Fixed and Random Effects

Josephine Jacobs, PhD
Liam Rose, PhD
April 3, 2019

Poll

How familiar are you with the concepts fixed and random effects?

1. Very familiar
2. Somewhat familiar
3. Not familiar at all

Overview

- Panel Data
- Panel Linear Regression Model
- Unobserved Heterogeneity
- Fixed Effects Model
- Random Effects Model
- Choosing between FE and RE
- Terminology

Panel Data

- **Panel data**: Repeated cross-sections of the same individuals (households, countries, etc.) over several time periods

Person	Year	Age	Sex	Income	Education
1	2010	45	F	$ 40,000	College
1	2011	46	F	$ 42,000	College
1	2012	47	F	$ 44,000	College
2	2010	53	M	$ 30,000	High school
2	2011	54	M	$ 30,000	High school
2	2012	55	M	$ 31,000	High school

Panel Linear Regression Model

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \varepsilon_{it}$$

- Y_{it} : outcome variable income for individual *i* at time *t*
- X_{it} : explanatory variable education for individual *i* at time *t*
- ε_{it}: error term for individual *i* at time *t*
  - ε contains all other factors besides education that determines income
- β_1: the change in income associated with a unit change in education
- In order for $\hat{\beta}_1$ to be an unbiased estimate of the casual effect of Education on Income, X must be exogenous

Exogeneity

- Assumption $E(\varepsilon_{it}| X_{it}) = 0$
  - Conditional mean of ε_{it} given X_{it} is zero
  - Implies that X_{it} and ε_{it} CANNOT be correlated

- X_{it} and ε_{it} are correlated when there is:
  - **Omitted variable bias**
  - Sample selection
  - Simultaneous causality

- If X_{it} and ε_{it} are correlated, then X is endogenous

Unobserved Heterogeneity

- If omitted factors directly effect both the outcome and explanatory variables, explanatory variables will be correlated with errors and regression coefficients will be biased

- Unobserved heterogeneity refers to omitted factors that remain constant over time

  - Individual level:
    - Demographics (e.g. race/ethnicity)
    - Family history
    - Innate abilities

  - State level
    - Geography
    - Demographic, educational, or religious composition

Unobserved Heterogeneity cont'd

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \varepsilon_{it}$$

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \boxed{\alpha_i + u_{it}}$$

- u_{it}: time varying error component
- α_i: individual time-invariant individual heterogeneity

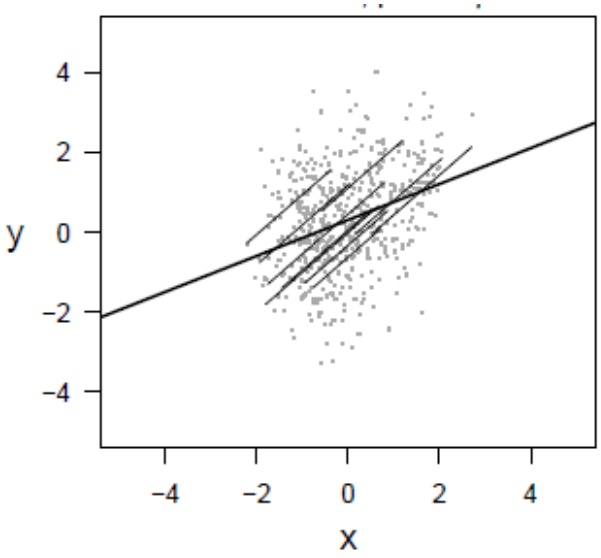
Unobserved Heterogeneity cont'd

- If $\text{cov}(X_{it}, \alpha_i) = 0$, then α is like any other unobserved factor that is not systematically related to Y in the error term
  - In our example, this is equivalent to stating motivation is not related to education.

