Lecture 33

Example: the perils of ignoring group effects

group-level

$$Y_{ij} = \beta_0 + u_i + \beta_1 X_i + \varepsilon_{ij} \qquad u_i \sim N(0, \sigma_u^2)$$

$X_i = 0$ or 1 (control vs. treatment) $\qquad \varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$

$H_0: \beta_1 = 0 \qquad H_A: \beta_1 \neq 0$

$X_1, \ldots, X_{m_0}$ control $\qquad X_{m_0+1}, \ldots, X_{\underbrace{m_0+m_1}_{=m}}$ treatment

n_i = # obs in group i

$N_1 = \sum_{i=m_0+1}^{m} n_i \qquad N_0 = \sum_{i=1}^{m_0} n_i \qquad N = N_0 + N_1$

Suppose we assume (incorrectly) that

$$Y_{ij} = \beta_0 + \beta_1 X_i + \varepsilon_{ij}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \qquad \hat{\beta}_1 = \frac{\sum_i \sum_j (X_i - \bar{X})(Y_{ij} - \bar{Y})}{\sum_i \sum_j (X_i - \bar{X})^2}$$

$$\hat{\beta}_1 = \frac{\sum_i \sum_j (X_i - \bar{X})(Y_{ij} - \bar{Y})}{\sum_i \sum_j (X_i - \bar{X})^2} = \frac{\sum_i \sum_i X_i Y_{ij} - \sum_i \sum_j X_i \bar{Y}}{\sum_i \sum_j (X_i - \bar{X})^2}$$

$$\sum_i \sum_j X_i \bar{Y} = \bar{Y} \sum_i \sum_j X_i = \bar{Y} N_1$$

$$\bar{Y} = \frac{N_1 \bar{Y}_1 + N_0 \bar{Y}_0}{N}$$

$\bar{Y}_1$ = mean of Y for treatment

$\bar{Y}_0$ = mean of Y for control

$$\sum_i \sum_j X_i Y_{ij} = N_1 \bar{Y}_1$$

$$\sum_i \sum_j (X_i - \bar{X})^2 = \sum_i \sum_i X_i - N (\bar{X})^2$$

$$= N_1 - N \left(\frac{N_1}{N}\right)^2 = N_1 - \frac{N_1^2}{N} = \frac{N_1 N_0}{N}$$

$$\hat{\beta}_1 = \frac{\bar{Y}_1 N_1 - \bar{Y} N_1}{N_1 N_0 / N} = (\bar{Y}_1 - \bar{Y}) \cdot \frac{N}{N_0} = \bar{Y}_1 \cdot \frac{N}{N_0} - \bar{Y}_1 \frac{N_1}{N_0} - \bar{Y}_0$$

$$= \bar{Y}_1 - \bar{Y}_0$$

$$Var(\hat{\beta}_1) = Var(\bar{Y}_1 - \bar{Y}_0) = Var(\bar{Y}_1) + Var(\bar{Y}_0)$$

$$Var(\hat{\beta}_1) = Var(\bar{Y}_1 - \bar{Y}_0) = Var(\bar{Y}_1) + Var(\bar{Y}_0)$$

Simplifying assumption: $n_i = n \quad \forall\ i$

$$\bar{Y}_1 = \frac{1}{m_1} \sum_{i=m_0+1}^{m} \bar{Y}_{i\cdot} \qquad \bar{Y}_{i\cdot} = \text{mean of } Y \text{ in group } 1$$

$$\bar{Y}_0 = \frac{1}{m_0} \sum_{i=1}^{m_0} \bar{Y}_{i\cdot}$$

$$Var(\bar{Y}_{i\cdot}) = \sigma_u^2 + \frac{\sigma_\varepsilon^2}{n}$$

$$Var(\bar{Y}_1) = \frac{1}{m_1}\left(\sigma_u^2 + \frac{\sigma_\varepsilon^2}{n}\right) \qquad Var(\bar{Y}_0) = \frac{1}{m_0}\left(\sigma_u^2 + \frac{\sigma_\varepsilon^2}{n}\right)$$

$$Var(\hat{\beta}_1) = \left(\sigma_u^2 + \frac{\sigma_\varepsilon^2}{n}\right)\left(\frac{m}{m_0 m_1}\right)$$

Our fitted model (incorrectly) thinks that the variance is

$$\hat{Var}(\hat{\beta}_1) = \frac{\frac{1}{N-2}\sum_i \sum_j (Y_{ij} - \hat{Y}_{ij})^2}{\sum_i \sum_j (X_i - \bar{X})^2}$$

$$\sum_i \sum_j (X_i - \bar{X})^2 \leftarrow \frac{N_1 N_0}{N}$$

Again assuming all $n_i = n$:

$$\text{numerator} = \frac{1}{N-2}\left(\underbrace{\sum_{i=1}^{m} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i.})^2}_{\approx m(n-1)\sigma_\varepsilon^2} + \underbrace{\sum_{i=1}^{m_0} n(\bar{Y}_{i.} - \bar{Y}_0)^2}_{n(m_0-1)(\sigma_u^2 + \frac{\sigma_\varepsilon^2}{n})} + \underbrace{\sum_{i=m_0+1}^{m} n(\bar{Y}_{i.} - \bar{Y}_1)^2}_{n(m_1-1)(\sigma_u^2 + \frac{\sigma_\varepsilon^2}{n})} \right)$$

$$\approx \left(\sigma_\varepsilon^2 + \sigma_u^2 + \frac{\sigma_\varepsilon^2}{n}\right)\left(\frac{m}{m_0 m_1}\right) \cdot \frac{1}{n}$$

Compare to the correct variance, which is $\approx \left(\sigma_u^2 + \frac{\sigma_\varepsilon^2}{n}\right)\left(\frac{m}{m_0 m_1}\right)$

If $\sigma_u^2 > 0$, then $\hat{Var}(\hat{\beta}_1)_{\text{incorrect}} \ll$ the variance

If $\sigma_u^2 = 0$, $\hat{Var}(\hat{\beta}_1)_{\text{incorrect}} \approx Var(\hat{\beta}_1)_{\text{correct}}$

Moral: If you have a group effect (i.e. $\sigma_u^2 > 0$, intra-class correlation > 0) you need to model the group effect

(ignoring group effect $\rightarrow$ inflated type I error for some hypothesis tests)

- problem is more severe as ICC increases

Class activity

https://sta712-f23.github.io/class_activities/ca_lecture_32.html