Advanced Dynamic Panel Data Methods

Perils of Unobserved Heterogeneity; Scope and Limitations of Panel Data Analysis

Sebastian Kripfganz

University of Exeter Business School, Department of Economics, Exeter, UK
www.kripfganz.de @Kripfganz

Universidad de Salamanca

July 17–19, 2023

Linear Panel Data Model

- Linear-in-coefficients model for the conditional mean:

$$E[y_{it}|\mathbf{X}_i] = \mathbf{x}_{it}'\beta$$

  - The outcome variable is y_{it}.
  - The K regressors x_{itk} are collected in the $K_x \times 1$ vector $\mathbf{x}_{it} = (x_{it1}, x_{it2}, \dots, x_{itK})'$.
  - $\mathbf{X}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT})'$ collects all observations for unit i in a $T \times K$ matrix. We assume $\text{rk}(E[\mathbf{X}_i'\mathbf{X}_i]) = K_x$ (no perfect multicollinearity).
  - $\beta = (\beta_1, \beta_2, \dots, \beta_K)'$ is a $K \times 1$ vector of (homogeneous) slope coefficients.
  - An intercept can (and generally should) be included in the model by setting $x_{it1} = 1$ for all i, t.

Linear Panel Data Model

- Object of interest:

$$\frac{\partial E[y_{it}|\mathbf{X}_i]}{\partial \mathbf{x}_{it}} = \beta$$

(or a single coefficient β_k, or a linear combination of coefficients).

- Linear (in coefficients) regression model:

$$y_{it} = \mathbf{x}'_{it}\beta + \varepsilon_{it}$$

  - ε_{it} is an error term that satisfies $E[\varepsilon_{it}|\mathbf{X}_i] = 0$.
- If $E[\varepsilon_{it}|\mathbf{X}_i] = 0$ holds – i.e., the model is correctly specified – the pooled ordinary least squares (OLS) estimator $\hat{\beta}_{POLS}$ is unbiased and consistent.

Heteroskedasticity

- (Conditional) homoskedasticity:

$$Var(y_{it}|\mathbf{X}_i) = Var(\varepsilon_{it}|\mathbf{X}_i) = \sigma_{\varepsilon}^2$$

- (Conditional) heteroskedasticity:

$$Var(y_{it}|\mathbf{X}_i) = Var(\varepsilon_{it}|\mathbf{X}_i) = \sigma_{\varepsilon,it}^2$$

  - Time-series heteroskedasticity: $\sigma_{\varepsilon,it}^2 = \sigma_{\varepsilon,t}^2$
  - Cross-sectional heteroskedasticity: $\sigma_{\varepsilon,it}^2 = \sigma_{\varepsilon,i}^2$
- Under homoskedasticity (and correct model specification), $\hat{\beta}_{POLS}$ is efficient. Under heteroskedasticity, it is no longer efficient (but remains unbiased/consistent). We usually do not explicitly model the conditional heteroskedasticity, but just compute "robust" standard errors.

Coefficient Heterogeneity

- Unit-specific intercepts:

$$E[y_{it}|\mathbf{X}_i; \alpha_i] = \mathbf{x}_{it}'\beta + \alpha_i$$

  - We will pay special attention to this case.

- Time-specific intercepts:

$$E[y_{it}|\mathbf{X}_i; \delta_t] = \mathbf{x}_{it}'\beta + \delta_t$$

  - We usually deal with such "time-fixed effects" by including $T-1$ dummy variables $d_s = \mathcal{I}(s=t)$, $s = 2, 3, \ldots, T$, as part of the regressors $\mathbf{x}_{it}$.
  - Beware the "dummy trap"; we cannot include a whole set of T time dummies if a common intercept $x_{it1} = 1$ is included because of perfect multicollinearity ($\sum_{s=1}^{T} d_s = 1$).

