Random Intercepts: Matrices

Paul E. Johnson\textsuperscript{1} \hspace{1cm} \textsuperscript{2}

\textsuperscript{1}Department of Political Science

\textsuperscript{2}Center for Research Methods and Data Analysis, University of Kansas

2015
Outline

1. Orientation: My Narrative on Random Intercepts
2. The Incidence Matrix
3. Variance-Covariance
   - Look at $Var(b)$ First
   - Individual Error Term $Var(\varepsilon)$
4. Matrix Tidbits
Outline

1. Orientation: My Narrative on Random Intercepts

2. The Incidence Matrix

3. Variance-Covariance
   - Look at $Var(b)$ First
   - Individual Error Term $Var(\varepsilon)$

4. Matrix Tidbits
My narrative of the multi-level model

Consider the usual “one error term” regression. A one level model

$$y_j = \beta_0 + \beta_1 X_{1j} + \epsilon_j, \; j = 1, \ldots, N$$

Maybe you add more predictors

$$y_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} + \ldots + \epsilon_j, \; j = 1, \ldots, N$$

Nothing surprising here, except that I’m using $j$ to index the data rows observations.
A Regression Genie enters the picture

- The regression genie comes along and adds an additional error term, \( b_i \).
- There are \( M \) different values of this new error term, which enter the picture like so:

\[
\begin{bmatrix}
  y_1 \\
  y_2 \\
  y_3 \\
  y_4 \\
  y_5 \\
  y_6 \\
  \vdots \\
  y_N \\
\end{bmatrix} = \begin{bmatrix} \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \vdots \\
  \beta_0
\end{bmatrix} + \begin{bmatrix}
  X_{11}\beta_1 \\
  X_{12}\beta_1 \\
  X_{13}\beta_1 \\
  X_{14}\beta_1 \\
  X_{15}\beta_1 \\
  X_{16}\beta_1 \\
  \vdots \\
  X_{1N}\beta_1
\end{bmatrix} + \cdots + \begin{bmatrix} \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \epsilon_4 \\
  \epsilon_5 \\
  \epsilon_6 \\
  \vdots \\
  \epsilon_N
\end{bmatrix} + \begin{bmatrix} b_1 \\
  b_1 \\
  b_1 \\
  b_2 \\
  b_2 \\
  b_2 \\
  \vdots \\
  b_M
\end{bmatrix}
\]  

- The new error \( b_i \) is same within groups of cases.
- This new error is sometimes called a “variance component”.
A Regression Genie enters the picture ...

- $b_1$ added to rows 1-3, $b_2$ added to rows 4-6, and so forth.
- We still assert that $X_1_i$ are “fixed”, the first column of errors $\epsilon_i$ are independent of the $b_i$.
- Two interesting interpretations, both helpful
  - Variance Components Interpretation: We have a composite error:
    
    $e_j = \epsilon_j + b_i$

    
    $y_j = \beta_0 + X_1_j \beta_1 + (\epsilon_j + b_i)$

  - Intercepts-as-outcomes: $b_i$ have the effect of creating different regression intercepts:

    
    $y_j = (\beta_0 + b_i) + X_1_j \beta_1 + \epsilon_j$
Outline

1. Orientation: My Narrative on Random Intercepts

2. The Incidence Matrix

3. Variance-Covariance
   - Look at $Var(b)$ First
   - Individual Error Term $Var(\varepsilon)$

4. Matrix Tidbits
Matrix view of $b_i$

When I write this:

$$\begin{bmatrix} b_1 \\ b_1 \\ b_1 \\ b_2 \\ b_2 \\ \vdots \\ b_M \end{bmatrix}$$

most methodologists will say “HOW GAUCHE”

We are closer to the research literature if we write

$$Z b = \begin{bmatrix} 1 & 0 & \cdots \\ 1 & 0 & \cdots \\ 0 & 1 & \cdots \\ 0 & 1 & \cdots \\ 0 & 0 & \cdots \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \end{bmatrix}$$
**Z: Design (or Incidence) Matrix**

\[ Z = \begin{bmatrix}
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 \\
0 & 1 & \ddots & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & \ddots & \ddots \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 
\end{bmatrix} \]

- **Z** is \( N \times M \) (one row per data row, one column for each row grouping).
- Only one non-zero element per row. The 1 indicates “membership” in column.
- Much of the pioneering work in mixed models was done by Charles Henderson, a researcher in animal breeding research. In that field, many scholars refer to **Z** as the “Incidence Matrix”.
The matrix algebra for a random effects model

