Quantitative Social Research II
Workshop 7: Hierarchical Data

Jose Pina-Sánchez

Workshop Aims

- Learn to identify hierarchical data
- Discuss the implications of violating the assumption of independence
- Modelling strategies considering lack of independence as a data nuisance problem to be adjusted
  - Robust standard errors
- Modelling strategies considering lack of independence as a substantively interesting process to be modelled
  - Multilevel modelling (aka hierarchical, random effects, or mixed effects models)

Workshop Aims: Recap

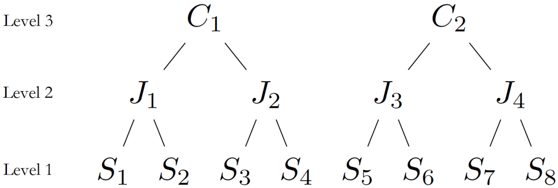
- The assumptions of the linear regression model:
  - Normality: $N \sim (0, Var(e))$
  - Homoskedasticity: $Var(e_i) = Var(e)$
  - **Independence:** $Cov(e_i, e_j) = 0$
  - No endogeneity: $Cov(X_i, e_i) = 0$
  - Perfectly measured variables
  - No missing data (other than missing at random)
  - No omitted relevant variables
  - No multicollinearity
  - Linearity

Hierarchical Data

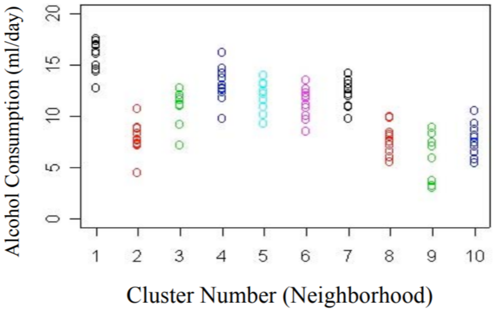
- When cases composing a sample can be grouped within clusters
  - E.g. students within modules within programs
  - You are not an independent sample of the University of Leeds student body
  - As a result of - or because you are - taking part in this module you share some commonalities (within-cluster correlation) that make you different from the population of students
  - Additional within correlations could be expected from being enrolled in a Sociology/ Criminology program
  - And the same applies to other students in different modules and programs
- Question: Can you think of any other examples of hierarchical data?
  - Interviewees within regions within countries
  - Sentences imposed by judges sitting in courts
  - Any instance where cluster sampling is used

Hierarchical Data

The Hierarchical Structure of Sentencing Data



Hierarchical Data



Source: Francesca Dominici

- Cases across this sample are not independent
- Cases within the same cluster are related to each other

Hierarchical Data: Notation

- We need different subscripts to distinguish units at different levels
  - For the case of sentencing data we have considered three levels
  - Court: $l = 1, 2, 3, ..., L$
  - Judge: $j = 1, 2, 3, ..., J_l$
  - Sentence: $i = 1, 2, 3, ..., I_{lj}$
- We will use these to identify values in our outcome, explanatory variables and residuals
  - $Y_{lji} = \beta_0 + \beta_k X_{klji} + \underbrace{e_{lji}}_{v_l + u_{lj} + \epsilon_{lji}}$
  - Notice how the residual term can now be partitioned to reflect the unobserved variability stemming from each level

The Assumption of Independence

- To estimate the standard errors of regression coefficients we use the variance covariance matrix
  - A matrix of the residuals' variances and covariances for each observation, for a simplified model of only $n = 3$ we have

$$\begin{pmatrix} var(e_1) & cov(e_1, e_2) & cov(e_1, e_3) \\ cov(e_2, e_1) & var(e_2) & cov(e_2, e_3) \\ cov(e_3, e_1) & cov(e_3, e_2) & var(e_3) \end{pmatrix}$$

The Assumption of Independence

- To estimate the standard errors of regression coefficients we use the variance covariance matrix
  - A matrix of the residuals' variances and covariances for each observation, for a simplified model of only $n = 3$ we have

$$\begin{pmatrix} var(e_1) & cov(e_1, e_2) & cov(e_1, e_3) \\ cov(e_2, e_1) & var(e_2) & cov(e_2, e_3) \\ cov(e_3, e_1) & cov(e_3, e_2) & var(e_3) \end{pmatrix}$$