Coefficient Heterogeneity

- Unit-specific slope coefficients:

$$E[y_{it}|\mathbf{X}_i; \beta_i] = \mathbf{x}'_{it}\beta_i$$

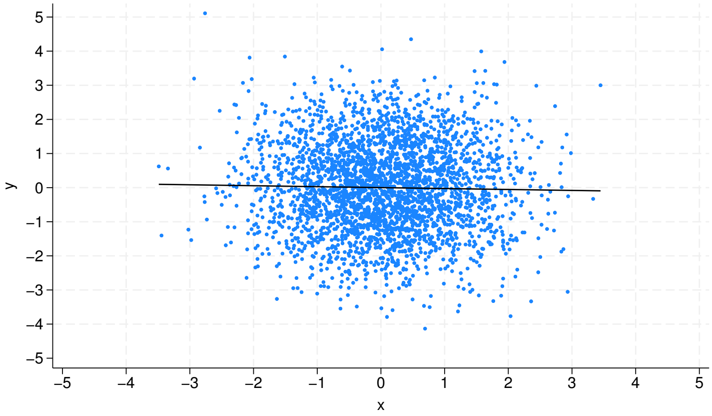
  - In the simplest form, such heterogeneity can be modeled with interaction terms if $\beta_i = f(\mathbf{x}_{it})$, which includes polynomials in $\mathbf{x}_{it}$. The resulting model can be written as a special case of the model with homogeneous slopes.
  - More generally, models with heterogeneous slopes fall into the class of "mixed-effects" models. We do not explicitly cover them here.
- Time-specific slope coefficients:

$$E[y_{it}|\mathbf{X}_i; \beta_t] = \mathbf{x}'_{it}\beta_t$$

  - Interaction terms between $\mathbf{x}_{it}$ and time dummies d_s are the easiest way to account for this type of heterogeneity.

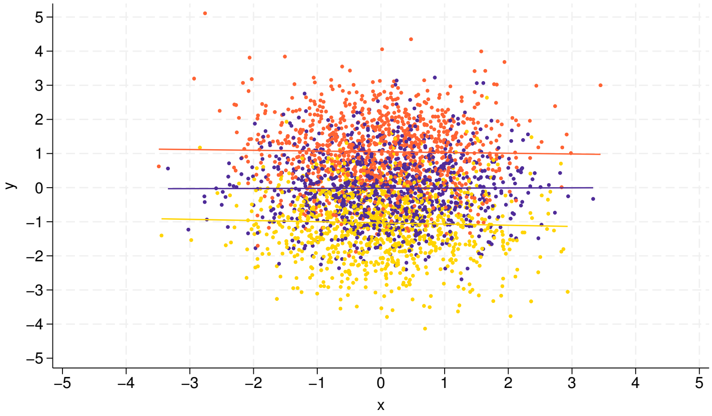
Heterogeneity (Unit-Specific Intercept)

- Simulated data, where $E[y_{it}|x_{it}] = E[y_{it}] = 0$ (i.e., $\beta = 0$):
  - $y_{it} = \alpha_i + \varepsilon_{it}$, where $\varepsilon_{it} \sim \mathcal{N}(0,1)$ and $\alpha_i \in \{-1, 0, 1\}$
  - $x_{it} = \kappa\alpha_i + \nu_{it}$, where $\nu_{it} \sim \mathcal{N}(0,1)$ and $\kappa = 0$



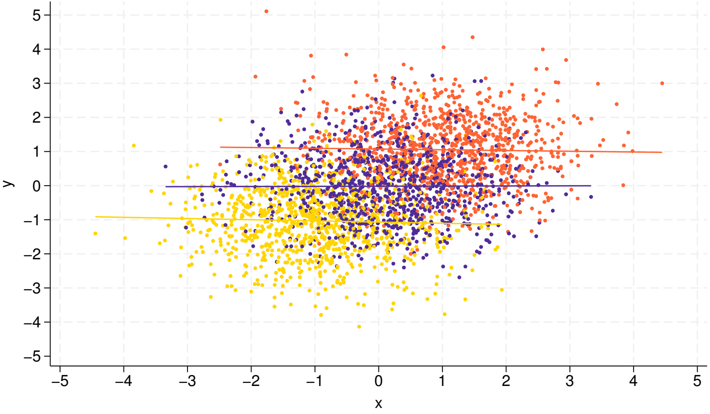
Heterogeneity (Unit-Specific Intercept)

- Simulated data, where $E[y_{it}|x_{it}] = E[y_{it}] = 0$ (i.e., $\beta = 0$):
  - $y_{it} = \alpha_i + \varepsilon_{it}$, where $\varepsilon_{it} \sim \mathcal{N}(0,1)$ and $\alpha_i \in \{-1, 0, 1\}$
  - $x_{it} = \kappa\alpha_i + \nu_{it}$, where $\nu_{it} \sim \mathcal{N}(0,1)$ and $\kappa = 0$