- If $\text{Cov}(X_{it}, \alpha_i) \neq 0$, putting α in the error term will be problematic

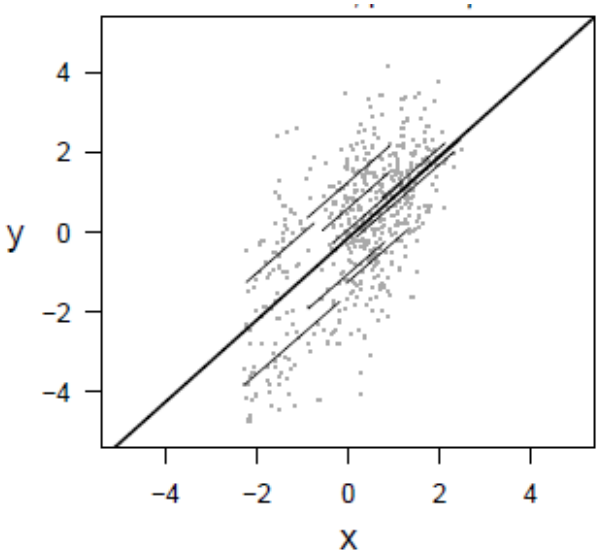
Unobserved Heterogeneity cont'd

A) Correlation =-0.7



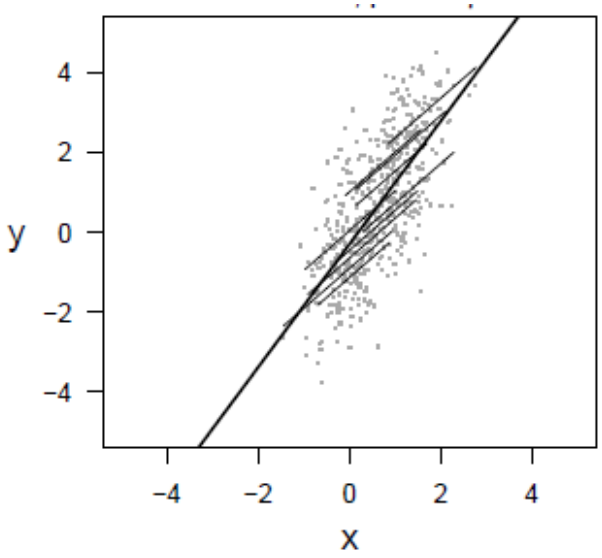
Pooled $\hat{\beta}$: 0.5

B) Correlation =0



Pooled $\hat{\beta}$: 1

C) Correlation =0.7



Pooled $\hat{\beta}$: 1.5

Source: Clark and Linzer, 2012

How can panel data help?

- Without an IV or additional data to control for these omitted factors, having repeated observations of the same units allows you to model α_i and control for unobserved, time-invariant factors

- Two standard approaches for modeling variation in α_i:
  - Fixed effects
  - Random effects

Fixed Effects Model

- Panel linear regression where the error term has two components:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + u_{it}$$

- In the fixed effects model, we replace the unobserved error component, α_i , with a set of **fixed** parameters $\mu_1 + \mu_2 + \mu_3 + \ldots + \mu_n$:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \mu_1 + \mu_2 + \mu_3 + \ldots + \mu_n + u_{it}$$

- Each unit has a unit-specific intercept that is estimated separately

- Designed to study causes of change *within* a unit

Fixed Effects Model

Least Squares Dummy Variable Estimator:

1. Create individual-specific dummy. For each observation, k:

$$D_{kit} = 0 \quad \text{if } k \neq i$$
$$D_{kit} = 1 \quad \text{if } k = i$$

2. Regress Y on the dummy variables and other explanatory variables:

$$Y_{it} = \beta_1 X_{it} + \mu_1 D_{1it} + \mu_2 D_{2it} + \ldots + \mu_n D_{nit} + u_{it}$$

Fixed Effects Model

Fixed Effects Estimator:

1. Determine the time-mean of Y, X, and ε

$$\bar{Y}_{\mathbf{i}} = \frac{1}{T}\sum_{t=1}^{T} Y_{it} \qquad \bar{\boldsymbol{X}}_{\mathbf{i}} = \frac{1}{T}\sum_{t=1}^{T} X_{it} \qquad \bar{\boldsymbol{u}}_{\mathbf{i}} = \frac{1}{T}\sum_{t=1}^{T} u_{it} \qquad \bar{\alpha}_{\mathbf{i}} = \frac{1}{T}\sum_{t=1}^{T} \alpha_{i}$$

Fixed Effects Model

Fixed Effects Estimator:

1. Determine the time-mean of Y, X, and ε

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it} \qquad \bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it} \qquad \bar{u}_i = \frac{1}{T}\sum_{t=1}^{T} u_{it} \qquad \bar{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T} \alpha_i$$

$$\bar{\alpha}_i = \alpha_i$$

Fixed Effects Estimator

Fixed Effects Estimator:

