Measuring ethnic segregation: a model based approach
Approaches to measuring segregation

- In the past, people have calculated indexes directly from data.
- Discussion has been about what index to use.
  - What properties should an index have?
- Recently an alternative approach has been developed: modelling.
  - Note: after modelling, option to calculate index of choice from results.
- We use this approach here.

**Examples of indexes**
- Dissimilarity (D)
- Gini
- Gorard
- Cowgill
- Non-white Ghetto
- Reproducibility
## Problems with index-only approach

### Problems
- Not based on the assumption of an underlying process: assumes observed values are the true values
- If we do assume an underlying process, there is random variability of the observed values around this.
- This means when there is no actual segregation, expected value of any index $\neq 0$: it is a function of the total number in each school and the total proportion in each category.

### Limitations
- Difficult or impossible to:
  - include explanatory variables (esp. at school or individual level)
  - measure segregation simultaneously at multiple levels (e.g. school and LEA or school and neighbourhood)
  - handle multiple categories (e.g. when measuring ethnic segregation)
The modelling approach

**nonwhite}_{ijk} \sim \text{Bin(total}_{ijk}, \pi_{ijk})

\text{logit}(\pi_{ijk}) = \beta_{0jk} \\
\beta_{0jk} = \beta_0 + \nu_k + u_{jk} \\
\nu_k \sim \text{N}(0, \sigma^2_v) \\
u_{jk} \sim \text{N}(0, \sigma^2_u) \\
\text{Var(nonwhite}_{ijk}|\pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk})/\text{total}_{ijk}

- We use $\sigma^2_u$ (variation in underlying proportions on the logit scale) to measure segregation at school level
  - A lot of variation in proportion non-White $\Rightarrow$ high segregation
  - Little variation in proportion non-White $\Rightarrow$ low segregation
- We use $\sigma^2_v$ to measure segregation at the LEA level
- The random LEA effect allows for different proportions nonwhite in different LEAs
- If we had few LEAs we could use LEA fixed effects instead (we would not then get a parameter to measure LEA segregation)

How does this avoid the problems?

Index only approach
- Not based on the assumption of an underlying process: assumes observed values are the true values.
- If we do assume an underlying process, there is random variability of the observed values around this.
- This means when there is no actual segregation, expected value of any index $\neq 0$: it is a function of the total number in each school and the total proportion in each category.

Model approach
- Model allows for binomial variability of observed value around true value $u_j$. Can get standard error for estimates. So allows inferences.
- $\sigma_u^2$ no longer a function of the total number in each school.
\( \sigma_u^2 \) no longer a function of the total number in each school because modelling regards the schools as the units of analysis. Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

- The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma^2_u$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- \( \sigma_u^2 \) no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

- The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

- The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

![Graphs for 3 and 7 schools](graph.png)

- The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma^2_u$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Modelling and segregation curves

- $\sigma_u^2$ no longer a function of the total number in each school because modelling regards the schools as the units of analysis.
- Here it differs from segregation curves.

The variance is different but segregation curves are the same.
Extensions to the model

Original model

\[
\text{nonwhite}_{ijk} \sim \text{Bin}(\text{total}_{ijk}, \pi_{ijk})
\]

\[
\logit(\pi_{ijk}) = \beta_{0jk}
\]

\[
\beta_{0jk} = \beta_0 + v_k + u_{jk}
\]

\[
v_k \sim \mathcal{N}(0, \sigma_v^2)
\]

\[
u_{jk} \sim \mathcal{N}(0, \sigma_u^2)
\]

\[
\text{Var}(\text{nonwhite}_{ijk} | \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk}) / \text{total}_{ijk}
\]

