BIOS 662 Fall 2018

Clustered Data

David Couper, Ph.D.

david_couper@unc.edu

or

couper@bios.unc.edu

https://sakai.unc.edu/portal

Correlated Data

- To this point, all methods have assumed data are iid
- What do we do if dependencies exist between observations?
- Typically, correlated data occur in clusters/groups
- Examples:
  - Repeated measures on individuals over time
  - Natural groupings of individuals (e.g., litters, schools)
- Can occur in observational or randomized studies; an example of the latter is a cluster randomized study

Cluster Randomized Studies

- Also known as *group allocation* designs
- Section 18.4 of the text (deals with correlation structures)
- Suppose we want to compare two school-based methods of smoking prevention in teenagers
- We may randomly assign interventions to schools, but measure smoking in children

Central Issue

- How do we do testing, estimation, sample size calculations, etc., taking into account that responses within a cluster/group (e.g., school) may not be independent?

- Use methods allowing for dependency (correlation) within groups but assuming independence (no correlation) between groups

Continuous Response Model

- Let Y_{ijk} = response of the k^{th} person in the j^{th} cluster at the i^{th} treatment level,

$$i = 1, 2, \ldots, I; \quad j = 1, 2, \ldots, J; \quad k = 1, 2, \ldots, K$$

- Let

$$\bar{Y}_{ij} = \frac{1}{K} \sum_{k=1}^{K} Y_{ijk}$$

- Assume:

$$E(Y_{ijk}) = \mu_i; \quad \text{Var}(Y_{ijk}) = \sigma^2$$

$$\text{Cov}(Y_{ijk}, Y_{ijk'}) = \rho\sigma^2; \quad \text{Cov}(Y_{ijk}, Y_{ij'k'}) = 0$$

Continuous Response Model

- Then

$$\begin{aligned}
\mathrm{Var}(\bar{Y}_{ij}) &= E(\bar{Y}_{ij}^2) - \mu_i^2 \\
&= K^{-2} E\left(\sum_{k=1}^{K} Y_{ijk}\right)^2 - \mu_i^2 \\
&= K^{-2} E\left(\sum_{k=1}^{K} Y_{ijk}^2 + \sum\sum_{k \neq k'} Y_{ijk} Y_{ijk'}\right) - \mu_i^2 \\
&= K^{-2}\Big(K\sigma^2 + K(K-1)\rho\sigma^2\Big) \\
&= \frac{\sigma^2}{K}\Big(1 + (K-1)\rho\Big)
\end{aligned}$$

Variance Inflation Factor (VIF)

- $\big(1 + (K-1)\rho\big)$ is the *variance inflation factor* (VIF)
- It measures the increase in the variance of the mean due to the within-subject correlation of measurements (ρ)
- VIF > 1 for $\rho > 0$ and $K > 1$
- Let

$$\bar{Y}_i = \frac{\sum_j \bar{Y}_{ij}}{J} = \frac{\sum_{j,k} Y_{ijk}}{JK}$$

- Then

$$\text{Var}(\bar{Y}_i) = \frac{\sigma^2}{JK}\text{VIF}$$

Continuous Response Model

- Suppose $I = 2$ and $n_1 = n_2 = JK$
- If we ignore the correlation within cluster

$$z_{\text{ignore}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma\sqrt{1/n_1 + 1/n_2}},$$

- Should instead use

$$z_{\text{true}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sigma\sqrt{(1/n_1 + 1/n_2) \cdot \text{VIF}}}$$

$$= \frac{z_{\text{ignore}}}{\sqrt{\text{VIF}}}$$

Effect of Correlation

- $|z_{\text{true}}| < |z_{\text{ignore}}|$ for $\rho > 0$ and $K > 1$
- Thus ignoring correlation will lead to inflated type I error
- Intuition: Naïve approach acts as if we have more information than we do

Sample Size When $I = 2$

- Sample size per arm

$$n = 2\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\Delta}\right)^2 \text{VIF}$$

  where $\Delta = |\mu_1 - \mu_2|/\sigma$

- If $\rho = 0$, then $\text{VIF} = 1$ and

$$n = 2\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\Delta}\right)^2$$

- If $\rho = 1$, then $\text{VIF} = K$

$$n = 2\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\Delta}\right)^2 \cdot K$$

- Typically $0.1 \leq \rho \leq 0.4$

Variance Inflation Factors

- Table 18.4 in the text:

	ρ				
K	0.001	0.01	0.02	0.05	0.1
2	1.001	1.01	1.02	1.05	1.10
5	1.004	1.04	1.09	1.20	1.40
10	1.009	1.09	1.18	1.45	1.90
100	1.099	1.99	2.98	5.95	10.90
1000	1.999	10.99	20.98	50.95	100.90

Concluding Remarks

- What if cluster/group sizes vary? Say $k = 1, \ldots, K_j$
- Use expected cluster size; cf. Manatunga, Hudgens, Chen (*Biometrical Journal*, 2001)
- Methods for analyzing clustered data include mixed models and generalized estimating equations (BIOS 762/3/7)