  - Under the assumption of homoskedasticity and independence

$$\begin{pmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_e^2 & 0 \\ 0 & 0 & \sigma_e^2 \end{pmatrix}$$

  - Under two level hierarchical data the diagonals will be equal to $\sigma^2_{\epsilon_i} + \sigma^2_{u_j}$ and the 0s equal to $\sigma_{uj,uj}$

The Assumption of Independence

- To estimate the standard errors of regression coefficients we use the variance covariance matrix
  - A matrix of the residuals' variances and covariances for each observation, for a simplified model of only $n = 3$ we have

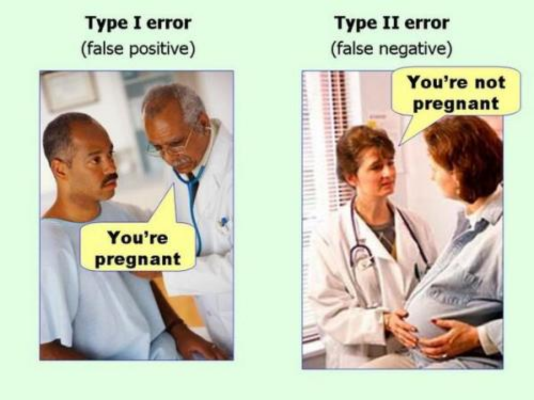
$$\begin{pmatrix} var(e_1) & cov(e_1, e_2) & cov(e_1, e_3) \\ cov(e_2, e_1) & var(e_2) & cov(e_2, e_3) \\ cov(e_3, e_1) & cov(e_3, e_2) & var(e_3) \end{pmatrix}$$

  - Under the assumption of homoskedasticity and independence

$$\begin{pmatrix} \sigma_e^2 & 0 & 0 \\ 0 & \sigma_e^2 & 0 \\ 0 & 0 & \sigma_e^2 \end{pmatrix}$$

  - Under two level hierarchical data the diagonals will be equal to $\sigma^2_{\epsilon_i} + \sigma^2_{u_j}$ and the 0s equal to $\sigma_{uj,uj}$
  - Assuming independence in the presence of hierarchical data will lead to a 'naive analysis'
  - Underestimated measures of uncertainty (smaller SEs, narrower CIs, higher chance of type I errors)

Type I & II Errors



Source: Paul Ellis *'Effect Sizes'*

Strategies to Adjust for Within-Cluster Correlation

- Three main choices, all with pros and cons

Strategies to Adjust for Within-Cluster Correlation

- Three main choices, all with pros and cons
- Robust standard errors (the 'sandwich estimator')
  - Each variance and covariance is estimated empirically $e_{ji} = Y_{ji} - \beta_0 - \beta_k X_{kji}$
  - Pros: provides robust SEs
  - Cons: within-cluster correlation as a data nuisance, i.e. we do not model and learn about these correlations

Strategies to Adjust for Within-Cluster Correlation

- Three main choices, all with pros and cons
- Robust standard errors (the 'sandwich estimator')
  - Each variance and covariance is estimated empirically

    $e_{ji} = Y_{ji} - \beta_0 - \beta_k X_{kji}$
  - Pros: provides robust SEs
  - Cons: within-cluster correlation as a data nuisance, i.e. we do not model and learn about these correlations
- Fixed effects models
  - Clusters are included in the model as dummy variables

    $Y_{ji} = \beta_0 + \beta_k X_{kji} + \beta_j X_{ji} + e_{ji}$
  - Pros: can model mean differences in the outcome by cluster

    which can be substantially interesting; e.g. which is the neighbourhood with higher alcohol consumption?

    can control for potential confounders; e.g. we might want to explore the effect of social class on alcohol consumption but this can be confounded by neighbourhood characteristics such as the number of pubs per km^2
  - Cons: overfitted models

Strategies to Adjust for Within-Cluster Correlation

- Multilevel modelling
  - The error term at each level is partitioned and modelled separately

$$Y_{ji} = \overbrace{\beta_0 + \beta_k X_{kji}}^{\text{fixed part}} + \overbrace{u_j + \epsilon_{ji}}^{\text{random part}}$$