Adding time

\[
\text{nonwhite}_{ijk} \sim \text{Bin}(\text{total}_{ijk}, \pi_{ijk})
\]

\[
\logit(\pi_{ijk}) = \beta_{0jk} + \beta_{1jk} \text{cohort}_{ijk}
\]

\[
\beta_{0jk} = \beta_0 + v_0k + u_{0jk}
\]

\[
\beta_{1jk} = \beta_1 + v_1k + u_{1jk}
\]

\[
\begin{bmatrix} v_0k \\ v_{1k} \end{bmatrix} \sim \mathcal{N}(0, \Omega_v), \quad \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01} \\ \sigma_{v01} & \sigma_{v1}^2 \end{bmatrix}
\]

\[
\begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim \mathcal{N}(0, \Omega_u), \quad \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix}
\]

\[
\text{Var}(\text{nonwhite}_{ijk} | \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk}) / \text{total}_{ijk}
\]

Adding response categories

\[
\text{proportion}_{ijkl} \sim \text{Multinomial}(\text{total}_{jkl}, \pi_{ijkl})
\]

\[
\log(\pi_{2kl} / \pi_{1kl}) = \beta_{0kl}
\]

\[
\log(\pi_{3kl} / \pi_{1kl}) = \beta_{1kl}
\]

\[
\beta_{0kl} = \beta_0 + f_0l + v_{0kl}
\]

\[
\beta_{1kl} = \beta_1 + f_1l + v_{1kl}
\]

\[
\begin{bmatrix} f_0l \\ f_{1l} \end{bmatrix} \sim \mathcal{N}(0, \Omega_f), \quad \Omega_f = \begin{bmatrix} \sigma_{f0}^2 & \sigma_{f01} \\ \sigma_{f01} & \sigma_{f1}^2 \end{bmatrix}
\]

\[
\begin{bmatrix} v_{0kl} \\ v_{1kl} \end{bmatrix} \sim \mathcal{N}(0, \Omega_v), \quad \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01} \\ \sigma_{v01} & \sigma_{v1}^2 \end{bmatrix}
\]

\[
\text{Cov}(y_{skl}, y_{rkl}) = \begin{cases} -\pi_{sijkl} \pi_{rijkl} / \text{total}_{jkl} & s \neq r \\ \pi_{sijkl}(1 - \pi_{sijkl}) / \text{total}_{jkl} & s = r \end{cases}
\]

\[
\text{Var}(\text{nonwhite}_{ijkl} | \pi_{ijkl}) = \pi_{ijkl}(1 - \pi_{ijkl}) / \text{total}_{ijkl}
\]

Adding other covariates

\[
\text{nonwhite}_{ijk} \sim \text{Bin}(\text{total}_{ijk}, \pi_{ijk})
\]

\[
\logit(\pi_{ijk}) = \beta_{0jk} + \beta_{1jk} x_{ijk}
\]

\[
\beta_{0jk} = \beta_0 + v_0k + u_{0jk}
\]

\[
\beta_{1jk} = \beta_1 + v_1k + u_{1jk}
\]

\[
\begin{bmatrix} v_0k \\ v_{1k} \end{bmatrix} \sim \mathcal{N}(0, \Omega_v), \quad \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01} \\ \sigma_{v01} & \sigma_{v1}^2 \end{bmatrix}
\]

\[
\begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim \mathcal{N}(0, \Omega_u), \quad \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix}
\]

\[
\text{Var}(\text{nonwhite}_{ijk} | \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk}) / \text{total}_{ijk}
\]
We now present some preliminary results from analyses using the model based approach:

- We used PLASC/ NPD data for 2002 to 2008.
- We looked at segregation by ethnicity over time.
- Ethnicity measured using three categories for simplicity (White, Black, Asian).
  - Asian includes Chinese.
  - All other ethnicities dropped from data.
- Response is proportion in each ethnic category in the cohort entering the school each year (not proportion in the whole school).
- Looking at secondary schools (so each cohort is age 11).
- Include schools in England only; the subset of the schools DCSF draws up league tables for which have min intake age 11.
- Drop cohorts with very few students or big change in proportions (> 25 percentage points).
Sample description

