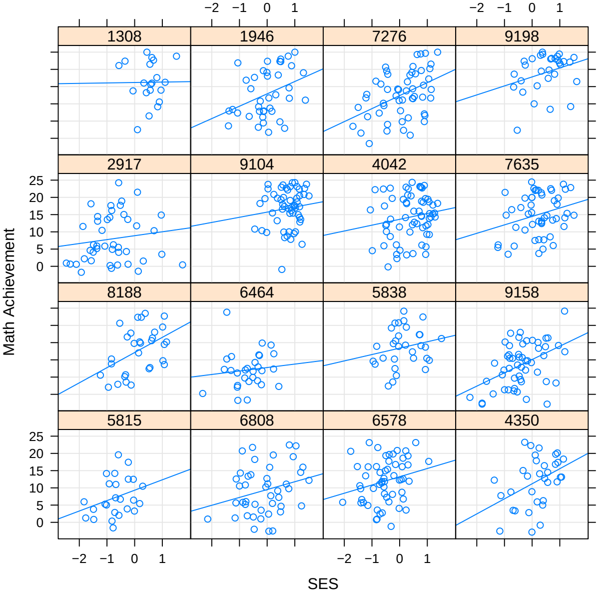
Multilevel Regression

Mark Andrews

Sept 14, 2016


1308
1946
7276
9198
2917
9104
4042
7635
8188
6464
5838
9158
5815
6808
6578
4350
Math Achievement
SES
−2
−1
0
1
0
5
10
15
20
25

Example: Multilevel model for reaction times

- Consider we have reaction time data from J subjects,

$$\{x_{j1}, x_{j2}, x_{j3} \dots x_{jn_j}\}_{j=1}^{J}.$$

- A simple multilevel model for this data might be:

$$x_{ji} \sim N(\mu_j, \sigma^2), \quad \text{for } i \in \{1 \dots n_j\},$$
$$\mu_j \sim N(\theta, \tau^2), \quad \text{for } j \in \{1 \dots J\}.$$

- In words, each x_{ji} is drawn from a Gaussian with mean μ_j and variance σ^2, and each μ_j is drawn from a Gaussian with mean θ and variance τ^2.

Example: Multilevel model for reaction times

- We can re-write $x_{ji} \sim N(\mu_j, \sigma^2)$ as

$$x_{ji} = \mu_j + \epsilon_{ji}, \quad \epsilon_{ji} \sim N(0, \sigma^2).$$

- We can re-write $\mu_j \sim N(\theta, \tau^2)$ as

$$\mu_j = \theta + \eta_j, \quad \eta_j \sim N(0, \tau^2).$$

- The multilevel model can be re-written

$$x_{ji} = \theta + \eta_j + \epsilon_{ji} \quad \epsilon_{ji} \sim N(0, \sigma^2), \eta_j \sim N(0, \tau^2).$$

- This is often termed a *random-effects* model.

Example: Multilevel model for reaction times

- In the model just described, there are three unknowns: θ, σ^2 and τ^2.
- Model estimation (fitting) estimates values for these variables.
- The variable θ denotes the global average reaction time.
- The variable σ^2 denotes the variance within any given subject.
- The variable τ^2 denotes the variance across subjects.

Example: Multilevel model for reaction times

- In the model just described, θ tells us the global average.
- The variance τ^2 tells us how much any given subject's average varies about θ.
- For example, 95% and 99% of the averages for individual subjects, will be in the ranges

$$\theta \pm 1.96 \times \tau, \quad \theta \pm 2.56 \times \tau,$$

  respectively.
- Likewise, 95% and 99% of any given subject's reaction times, i.e. x_{ji}, will be in the ranges

$$\theta + \eta_j \pm 1.96 \times \sigma, \quad \theta + \eta_j \pm 2.56 \times \sigma.$$

Example: Multilevel model for reaction times

- A model, such as the previous one, would be specified as follows in the `lmer` program in R:

  ```
  lmer(latency ~ 1 + (1|subject))
  ```

  where `latency` is reaction time, and `subject` is a categorical variable that indicates the identity of the subject.
- In other words, here we are saying that reaction time is modelled as a variation around a global average plus an individual average.