  - That's why MLM are often called mixed or random effects models, and why we called fixed effects models that way

Strategies to Adjust for Within-Cluster Correlation

- Multilevel modelling
  - The error term at each level is partitioned and modelled separately

$$Y_{ji} = \overbrace{\beta_0 + \beta_k X_{kji}}^{\text{fixed part}} + \overbrace{u_j + \epsilon_{ji}}^{\text{random part}}$$

  - That's why MLM are often called mixed or random effects models, and why we called fixed effects models that way
  - Pros: if modelled properly can provide robust SEs

    Allows modelling variability between and within clusters:

    E.g.1 Are there between court inconsistencies in sentencing?

    E.g.2 Are differences in happiness due to differences across countries or individuals?

Strategies to Adjust for Within-Cluster Correlation

- Multilevel modelling
  - The error term at each level is partitioned and modelled separately

    $$Y_{ji} = \overbrace{\beta_0 + \beta_k X_{kji}}^{\text{fixed part}} + \overbrace{u_j + \epsilon_{ji}}^{\text{random part}}$$

  - That's why MLM are often called mixed or random effects models, and why we called fixed effects models that way
  - Pros: if modelled properly can provide robust SEs

    Allows modelling variability between and within clusters:

    E.g.1 Are there between court inconsistencies in sentencing?

    E.g.2 Are differences in happiness due to differences across countries or individuals?
  - Cons: don't control for cluster units with a potential confounding effect

    Invoke further assumptions:

    $u_j \sim N(0, \sigma_u)$; $cov(u_j, u_{j'}) = 0$

    $\epsilon_j \sim N(0, \sigma_\epsilon)$; $cov(\epsilon_j, \epsilon_{j'}) = 0$

    $cov(\epsilon_{ji}, uj) = 0$

Random Intercepts

- The simplest form of MLM
  - Allows for the intercept to vary across clusters
  - For the case of a 2-level MLM with one explanatory variable could be expressed as

$$Y_{ji} = \overbrace{\beta_0 j}^{\beta_0 + u_j} + \beta_1 X_{ji} + \epsilon_{ji}$$

  - Invokes the same assumptions listed in the previous slide

Random Intercepts

- The simplest form of MLM
  - Allows for the intercept to vary across clusters
  - For the case of a 2-level MLM with one explanatory variable could be expressed as

$$Y_{ji} = \overbrace{\beta_0 j}^{\beta_0 + u_j} + \beta_1 X_{ji} + \epsilon_{ji}$$

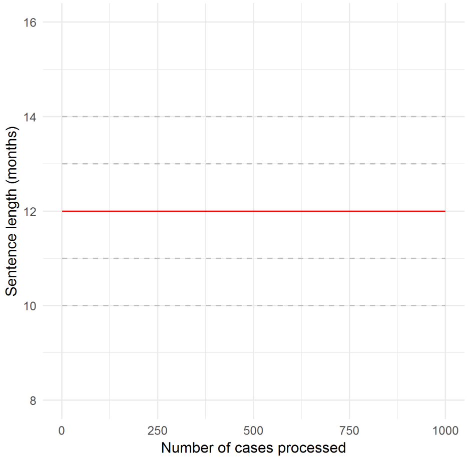
  - Invokes the same assumptions listed in the previous slide
- Can be used to estimate the intracluster correlation coefficient
  - $ICC = \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}$
  - The proportion of unobserved variability in the outcome variable (i.e. residual variability) stemming from level 2, E.g. the extent to which sentencing disparities are due to between judge differences as oppose to case level differences
  - Can also be understood as the correlation between observations from the same cluster, E.g. the similarities between sentences imposed by the same judge