Number of units
146 LEAs
3,176 schools
3,552,319 pupils

Schools and LEAs
Mean number of schools per LEA is 22; maximum 103
minimum 1

Cohorts
- Mean cohort size is 170; maximum 705, minimum 15
- Some schools don’t have all cohorts 2002-2008
- Some schools have cohorts entirely of one ethnicity

Percentages of total sample

<table>
<thead>
<tr>
<th></th>
<th>Non</th>
<th>FSM</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>76</td>
<td>13</td>
<td>89</td>
</tr>
<tr>
<td>Black</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Asian</td>
<td>5</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>84</td>
<td>16</td>
<td>100</td>
</tr>
</tbody>
</table>

Proportion in each cohort

<table>
<thead>
<tr>
<th></th>
<th>Min</th>
<th>Mean</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>0</td>
<td>0.88</td>
<td>1</td>
</tr>
<tr>
<td>Black</td>
<td>0</td>
<td>0.05</td>
<td>1</td>
</tr>
<tr>
<td>Asian</td>
<td>0</td>
<td>0.08</td>
<td>1</td>
</tr>
<tr>
<td>FSM</td>
<td>0</td>
<td>0.17</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Results

\[ \text{proportion}_{ijkl} \sim \text{Multinomial}(\text{total}_{jkl}, \pi_{ijkl}) \]

\[
\log(\pi_{2jkl} / \pi_{1jkl}) = \beta_{0kl} + \beta_{2kl}(\text{cohort-2002})_{jkl}
\]

\[
\log(\pi_{3jkl} / \pi_{1jkl}) = \beta_{1kl} + \beta_{3kl}(\text{cohort-2002})_{jkl}
\]

\[
\begin{align*}
\beta_{0kl} &= -4.573(0.161) + f_{0l} + v_{0kl} \\
\beta_{1kl} &= -3.815(0.027) + f_{1l} + v_{1kl} \\
\beta_{2kl} &= 0.081(0.007) + f_{2l} + v_{2kl} \\
\beta_{3kl} &= 0.103(0.004) + f_{3l} + v_{3kl}
\end{align*}
\]

\[
\begin{bmatrix}
    f_{0l} \\
    f_{1l} \\
    f_{2l} \\
    f_{3l}
\end{bmatrix}
\sim N(0, \Omega_f), \quad \Omega_f =
\begin{bmatrix}
    5.161(0.635) & 2.946(0.372) \\
    3.380(0.452) & 0.005(0.001) \\
    -0.005(0.017) & -0.008(0.006) \\
    -0.012(0.008) & 0.000(0.000)
\end{bmatrix}
\]

\[
\begin{bmatrix}
    v_{0kl} \\
    v_{1kl} \\
    v_{2kl} \\
    v_{3kl}
\end{bmatrix}
\sim N(0, \Omega_v), \quad \Omega_v =
\begin{bmatrix}
    1.244(0.044) & 2.037(0.061) \\
    0.976(0.041) & 0.010(0.001) \\
    -0.022(0.004) & -0.032(0.004) \\
    -0.006(0.003) & 0.004(0.000)
\end{bmatrix}
\]

\[
\text{Cov}(y_{sjkl}, y_{rjkl}) = \begin{cases} 
-\pi_{sjkl} \pi_{rjkl} / \text{total}_{jkl} & s \neq r \\
\pi_{sjkl} (1 - \pi_{sjkl}) / \text{total}_{jkl} & s = r
\end{cases}
\]
Segregation over time