- Notice that it looks like an R “model matrix” for a predictor variable that is a factor.

```r
x <- gl(5, 3, labels = LETTERS[1:5])
model.matrix(~ x - 1)
```

<table>
<thead>
<tr>
<th></th>
<th>xA</th>
<th>xB</th>
<th>xC</th>
<th>xD</th>
<th>xE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
attr("assign")
The matrix algebra for a random effects model ...

\[
\begin{bmatrix}
1 & 1 & 1 & 1 & 1 \\
\text{attr(, "contrasts")}
\end{bmatrix}
\]

attr(, "contrasts")$x$

[1] "contr.treatment"
The matrix algebra for a random effects model

\[ y = X\beta + Zb + \epsilon \]  

(2)

- This is the “state of the art” description of the model (Laird and Ware, Bates, etc).
- Notational variations you might find
  - The letter \( u \) instead of \( b \)
  - The letter \( \alpha \) instead of \( \beta \)
- Some authors will write out the model for the separate groups. Depending on the structure of \( Z \), this may make for a more compact presentation (Laird and Ware).

\[ y_i = X_i\beta + Z_ib_i + \epsilon_i, \text{i varies across groups} \]  

(3)

- \( y_i \) and \( \epsilon_i \) are VECTORS, with \( n_i \) observations
- \( X_i \) is an \( n_i \times p \) matrix, \( \beta \) is \( p \times 1 \)
The matrix algebra for a random effects model ...

- Need a formal statement of the variance-covariance matrices for $Zb$ and $\epsilon$.
  - Literature not uniform on symbols. Will deal with that later.
If you don’t want to use matrices, look what you have to do

- List $i$ groups of observations (respondents, whatever a “row” might be).
- $n_i$ is the number of people in the $i$’th group.
- Re-label the outcome variable $y_{ij}$, so now $i$ counts up from 1 to $n_i$ within each group.
- Re-label $X_{1i}$ as $X_{1ij}$, again reflecting the fact that each group & individual respondent is unique.
- Write the regression

$$y_{ij} = X_{ij}\beta + b_i + \epsilon_{ij}, i = 1, \ldots, M, j = 1, \ldots, n_i$$

(4)
If you don’t want to use matrices, look what you have to do ...

- My usual tendency is to think of $i$ as the subscript for individual rows, and $j$ is for groups. A few books follow that tradition, but, by far, the large majority of the authors in this area use $i$ for the grouping variable, and $j$ for the individual scores. So I’m changing notation.

- When he visited KU in November, 2013, Professor Bates recommended we avoid the “subscript fest” that ensues when we try to write the next phase without using matrix algebra.
Writing subscripts into the line-by-line view of regression

\[
\begin{bmatrix}
y_{11} \\
y_{12} \\
y_{13} \\
y_{21} \\
y_{22} \\
y_{23} \\
\vdots \\
y_{mnj}
\end{bmatrix}
= 
\begin{bmatrix}
\beta_0 \\
\beta_0 \\
\beta_0 \\
\beta_0 \\
\beta_0 \\
\beta_0 \\
\vdots \\
\beta_0
\end{bmatrix}
+ 
\begin{bmatrix}
X_{11} \beta_1 \\
X_{12} \beta_1 \\
X_{13} \beta_1 \\
X_{21} \beta_1 \\
X_{22} \beta_1 \\
X_{23} \beta_1 \\
\vdots \\
X_{mnj} \beta_1
\end{bmatrix}
+ \cdots + 
\begin{bmatrix}
\epsilon_{11} \\
\epsilon_{12} \\
\epsilon_{13} \\
\epsilon_{21} \\
\epsilon_{22} \\
\epsilon_{23} \\
\vdots \\
\epsilon_{mnj}
\end{bmatrix}
+ 
\begin{bmatrix}
b_1 \\
b_1 \\
b_1 \\
b_2 \\
b_2 \\
b_2 \\
\vdots \\
b_m
\end{bmatrix}
\] (5)
Re-arrange the $b_i$ as separate columns

