# Notes on Parametric Power for Clustered Randomization

Prepared by Mauricio Cáceres

February 9, 2017

## 1 MDE and Sample Size with a Cluster Design

Recall Equation (1); without controls we have

$$Y_{ij} = \alpha + \beta T_j + \varepsilon_{ij}, \tag{1}$$

with $E(\varepsilon_{ij}\varepsilon_{i'j}) \neq 0$ and the randomization $T_j$ at the cluster level ($j = 1, \dots, J$ clusters and $n_j$ individuals per cluster, $\sum_{j=1}^{J} n_j = N$). Following Duflo et al. (2007, p. 3921-2), suppose we can additively decompose the error term $\varepsilon_{ij}$ as

$$Y_{ij} = \alpha + \beta T_j + v_j + u_{ij}, \tag{2}$$

with $v_j \overset{iid}{\sim} (0, \sigma_v^2), u_{ij} \overset{iid}{\sim} (0, \sigma_u^2)$. Duflo et al. provide a formula for the case when we have equal cluster sizes, but not the case when cluster sizes vary. Let us further assume $n_j \overset{iid}{\sim} (\mu_n, \sigma_n^2)$ and $\rho \equiv \sigma_v^2 / \sigma_\varepsilon^2$ so that $E(\varepsilon_{ij}\varepsilon_{i'j}) = \rho\sigma_\varepsilon^2$. For significance level $\alpha$, power $\kappa$, and proportion randomized $P$ we have

$$\begin{aligned} J &= \left(\frac{t_{1-\kappa} + t_{\alpha/2}}{MDE}\right)^2 \frac{DE \cdot \sigma_\varepsilon^2}{\mu_n P(1-P)} = N_0 \frac{DE}{\mu_n} \\ MDE &= \left|t_{1-\kappa} + t_{\alpha/2}\right| \sqrt{\frac{DE \cdot \sigma_\varepsilon^2}{J \mu_n P(1-P)}} = MDE_0 \cdot \sqrt{DE}, \end{aligned} \tag{3}$$

where $MDE_0, N_0$ are the MDE and sample size required if the model used individual data and $DE$ is the so-called design effect (DE) or variance inflation factor (VIF):

$$DE = \begin{cases} 1 + \rho(\mu_n - 1) & \sigma_n^2 = 0 \\ 1 + \rho((\sigma_n^2/\mu_n^2 + 1)\mu_n - 1) & \sigma_n^2 > 0 \end{cases}. \tag{4}$$

The formulas above are a slight modification of equations (1) and (3) in Manatunga et al. (2001) to account for the case when the proportion randomized $P \neq 0.5$. Note we're leveraging the fact that testing the significance of $\widehat{\beta}_{OLS}$ is equivalent to a paired $t$-test in this case. If $Y_{ij}$ is also binary then we need to adjust the variance. Equation (3) in Kong et al. (2003) gives

$$J = \left(\frac{t_{1-\kappa} + t_{\alpha/2}}{MDE}\right)^2 \frac{DE}{\mu_n P(1-P)} \Big(\mu_T(1-\mu_T)(1-P) + \mu_C(1-\mu_C)P\Big). \tag{5}$$

Again, we modify the formula slightly so we account for $P \neq 0.5$. Note that the variance of $\widehat{\beta}_{OLS}$ is

$$
\begin{aligned}
V_{\widehat{\beta}} &= \widehat{V}_T + \widehat{V}_C \\
V_k &= \frac{\sum_{j=1}^{J} n_{ik}\left(1 + (n_{jk} - 1)\rho\right)}{\left(\sum_{j=1}^{J} n_{jk}\right)^2} \sigma_\varepsilon^2 \qquad k = T, C.
\end{aligned} \tag{6}
$$

$JV_{\widehat{\beta}} \xrightarrow{P} \sigma_\varepsilon^2 DE/\mu_n$ as $J \to \infty$, meaning we can estimate $DE$ if the cluster sizes are known, given an estimate of $\rho$. Kong et al. suggests using an ANOVA-based estimate. Intuitively, from equation (2) we see that a random effects model $Y_{ij} = \alpha + v_j + u_{ij}$ is the true model under the null of $H_0 : \beta = 0$; then $\rho = \sigma_v^2/(\sigma_u^2 + \sigma_v^2)$.

# 2 The Effect of Covariates

Adding controls has the effect of absorbing some of the variation in the error terms, thereby reducing $\sigma_\varepsilon^2$ and $\rho$ and improving the precision of our estimates. The proportion of the unexplained variation absorbed by the covariates will be $R^2$, meaning we can adjust our parametric estimates to account for covariates by multiplying the variance by $1 - R^2$.

For the intra-cluster correlation $\rho$ there is no trivial way of accounting for an arbitrary number of covariates—however, it is possible to make a simple adjustment for any one covariate. We follow the approach outlined in Stanish and Taylor (1983) and adjust the ANOVA-based estimate for $\rho$ to account for the lag of the outcome variable. Taking $MSB$ and $MSW$ as they are usually defined, we have

$$
\begin{aligned}
n &= \frac{1}{J-1}\left[N - \sum_j n_j^2/N\right] \\
k &= \frac{1}{J-1}\left[\frac{\sum_j n_j^2(\overline{x}_j - \overline{x})^2}{\sum_j \sum_i (x_{ij} - \overline{x})^2}\right] \\
\widehat{\rho} &= \frac{MSB - MSW}{MSB + (n - k - 1)MSW},
\end{aligned}
$$

where the unadjusted $\widehat{\rho}$ would simply use $k = 0$.

# References


Duflo, E., Glennerster, R., and Kremer, M. (2007). Chapter 61 – Using Randomization in Development Economics Research: A Toolkit. In *Handbook of Development Economics*, volume 4, pages 3895–3962.

Kong, S.-H., Ahn, C. W., and Jung, S.-H. (2003). Sample Size Calculation for Dichotomous Outcomes in Cluster Randomization Trials with Varying Cluster Size. *Drug Information Journal*, 37(1):109–114.

Manatunga, A. K., Hudgens, M. G., and Chen, S. (2001). Sample Size Estimation in Cluster Randomized Studies with Varying Cluster Size. *Biometrical Journal*, 43(1):75–86.

Stanish, W. M. and Taylor, N. (1983). Estimation of the Intraclass Correlation Coefficient for the Analysis of Covariance Model. *The American Statistician*, 37(3):221–224.