Year

Variance

LA level

School level

Black

Asian
Modelling segregation by ethnicity and FSM

**Long version**

\[ \text{proportion}_{ijkl} \sim \text{Multinomial}(\text{total}_{jkl}, \pi_{ijkl}) \]

\[
\begin{align*}
\log(\pi_{1jkl}/\pi_{0jkl}) &= \beta_{0kl} \\
\log(\pi_{2jkl}/\pi_{0jkl}) &= \beta_{1kl} \\
\log(\pi_{3jkl}/\pi_{0jkl}) &= \beta_{2kl} \\
\log(\pi_{4jkl}/\pi_{0jkl}) &= \beta_{0kl} + \beta_{2kl} \\
\log(\pi_{5jkl}/\pi_{0jkl}) &= \beta_{1kl} + \beta_{2kl} \\
\end{align*}
\]

\[ \begin{bmatrix} f_0l \\ f_1l \\ f_2l \end{bmatrix} \sim N(0, \Omega_f), \quad \Omega_f = \begin{bmatrix} \sigma_{f0}^2 & \sigma_{f1}^2 & \sigma_{f2}^2 \\ \sigma_{f01} & \sigma_{f1} & \sigma_{f2}^2 \\ \sigma_{f02} & \sigma_{f12} & \sigma_{f2}^2 \end{bmatrix} \]

\[ \begin{bmatrix} v_{0kl} \\ v_{1kl} \\ v_{2kl} \end{bmatrix} \sim N(0, \Omega_v), \quad \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v1}^2 & \sigma_{v2}^2 \\ \sigma_{v01} & \sigma_{v1} & \sigma_{v2}^2 \\ \sigma_{v02} & \sigma_{v12} & \sigma_{v2}^2 \end{bmatrix} \]

\[ \text{Cov}(y_{sjkl}, y_{rjkl}) = \begin{cases} 
-\pi_{sjkl}\pi_{rjkl}/\text{total}_{jkl} & s \neq r \\
\pi_{sjkl}(1 - \pi_{sjkl})/\text{total}_{jkl} & s = r 
\end{cases} \]

where Black(i) = 1 for response categories BlackNonFSM and BlackFSM and 0 for the other categories and similarly for Asian(i) and FSM(i)

**Condensed version**

\[ \text{proportion}_{ijkl} \sim \text{Multinomial}(\text{total}_{jk}, \pi_{ijkl}) \]

\[
\begin{align*}
\log(\pi_{ijkl}/\pi_{ijkl}) &= \beta_{0kl} \text{Black}(i) + \beta_{1kl} \text{Asian}(i) + \beta_{2kl} \text{FSM}(i) \\
\beta_{0kl} &= \beta_0 + f_{0l} + v_{0kl} \\
\beta_{1kl} &= \beta_1 + f_{1l} + v_{1kl} \\
\beta_{2kl} &= \beta_2 + f_{2l} + v_{2kl} \\
\end{align*}
\]

\[ \begin{bmatrix} f_{0l} \\ f_{1l} \\ f_{2l} \end{bmatrix} \sim N(0, \Omega_f), \quad \Omega_f = \begin{bmatrix} \sigma_{f0}^2 & \sigma_{f1}^2 & \sigma_{f2}^2 \\ \sigma_{f01} & \sigma_{f1} & \sigma_{f2}^2 \\ \sigma_{f02} & \sigma_{f12} & \sigma_{f2}^2 \end{bmatrix} \]

\[ \begin{bmatrix} v_{0kl} \\ v_{1kl} \\ v_{2kl} \end{bmatrix} \sim N(0, \Omega_v), \quad \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v1}^2 & \sigma_{v2}^2 \\ \sigma_{v01} & \sigma_{v1} & \sigma_{v2}^2 \\ \sigma_{v02} & \sigma_{v12} & \sigma_{v2}^2 \end{bmatrix} \]

\[ \text{Cov}(y_{sjkl}, y_{rjkl}) = \begin{cases} 
-\pi_{sjkl}\pi_{rjkl}/\text{total}_{jkl} & s \neq r \\
\pi_{sjkl}(1 - \pi_{sjkl})/\text{total}_{jkl} & s = r 
\end{cases} \]

We have exactly the same pattern of coefficients for time
Modelling segregation by ethnicity and FSM