\[
\begin{bmatrix}
  y_{11} \\
  y_{12} \\
  y_{13} \\
  y_{21} \\
  y_{22} \\
  y_{23} \\
  \vdots \\
  y_{Mn_M}
\end{bmatrix}
= \begin{bmatrix}
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \vdots \\
  \beta_0
\end{bmatrix}
+ \begin{bmatrix}
  X_{111} \beta_1 \\
  X_{112} \beta_1 \\
  X_{113} \beta_1 \\
  X_{121} \beta_1 \\
  X_{122} \beta_1 \\
  X_{123} \beta_1 \\
  \vdots \\
  X_{MnM} \beta_1
\end{bmatrix}
+ \ldots + \begin{bmatrix}
  \epsilon_{11} \\
  \epsilon_{12} \\
  \epsilon_{13} \\
  \epsilon_{21} \\
  \epsilon_{22} \\
  \epsilon_{23} \\
  \vdots \\
  \epsilon_{MnM}
\end{bmatrix}
+ \begin{bmatrix}
  b_1 \\
  b_1 \\
  b_1 \\
  0 \\
  0 \\
  0 \\
  \vdots \\
  0
\end{bmatrix}
\]

(6)
Prepares us to write it like this

\[
\begin{bmatrix}
  y_{11} \\
  y_{12} \\
  y_{13} \\
  y_{21} \\
  y_{22} \\
  y_{23} \\
  \vdots \\
  y_{MnM}
\end{bmatrix} =
\begin{bmatrix}
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \vdots \\
  \beta_0
\end{bmatrix} +
\begin{bmatrix}
  X_{11} \beta_1 \\
  X_{12} \beta_1 \\
  X_{13} \beta_1 \\
  X_{21} \beta_1 \\
  X_{22} \beta_1 \\
  X_{23} \beta_1 \\
  \vdots \\
  X_{MnM} \beta_1
\end{bmatrix} + \ldots +
\begin{bmatrix}
  \epsilon_{11} \\
  \epsilon_{12} \\
  \epsilon_{13} \\
  \epsilon_{21} \\
  \epsilon_{22} \\
  \epsilon_{23} \\
  \vdots \\
  \epsilon_{MnM}
\end{bmatrix} +
\begin{bmatrix}
  1 \\
  0 \\
  0 \\
  0 \\
  1 \\
  0 \\
  \vdots \\
  0 \\
  0 \\
  1
\end{bmatrix} \begin{bmatrix}
  b_1 \\
  b_2 \\
  b_3 \\
  \vdots \\
  b_M
\end{bmatrix}
\]
If you want another level, you need another subscript

Suppose the lowest level observed is the individual respondent. Individuals are nested within cities. Cities are nested within states. Using subscript notation, we would need $y_{ijk}$ and some rule for deciding which is which. We’d need random effects separate for each level.
If you want another level, you need another subscript

\[
\begin{bmatrix}
  y_{111} \\
  y_{112} \\
  y_{113} \\
  y_{121} \\
  y_{122} \\
  y_{123} \\
  \vdots \\
  y_{kmnj}
\end{bmatrix}
= 
\begin{bmatrix}
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \beta_0 \\
  \vdots \\
  \beta_0
\end{bmatrix}
+ 
\begin{bmatrix}
  X_{111} \beta_1 \\
  X_{112} \beta_1 \\
  X_{113} \beta_1 \\
  X_{121} \beta_1 \\
  X_{122} \beta_1 \\
  X_{123} \beta_1 \\
  \vdots \\
  X_{imnj} \beta_1
\end{bmatrix}
+ \ldots +
\begin{bmatrix}
  \epsilon_{111} \\
  \epsilon_{112} \\
  \epsilon_{113} \\
  \epsilon_{121} \\
  \epsilon_{122} \\
  \epsilon_{123} \\
  \vdots \\
  \epsilon_{kmnj}
\end{bmatrix}
\]

\[
= 
\begin{bmatrix}
  b_{11} \\
  b_{11} \\
  b_{11} \\
  0 \\
  0 \\
  0 \\
  \vdots \\
  0
\end{bmatrix}
+ 
\begin{bmatrix}
  0 \\
  0 \\
  0 \\
  b_{12} \\
  b_{12} \\
  b_{12} \\
  \vdots \\
  b_{km}
\end{bmatrix}
+ \ldots +
\begin{bmatrix}
  b_1 \\
  b_1 \\
  b_1 \\
  0 \\
  0 \\
  0 \\
  \vdots \\
  0
\end{bmatrix}
\]

The individual random effect \( e_{ijk} \), the lower-level group effect \( b_{jk} \), and the upper level group effect \( b_i \).
Join the separate Incidence Matrices side by side

Join all of the incidence matrices together into a giant block of $Z$

$$Z = \begin{bmatrix}
1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 0 & 1 \\
0 & 1 & \ddots & 0 & 1 \\
0 & 1 & 0 & 0 & 1 \\
0 & 1 & 0 & 0 & 1 \\
0 & 0 & \ddots & \ddots & 1 \\
0 & 0 & 1 & 1 & \ddots \\
0 & 0 & 1 & \ddots & \ddots \\
0 & 0 & 1 \\
0 & 0 & 1 \\
\end{bmatrix}$$
Join the separate Incidence Matrices side by side ...