Random Intercepts

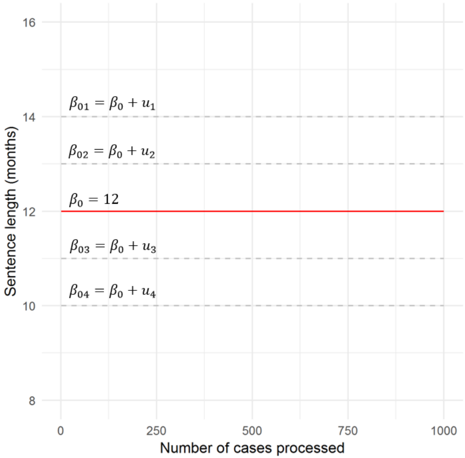
- The simplest form of MLM
  - Allows for the intercept to vary across clusters
  - For the case of a 2-level MLM with one explanatory variable could be expressed as

    $$Y_{ji} = \overbrace{\beta_0 j}^{\beta_0 + u_j} + \beta_1 X_{ji} + \epsilon_{ji}$$
  - Invokes the same assumptions listed in the previous slide
- Can be used to estimate the intracluster correlation coefficient
  - $ICC = \dfrac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}$
  - The proportion of unobserved variability in the outcome variable (i.e. residual variability) stemming from level 2, E.g. the extent to which sentencing disparities are due to between judge differences as oppose to case level differences
  - Can also be understood as the correlation between observations from the same cluster, E.g. the similarities between sentences imposed by the same judge
- Can be extended to 3 or more levels
  - $Y_{lji} = \overbrace{\beta_0 lj}^{\beta_0 + v_l + u_{lj}} + \beta_1 X_{lji} + \epsilon_{lji}$

Random Intercepts



Random Intercepts



Random Slopes

- The RIs model can be extended by allowing between cluster variability around the intercept but also around specific slopes

Random Slopes

- The RIs model can be extended by allowing between cluster variability around the intercept but also around specific slopes
  - For the case of a 2-level MLM with one explanatory variable could be expressed as

$$Y_{ji} = \overbrace{\beta_{0j}}^{\beta_0 + u_{0j}} + \overbrace{\beta_{1j}}^{\beta_1 + u_{1j}} X_{ji} + \epsilon_{ji}$$

Random Slopes

- The RIs model can be extended by allowing between cluster variability around the intercept but also around specific slopes
  - For the case of a 2-level MLM with one explanatory variable could be expressed as

$$Y_{ji} = \overbrace{\beta_0 j}^{\beta_0 + u_{0j}} + \overbrace{\beta_{1j}}^{\beta_1 + u_{1j}} X_{ji} + \epsilon_{ji}$$

  - As before, level-1 and level-2 residuals are assumed to be

$$u_{0j} \sim N(0, \sigma_{u0}) \ ; \ cov(u_{0j}, u_{0j'}) = 0$$
$$u_{1j} \sim N(0, \sigma_{u1}) \ ; \ cov(u_{1j}, u_{1j'}) = 0$$
$$\epsilon_j \sim N(0, \sigma_\epsilon) \ ; \ cov(\epsilon_j, \epsilon_{j'}) = 0$$

Random Slopes

- The RIs model can be extended by allowing between cluster variability around the intercept but also around specific slopes
  - For the case of a 2-level MLM with one explanatory variable could be expressed as

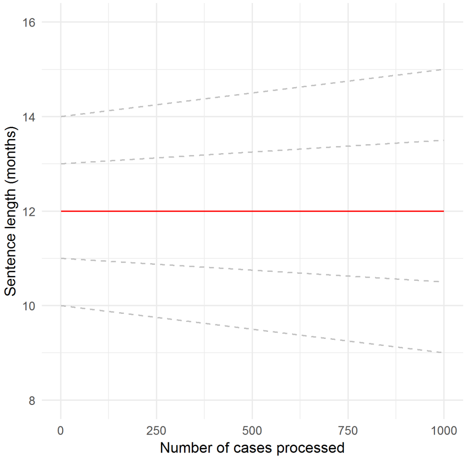
$$Y_{ji} = \overbrace{\beta_0 j}^{\beta_0 + u_{0j}} + \overbrace{\beta_{1j}}^{\beta_1 + u_{1j}} X_{ji} + \epsilon_{ji}$$