Long version

\[ \text{proportion}_{ijkl} \sim \text{Multinomial}(\text{total}_{jkl}, \pi_{ijkl}) \]

\[ \log(\pi_{1jkl}/\pi_{0jkl}) = \beta_{0kl} \]
\[ \log(\pi_{2jkl}/\pi_{0jkl}) = \beta_{1kl} \]
\[ \log(\pi_{3jkl}/\pi_{0jkl}) = \beta_{2kl} \]
\[ \log(\pi_{4jkl}/\pi_{0jkl}) = \beta_{0kl} + \beta_{2kl} \]
\[ \log(\pi_{5jkl}/\pi_{0jkl}) = \beta_{1kl} + \beta_{2kl} \]

\[ \beta_{0kl} = \beta_0 + f_{0l} + v_{0kl} \]
\[ \beta_{1kl} = \beta_1 + f_{1l} + v_{1kl} \]
\[ \beta_{2kl} = \beta_2 + f_{2l} + v_{2kl} \]

\[ \begin{bmatrix} f_{0l} \\ f_{1l} \\ f_{2l} \end{bmatrix} \sim N(0, \Omega_f), \quad \Omega_f = \begin{bmatrix} \sigma_{f0}^2 & \sigma_{f01}^2 & \sigma_{f02}^2 \\ \sigma_{f01}^2 & \sigma_{f1}^2 & \sigma_{f12}^2 \\ \sigma_{f02}^2 & \sigma_{f12}^2 & \sigma_{f2}^2 \end{bmatrix} \]

\[ \begin{bmatrix} v_{0kl} \\ v_{1kl} \\ v_{2kl} \end{bmatrix} \sim N(0, \Omega_v), \quad \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01}^2 & \sigma_{v02}^2 \\ \sigma_{v01}^2 & \sigma_{v1}^2 & \sigma_{v12}^2 \\ \sigma_{v02}^2 & \sigma_{v12}^2 & \sigma_{v2}^2 \end{bmatrix} \]

\[ \text{Cov}(y_{sjkl}, y_{rjkl}) = \begin{cases} -\pi_{sjkl}\pi_{rjkl}/\text{total}_{jkl} & s \neq r \\ \pi_{sjkl}(1 - \pi_{sjkl})/\text{total}_{jkl} & s = r \end{cases} \]

Condensed version

\[ \text{proportion}_{ijkl} \sim \text{Multinomial}(\text{total}_{jk}, \pi_{ijkl}) \]

\[ \log(\pi_{ijkl}/\pi_{ijkl}) = \beta_{0kl}\text{Black}_i + \beta_{1kl}\text{Asian}_i + \beta_{2kl}\text{FSM}_i \]
\[ + \beta_{5kl}\text{Black.cohort}_i + \beta_{6kl}\text{Asian.cohort}_i \]
\[ + \beta_{7kl}\text{FSM.cohort}_i \]

where Black\((i)\) = 1 for response categories BlackNonFSM and BlackFSM and 0 for the other categories and similarly for Asian\((i)\) and FSM\((i)\)

We have exactly the same pattern of coefficients for time
Segregation over time

<table>
<thead>
<tr>
<th>Year</th>
<th>LA level</th>
<th>School level</th>
<th>BlackNon</th>
<th>AsianNon</th>
<th>WhiteFSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Segregation over time

<table>
<thead>
<tr>
<th>Year</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td></td>
</tr>
<tr>
<td>2003</td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td></td>
</tr>
<tr>
<td>2005</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td></td>
</tr>
</tbody>
</table>

LA level
School level
BlackNon
AsianNon
WhiteFSM
BlackFSM
AsianFSM
Segregation over time

Year

Variance

LA level
School level
BlackNon
AsianNon
WhiteFSM
BlackFSM
AsianFSM
Further work to be done

Check appropriateness of model
- Check assumption that there are no interaction effects between ethnicity and FSM status
- Is it problematic that some schools never have any students in some response categories?
- Should we be fitting time (cohort) as a polynomial?

