Lecture 11: Repeated Observations I

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

March 25, 2019

Repeated observations and causal effects

"Repeated observations" (panel data, time-series cross section data, spatially clustered data, etc.) allow new opportunities:

Repeated observations and causal effects

"Repeated observations" (panel data, time-series cross section data, spatially clustered data, etc.) allow new opportunities:

- Dynamics effects
  - Growth curves and trajectories
  - Dynamic treatment regimes
- Identification strategies
  - Fixed effects
  - Difference in differences

Fixed effects

- Start with a simple one-way FE cases.
- Units $i = 1, ..., N$ from large population.
- For each i observe (Y_{it}, D_{it}) for time periods: $t = 1, ..., T$.
- Outcome DGP:

$$Y_{it} = \mu + \rho D_{it} + U_i + \epsilon_{it}.$$

- Assuming homogenous effect, ρ.
- Suppose $\epsilon_{it} \sim$ [white noise] but $D_{it} = \gamma_0 + \gamma_1 U_i + \nu_{it}$, ν_{it} also white noise.
- OLS of Y on D has expectation $\rho + \gamma_1(\text{Var}[U]/\text{Var}[D])$.
- Omitted variable bias.

Fixed effects

- Try something else—take one i at a time $\Rightarrow$ mini datasets:

$$
\begin{aligned}
Y_{i1} &= \mu + \rho D_{i1} + U_i + \epsilon_{i1} \\
Y_{i2} &= \mu + \rho D_{i2} + U_i + \epsilon_{i2} \\
&\vdots \\
Y_{iT} &= \mu + \rho D_{iT} + U_i + \epsilon_{iT}
\end{aligned}
$$

  and OLS Y on D (for cases where $\text{Var}\left[D_{it}|i\right] > 0$).
- $(\mu + U)$ is constant, ϵ is random noise.
- Yields coef ρ_i with expected value ρ (homog. effects).
- Aggregate over all i and you get a more precise estimate of ρ.
- Maximal precision: weight proportional to $\text{Var}\left[D_{it}|i\right]$.
- "Within" regression.

Fixed effects

- Another idea, center on unit (i-specific) means:

$$Y_{it} - \bar{Y}_i = (\mu - \mu) + \rho(D_{it} - \bar{D}_i) + (U_i - U_i) + (\epsilon_{i1} - \bar{\epsilon}_i)$$
$$Y_{it} - \bar{Y}_i = \rho(D_{it} - \bar{D}_i) + (\epsilon_{it} - \bar{\epsilon}_i)$$
$$Y^*_{it} = \rho D^*_{it} + \epsilon^*_{it}$$

- OLS of Y^* on D^* has expected value ρ.
- "Sweep" transformation.

Fixed effects

- Now note the following:
- First, from lecture 3, recall using OLS to fit

$$Y_{it} = \mu + \rho_i D_i + \sum_{j=1}^{N} I(i = j) + \epsilon_{it}$$

yields coefficient on D that converges to

$$\rho_R = \frac{\sum_i \rho_i \text{Var}\left[D_{it}|i\right] \Pr[i]}{\sum_i \text{Var}\left[D_{it}|i\right] \Pr[i]}.$$

- Same as "within" regression above.
- Second, apply FWL to this dummy variable regression:
- Residualizing wrt $1(i = j)$ is subtracting off mean values for unit j, leave other units untouched.
- Same as "sweep" transformation above.

Fixed effects

- Thus, the following are algebraically equivalent:
  - Dummy variable OLS with $1(i = j)$,
  - Variance weighted average of coefs. from "within" OLS by i,
  - OLS after "sweep" transformation with i-specific means.
- This is "one-way fixed effects" regression.
- Addresses "time-invariant" confounders for DGPs like

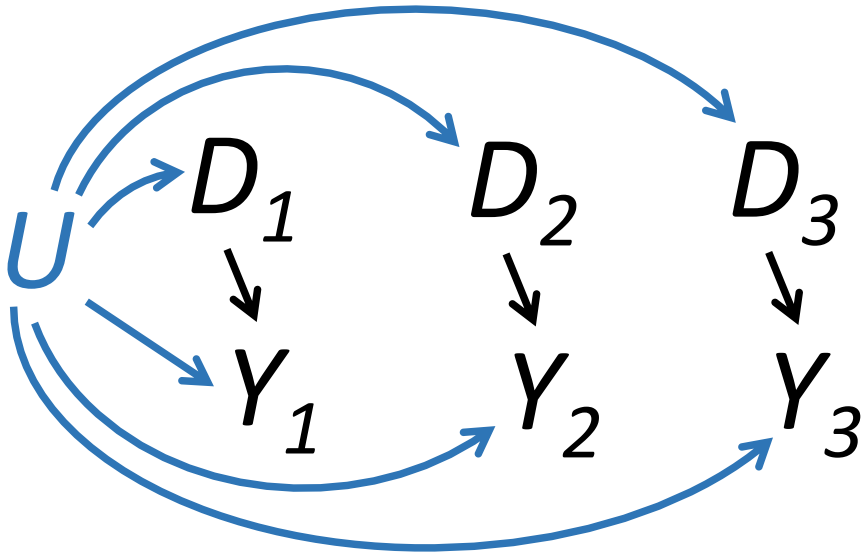
$$Y_{it} = \mu + \rho D_{it} + X_{it}\beta + U_i + \epsilon_{it}$$
$$D_{it} = \gamma_0 + \gamma_1 U_i + X_{it}\lambda + \nu_{it}.$$