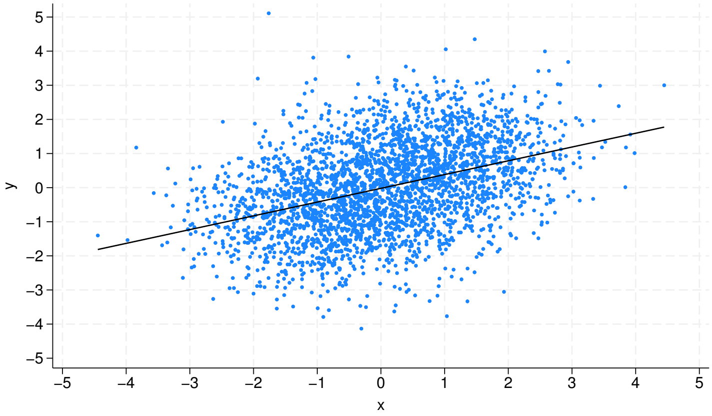
Heterogeneity (Unit-Specific Intercept)

- Simulated data, where $E[y_{it}|x_{it}] = E[y_{it}] = 0$ (i.e., $\beta = 0$):
  - $y_{it} = \alpha_i + \varepsilon_{it}$, where $\varepsilon_{it} \sim \mathcal{N}(0,1)$ and $\alpha_i \in \{-1, 0, 1\}$
  - $x_{it} = \kappa\alpha_i + \nu_{it}$, where $\nu_{it} \sim \mathcal{N}(0,1)$ and $\kappa = 1$



Heterogeneity (Unit-Specific Intercept)

- Simulated data, where $E[y_{it}|x_{it}] = E[y_{it}] = 0$ (i.e., $\beta = 0$):
  - $y_{it} = \alpha_i + \varepsilon_{it}$, where $\varepsilon_{it} \sim \mathcal{N}(0,1)$ and $\alpha_i \in \{-1, 0, 1\}$
  - $x_{it} = \kappa\alpha_i + \nu_{it}$, where $\nu_{it} \sim \mathcal{N}(0,1)$ and $\kappa = 1$



Heterogeneity (Unit-Specific Intercept)

- If the regressors $\mathbf{x}_{it}$ are uncorrelated with the unit-specific intercept α_i, $\hat{\beta}_{POLS}$ is still unbiased/consistent for

$$\frac{\partial E[y_{it}|\mathbf{X}_i]}{\partial \mathbf{x}_{it}} = \frac{\partial E[y_{it}|\mathbf{X}_i; \alpha_i]}{\partial \mathbf{x}_{it}} = \beta$$

  but no longer efficient.

- The “random-effects” (RE) estimator $\hat{\beta}_{RE}$ is efficient (under homoskedasticity) as it accounts for the serial correlation in the error term, which is due to the time-invariant nature of the unit-specific error component α_i:

$$Cov(\alpha_i + \varepsilon_{it}, \alpha_i + \varepsilon_{is}) = Var(\alpha_i)$$

  for $s \neq t$, under the usual assumption that the idiosyncratic error component ε_{it} is serially uncorrelated – i.e, $Cov(\varepsilon_{it}, \varepsilon_{is}) = 0$ for $s \neq t$.

Heterogeneity (Unit-Specific Intercept)

- If the regressors $\mathbf{x}_{it}$ are correlated with the unit-specific intercept α_i, failure to account for the latter leads to biased/inconsistent estimation of the coefficients $\boldsymbol{\beta}$.
  - In the linear regression model

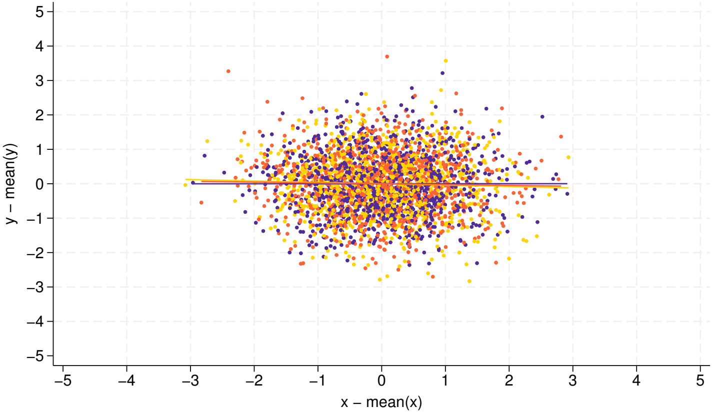
$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i + \varepsilon_{it}$$