Check results sensible
- Check have run MCMC long enough
- Check predicted confidence intervals for proportions

Check robustness
- Compare results to models fitted to each cohort separately
- Compare results to model fitted to selected LEAs
- Check sensitivity of results to definition of ethnic categories
Including time

\[ \text{nonwhite}_{ijk} \sim \text{Bin}(\text{total}_{ijk}, \pi_{ijk}) \]

\[ \text{logit}(\pi_{ijk}) = \beta_{0jk} + \beta_{1jk} \text{cohort}_{ijk} \]

\[ \beta_{0jk} = \beta_0 + v_0 + u_{0jk} \]

\[ \beta_{1jk} = \beta_1 + v_1 + u_{1jk} \]

\[
\begin{bmatrix}
  v_{0k} \\
  v_{1k}
\end{bmatrix} \sim \text{N}(0, \Omega_{\nu}), \quad \Omega_{\nu} = \begin{bmatrix}
  \sigma_{v0}^2 \\
  \sigma_{v01} & \sigma_{v1}^2
\end{bmatrix}
\]

\[
\begin{bmatrix}
  u_{0jk} \\
  u_{1jk}
\end{bmatrix} \sim \text{N}(0, \Omega_{u}), \quad \Omega_{u} = \begin{bmatrix}
  \sigma_{u0}^2 \\
  \sigma_{u01} & \sigma_{u1}^2
\end{bmatrix}
\]

\[ \text{Var} (\text{nonwhite}_{ijk} | \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk}) / \text{total}_{ijk} \]

- **Cohort** was already specified as a level
- We add cohort as an explanatory variable, including
  - a fixed effect (lets the overall proportion change with time)
  - a random effect at the LEA level (allows changing segregation)
  - a random effect at the school level (allows changing segregation)
- Other options are to put in a polynomial or set of dummies
- Can mix and match e.g. dummies in fixed part but linear in random part
Including time

\[ \text{nonwhite}_{ijk} \sim \text{Bin}(\text{total}_{ijk}, \pi_{ijk}) \]
\[ \logit(\pi_{ijk}) = \beta_{0jk} + \beta_{1jk} \text{cohort}_{ijk} \]
\[ \beta_{0jk} = \beta_0 + v_{0k} + u_{0jk} \]
\[ \beta_{1jk} = \beta_1 + v_{1k} + u_{1jk} \]

\[ \begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim \mathcal{N}(0, \Omega_v), \quad \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01} \\ \sigma_{v01} & \sigma_{v1}^2 \end{bmatrix} \]

\[ \begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} \sim \mathcal{N}(0, \Omega_u), \quad \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01} \\ \sigma_{u01} & \sigma_{u1}^2 \end{bmatrix} \]

\[ \text{Var}(\text{nonwhite}_{ijk} | \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk}) / \text{total}_{ijk} \]

- **Cohort** was already specified as a level
- We add **cohort** as an explanatory variable, including
  - a fixed effect (lets the overall proportion change with time)
  - a random effect at the LEA level (allows changing segregation)
  - a random effect at the school level (allows changing segregation)
- Other options are to put in a polynomial or set of dummies
- Can mix and match e.g. dummies in fixed part but linear in random part
Including time

