Quantitative Social Research II
Workshop 8: 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

- Assumptions in the linear regression model ($Y = \alpha + \beta_k X_k + e$):
  - normality: residuals are normally distributed
  - homoskedasticity: the variance of the residuals is constant
  - **independence**: residuals are independent of each other
  - no multicollinearity
  - perfectly measured variables
  - no missing data (other than missing at random)
  - no unobserved confounders: we control for all common causes of X_1 and Y
  - no reverse causality: Y does not cause X_1
  - linearity: the effect of X_1 on Y is the same across the range of X_1

Hierarchical Data

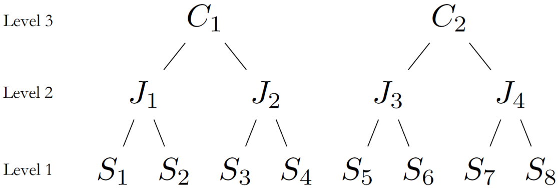
- When cases composing a sample can be grouped within clusters
  - e.g. students within modules within programs
  - this class is not an independent sample of the University of Leeds student body
  - as a result of - or because you are - taking part in this module you share some commonalities (within-cluster correlation) that make you different from other 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?

Hierarchical Data

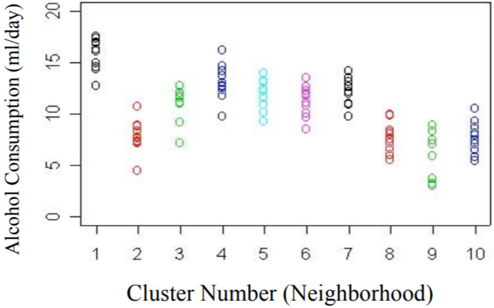
- When cases composing a sample can be grouped within clusters
  - e.g. students within modules within programs
  - this class is not an independent sample of the University of Leeds student body
  - as a result of - or because you are - taking part in this module you share some commonalities (within-cluster correlation) that make you different from other 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

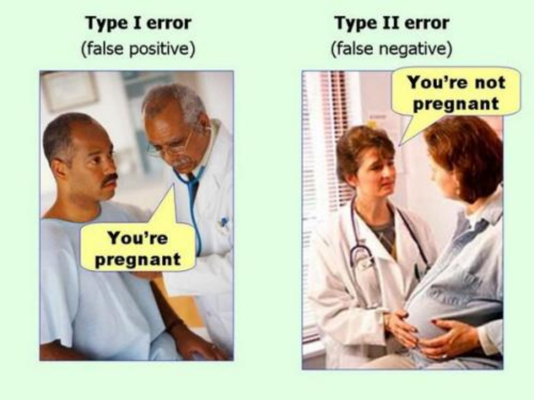
$$\begin{pmatrix} var(e_1) & cov(e_1, e_2) & cov(e_1, e_3) \\ cov(e_2, e_1) & var(e_2) & cov(e_2, e_3) \\ cov(e_3, e_1) & cov(e_3, e_2) & var(e_3) \end{pmatrix}$$

  - 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 'naive' findings
  - 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 is seen 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 is seen 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 cluster-level confounders; e.g. we might want to explore the effect of social class on alcohol consumption, which can be confounded by type of neighbourhood
  - cons: can lead to overfitted models

Strategies to Adjust for Within-Cluster Correlation

- Multilevel modelling (MLM)
  - 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 (MLM)
  - 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 (MLM)
  - 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-level confounders

    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

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

  - 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

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

  - 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 proportion of sentencing disparities due to between judge 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

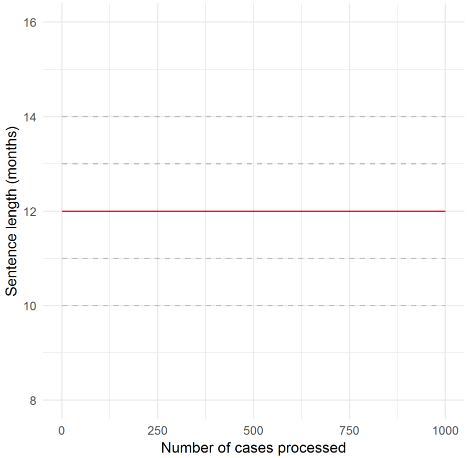
$$Y_{ji} = \overbrace{\beta_0 j}^{\beta_0 + u_j} + \beta_1 X_{1ji} + \epsilon_{ji}$$

