs. Beck & Katz), and that other approaches (e.g., RE or OLS with adequate controls) are "better."
- For *causal inference* on D, we don't care about the baby:
  - The point of the regression is to estimate the effect of D_{it}.
  - If FE addresses confounding due to within-stratum- or time-invariant X_i without having to estimate a coefficient for X_i, then that's great!
  - If the *treatment* of interest does not vary over t, then obviously FE is irrelevant altogether!

Regressors that are constant within strata

- This has led some to conclude that FE "throws the baby out with the bath water" (cf. Green et al. vs. Beck & Katz), and that other approaches (e.g., RE or OLS with adequate controls) are "better."
- For *causal inference* on D, we don't care about the baby:
  - The point of the regression is to estimate the effect of D_{it}.
  - If FE addresses confounding due to within-stratum- or time-invariant X_i without having to estimate a coefficient for X_i, then that's great!
  - If the *treatment* of interest does not vary over t, then obviously FE is irrelevant altogether!
- Such arguments are relevant when we are trying to create a *predictive model* that accounts for variation in *both* within-stratum- or time-invariant factors *and* within-stratum- or time-varying factors.

Clustering standard errors by FE strata

- Recall that we cluster to account for dependencies in the *treatment*.
- If treatments are assigned randomly within FE strata (even if treatment probabilities/distributions differ from stratum-to-stratum), no need to cluster by strata.
- If treatment assignment within strata exhibits serial dependence, or "contagion"-based dependence (whether positive or negative), then you want to cluster on the stratum indicators.
- Clustering in multiple directions can be handled by multi-way cluster robust (Cameron et al. 2011); for dyadic data, see Aronow et al. (2015).
- NB: `reghdfe` command in Stata uses the correct degrees-of-freedom adjustment when FE strata and clusters coincide (see http://scorreia.com/software/reghdfe/). Usual `areg`, `xtreg`, and R commands are overconservative.

Lag specifications

We may want to account for either (i) effects of D_{it} into future periods or (ii) possibility that D_{it} is endogenous to past Y_{it} or X_{it}, which also affect current Y_{it}.

Lag specifications

We may want to account for either (i) effects of D_{it} into future periods or (ii) possibility that D_{it} is endogenous to past Y_{it} or X_{it}, which also affect current Y_{it}.

- Consider one-period autoregressive distributed lag (ADL) model:

  $$Y_{it} = \mu + \alpha_i + \lambda_t + \pi Y_{i,t-1} + \rho D_{it} + \rho_{-1} D_{i,t-1} + X_{it}'\beta + X_{i,t-1}'\beta_{-1} + \varepsilon_{it},$$

  where ε_{it} is exogenous to D_{it} and $D_{i,t-1}$ conditional on the other regressors. (Deeper lags are conceivable of course.)

Lag specifications

We may want to account for either (i) effects of D_{it} into future periods or (ii) possibility that D_{it} is endogenous to past Y_{it} or X_{it}, which also affect current Y_{it}.

- Consider one-period autoregressive distributed lag (ADL) model:

  $$Y_{it} = \mu + \alpha_i + \lambda_t + \pi Y_{i,t-1} + \rho D_{it} + \rho_{-1} D_{i,t-1} + X_{it}'\beta + X_{i,t-1}'\beta_{-1} + \varepsilon_{it},$$

  where ε_{it} is exogenous to D_{it} and $D_{i,t-1}$ conditional on the other regressors. (Deeper lags are conceivable of course.)
- With small T, FE methods above result in biased $\hat{\pi}$, which can propagate to other estimates. This "Nickell bias" arises because $\varepsilon_{it} - \bar{\varepsilon}_i$ contains $\varepsilon_{i,t-1}$, which is part of $Y_{i,t-1}$. Disappears as T gets large. cf. MHE for strategies when T is small.

Lag specifications

We may want to account for either (i) effects of D_{it} into future periods or (ii) possibility that D_{it} is endogenous to past Y_{it} or X_{it}, which also affect current Y_{it}.

- Consider one-period autoregressive distributed lag (ADL) model:

$$Y_{it} = \mu + \alpha_i + \lambda_t + \pi Y_{i,t-1} + \rho D_{it} + \rho_{-1} D_{i,t-1} + X'_{it}\beta + X'_{i,t-1}\beta_{-1} + \varepsilon_{it},$$

  where ε_{it} is exogenous to D_{it} and $D_{i,t-1}$ conditional on the other regressors. (Deeper lags are conceivable of course.)

- With small T, FE methods above result in biased $\hat{\pi}$, which can propagate to other estimates. This “Nickell bias” arises because $\varepsilon_{it} - \bar{\varepsilon}_i$ contains $\varepsilon_{i,t-1}$, which is part of $Y_{i,t-1}$. Disappears as T gets large. cf. MHE for strategies when T is small.
- If ε_{it} contains serial correlation despite the inclusion of $Y_{i,t-1}$, then we again have bias on π, and there’s basically nothing that you can do about it.