2. "Within" transformation (time-demean data) and regress time-demeaned data:

$$Y_{it} - \bar{Y}_i = \beta_1(X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i) + (\alpha_i - \bar{\alpha}_i)$$

This is equivalent to a dummy variable for each individual *i*

Fixed Effects Estimator

Fixed Effects Estimator:

2. "Within" transformation (time-demean data) and regress time-demeaned data:

$$Y_{it} - \bar{Y}_i = \beta_1(X_{it} - \bar{X}_i) + (u_{it} - \bar{u}_i) + (\alpha_i - \bar{\alpha}_i)$$

$(\alpha_i - \bar{\alpha}_i) \rightarrow 0$

In Stata, we can use *xtreg, fe*

This is equivalent to reg y x i.variable

Pros and Cons

- **Pro**: Will produce unbiased estimate of coefficient when $Cov(X_{it}, \alpha_i) \neq 0$
- **Con**: Estimation of time-invariant explanatory variables or variables that change very little over time is not possible
- **Con**: Those estimates can be subject to high sample-to-sample variability when:
  - Few observations per unit
  - X does not vary much **within** each unit relative to the variation in Y
- **Con**: Out-of-sample predictions not possible

FE Example

- Oberg (2016) assessed association between labor induction and Autism Spectrum Disorder (ASD).
- 1992-2005 Swedish register data, including linked population registers with familial relations
  - 1,362,950 births (22,077 diagnosed with ASD)
- Within siblings comparison

FE Example Cont'd

- Observables:
  - Birth year, parity
  - Maternal: age at birth, education, country of origin, BMI in early pregnancy, other health factors
- Unobservables:
  - Some environmental factors and all genetic factors shared within families
  - Look at variation *within* maternal sibling pairs with discordance with respect to induction
    - FE to allow the underlying hazard to vary between mothers, so comparison is within siblings only

FE Example Cont'd

Sample	No. of Women	Hazard Ratio (95% CI)			
		Model 1: Baseline	Model 2: +Stable Maternal	Model 3: +Birth-Specific	Model 4: Within Siblings
Complete case	1 117 220	1.32 (1.27-1.38)	1.31 (1.26-1.37)	1.19 (1.13-1.24)	0.99 (0.88-1.10)

[a] Model 1, baseline, adjusted for birth year, parity, and maternal age at birth; model 2 added maternal education and country of origin; model 3 added smoking and body mass index in early pregnancy, gestational diabetes or gestational hypertension, preeclampsia, chorioamnionitis, urogenital infection, intrauterine growth, premature ruptures of membranes, postterm gestation, multiple gestation, and high-risk pregnancy; and model 4 adjusted for all factors shared by maternal siblings, and all measured birth-specific covariates.

Source: Oberg, 2016

- Positive and significant association when sibling-specific characteristics are not accounted for
- Attenuated by additional covariates, but still significant

FE Example Cont'd

Sample	No. of Women	Hazard Ratio (95% CI)			
		Model 1: Baseline	Model 2: +Stable Maternal	Model 3: +Birth-Specific	Model 4: Within Siblings
Complete case	1 117 220	1.32 (1.27-1.38)	1.31 (1.26-1.37)	1.19 (1.13-1.24)	0.99 (0.88-1.10)

[a] Model 1, baseline, adjusted for birth year, parity, and maternal age at birth; model 2 added maternal education and country of origin; model 3 added smoking and body mass index in early pregnancy, gestational diabetes or gestational hypertension, preeclampsia, chorioamnionitis, urogenital infection, intrauterine growth, premature ruptures of membranes, postterm gestation, multiple gestation, and high-risk pregnancy; and model 4 adjusted for all factors shared by maternal siblings, and all measured birth-specific covariates.