And stack the random effects into one long column

\[ Zb = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & \ddots & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_{11} \\ b_{12} \\ b_{13} \\ \vdots \\ b_1 \\ \vdots \\ b_k \end{bmatrix} \] (10)
Join the separate Incidence Matrices side by side ...
Outline

1. Orientation: My Narrative on Random Intercepts
2. The Incidence Matrix
3. Variance-Covariance
   - Look at $\text{Var}(b)$ First
   - Individual Error Term $\text{Var}(\epsilon)$
4. Matrix Tidbits
Variance Components

- Looking at one row in isolation, we see the “unobserved error”, which is referred to by the letter $e$, is a combination of the error found at the individual and group levels.

\[ e_{ijk} = b_{ij} + b_j + \epsilon_{ijk} \]  \hspace{1cm} (11)

- Here, $i$ is the “state” level, $j$ is the city level, and $k$ is the individual level.

- The Variance of that composite error term is, conceivably, horrible and impossibly complicated. If you start with the “one giant stack of random effects” representation in (10), the variance-covariance matrix would be

\[ \text{Var}(b + \epsilon) = \text{Var}(b) + \text{Var}(\epsilon) + 2 \cdot \text{Cov}(b, \epsilon) \]
The art and mystery of multi-level modeling is figuring out which of those elements we are going to try to estimate, and which we will assume are 0.

In very few (almost none) of the articles, will you find any tolerance for the idea that $\text{Cov}(b, \epsilon) \neq 0$. It is almost always assumed they are uncorrelated, independent of each other.

After assuming they are uncorrelated with one another, we are still left with the possibility that $\text{Var}(b)$ and $\text{Var}(\epsilon)$ are stupendously complicated things.

We get out of jail there by making 2 sorts of simplifying assumptions.

1. Set big sections of those variance matrices to 0,
2. Assume redundant structures within the main diagonal blocks of the matrix.
Simplifying the Variance Assumptions

- Focus first on simplifying the variance assumed among the $b$'s.
- In the multi-level model, the random effects hit rows in clumps.
  - $b_1$ hits rows 1-3, $b_2$ hits rows 4-6. For 6 rows of data
  - The contribution to $y$ from that part is

$$Zb = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$
Variance of \( Zb \)

\[
\text{Var}(Zb) = Z \times \text{Var}(b) \times Z^T
\]

- By analogy, for a single random variable, \( b_1 \) (one draw from the process that creates the group level \( b_i \)), consider \( k \) a constant, then

\[
\text{Var}(k \cdot b_1) = k^2 \text{Var}(b_1) = k \text{ Var}(b_1) k \tag{12}
\]

- if \( Z \) is “balanced”, the pattern is easy to see: \( ZZ^T \) is block diagonal
Simplifying the Variance Assumptions ...

\[ ZZ^T = \begin{bmatrix}
1 & 0 & 0 \\
1 & 0 & 0 \\
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
0 & 0 & 1 \\
\end{bmatrix} \begin{bmatrix}
1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\
\end{bmatrix} \begin{bmatrix}
1 & 1 & 1 \\
1 & 1 & 1 \\
1 & 1 & 1 \\
\end{bmatrix} \]

When we calculate the Variance of $Zb$, we slide a square variance matrix into the middle of that.
Simplifying the Variance Assumptions ...

Suppose we simplify $\text{Var}(Zb)$ by assuming the errors that hit each group are independent of one another:

$$\text{Var}(b) = \begin{bmatrix} \sigma^2_{b_1} & 0 \\ 0 & \sigma^2_{b_2} \end{bmatrix}$$

Within one group of rows, the random effect is completely unrelated to the effect within another block of rows.

