Statistics and Probability Letters 78 (2008) 2654–2659

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Bias of the regression estimator for experiments using clustered random assignment

Joel A. Middleton *

Yale University, Department of Political Science, 77 Prospect Street, New Haven, CT 06511, United States

ARTICLE INFO

Article history:
Received 2 July 2007
Received in revised form 11 March 2008
Accepted 14 March 2008
Available online 22 March 2008

ABSTRACT

This paper shows that regression may be biased for cluster randomized experiments. For one application bias tends to zero when the number of clusters is large but for another, regression is not consistent. Results underscore Freedman's [Freedman D.A., 2008. On regression adjustments to experimental data. Advances in Applied Mathematics 40, 180–193] insight that randomization does not justify regression assumptions.

1. Introduction

The randomization model (also known as the Neyman nonparametric model, counterfactual model, potential outcomes and Rubin's Causal Model) is a nonparametric model for experimental inference (see for example Dabrowska and Speed (1990), Little and Rubin (2000), Maldanado and Greenland (2002) and Rubin (1990)). Under the randomization model each subject has two potential responses, one if treated and the other if not treated.[1]

Angrist et al. (1996) use the model to justify the use of instrumental variables regression for the estimation of causal effects. While their work is sometimes interpreted to say that randomization ensures the independence of the treatment indicator and disturbances, Freedman (2008, Section 1, p. 181) shows that for regression *with* covariate adjustment the treatment indicator and disturbances "will generally be strongly related" and regression may be biased. Freedman concludes with the important insight that "randomization does not justify regression assumptions".

Underscoring Freedman's insight, this paper demonstrates that bias can also arise when regression is applied to data from cluster randomized experiments *even without covariate adjustment*. Clustered randomization is used when subjects cannot be assigned to treatment conditions independent of others with whom they are associated (e.g. by village, school, or group affiliation), a situation common in field experimentation.

An expression for the bias of regression is derived for two approaches to data analysis: *individual level regression* (ILR) and *collapsed data regression* (CDR). For ILR, bias is small when the number of clusters is large and the estimator is consistent. Though bias is not present if there is no relationship between cluster size and individual level outcomes, a relationship between size and outcomes may be present in any number of real-world cases.[2] Recognition of this problem has motivated parametric workarounds on the one hand (Panageas et al., 2007; Williamson et al., 2003, 2007) and, on the other hand, a randomization-based solution (Middleton, 2007).

To implement CDR, the average outcome is taken at the cluster level and regressed on the treatment indicator. The prospect of ecological fallacy dissuades some from using CDR. Nonetheless Donner and Klar (2000, pp. 82–83) say that

* Tel.: +1 202 431 1412.
E-mail address: joel.middleton@yale.edu.

[1] For clarity, this paper considers only the case with one control group and one treatment group, however, this generalizes to multi-arm trials also.

[2] Middleton (2007) provides an example from a field experiment where voting rates are related to size of city block, Williamson et al. (2007) demonstrate that rate of condom use is related to the total number of sex acts per individual and Panageas et al. (2007) show that number of teeth is inversely related to the health of the remaining teeth.

0167-7152/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2008.03.008

randomization protects it against ecological fallacy and Arceneaux (2005) concludes that CDR is appropriate, though less desirable than ILR for cluster randomized experiments. However these authors implicitly assume that cluster size is ancillary. In Section 4, analysis proves that CDR is generally biased and inconsistent when cluster size and outcomes are related.

2. Notation

This section establishes relevant notation. Under the nonparametric randomization model each cluster, $i = 1, 2, 3, 4 \ldots . M$ is assigned to treatment or control. Let C_{ij} be the response of the ith individual in the jth cluster if the cluster is assigned to control and T_{ij} be the response of the ith individual in the jth cluster if the cluster is assigned to treatment. Let N_j be the number of individuals in the jth cluster. Now define $C_j = \sum_{i=1}^{N_j} C_{ij}$ to be the *sum* of responses of the individuals in jth cluster if it is assigned to control and $T_j = \sum_{i=1}^{N_j} T_{ij}$ be the *sum* of the responses of the individuals in the jth cluster if it is assigned to treatment. For each individual only one of the two possible responses, C_{ij} or T_{ij}, may be observed, and likewise, since individuals are assigned to treatment conditions in clusters, for any given cluster only one of the possible responses, C_j or T_j, may be observed. The quantity of interest is the average individual level treatment effect. Formally the average individual level treatment effect, Δ, is defined as

$$\Delta = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j}(T_{ij} - C_{ij})}{\sum_{j=1}^{M}\sum_{i=1}^{N_j} 1} = \frac{\tau - \kappa}{\eta} \tag{1}$$

where $\eta = \frac{1}{M}\sum_{j=1}^{M} N_j$ is the average number of individuals per cluster across the M clusters, $\tau = \frac{1}{M}\sum_{j=1}^{M} T_j$ is the average cluster total across the M clusters when treated, and $\kappa = \frac{1}{M}\sum_{j=1}^{M} C_j$ is the average cluster total across the M clusters when not treated.