with combined error term $\alpha_i + \varepsilon_{it}$, the unit-specific error component α_i acts as omitted variable because $E[\alpha_i + \varepsilon_{it}|\mathbf{X}_i] = E[\alpha_i|\mathbf{X}_i] \neq 0$, and

$$\frac{\partial E[y_{it}|\mathbf{X}_i]}{\partial \mathbf{x}_{it}} = \underbrace{\frac{\partial E[y_{it}|\mathbf{X}_i; \alpha_i]}{\partial \mathbf{x}_{it}}}_{=\boldsymbol{\beta}} + \underbrace{\frac{\partial E[\alpha_i|\mathbf{X}_i]}{\partial \mathbf{x}_{it}}}_{\neq \mathbf{0}}$$

Heterogeneity (Unit-Specific Intercept)

- Re-centering of y_{it} and x_{it}:
  - Deviations from unit-specific means, $y_{it} - \bar{y}_i$ and $x_{it} - \bar{x}_i$, where $\bar{y}_i = \frac{1}{T}\sum_{s=1}^{T} y_{is}$ and $\bar{x}_i = \frac{1}{T}\sum_{s=1}^{T} x_{is}$



Heterogeneity (Unit-Specific Intercept)

- The "fixed-effects" (FE) estimator $\hat{\beta}_{FE}$ is based on the de-meaned regression

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\beta + \varepsilon_{it} - \bar{\varepsilon}_i$$

  where the unit-specific error component α_i is dropped due to its time invariance, and thus no longer causes an omitted-variables bias.
    - This requires strict exogeneity of the regressors $\mathbf{x}_{it}$ – i.e., $E[\varepsilon_{it}|\mathbf{X}_i] = E[\varepsilon_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT}] = 0$ (as opposed to $E[\varepsilon_{it}|\mathbf{x}_{it}] = 0$ under contemporaneous exogeneity).
- Alternatively, the first-difference (FD) estimator $\hat{\beta}_{FD}$ achieves the same goal by using the transformed regression

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\beta + \varepsilon_{it} - \varepsilon_{i,t-1}$$

Dynamic Models (Lagged Dependent Variable)

- Linear model with state dependence (lagged dependent variable):

$$E[y_{it}|y_{i,t-1}, \mathbf{X}_i; \alpha_i] = \lambda y_{i,t-1} + \mathbf{x}'_{it}\beta + \alpha_i$$

with regression analogue

$$y_{it} = \lambda y_{i,t-1} + \mathbf{x}'_{it}\beta + \alpha_i + \varepsilon_{it}$$

  - Notice that the model is only well defined for $t = 2, 3, \dots, T$ because y_{i0} is unobserved. This reduces the effective number of observations by 1 for every unit.
  - A lagged dependent variable can be motivated by habit formation and other reasons for partial-adjustment processes.

Dynamic Models (Lagged Dependent Variable)

- Reparameterization of the regression model in error correction form:

$$y_{it} - y_{i,t-1} = \underbrace{(\lambda - 1)}_{\substack{\text{(negative)}\\ \text{speed of}\\ \text{adjustment}}} \underbrace{\left(y_{i,t-1} - \mathbf{x}_{i,t-1}' \frac{\beta}{1-\lambda} - \frac{\alpha_i}{1-\lambda} \right)}_{\substack{\text{deviation from long-run equilibrium}\\ E[y_{it}|\mathbf{X}_i;\alpha_i] = \mathbf{x}_{it}' \frac{\beta}{1-\lambda} + \frac{\alpha_i}{1-\lambda}}} + (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\beta + \varepsilon_{it}$$

- Objects of interest:
  - Short-run effects:

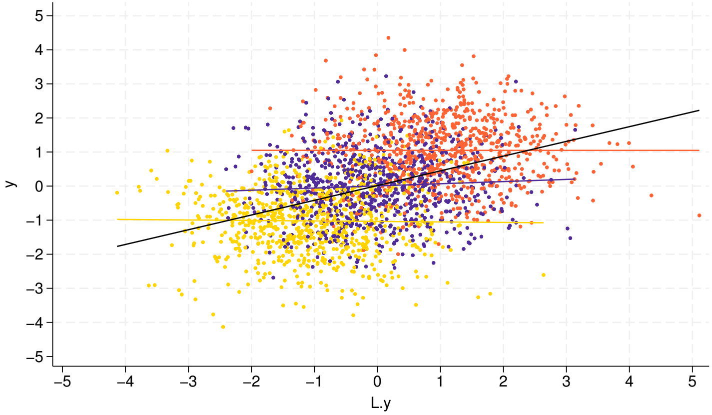
$$\frac{\partial E[y_{it}|y_{i,t-1}, \mathbf{X}_i; \alpha_i]}{\partial \mathbf{x}_{it}} = \beta$$