Just to be obnoxious about it, consider 6 rows of data (2 groups of 3 each)

$$b = \begin{bmatrix} b_1 \\ b_1 \\ b_2 \\ b_2 \\ b_2 \\ b_2 \end{bmatrix}$$
Simplifying the Variance Assumptions ...

\[ Z\text{Var}(b)Z^T = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma^2_{b_1} & 0 & 0 \\ 0 & \sigma^2_{b_2} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \sigma^2_{b_1} & \sigma^2_{b_1} & \sigma^2_{b_1} & 0 & 0 & 0 \\ \sigma^2_{b_1} & \sigma^2_{b_1} & \sigma^2_{b_1} & 0 & 0 & 0 \\ \sigma^2_{b_1} & \sigma^2_{b_1} & \sigma^2_{b_1} & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_{b_2} & \sigma^2_{b_2} & \sigma^2_{b_2} \\ 0 & 0 & 0 & \sigma^2_{b_2} & \sigma^2_{b_2} & \sigma^2_{b_2} \\ 0 & 0 & 0 & \sigma^2_{b_2} & \sigma^2_{b_2} & \sigma^2_{b_2} \end{bmatrix} \]
Simplifying the Variance Assumptions ...

- It often (almost always) happens that we don’t have any good reason to expect that the variance differs across groups. Thus, the model is simplified to have the same variance parameter within all groups:

\[
\begin{bmatrix}
\sigma_b^2 & \sigma_b^2 & \sigma_b^2 & 0 & 0 & 0 \\
\sigma_b^2 & \sigma_b^2 & \sigma_b^2 & 0 & 0 & 0 \\
\sigma_b^2 & \sigma_b^2 & \sigma_b^2 & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\
0 & 0 & 0 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\
0 & 0 & 0 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\
\end{bmatrix}
\]

- When most people talk about estimating the variance of a random effect, they have in mind the estimation of $\sigma_b^2$.

- Different symbols I’ve seen used for $\text{Var}(b)$
  - $\Psi$ Pinheiro and Bates (2000)
  - $R$ Henderson
  - In Snijder & Bosker, it is $T$ with individual variance elements $\tau_i^2$ and covariances $\tau_{ij'}$
Look at the error term column for a moment

- Why don’t we have an incidence matrix on the error term as well?
- We could. The column of ordinary errors would be written out with the identity matrix as the incidence matrix

\[
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4 \\
\epsilon_5 \\
\epsilon_6 \\
\vdots \\
\epsilon_N
\end{bmatrix}
= 
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4 \\
\epsilon_5 \\
\epsilon_6 \\
\vdots \\
\epsilon_N
\end{bmatrix}
\]

- That’s just multiplying a vector by 1, but...
  - it brings to the forefront the simplifying nature of our assumptions, if we think of the matrix on the left as a variance-covariance matrix.
Look at the error term column for a moment ...

- We’ve assumed that \( \varepsilon_1 \) is uncorrelated with \( \varepsilon_2 \), for example.
- I introduce another letter for your entertainment.
  - \( u \) is a a standard normal random variable (all \( u \) drawn from same Normal).
  - The error term in the regression is a “rescaled” version of \( u \).

\[
\begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\varepsilon_3 \\
\varepsilon_4 \\
\varepsilon_5 \\
\vdots \\
\varepsilon_N
\end{bmatrix} =
\begin{bmatrix}
\sigma_{11} & 0 & & & & \\
0 & \sigma_{22} & & & & \\
& & \sigma_{33} & & & \\
& & & \sigma_{44} & & \\
& & & & \sigma_{55} & \\
0 & 0 & & & & \sigma_{11}
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_2 \\
u_3 \\
u_4 \\
u_5 \\
\vdots \\
u_N
\end{bmatrix}
\]
Look at the error term column for a moment ...

- Ordinary least squares with homoskedastic errors implies all of the elements on the diagonal are identical

\[
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4 \\
\epsilon_5 \\
\vdots \\
\epsilon_N
\end{bmatrix} = \begin{bmatrix}
\sigma^2 & 0 & 0 \\
0 & \sigma^2 & 0 \\
0 & 0 & \sigma^2 \\
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4 \\
\epsilon_5 \\
\vdots \\
\epsilon_N
\end{bmatrix} \begin{bmatrix}
u_1 \\
v_2 \\
v_3 \\
v_4 \\
v_5 \\
\vdots \\
u_N
\end{bmatrix}
\]

- If the error draws are correlated with each other, then that matrix gets a lot more intimidating. Write in all those cells.
Look at the error term column for a moment ...