  - invokes the same assumptions listed in the previous slide

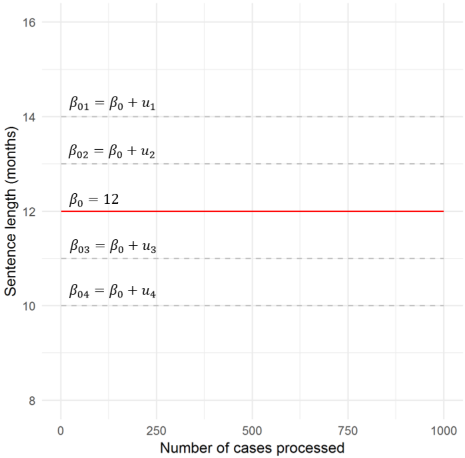
- Can be used to estimate the intracluster correlation coefficient
  - $ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\epsilon^2}$
  - the proportion of unobserved variability in the outcome variable (i.e. residual variability) stemming from level 2, e.g. the proportion of sentencing disparities due to between judge 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
  - $Y_{lji} = \overbrace{\beta_0 lj}^{\beta_0 + v_l + u_{lj}} + \beta_1 X_{1lji} + \epsilon_{lji}$

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

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

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

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

  - 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

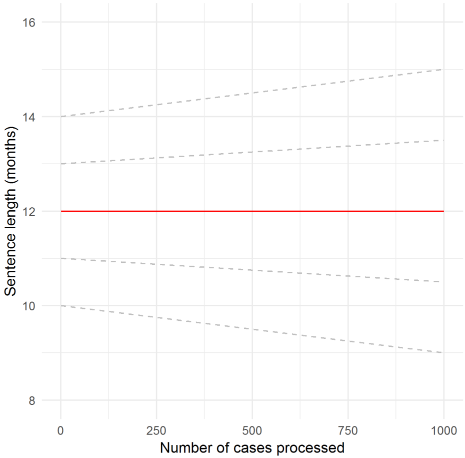
$$Y_{ji} = \overbrace{\beta_0 j}^{\beta_0 + u_{0j}} + \overbrace{\beta_{1j}}^{\beta_1 + u_{1j}} X_{1ji} + \epsilon_{ji}$$

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

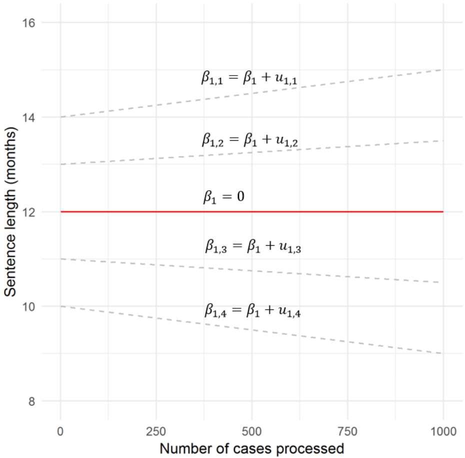
$$u_{0j} \sim N(0, \sigma_{u0}) \; ; \; cov(u_{0j}, u_{0j'}) = 0$$
$$u_{1j} \sim N(0, \sigma_{u1}) \; ; \; cov(u_{1j}, u_{1j'}) = 0$$
$$\epsilon_j \sim N(0, \sigma_\epsilon) \; ; \; cov(\epsilon_j, \epsilon_{j'}) = 0$$

  - 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

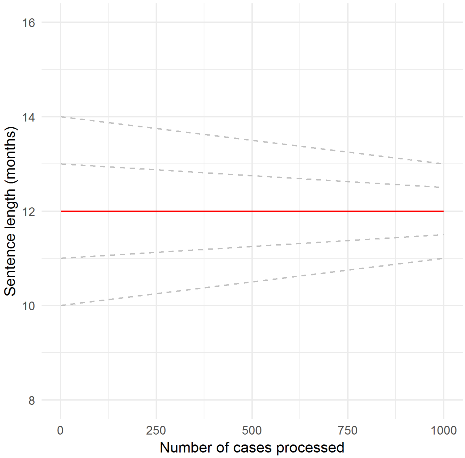
Random Slopes (+cov)



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 → 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 → 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