  - As before, level-1 and level-2 residuals are assumed to be

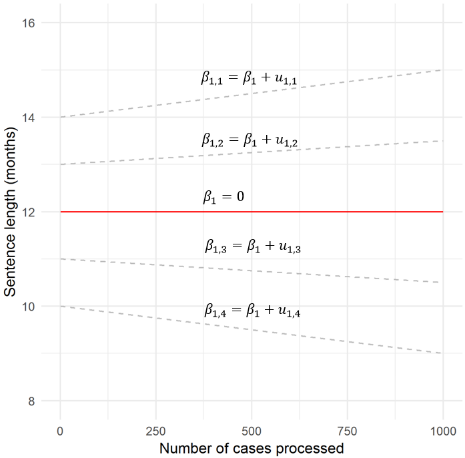
$$u_{0j} \sim N(0, \sigma_{u0}) \; ; \; cov(u_{0j}, u_{0j'}) = 0$$
$$u_{1j} \sim N(0, \sigma_{u1}) \; ; \; cov(u_{1j}, u_{1j'}) = 0$$
$$\epsilon_j \sim N(0, \sigma_\epsilon) \; ; \; cov(\epsilon_j, \epsilon_{j'}) = 0$$

  - However, now we might be interested in exploring whether $cov(u_{0j}, u_{1j'}) \neq 0$
  - If positive the slopes will diverge, i.e. higher intercepts are associated with higher slopes and vice versa
  - If negative the slopes will converge, i.e. higher intercepts are associated with lower slopes and vice versa

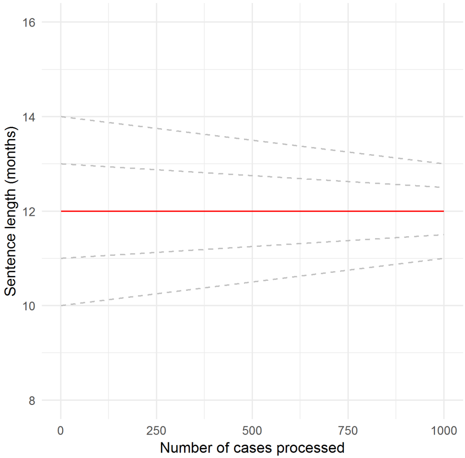
Random Slopes (+cov)



Random Slopes (+cov)



Random Slopes (-cov)



Recap

- In the presence of hierarchical data the assumption of independence does not hold
  - Measures of uncertainty will tend to be underestimated $\rightarrow$ type I errors

Recap

- In the presence of hierarchical data the assumption of independence does not hold
  - Measures of uncertainty will tend to be underestimated $\rightarrow$ type I errors
- We have covered the three main adjustment strategies
  - Robust standard errors
  - Fixed effects
  - Multilevel modelling

Recap

- In the presence of hierarchical data the assumption of independence does not hold
  - Measures of uncertainty will tend to be underestimated → type I errors
- We have covered the three main adjustment strategies
  - Robust standard errors
  - Fixed effects
  - Multilevel modelling
- Robust standard errors (the 'sandwich estimator')
  - Provide unbiased measures of uncertainty (also in the presence of heteroskedasticity)
  - Doesn't control for systematic difference between clusters (potential confounders)
  - Doesn't tell us anything about between/within cluster variability
  - To be used when cluster variability is not of interest (considered a data nuisance)

Recap

- Fixed effects
  - Partially adjust SEs while controlling for systematic differences between clusters
  - To be used when confounders are a serious concern
  - Can be used to compare means across clusters but not great at assessing variability
  - If the number of clusters is large will risk overfitting the model

Recap

- Fixed effects
  - Partially adjust SEs while controlling for systematic differences between clusters
  - To be used when confounders are a serious concern
  - Can be used to compare means across clusters but not great at assessing variability
  - If the number of clusters is large will risk overfitting the model
- Multilevel modelling
  - More flexible than fixed effects models at adjusting SEs
  - Does not control for systematic differences between clusters
  - Allow exploring susbtantive questions related to between/within cluster variability

Recap

- Fixed effects
  - Partially adjust SEs while controlling for systematic differences between clusters
  - To be used when confounders are a serious concern
  - Can be used to compare means across clusters but not great at assessing variability
  - If the number of clusters is large will risk overfitting the model
- Multilevel modelling
  - More flexible than fixed effects models at adjusting SEs
  - Does not control for systematic differences between clusters
  - Allow exploring susbtantive questions related to between/within cluster variability
- To learn more about multilevel modelling
  - Read Goldstein (1995) Chapter 2
  - And watch the online course from Brunton-Smith (2019)
  - Sign up for the LEMMA online course