Methods Meetup – 1st Session, MT 2023
Violet Butler Room, Barnett House, Oxford

1) Vaisey, S. & Miles, A. (2017) 'What you can – and can't – do with three-wave panel data', *Sociological Methods & Research*
2) Bell, A., Fairbrother, M. & Jones, K. (2019) 'Fixed and random effects models: making an informed choice', *Quality & Quantity*

Kun Lee

DPhil Candidate in Social Policy

University of Oxford

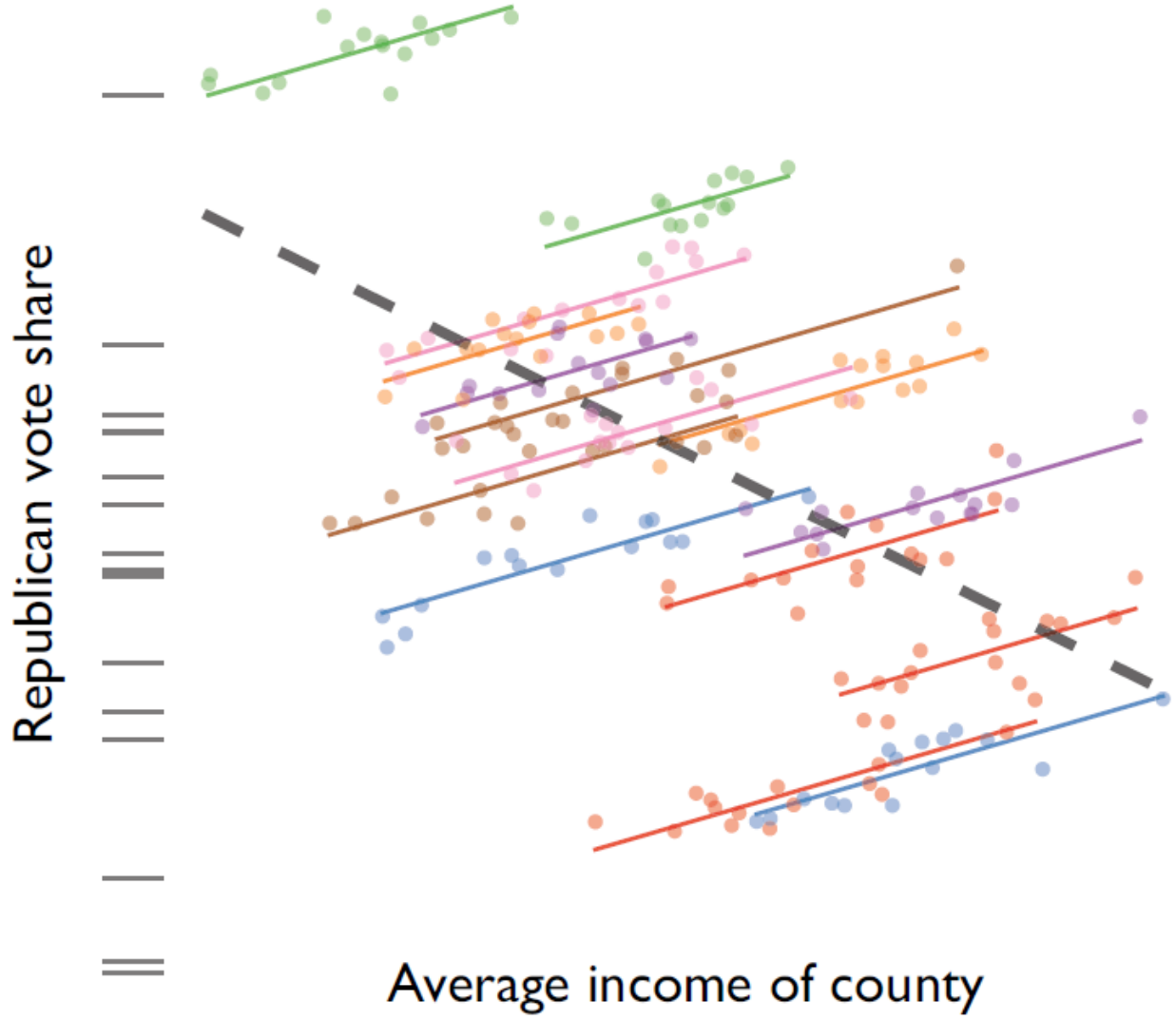
What is panel/longitudinal/TSCS data?