\[
\begin{bmatrix}
\epsilon_1 \\
\epsilon_2 \\
\epsilon_3 \\
\epsilon_4 \\
\epsilon_5 \\
\vdots \\
\epsilon_N \\
\end{bmatrix}
= 
\begin{bmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{1(N-1)} & \sigma_{1N} \\
\sigma_{12} & \sigma_{22} & & & & \\
\sigma_{13} & \sigma_{33} & & & & \\
\sigma_{14} & & & & & \\
\vdots & & & \sigma_{44} & & \\
\sigma_{(N-1)1} & & & \sigma_{(N-1)N} & & \\
\sigma_{N1} & \sigma_{N2} & \sigma_{N3} & \sigma_{N(N-1)} & \sigma_{NN} & \\
\end{bmatrix}
\begin{bmatrix}
u_1 \\
u_2 \\
u_3 \\
u_4 \\
u_5 \\
\vdots \\
u_N \\
\end{bmatrix}
\]
Interesting property about random variables

I realize now that I’ve forgotten to mention this.

- Suppose we want a Normal variable with a mean of 0 and variance $\sigma^2$: $\varepsilon_i \sim N(0, \sigma^2)$
- In the olden days, we had random number generators that were only giving out standard normal variables, $u_i \sim N(0, 1)$.
- So we create an $N(0, \sigma^2)$ by multiplying $\sigma \cdot u_i$
- Result: Any normally distributed variable can be produced by re-scaling a standard normal variable.
- The same is true if $u$ is a vector of draws from a standard MVN. The only difference is that the weight applied to re-scale from $u$ is a matrix, rather than a single coefficient.
Thinking about regression differently

- Suppose the garden variety model. Predictor columns named $X_1$, $X_2$, $X_3$, and so forth:

$$y_i = \beta_0 + \beta_1 X_1i + \beta_2 X_2i + \ldots + \epsilon_i$$  \hspace{1cm} (13)

- All of the error terms are drawn from the “same distribution” in which
  - the standard deviation is some constant, $\sigma_\epsilon$.
  - expected value is zero, $E[\epsilon_i] = 0$
  - We often assert $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, but that’s not strictly necessary.

Conjecture:

That model is identical to this model

$$y_i = \beta_0 + \beta_1 X_1i + \beta_2 X_2i + \ldots + \sigma_\epsilon u_i, \ u_i \sim N(0, 1)$$
Thinking about regression differently ...

- Notice what’s different? The standard deviation of the error term appears as a “scale multiplier”
- if $\sigma_{\epsilon}$ shrinks to 0, then the model degenerates to $X_i \beta$.
- We could think of the standard deviation of the error term as “just another slope coefficient” in a linear model.
- Conceptual benefit? $\sigma$ is no longer an implicit parameter that students can ignore
- Of course, $\sigma_{\epsilon}$ can’t be estimated in the same way as $\beta_j$, since don’t have observed $u_i$.
- Perhaps I mean to say that I don’t believe in regressions with error terms.
  - All of my regressions have random intercepts(?)

$$y_i = b_j + \beta_0 + \beta_1 X1_i$$
Thinking about regression differently ...

- Or perhaps all regression models are really “mixed effects” models

\[ y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \sigma \varepsilon \]

\[
\begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & \ddots & 0 \\
0 & 0 & 0 & 0 & 1 \\
\end{bmatrix}
\begin{pmatrix}
u_1 \\
u_2 \\
u_3 \\
\vdots \\
u_N \\
\end{pmatrix}
\] (14)
What’s different in \( \text{Var}(\varepsilon_{ij}) \) compared to \( \text{Var}(b) \)?

Recall that each individual row gets its own \( \varepsilon \)

- It does not matter to me if you index the rows 1, \ldots, \( N \) or if you break the index into subgroups like \( \{\varepsilon_{11}, \varepsilon_{12}, \ldots, \varepsilon_{1n_1}, \varepsilon_{21}, \varepsilon_{22}, \ldots, \varepsilon_{2n_2}, \ldots\} \). (The number within each group is \( n_i \))

- In the worst case scenario, each error term is drawn from a unique distribution and it covaries with each of the others in a unique way

\[
\begin{bmatrix}
\sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} \\
\sigma_{12} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} \\
\sigma_{13} & \sigma_{23} & \sigma_3^2 & \sigma_{34} & \sigma_{35} & \sigma_{36} \\
\sigma_{14} & \sigma_{24} & \sigma_{34} & \sigma_4^2 & \sigma_{45} & \sigma_{46} \\
\sigma_{15} & \sigma_{25} & \sigma_{35} & \sigma_{45} & \sigma_5^2 & \sigma_{56} \\
\sigma_{16} & \sigma_{26} & \sigma_{36} & \sigma_{46} & \sigma_{56} & \sigma_6^2
\end{bmatrix}
\]

- We can’t tolerate that, so assume that the errors within one group are separate from the other.
What's different in $\text{Var}(\varepsilon_{ij})$ compared to $\text{Var}(b)$? ...