Source: Oberg, 2016

- With the inclusion of maternal sibling fixed effects, labor induction no longer associated with offspring Autism Spectrum Disorder
- Unobserved genetic and family level characteristics may have been unaccounted for in Models 1-3

Multiple Fixed Effects

- Can include fixed effects on more than one dimension.
- Example: a fixed effect for a person, and a fixed effect for a year:

$$Income_{it} = b_0 + b_1 Education_{it} + \boldsymbol{Person_i} + \boldsymbol{Year_t} + e_{it}$$

Fixed Effects as Difference Differences

- Difference in differences is just a fixed effects regression:

$$y_{it} = b_0 + b_1\boldsymbol{post_t} + b_2\boldsymbol{group_i} + b_3\boldsymbol{post_t} * \boldsymbol{group_i} + e_{it}$$

for individual *i* in time *t*

- Generalized:

$$Y_{ist} = b_0 + b_2 post_{st} + \boldsymbol{A_s} + \boldsymbol{B_t} + \boldsymbol{C_i} + e_{ist}$$

for individual *i* in group *s* in time *t*

Random Effects Model

- If you can assume that $Cov(X_{it}, \alpha_i) = 0$, do not need to use FE, BUT you cannot simply run a pooled OLS

- This would create issues with **serial correlation**, where the correlation between the error term at one time (t) is correlated with the error term at some other point in time (s):

$$Cov(\alpha_{it}, \alpha_{is}) \neq 0$$

Serial Correlation

- When there is serial correlation, this implies that the OLS estimator will be **inefficient**
  - Standard errors can be underestimated
  - Implications for hypothesis testing

Random Effects Model

- Instead of FE, we can use a technique that is more efficient that FE, but that accounts for unobserved heterogeneity: Random Effects

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \alpha_i + u_{it}$$

- RE assumes that α is a random quantity sampled from a probability distribution (often normal distribution) with mean 0 and variance σ^2
  - Compromise between fixed-effects (within estimator) and a pooled OLS (between estimator)

Random Effects Estimator

- Transforms the fixed effects system with an inverse variance weight, λ:

$$\lambda = 1 - \sqrt{\left(\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}\right)}$$

σ_u^2: variance of u_{it}

σ_α^2: variance of α_i

- Use λ to quasi-time demean the system
  - Take off a fraction of the time demeaned values:

$$Y_{it} - \lambda\bar{Y}_i = \beta_0(1-\lambda) + \beta_1(X_{it} - \lambda\bar{X}_i) + (u_{it} - \lambda\bar{u}_i) + (\alpha_i - \lambda\bar{\alpha}_i)$$

Random Effects Intuition

- $0 \leq \lambda \leq 1$
- When $\sigma_\alpha^2 = 0$, $\lambda = 0$ and RE is equal to pooled OLS
  - Variation in α_i does not comprise significant portion of the error term, and it can be ignored
- When $\sigma_\alpha^2 \to \infty$, $\lambda = 1$ and RE is equal FE
  - Variation in α_i comprises a significant portion of the error term, and it cannot be ignored and RE tries to remove as much of this effect as possible

Random Effects Intuition

- Groups with outlying unit effects will have their α_i shrunk back towards the mean α which brings $\hat{\beta}$ closer to the pooled OLS estimate and further from the FE
  - Effect will be greatest for units containing fewer observations and when estimates of variance of α_i are close to zero

Operationalizing Random Effects

- Operationalized in two stages:
  1. Obtain an estimate of λ ($\hat{\lambda}$) by estimating σ_u^2 and σ_α^2
     - Obtained by estimating a FE or OLS regression
  2. Substitute $\hat{\lambda}$ to transform the system and run OLS:

$$Y_{it} - \hat{\lambda}\,\bar{Y}_i = \beta_0(1-\hat{\lambda}) + \beta_1(X_{it} - \hat{\lambda}\,\bar{X}_i) + (u_{it} - \hat{\lambda}\,\bar{u}_i) + (\alpha_i - \hat{\lambda}\,\bar{\alpha}_i)$$

In stata, we can use *xtreg, re*

Pros and Cons

- **Pro**: Can constrain the variance of β estimates
  - This leads to estimates that are closer, on average, to the true value in any particular sample
- **Pro**: Can include time-invariant covariates in the model
- **Pro**: Take into account unreliability associated with estimates from small samples within units
- **Con**: Will likely introduce bias in estimates of β
  - The greater the correlation between X_{it} and α_i, the greater the bias in estimates of β
- **Con**: Don't actually estimate α_i (α treated as random variables)

Poll

From an econometrics standpoint, when is it appropriate to use random effects in place of fixed effects?

1. When the unobserved unit-specific factors, α_i, are **NOT** correlated with the covariates in the model.

2. When the unobserved unit-specific factors, α_i, are correlated with the covariates in the model.

3. The models can be used interchangeably

Poll

From an econometrics standpoint, when is it appropriate to use random effects in place of fixed effects?