  - Long-run effects:

$$\frac{\partial E[y_{it}|\mathbf{X}_i; \alpha_i]}{\partial \mathbf{x}_{it}} = \frac{\beta}{1-\lambda}$$

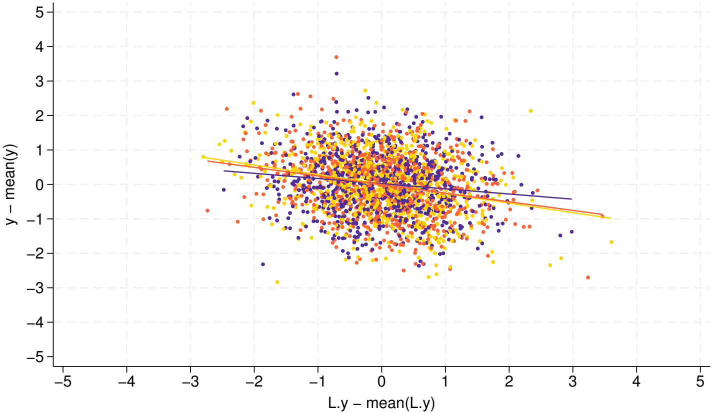
Dynamic Models (Lagged Dependent Variable)

- Simulated data, where $E[y_{it}|y_{i,t-1}] = E[y_{it}] = 0$ (i.e., $\lambda = 0$):
  - $y_{it} = \alpha_i + \varepsilon_{it}$, where $\varepsilon_{it} \sim \mathcal{N}(0,1)$ and $\alpha_i \in \{-1, 0, 1\}$
  - $N = 600$, $T = 5$



Dynamic Models (Lagged Dependent Variable)

- Re-centering of y_{it} and $y_{i,t-1}$:
  - Deviations from unit-specific means, $y_{it} - \bar{y}_i$ and $y_{i,t-1} - \bar{y}_{i,-1}$, where $\bar{y}_i = \frac{1}{T-1}\sum_{s=2}^{T} y_{is}$ and $\bar{y}_{i,-1} = \frac{1}{T-1}\sum_{s=2}^{T} y_{i,s-1}$



Dynamic Models (Lagged Dependent Variable)

- The lagged dependent variable $y_{i,t-1}$ is correlated with the unit-specific error component α_i by construction of the model:

$$y_{it} = \lambda y_{i,t-1} + \mathbf{x}_{it}'\beta + \alpha_i + \varepsilon_{it}$$

  - $y_{i,t-1}$ is not strictly exogenous, but only sequentially exogenous/weakly exogenous/predetermined – i.e., $E[\varepsilon_{it}|y_{i1}, y_{i2}, \dots, y_{i,t-1}, \mathbf{X}_i; \alpha_i] = 0$

Dynamic Models (Lagged Dependent Variable)

- The FE estimator does not successfully eliminate the bias/inconsistency (Nickell, 1981) because $y_{i,t-1} - \bar{y}_{i,-1}$ is correlated with $\varepsilon_{it} - \bar{\varepsilon}_i$ (when T is small) in the regression
$$y_{it} - \bar{y}_i = \lambda(y_{i,t-1} - \bar{y}_{i,-1}) + (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\beta + \varepsilon_{it} - \bar{\varepsilon}_i$$
- The first-difference estimator is biased/inconsistent as well because $y_{i,t-1} - y_{i,t-2}$ is correlated with the transformed error term $\varepsilon_{it} - \varepsilon_{i,t-1}$ (for any T) in the regression
$$y_{i,t-1} - y_{i,t-1} = \lambda(y_{i,t-1} - y_{i,t-2}) + (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\beta + \varepsilon_{it} - \varepsilon_{i,t-1}$$
- Similar problems occur also without a lagged dependent variable when $\mathbf{x}_{it}$ violates the strict-exogeneity assumption.