$\begin{bmatrix}
\sigma_1^2 & \sigma_{12} & \sigma_{13} & 0 & 0 & 0 \\
\sigma_{12} & \sigma_2^2 & \sigma_{23} & 0 & 0 & 0 \\
\sigma_{13} & \sigma_{23} & \sigma_3^2 & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma_4^2 & \sigma_{54} & \sigma_{64} \\
0 & 0 & 0 & \sigma_{54} & \sigma_5^2 & \sigma_{56} \\
0 & 0 & 0 & \sigma_{64} & \sigma_{65} & \sigma_6^2
\end{bmatrix}$

- In the 'usual' cross sectional multi-level regression, the errors experienced by each row are uncorrelated, and then

$\begin{bmatrix}
\sigma_1^2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \sigma_2^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \sigma_3^2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma_4^2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_5^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \sigma_6^2
\end{bmatrix}$
What’s different in $\text{Var}(\varepsilon_{ij})$ compared to $\text{Var}(b)$? ...

- If we assume 1) errors across individuals don’t covary, we make a very tractable variance matrix (Weighted Least Squares).

- If the groups of rows are thought of as repeat observations on particular individuals, we don’t get that simplification. Now I have to write the error vector with 2 indexes:

$$\varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{23} \end{bmatrix}$$
What’s different in $\text{Var}(\varepsilon_{ij})$ compared to $\text{Var}(b)$? ...

- It would be a little unusual to find the assumption that the variance matrix of each group is unique

\[
\begin{pmatrix}
\sigma_{11}^2 & \sigma_{11,12} & \sigma_{11,13} & 0 & 0 & 0 \\
\sigma_{11,12} & \sigma_{12}^2 & \sigma_{23,13} & 0 & 0 & 0 \\
\sigma_{11,13} & \sigma_{23,13} & \sigma_{13}^2 & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma_{21}^2 & \sigma_{21,22} & \sigma_{21,23} \\
0 & 0 & 0 & \sigma_{21,22} & \sigma_{22}^2 & \sigma_{22,23} \\
0 & 0 & 0 & \sigma_{21,23} & \sigma_{22,23} & \sigma_{23}^2
\end{pmatrix}
\]

Note the subscripts are pairs 11, 12 is correlation between group 1, observation 1 and group 1, observation 2.

- Instead, we might assume that the variance of the errors with all groups is an identical matrix,
What’s different in $\text{Var}(\epsilon_{ij})$ compared to $\text{Var}(b)$? ...

\[
\begin{bmatrix}
\sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} & 0 & 0 & 0 \\
\sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} & 0 & 0 & 0 \\
\sigma_{1,3} & \sigma_{2,3} & \sigma_3^2 & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma_1^2 & \sigma_{1,2} & \sigma_{1,3} \\
0 & 0 & 0 & \sigma_{1,2} & \sigma_2^2 & \sigma_{2,3} \\
0 & 0 & 0 & \sigma_{1,3} & \sigma_{2,3} & \sigma_3^2
\end{bmatrix}
\]

- The Kroenecker product from matrix algebra, $\otimes$, is a compact way to represent that idea. If you name that $3 \times 3$ block $\Sigma$, that would be

\[ I \otimes \Sigma \]
Outline

1. Orientation: My Narrative on Random Intercepts
2. The Incidence Matrix
3. Variance-Covariance
   - Look at $Var(b)$ First
   - Individual Error Term $Var(\varepsilon)$
4. Matrix Tidbits
More Complicated Z’s

With a multi-effect Z matrix, pattern is more difficult to see. I’ve decided to make group A from 12 rows and 4 columns (4 sets of 3), and group B is 12 rows, but from 3 sets of 4

\[
Z \times Z^T = \begin{bmatrix}
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & 1
\end{bmatrix} \times \begin{bmatrix}
1 & 1 & 1 & 0 & 0 & \ldots \\
0 & 0 & 0 & 1 & 1 & 1 & \ldots \\
0 & 0 & 0 & 0 & 0 & 0 & \ldots \\
0 & 0 & 0 & 0 & 0 & 0 & \ldots \\
1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & \ldots \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & \ldots \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & \ldots
\end{bmatrix}
\]
More Complicated Z’s ...