1. **When the unobserved unit-specific factors, α_i, are NOT correlated with the covariates in the model.**

2. When the unobserved unit-specific factors, α_i, are correlated with the covariates in the model.

3. The models can be used interchangeably

Choosing between FE and RE

- Hausman test
  - Measure of the difference between the FE estimate and the RE estimate
  - H_0: coefficients estimated by the RE estimator are the same as the ones estimated by the FE estimator
  - Rejection of null hypothesis: the two models are different, and reject the random effects model in favor of fixed effects

Choosing between FE and RE

- Hausman test drawbacks:
  - A rejection of the null hypothesis may be because the test does not have sufficient statistical power to detect departures from the null
  - With FE and RE there is a tradeoff between bias reduction and variance reduction – Hausman does not help in evaluating this tradeoff

Choosing between FE and RE

- Clark and Linzer (2012) suggest 3 considerations:

1. Extent to which variation in explanatory variable is primarily within unit as opposed to across units

2. Amount of data one has (# of units and observations per unit

3. Goal of modeling exercise

Choosing between FE and RE

- When variation is primarily within units:
  - Decide based on purposes of research : Any bias in slope parameter with RE is more than compensated for by increase in estimate efficiency

- When variation is primarily across units
  - Depends on the amount of data and the underlying level of correlation between unit effects and regressors