Omitted Dynamics

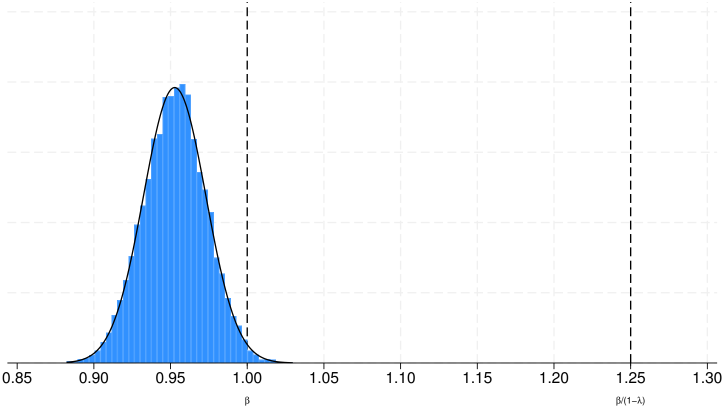
- Given the difficulties of estimating a (short-T) dynamic panel data model, applied researchers often decide to stick to a static panel data model instead. However, estimating

$$y_{it} = \mathbf{x}_{it}'\mathbf{b} + a_i + e_{it}$$

with the FE estimator $\hat{\mathbf{b}}_{FE}$ generally yields biased estimates of the short-run effects $\frac{\partial E[y_{it}|y_{i,t-1},\mathbf{X}_i;\alpha_i]}{\partial \mathbf{x}_{it}} = \beta$ when the true data-generating process is dynamic.
  - Even worse, many researchers using a static model specification implicitly aim to estimate long-run "equilibrium" effects $\frac{\partial E[y_{it}|\mathbf{X}_i;\alpha_i]}{\partial \mathbf{x}_{it}} = \frac{\beta}{1-\lambda}$.

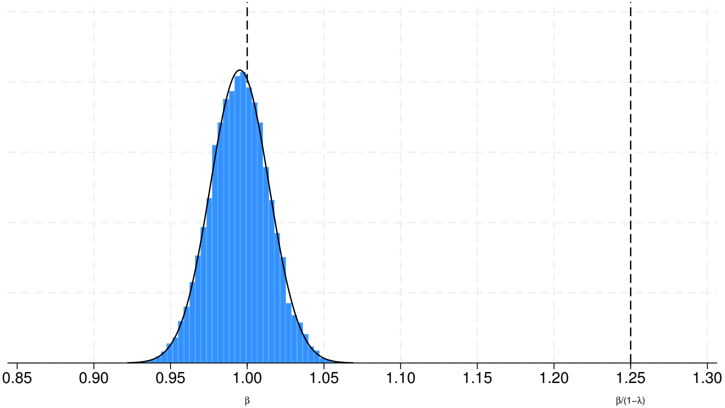
Omitted Dynamics

- Static FE estimator based on 10,000 replications:
  - $y_{it} = \lambda y_{i,t-1} + \beta x_{it} + \alpha_i + \varepsilon_{it}$, where $x_{it} \sim \mathcal{N}(0,1)$, $\varepsilon_{it} \sim \mathcal{N}(0,1)$, $\alpha_i \in \{-1,0,1\}$, $\lambda = 0.2$, and $\beta = 1$; $N = 600$, $T = 5$



Omitted Dynamics

- Static FE estimator based on 10,000 replications:
  - $y_{it} = \lambda y_{i,t-1} + \beta x_{it} + \alpha_i + \varepsilon_{it}$, where $x_{it} \sim \mathcal{N}(0,1)$, $\varepsilon_{it} \sim \mathcal{N}(0,1)$, $\alpha_i \in \{-1,0,1\}$, $\lambda = 0.2$, and $\beta = 1$; $N = 60$, $T = 50$



Interim Conclusion

- Summary statistics and visual data representation are useful for initial data exploration, but should be treated with suspicion. Unobserved heterogeneity can lead to erroneous conclusions.
- Similarly, findings from multivariate regression analysis (pooled OLS) can be misleading.
- Misspecification can be hard to detect. Even after omitting unit-specific intercepts, the implied regression residuals might appear independently normally distributed.
- Be clear about your objects of interest. Use guidance from economic theory or policy objectives when deciding about the econometric model specification.
- Do not match the model to the estimator, but find the appropriate estimator for the chosen model.
  - Understand the difference between a model and an estimator.
- Simulations can help to uncover properties of estimators.