\[
\begin{align*}
\text{nonwhite}_{ijk} & \sim \text{Bin}(\text{total}_{ijk}, \pi_{ijk}) \\
\logit(\pi_{ijk}) &= \beta_{0jk} + \beta_{1jk}\text{cohort}_{ijk} \\
\beta_{0jk} &= \beta_0 + \nu_{0k} + u_{0jk} \\
\beta_{1jk} &= \beta_1 + \nu_{1k} + u_{1jk} \\
\begin{bmatrix} \nu_{0k} \\ \nu_{1k} \end{bmatrix} &\sim \mathcal{N}(0, \Omega_\nu), \quad \Omega_\nu = \begin{bmatrix} \sigma_{\nu 0}^2 & \sigma_{\nu 01} \\ \sigma_{\nu 01} & \sigma_{\nu 1}^2 \end{bmatrix} \\
\begin{bmatrix} u_{0jk} \\ u_{1jk} \end{bmatrix} &\sim \mathcal{N}(0, \Omega_u), \quad \Omega_u = \begin{bmatrix} \sigma_{u 0}^2 & \sigma_{u 01} \\ \sigma_{u 01} & \sigma_{u 1}^2 \end{bmatrix} \\
\text{Var}(\text{nonwhite}_{ijk} | \pi_{ijk}) &= \pi_{ijk}(1 - \pi_{ijk})/\text{total}_{ijk}
\end{align*}
\]

- **Cohort** was already specified as a level
- We add **cohort** as an explanatory variable, including
  - a fixed effect (lets the overall proportion change with time)
  - a random effect at the LEA level (allows changing segregation)
  - a random effect at the school level (allows changing segregation)
- Other options are to put in a polynomial or set of dummies
- Can mix and match e.g. dummies in fixed part but linear in random part
Including time

\[
\text{nonwhite}_{ijk} \sim \text{Bin}(\text{total}_{ijk}, \pi_{ijk})
\]

\[
\logit(\pi_{ijk}) = \beta_{0jk} + \beta_{1jk} \text{cohort}_{ijk}
\]

\[
\beta_{0jk} = \beta_0 + v_0k + u_{0jk}
\]

\[
\beta_{1jk} = \beta_1 + v_{1k} + u_{1jk}
\]

\[
\begin{bmatrix}
 v_{0k} \\
 v_{1k}
\end{bmatrix} \sim \text{N}(0, \Omega_v), \quad \Omega_v = \begin{bmatrix}
 \sigma^2_{v0} & \sigma^2_{v01} \\
 \sigma^2_{v01} & \sigma^2_{v1}
\end{bmatrix}
\]

\[
\begin{bmatrix}
 u_{0jk} \\
 u_{1jk}
\end{bmatrix} \sim \text{N}(0, \Omega_u), \quad \Omega_u = \begin{bmatrix}
 \sigma^2_{u0} & \sigma^2_{u01} \\
 \sigma^2_{u01} & \sigma^2_{u1}
\end{bmatrix}
\]

\[
\text{Var(\text{nonwhite}_{ijk}|\pi_{ijk})} = \pi_{ijk}(1 - \pi_{ijk})/\text{total}_{ijk}
\]

- **Cohort** was already specified as a level
- We add **cohort** as an explanatory variable, including
  - a fixed effect (lets the overall proportion change with time)
  - a random effect at the LEA level (allows changing segregation)
  - a random effect at the school level (allows changing segregation)
- Other options are to put in a polynomial or set of dummies
- Can mix and match e.g. dummies in fixed part but linear in random part
Including time

- **Cohort** was already specified as a level
- We add **cohort** as an explanatory variable, including
  - a fixed effect (lets the overall proportion change with time)
  - a random effect at the LEA level (allows changing segregation)
  - a random effect at the school level (allows changing segregation)
- Other options are to put in a polynomial or set of dummies
- Can mix and match e.g. dummies in fixed part but linear in random part
Including other covariates

\[ \text{nonwhite}_{ijk} \sim \text{Bin}(\text{total}_{ijk}, \pi_{ijk}) \]

\[ \text{logit}(\pi_{ijk}) = \beta_{0jk} + \beta_{1jk}x_{ijk} \]

\[ \beta_{0jk} = \beta_0 + v_{0k} + u_{0jk} \]

\[ \beta_{1jk} = \beta_1 + v_{1k} + u_{1jk} \]

\[ \begin{bmatrix} v_{0k} \\ v_{1k} \end{bmatrix} \sim N(0, \Omega_v), \quad \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01}^2 \\ \sigma_{v01}^2 & \sigma_{v1}^2 \end{bmatrix} \]

\[ \begin{bmatrix} u_{0ijk} \\ u_{1ijk} \end{bmatrix} \sim N(0, \Omega_u), \quad \Omega_u = \begin{bmatrix} \sigma_{u0}^2 & \sigma_{u01}^2 \\ \sigma_{u01}^2 & \sigma_{u1}^2 \end{bmatrix} \]

\[ \text{Var} (\text{nonwhite}_{ijk} | \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk})/\text{total}_{ijk} \]