Source: Clark and Linzer, 2012

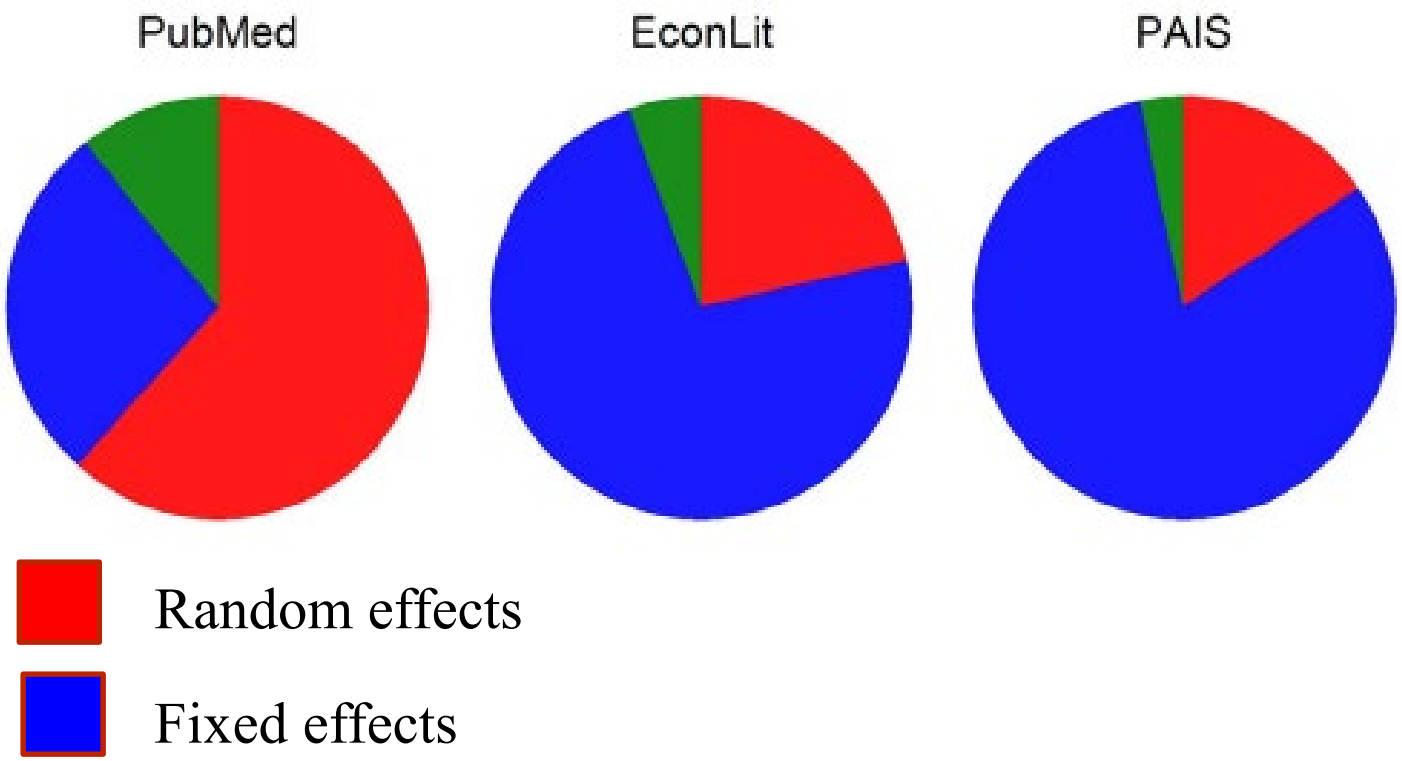
Choosing between FE and RE

Choosing between FE and RE when variation is primarily across units

		Observations per Unit	
# Units		**Fe (≤5)**	**Many**
	Few (≤10)	RE	RE is correlation is low; FE otherwise
	Many	RE if correlation is low; FE otherwise	FE unless correlation is close to zero

Source: Clark and Linzer, 2012

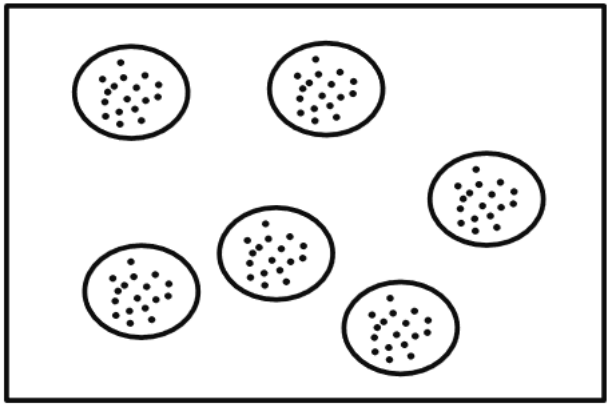
Choosing between FE and RE



Source: Dieleman and Templin, 2014

Mixed Models

- It is also possible to include both fixed and random effects.
  - This is most often done with clinical trials where the outcome is a repeated measurement (ex: the same memory test repeated over the course of a treatment)
  - Observations are correlated within unit, but not across group



Mixed Models

- Matrix notation:

$$\boldsymbol{y_i = X_i B + Z_i u + e}$$

where y is the outcome, **X** contains the fixed effects, and **Z** contains the random effects

- In practice, use "xtmixed" in stata and "lme4" in R.
  - Example: The effect of a treatment (treat) on test score (tscore) with a fixed effect for a week and a random effect for site (site_id)

```
xtmixed tscore treat week || site_id
```

FE and RE Terminology

Variable definitions:

"Fixed effects are constant across individuals, and random effects vary" (Kreft and Deleeuw, 1998)

"Effects are fixed if they are interesting in themselves or random if there is interest in the underlying population" (Searle, Casella, and McCulloch, 1992)

"When a sample exhausts the population, the corresponding variable is *fixed*; when the sample is a small (i.e., negligible) part of the population the corresponding variable is *random*." (Green and Turkey, 1960)

"If an effect is assumed to be a realized value of a random variable, it is called a random effect" (LaMotte, 1983)

Source: Gelman, 2005

References

Baltagi, B.H. (2001). *Econometric Analysis of Panel Data*. New York, NY: John Wiley & Sons Ltd.

Clark, T.S. & Linzer, D.A. (2014). Should I use Fixed or Random Effects? *Political Science Research Methods*, 3(2), 399-408.

Dieleman, J.L. & Templin, T (2014). Random-Effects, Fixed-Effects and the Within-Between Specification for Clustered Data in Observational Health Studies: A simulation study. *PLOS ONE*, 9(10), 1-17.

Gelman, A. (2005). Discussion Paper: Analysis of variance – why it is more important than ever. *The Annals of Statistics*, 33(1), 1-53.

Oberg, S.A., D'Onofrio, B.M., Richert, M.E., Hernandez-Diaz, S., Eckner, J.L., Almqvist, C., et al. (2016). Association of Labor Induction with Offspring Risk of Autism Spectrum Disorder, *JAMA Pediatrics*, 179(9), 1-7.

Setodji, C.C. & Schwartz, M. (2013). Fixed-effect or Random-effect Models? What are the key inference issues? *Medical Care*, 51(1), 25-28.

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data*. London, England: The MIT Press.