Lag specifications

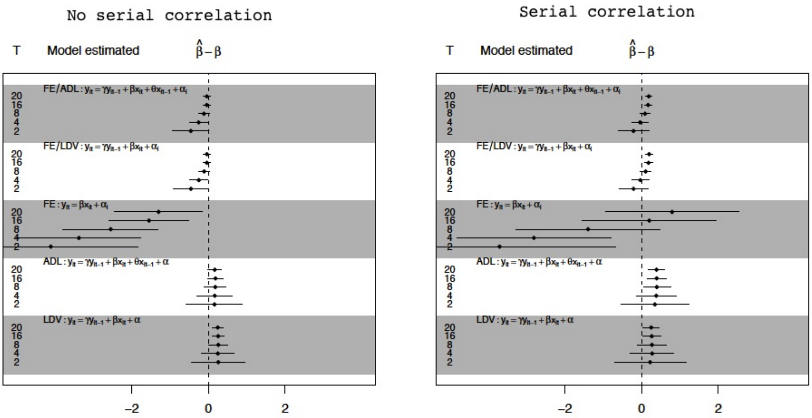
We may want to account for either (i) effects of D_{it} into future periods or (ii) possibility that D_{it} is endogenous to past Y_{it} or X_{it}, which also affect current Y_{it}.

- Consider one-period autoregressive distributed lag (ADL) model:

$$Y_{it} = \mu + \alpha_i + \lambda_t + \pi Y_{i,t-1} + \rho D_{it} + \rho_{-1} D_{i,t-1} + X_{it}' \beta + X_{i,t-1}' \beta_{-1} + \varepsilon_{it},$$

  where ε_{it} is exogenous to D_{it} and $D_{i,t-1}$ conditional on the other regressors. (Deeper lags are conceivable of course.)
- With small T, FE methods above result in biased $\hat{\pi}$, which can propagate to other estimates. This "Nickell bias" arises because $\varepsilon_{it} - \bar{\varepsilon}_i$ contains $\varepsilon_{i,t-1}$, which is part of $Y_{i,t-1}$. Disappears as T gets large. cf. MHE for strategies when T is small.
- If ε_{it} contains serial correlation despite the inclusion of $Y_{i,t-1}$, then we again have bias on π, and there's basically nothing that you can do about it.

Lag specifications



- In these sims, both unit FEs and $Y_{i,t-1}$ needed for identification.
- Shows decay in Nickell bias and irremovable bias due to LDV and serial correlation.

Lag specifications

- Assuming the model is *correct and identified*, ADL lends itself to dynamic interpretations (cf. DeBoef & Keele, 2008).

$$Y_{it} = \mu + \alpha_i + \lambda_t + \pi Y_{i,t-1} + \rho D_{it} + \rho_{-1} D_{i,t-1} + X'_{it}\beta + X'_{i,t-1}\beta_{-1} + \varepsilon_{it}$$

- ρ represents the *immediate* effect of D_{it} on Y_{it}.
- Effect of change in D_{it} after one period is,

$$\frac{\partial Y_{i,t+1}}{\partial D_{it}} = \pi \frac{\partial Y_{i,t}}{\partial D_{it}} + \rho_{-1} = \pi\rho + \rho_{-1}$$

- After two periods, $\frac{\partial Y_{i,t+2}}{\partial D_{it}} = \pi^2\rho + \pi\rho_{-1}$.
- Assuming $|\pi| < 1$, after ≈infinite periods, the long-run effect of a treatment change in period t on future outcomes is $\frac{\rho+\rho_{-1}}{1-\pi}$.

Remarks

- Huge literature on panel, TSCS, and other FE models.
- A lot more than one could do using unit-specific time trends, first differences, forward deviations, error correction specifications, dynamic panel models and panel instruments, and so on (cf. MHE for some nice applied examples).
- Full gamut of time series techniques could also be brought to bear here.
- Efficiency gains are possible by using multilevel models or other types models that "borrow strength" across strata (covered in Quant III).

Remarks

- That being said, unleashing a larger arsenal does not necessarily result in more credible, much less interpretable, estimates.
- The models here sometimes obscure issues such as post-treatment biases and effect heterogeneity that may lead to misguided inference.
- Beware of “mechanical identification”...

Remarks

- That being said, unleashing a larger arsenal does not necessarily result in more credible, much less interpretable, estimates.
- The models here sometimes obscure issues such as post-treatment biases and effect heterogeneity that may lead to misguided inference.
- Beware of “mechanical identification”...

3. Any pet peeves with submissions or with referees that it would be good for people to avoid?

Unfortunately yes. Our main two criteria in selecting papers for publication are rigorous identification and policy relevance. The two go together as we cannot have credible policy recommendations without strong causal inference. Too many of the submitted papers offer simple "determinants" that are partial correlates with no causal value, and yet are the basis for bold policy recommendations, sometimes of first order of importance for development practice. This includes a large number of cross-country panel regressions with only mechanical, and hence not credible, identification, and yet eventually huge claims of policy implications. Regarding policy relevance, papers too often address issues of nth order of