Examples

- Is there more school segregation in LEAs with greater levels of deprivation? → add IMD or IDACI
- Is the segregation such that the more ethnically diverse schools are also the poorer quality schools? → include measure of school quality
- How much segregation remains after controlling for differences in intake ability of pupils? → add pupils’ prior achievement

Exactly same form as model adding time

Can add covariates at individual, cohort, school or LEA level

Covariates can be continuous or categorical
Extending to more response categories

\[ \text{proportion}_{ijkl} \sim \text{Multinomial}(\text{total}_{jkl}, \pi_{ijkl}) \]

\[
\log(\pi_{2jkl}/\pi_{1jkl}) = \beta_{0kl} \\
\log(\pi_{3jkl}/\pi_{1jkl}) = \beta_{1kl}
\]

\[
\beta_{0kl} = \beta_0 + f_{0l} + v_{0kl} \\
\beta_{1kl} = \beta_1 + f_{1l} + v_{1kl}
\]

\[
\begin{bmatrix} f_{0l} \\ f_{1l} \end{bmatrix} \sim N(0, \Omega_f), \quad \Omega_f = \begin{bmatrix} \sigma_{f0}^2 & \sigma_{f01}^2 \\ \sigma_{f01}^2 & \sigma_{f1}^2 \end{bmatrix}
\]

\[
\begin{bmatrix} v_{0kl} \\ v_{1kl} \end{bmatrix} \sim N(0, \Omega_v), \quad \Omega_v = \begin{bmatrix} \sigma_{v0}^2 & \sigma_{v01}^2 \\ \sigma_{v01}^2 & \sigma_{v1}^2 \end{bmatrix}
\]

\[
\text{Cov}(y_{sjkl}, y_{rjkl}) = \begin{cases} 
-\pi_{sjkl}\pi_{rjkl}/\text{total}_{jkl} & s \neq r \\
\pi_{sjkl}(1-\pi_{sjkl})/\text{total}_{jkl} & s = r
\end{cases}
\]

- We add response categories by moving to a multinomial model
- Each category has a separate intercept and a separate variance
- So we have a separate measure of segregation for each category

Our segregation measures are now the variances of the log odds for the respective categories
We pick a reference category: we are measuring segregation of the other categories from this category
We also estimate covariances between the log odds for each pair of categories
(In theory,) can have as many response categories as we want
Testing assumptions

No interaction effects

Full (saturated) model is

\[
\log(\frac{\pi_{ijkl}}{\pi_{ijkl}}) = \beta_{0kl}Black_i + \beta_{1kl}Asian_i + \beta_{2kl}FSM_i \\
+ \beta_{3kl}Black.FSM_i + \beta_{4kl}Asian.FSM_i \\
+ \beta_{5kl}Black.cohort_i + \beta_{6kl}Asian.cohort_i + \beta_{7kl}FSM.cohort_i \\
+ \beta_{8kl}Black.FSM.cohort_i + \beta_{9kl}Asian.FSM.cohort_i
\]

Need to check all the extra fixed and random effects in this model are not important

Schools with zero proportions

If school \( k \) in LEA \( l \) never has any students who fall into response category \( i \), then for all cohorts \( j \)

\[
\pi_{ijkl} = 0
\]

\[
\Rightarrow \log(\frac{\pi_{ijkl}}{\pi_{0jkl}}) = \log(0) = -\infty
\]

Therefore perhaps we need to fit a mixture model.