- "Repeated observation of same cross-section units" vs "pooled time-series of multiple units"
- Why so useful? Observing social processes over time; compare between individuals (Treat vs Control groups); large sample size
- Different terminology depending on data structure & discipline
  - Large N, short T vs. small N, long T
  - Micro (individual, household) vs **macro** (country, state, region…) data
  - Microeconomics/sociology vs macroeconomics/CPE
- Different modelling strategies, c.f. Law of large numbers

Problem: Omitted Variable Bias in OLS

Source: Adolph (2021)
Example from Gelman (2008)

Problem: Omitted Variable Bias in OLS



Source: Adolph (2021)
Example from Gelman (2008)

Dealing with "Unobserved Heterogeneity"

- Unobserved "time-constant" characteristics: personality, genetic trait
  - Macro: culture, social norms, entrenched institutions

$$y_{it} = \mu_t + \mathbf{x}'_{\mathbf{it}}\beta + \mathbf{z}'_{\mathbf{i}}\gamma + \upsilon_i + \epsilon_{it}, \quad i = 1, \ldots, N, \quad t = 1, 2, 3. \qquad (1)$$

- Two approaches to deal with this: fixed effect (FE) vs random effect (RE) models
- FE: mean differencing or including dummy variables
- RE: treat as errors (normally distributed, uncorrelated with X_{it})
- "The only difference between RE and FE lies in the assumption they make…" (Vaisey & Miles, 2017: p. 47)

Cautions with FE Models (Vaisey & Miles, 2017)

- FE should be preferred because unobserved characteristics being uncorrelated with X_{it} is generally unrealistic in social science
  - Church attendance (X) & anti-abortion (Y): one's given personality correlated with X & Y

1) *Endogenous selection* ($Y_{t-1} \rightarrow X_t \rightarrow Y_t$): your anti-abortion attitude (Y) this year will make you attend church (X) more next year (c.f. reverse causality)
2) *Unequal underlying trajectories*: those who are planning to get married (X) in the next few months are likely to report increasing happiness (Y)
   - Effect of marriage vs. effect of having good couple relationship?
   - c.f. Parallel trends assumption in diff-in-diff models
3) Using lagged predictors ($X_{i,t-1}$) is not a panacea (often creates downward bias or sometimes flips the sign)

Cost of FE models

- FE models control all time-constant variables, but at huge costs
  - Erase the 'level' information (e.g., contextual effects or "between" effects)
  - Cannot estimate the effect of other time-constant variables (e.g., race, gender)
  - Statistical cost: if your X_{it} does not change much over time (e.g., hhsize, democracy)
  - When you have large N, short T (e.g., 500 * 3)
  - Note: *Hausman* test only gives you whether within & between effects are different

Arguments from Bell et al. (2019)

- RE: gives you a weighted average of within & between-unit (or level-2) effect
  - May not or may make sense (i.e., effects of no. of children on poverty risk)
- Therefore, one should always prefer REWB or Mundlak formation

$$y_{it} = \beta_0 + \beta_{1W}(x_{it} - \bar{x}_i) + \beta_{2B}\bar{x}_i + \beta_3 z_i + (\upsilon_i + \epsilon_{it}).\ y_{it} = \beta_0 + \beta_{1W}x_{it} + \beta_{2C}\bar{x}_i + \beta_4 z_i + (\upsilon_i + \epsilon_{it}).$$

- You need to include random slope (variation of effects between clusters)
  - Failure to do so will produce anti-conservative SEs
  - However, also note the trade-off between model flexibility & analytical parsimony

Group Discussion: Which one should I use? FE, RE, REWB, or Mundlak?

- We have a hypothetical survey data of political attitude in England. Our main interest is the relationship between anti-immigrant attitude (scaled 0-10) and being positive towards Tory (scaled 0-10). In the dataset, there are 500 individuals surveyed over 4 waves (no missing data). Individuals are also nested within 35 districts.
  - Choose any models (FE, RE, Mundlak…) that you prefer with this data structure
  - What is your research question? How would you frame it with your chosen model?
  - What would be the limitation or cost of using your model
  - Instead of this dataset, you can also consider your own data & question from your current project
- What does it mean that your estimates are 'biased'? Discuss the concept with your own research questions.
  - What is your parameter of interest or *estimand* (Lundberg et al. 2021)?
  - Does your question imply causal direction? If not, can it still be 'biased'?