- Because we joined together 2 incidence matrices, we now no longer enforce the idea that each row only has one non-zero value.
- At this point, I lose patience with writing out matrices by hand, so I’ll show you some R output of same

```r
Z <- matrix(0, ncol = 7, nrow = 12)
Z[1:3,1] <- 1
Z[4:6,2] <- 1
Z[7:9,3] <- 1
Z[10:12,4] <- 1
Z[1:4,5] <- 1
Z[5:8,6] <- 1
Z[9:12,7] <- 1
Z
```
More Complicated Z’s ...

$$Z \otimes {}^\top t(Z)$$
More Complicated Z’s ...

<table>
<thead>
<tr>
<th></th>
<th>[ ,1 ]</th>
<th>[ ,2 ]</th>
<th>[ ,3 ]</th>
<th>[ ,4 ]</th>
<th>[ ,5 ]</th>
<th>[ ,6 ]</th>
<th>[ ,7 ]</th>
<th>[ ,8 ]</th>
<th>[ ,9 ]</th>
<th>[ ,10 ]</th>
<th>[ ,11 ]</th>
<th>[ ,12 ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1 ,]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[2 ,]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[3 ,]</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[4 ,]</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[5 ,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[6 ,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[7 ,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[8 ,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>[9 ,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>[10 ,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>[11 ,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>[12 ,]</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
More Complicated Z’s …

However, we usually suppose the errors in the 2 levels are independent, they don’t have covariance, so we really only need a diagonal matrix for groups A and B, say

\[
\text{Var}(b) = \begin{bmatrix}
\sigma_{bA}^2 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \sigma_{bA}^2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \sigma_{bA}^2 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma_{bA}^2 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \sigma_{bB}^2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \sigma_{bB}^2 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \sigma_{bB}^2
\end{bmatrix}
\]

I’ll set \(\sigma_{A}^2 = 2\) and \(\sigma_{B}^2 = 33\) so we can see the effect on \(Z\text{Var}(b)Z^T\)

\[
\begin{align*}
\text{Vb} & \leftarrow \text{matrix}(0, \text{ncol} = 7, \text{nrow} = 7) \\
\text{diag(Vb)}[1:4] & \leftarrow 2 \\
\text{diag(Vb)}[5:7] & \leftarrow 33 \\
\text{Vb}
\end{align*}
\]
More Complicated Z’s ...

\[
\begin{array}{cccccccc}
[1,1] & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\
[2,1] & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\
[3,1] & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\
[4,1] & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\
[5,1] & 0 & 0 & 0 & 0 & 33 & 0 & 0 \\
[6,1] & 0 & 0 & 0 & 0 & 0 & 33 & 0 \\
[7,1] & 0 & 0 & 0 & 0 & 0 & 0 & 33 \\
\end{array}
\]

\[
\begin{array}{cccccccc}
[1,1] & 35 & 35 & 35 & 33 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
[2,1] & 35 & 35 & 35 & 33 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
[3,1] & 35 & 35 & 35 & 33 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
[4,1] & 33 & 33 & 33 & 35 & 2 & 2 & 0 & 0 & 0 & 0 & 0 \\
[5,1] & 0 & 0 & 0 & 2 & 35 & 35 & 33 & 33 & 0 & 0 & 0 \\
[6,1] & 0 & 0 & 0 & 2 & 35 & 35 & 33 & 33 & 0 & 0 & 0 \\
[7,1] & 0 & 0 & 0 & 0 & 33 & 33 & 35 & 35 & 2 & 0 & 0 \\
[8,1] & 0 & 0 & 0 & 0 & 33 & 33 & 35 & 35 & 2 & 0 & 0 \\
[9,1] & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 2 & 35 & 33 & 33 & 33 \\
[10,1] & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 33 & 35 & 35 & 35 \\
[11,1] & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 33 & 35 & 35 & 35 \\
\end{array}
\]
More Complicated Z’s ...

\[
\begin{bmatrix}
12 \\
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 33 & 35 & 35 & 35
\end{bmatrix}
\]