Example: Multiple drivers, multiple cars

- Let's say we want to measure the mpg of a given model of car (e.g. a Porsche 911).
- Because any one car could vary from others of the same model, we have K different examples of this model of car.
- Likewise, because any one driver could affect the recorded mpg of the car he drives, we have J different drivers.
- We get each of the J drivers to drive each of the K cars, and record the mpg as

$$y_{jk} = \text{mpg for driver } j, \text{ car } k.$$

Example: Multiple drivers, multiple cars

- A multilevel model for this mpg experiment could be

$$
\begin{aligned}
y_{jk} &\sim N(\mu_j + \nu_k, \sigma^2), \\
\mu_j &\sim N(\phi, \tau^2) \\
\nu_k &\sim N(\psi, \upsilon^2)
\end{aligned}
$$

which would work out as

$$y_{jk} = \underbrace{\theta}_{\phi+\psi} + \eta_j + \zeta_k + \epsilon_{jk},$$

with

$$\eta_j \sim N(0, \tau^2),\ \zeta_k \sim N(0, \upsilon^2),\ \epsilon_{jk} \sim N(0, \sigma^2).$$

Example: Multiple drivers, multiple cars

- In this example, we have three sources of variation

$$y_{jk} = \theta + \underbrace{\eta_j}_{\text{within driver}} + \underbrace{\zeta_k}_{\text{within car}} + \underbrace{\epsilon_{jk}}_{\text{within trial}},$$

  where τ^2 gives the within driver variance, υ^2 gives the within car variation, and σ^2 gives within trial variation.
- The variable θ provides the average mpg for the car model (i.e. the Porsche 911)
- *Your mileage may vary*: The variables τ^2, υ^2 and σ^2 provide measures of the relative variation across in mpg drivers, cars and trials, respectively.

Example: Multiple drivers, multiple cars

- The mpg model would be specified as follows in the `lmer`:

  `lmer(mpg ~ 1 + (1|driver) + (1|car))`

  where `mpg` is continuous variable, and `driver` and `car` are categorical variables that indicate the identity of the driver and car, respectively.

Example: Mathematical achievement and socioeconomic status

- In this problem, we have J schools. Within each school, we have n_j students.
- For student i in school j, their ses score is x_{ji} and their mathematical achievemnt score is y_{ji}.
- A multilevel model for this data is

$$
\begin{aligned}
y_{ji} &\sim N(\alpha_j + \beta_j x_{ji}, \sigma^2), \\
\alpha_j &\sim N(a, \tau_a^2), \\
\beta_j &\sim N(b, \tau_b^2).
\end{aligned}
$$

Example: Mathematical achievement and socioeconomic status

- The model

$$
\begin{aligned}
y_{ji} &\sim N(\alpha_j + \beta_j x_{ji}, \sigma^2),\\
\alpha_j &\sim N(a, \tau_a^2),\\
\beta_j &\sim N(b, \tau_b^2),
\end{aligned}
$$

can be re-written

$$y_{ji} = a + b x_{ji} + \eta_j + \zeta_j x_{ji} + \epsilon_{ji},$$

where

$$\eta_j \sim N(0, \tau_a^2),\ \zeta_j \sim N(0, \tau_b^2),\ \epsilon_j \sim N(0, \sigma^2).$$

Example: Mathematical achievement and socioeconomic status

- We can look at this model as

$$y_{ji} = \underbrace{a + bx_{ji}}_{\text{general model}} + \underbrace{\eta_j + \zeta_j x_{ji}}_{\text{school-level model}} + \epsilon_{ji}.$$

- For any given school, the model can be viewed as

$$y_{ji} = \underbrace{(a + \eta_j)}_{\alpha_j} + \underbrace{(b + \zeta_j)}_{\beta_j} x_{ji} + \epsilon_{ji}.$$

- In other words, any given school's regression model is a variation on a general regression model.

Example: Mathematical achievement and socioeconomic status

- In the model just described, a and b are the general regression coefficients.
- The variance τ_a^2 tells us how much variation in the intercept term there is across schools. The variance τ_b^2 tells us how much variation in the slope term there is across schools.
- For example, 95% and 99% of the intercepts for individual schools will be in the ranges

$$a \pm 1.96 \times \tau_a, \quad a \pm 2.56 \times \tau_a,$$

respectively. Likewise, 95% and 99% of the slope terms for schools will be in the ranges

$$b \pm 1.96 \times \tau_b, \quad b \pm 2.56 \times \tau_b.$$

Example: Mathematical achievement and socioeconomic status

- The ses regression model would be specified as follows in the `lmer`:

  ```
  lmer(math ~ 1 + ses + (1|school) + (0 + ses|school))
  ```

  where `math` is continuous variable, and `school` is a categorical variable that indicates the identity of the school[1]

[1] Writing `(1 + ses|school)` would imply correlated slopes and intercepts.

Why multilevel models?

- When data occurs in groups, multilevel models should always be considered.
- Multilevel models provide a macro/micro perspective on the data: The show the general pattern across all groups, and show how each individual group varies about this general pattern.
- By appropriately indentifying different sources of variation, the general patterns in the data are more accurately inferred.
- Even estimates for an individual group are improved by a process of *strength-sharing*. In other words, knowing the general pattern in the data facilitates the inference of the patterns for individual