(X_{it}, assumed measured, added for more generality).

Issues with fixed effects estimation

- DGPs for which FE identifies and those for which it doesn't (DAGs).
- Interpretation of effects (potential outcomes).
- Estimation and inference.

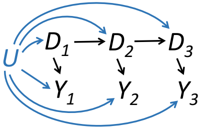
DGPs and FE identification

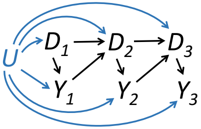


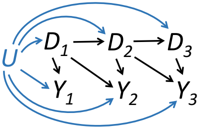
FE on a DAG:

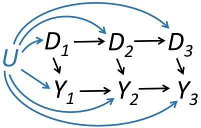
- Partial out U via sweep – implies "partial" conditioning on lags and leads.
- Collapse time periods.

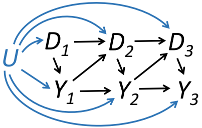
DGPs and FE identification











Interpretation of effects

Classic FE model in potential outcomes:

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

Interpretation of effects

Classic FE model in potential outcomes:

- Sps. $D_{it} = 0, 1$ 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.

Interpretation of effects

Classic FE model in potential outcomes:

- Sps. $D_{it} = 0, 1$ 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.

Interpretation of effects

Classic FE model in potential outcomes:

- Sps. $D_{it} = 0, 1$ 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
  $Y_{it} = D_{it} Y_{1it} + (1 - D_{it}) Y_{0it} = Y_{0it} + D_{it}(Y_{1it} - Y_{0it})$.
- U_i is vector of "time-invariant" attributes of i.

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

Interpretation of 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:

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

  This satisfied under CIA conditional on U_i and X_{it}.

Interpretation of 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:

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

  This satisfied under CIA conditional on U_i and X_{it}.

- Assumption 2: Linearity:

$$\mathrm{E}\left[Y_{0it}|U_i, X_{it}\right] = \mu + U_i + X_{it}'\beta.$$

- Assumption 3: Constant additive effects:

$$\mathrm{E}\left[Y_{1it}|U_i, X_{it}\right] = \mathrm{E}\left[Y_{0it}|U_i, X_{it}\right] + \rho.$$

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

Interpretation of effects

- What if effects are heterogenous—e.g., each unit has its own average per-period effect, ρ_i?
- Recall:

$$\rho_R = \frac{\sum_i \rho_i \text{Var}\left[D_{it}|i\right] \Pr[i]}{\sum_i \text{Var}\left[D_{it}|i\right] \Pr[i]}.$$

- Already discussed in lecture 3: ρ_R is not the ATT, ATE, or ATC because of weighting by $\text{Var}\left[D_{it}|i\right]$ rather than just $Pr[i]$.
- Solutions:
  - Stratified (matching) estimator.
  - Centered-interaction models (Imbens & Wooldridge, 2009, p. 28 – see simulation).

Estimation and inference

- Can account for U_i through sweep transformation.
- 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,

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

- 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.

Estimation and inference

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

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

Estimation and inference

- 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,

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

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

Estimation and inference

- 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,

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

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

Sources of Confusion

- Regressors that do not vary within strata or units and FE.
- Clustering standard errors by FE strata.
- Lags with FE.

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.
- 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:

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

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:

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

  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 $\epsilon_{it} - \bar{\epsilon}_i$ contains $\epsilon_{i,t-1}$, which is part of $Y_{i,t-1}$. Disappears as T gets large. cf. MHE for strategies when T is small.

