Linear Statistical Models

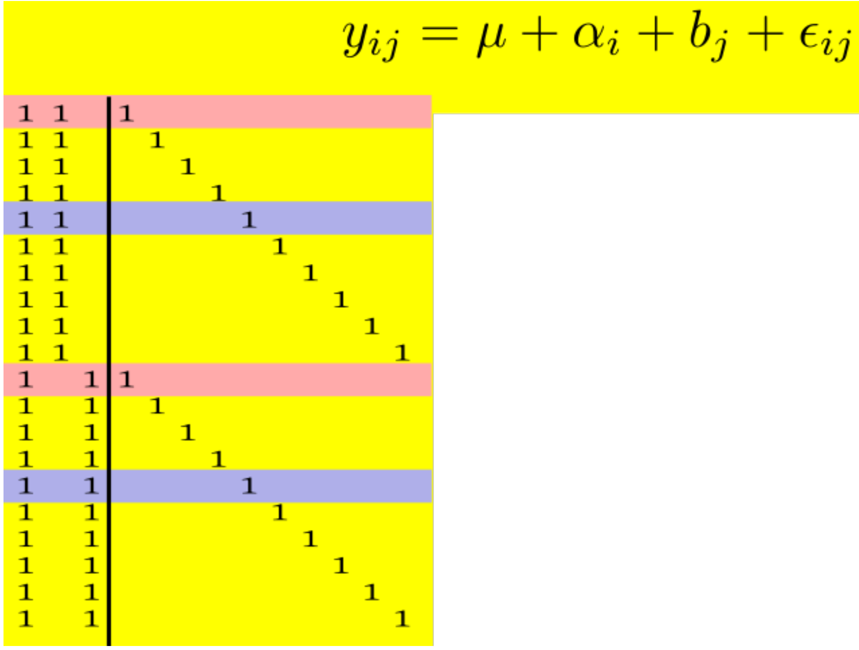
Video 145: Grouped Data

Anushka De
Roll: BS2042

To understand the concept of **grouping**, we proceed in the following way:

Importance of random effect model/ mixed effect model:
Because out of the two given inputs, one is random. In our example the *tablet* input is random. In the sense that the particular tablet used in our experiment are of no particular concern, that is the reason mixed effects model are being used.

R starts by identifying that input. In this case it is *tablet*, with subscript j. Wherever that subscript will occur the corresponding coefficient will become random.



$$y_{ij} = \mu + \alpha_i + b_j + \epsilon_{ij}$$

$$\begin{array}{ccc|cccccccccc}
1 & 1 & & 1 & & & & & & & & & \\
1 & 1 & & & 1 & & & & & & & & \\
1 & 1 & & & & 1 & & & & & & & \\
1 & 1 & & & & & 1 & & & & & & \\
1 & 1 & & & & & & 1 & & & & & \\
1 & 1 & & & & & & & 1 & & & & \\
1 & 1 & & & & & & & & 1 & & & \\
1 & 1 & & & & & & & & & 1 & & \\
1 & 1 & & & & & & & & & & 1 & \\
1 & 1 & & & & & & & & & & & 1 \\
1 & & 1 & 1 & & & & & & & & & \\
1 & & 1 & & 1 & & & & & & & & \\
1 & & 1 & & & 1 & & & & & & & \\
1 & & 1 & & & & 1 & & & & & & \\
1 & & 1 & & & & & 1 & & & & & \\
1 & & 1 & & & & & & 1 & & & & \\
1 & & 1 & & & & & & & 1 & & & \\
1 & & 1 & & & & & & & & 1 & & \\
1 & & 1 & & & & & & & & & 1 & \\
1 & & 1 & & & & & & & & & & 1
\end{array}$$

Figure 1: *Matrix with groups highlighted in same color*

In our example, there are 10 groups. For better understanding two such groups have been highlighted in red and blue respectively.
In the red group, b_1 coefficient plays a role and so both the red rows occur due to *Tablet 1*. Similarly the blue rows occur due to *Tablet 5*.

So instead of saying there are 20 rows, **R** likes to think of it as 10 pairs of numbers.

Each tablet is called a **group**. This is an example of a grouped data.

*Note that this terminology is by no means standard, it is just used by **R**.*

R suggests to look at each group separately.
Consider the red group first. The fixed part consists of the numbers $1, 1, 0$ and $1, 0, 1$ respectively. Similar thing happens for the blue part.

It can be easily observed that irrespective of the tablet chosen if we restrict our attention to only the corresponding 2 rows, we have the exact sub-matrix. In this example the sub-matrix is $\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$. The reason being that the fixed effects part does not involve the subscript j.

Consider the random effects part, it can be seen that for the red part only b_1 plays any role. So for the red group, b_2 to b_{10} are just absurd. So if we focus our attention on only one random effect the matrix is simply $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.
Note that the position of 1s in the matrix is the random effects part keeps changing but for every group, there is only one particular random effect coefficient to consider. With respect to that random effect coefficient the matrix is always $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

This is bound to be the case in every situation. Because had this not been the case it would imply separate importance is being given to different tablet, which is not desired.

So the focus is only on the $j - th$ group. Hence the sub-vector $\overrightarrow{y_j}$ will have the corresponding model:

$$\boxed{\overrightarrow{y_j} = \tilde{X}\overrightarrow{\beta} + \tilde{Z}\overrightarrow{b_j} + \overrightarrow{\epsilon_j}}$$

where, $\tilde{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$, $\tilde{Z} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and *the corresponding error vector being restricted to the subscript* j.

The larger matrices **X** and **Z** have the special structure as they are made by repeating the smaller matrices $\tilde{X}$ and $\tilde{Z}$ again and again respectively. The X-thing being repeated one below other and the Z-thing being repeated in a diagonal manner.

R suggests to model the smaller matrices $\tilde{X}$ and $\tilde{Z}$ while specifying the design matrix rather than the larger matrices **X** and **Z**. The basic advantage of this model is that **R** forces the user to write by identifying the cause behind the genesis of the random effect and then follow along that line. Whenever that particular subscript j is present, it is made *random*. This results in writing meaningful models.