Lag specifications

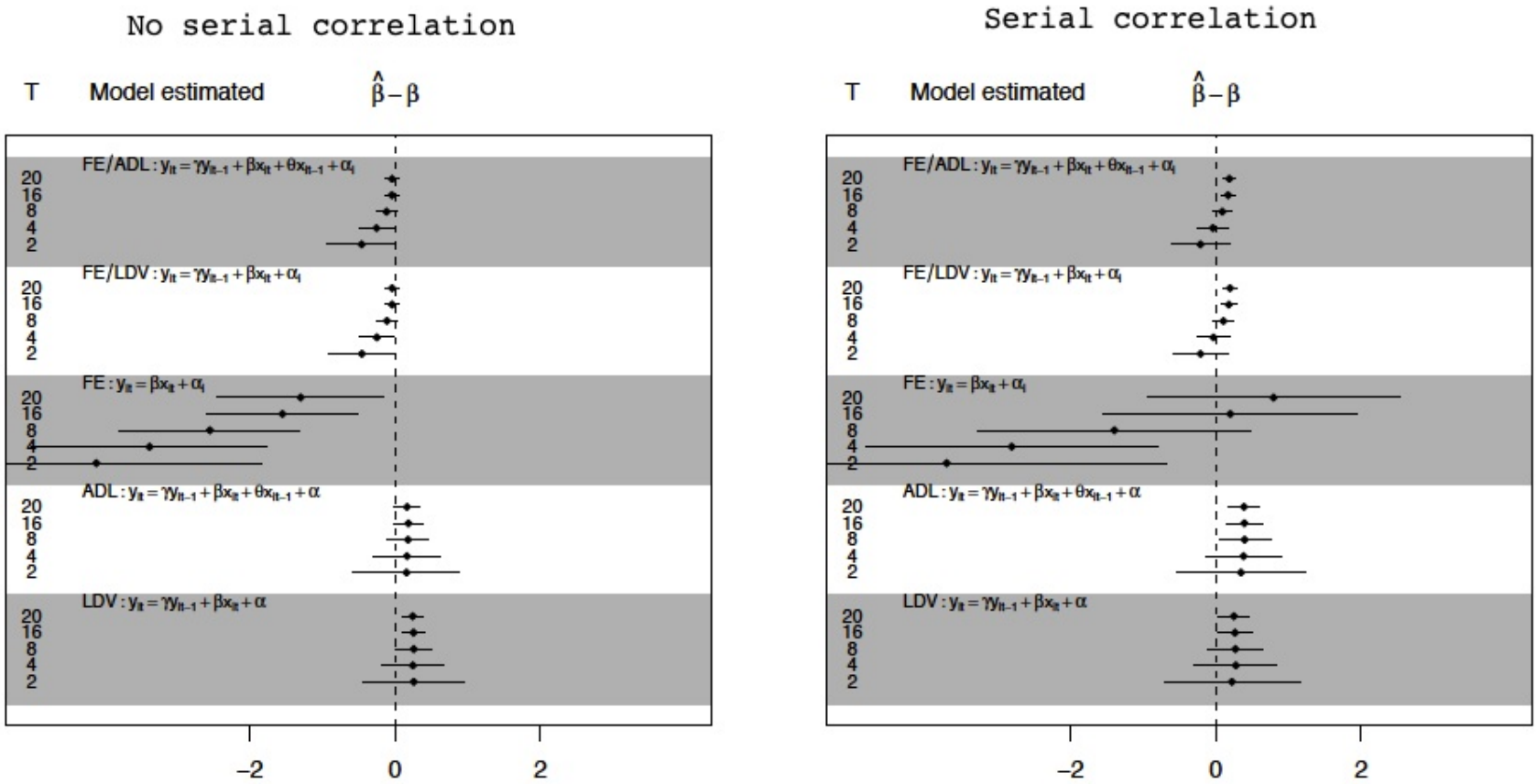
We may want to account for either (i) effects of D_{it} into future periods or (ii) possibility that D_{it} is endogenous to past Y_{it} or X_{it}, which also affect current Y_{it}.

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

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

  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 $\epsilon_{it} - \bar{\epsilon}_i$ contains $\epsilon_{i,t-1}$, 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).

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

- ρ represents the *immediate* effect of D_{it} on Y_{it}. Then,

$$\begin{aligned}
\frac{\partial Y_{i,t}}{\partial D_{it}} &= \rho \\
\frac{\partial Y_{i,t+1}}{\partial D_{it}} &= \pi \frac{\partial Y_{i,t}}{\partial D_{it}} + \rho_{-1} = \pi\rho + \rho_{-1} \\
\frac{\partial Y_{i,t+2}}{\partial D_{it}} &= \pi^2\rho + \pi\rho_{-1} \\
\frac{\partial Y_{i,t+k}}{\partial D_{it}} &= \pi^k\rho + \pi^{k-1}\rho_{-1}
\end{aligned}$$

- With $|\pi| < 1$, as $k \to \infty$, accumulated effect of ∂D_{it} 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).