Let X_j be the assignment variable for the jth each cluster: $X_j = 1$ if cluster j is assigned to treatment and $X_j = 0$ if cluster j is assigned to control. Assume that a fixed number m clusters are selected for inclusion in the treatment group such that $0 < m < M$.[3] The remaining $M - m$ clusters are assigned to control.

For the jth cluster the observed response is

$$Y_j = X_j T_j + (1 - X_j) C_j.$$

Note that the treatment indicator X_j is the only stochastic component in this model. For a given experimental allocation, the average of sums across clusters assigned to treatment can be written as

$$\bar{Y}_t = \frac{\sum_{j=1}^{M} X_j T_j}{\sum_{j=1}^{M} X_j}$$

and likewise the average of sums across clusters assigned to the control can be written as

$$\bar{Y}_c = \frac{\sum_{j=1}^{M} (1 - X_j)\, C_j}{\sum_{j=1}^{M} (1 - X_j)}.$$

Average cluster size in the treatment and control groups are

$$\bar{N}_t = \frac{\sum_{j=1}^{M} X_j N_j}{\sum_{j=1}^{M} X_j}$$

and

$$\bar{N}_c = \frac{\sum_{j=1}^{M} (1 - X_j)\, N_c}{\sum_{j=1}^{M} (1 - X_j)},$$

[3] Experimentalists often chose m such that $m = M - m$, i.e. there are equal number of clusters in treatment and control groups. However, there are cases in field research where unequal number of treatment and control clusters is preferable (see for example Middleton (2007) or Zucker et al. (1995)). Allowing for the possibility of unequal treatment and control groups is a more general framework and it is noteworthy that, as will be shown, choice of m affects the bias of the ILR estimator.

respectively. And finally, the number of clusters in treatment and control are

$$m = \sum_{j=1}^{M} X_j$$

and

$$M - m = \sum_{j=1}^{M} (1 - X_j),$$

respectively.

3. Individual level regression (ILR) for randomized experiments

Define the individual level regression (ILR) as the simple linear regression estimator achieved when regressing individual level outcomes on the treatment indicator. Freedman (2008) has shown that the simple linear regression estimator is equivalent to the difference between the averages of the treatment and control groups. Using this result and applying the notation for clustering as defined in the section above, we can write the individual level regression (ILR) estimator as

$$\begin{aligned}\hat{B}_{ILR} &= \frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j} T_{ij}X_j}{\sum_{j=1}^{M}\sum_{i=1}^{N_j} 1 \cdot X_j} - \frac{\sum_{j=1}^{M}\sum_{i=1}^{N_j} C_{ij}(1 - X_j)}{\sum_{j=1}^{M}\sum_{i=1}^{N_j} 1 \cdot (1 - X_j)} \\ &= \frac{\sum_{j=1}^{M} T_j X_j}{\sum_{j=1}^{M} N_j X_j} - \frac{\sum_{j=1}^{M} C_j (1 - X_j)}{\sum_{j=1}^{M} N_j (1 - X_j)} \\ &= \left(\frac{\bar{Y}_t}{\bar{N}_t}\right) - \left(\frac{\bar{Y}_c}{\bar{N}_c}\right). \end{aligned} \tag{2}$$

3.1. Bias of ILR

Theorem 1. *The bias of the ILR estimator for cluster randomized data is*

$$\text{Bias}\left(\hat{B}_{ILR}\right) = \frac{1}{\eta}\left[\text{Cov}\left(\frac{\bar{Y}_c}{\bar{N}_c}, \bar{N}_c\right) - \text{Cov}\left(\frac{\bar{Y}_t}{\bar{N}_t}, \bar{N}_t\right)\right] \tag{3}$$

where η, $\bar{Y}_c$, $\bar{Y}_t$, $\bar{N}_c$ *and* $\bar{N}_t$ *are defined as in Section 2 and* $\text{Cov}(\circ, \circ)$ *represents the covariance operator.*

Proof of Theorem 1. Given two random variables u and v, where v is never zero and $E(v)$ is never zero, the expected value of the ratio u/v can be written

$$E\left(\frac{u}{v}\right) = \frac{1}{\mu_v}\left[\mu_u - \text{Cov}\left(\frac{u}{v}, v\right)\right], \tag{4}$$

where μ_u and μ_v are the first moments for u and v, respectively (see for example Hartley and Ross (1954)). Therefore, taking the expected value of $\hat{B}_{ILR}$ in Eq. (2), then applying Eq. (4) yields

$$\begin{aligned} E\hat{B}_{ILR} &= E\left(\frac{\bar{Y}_t}{\bar{N}_t}\right) - E\left(\frac{\bar{Y}_c}{\bar{N}_c}\right), \\ &= \frac{1}{\eta}\left[\tau - \text{Cov}\left(\frac{\bar{Y}_t}{\bar{N}_t}, \bar{N}_t\right)\right] - \frac{1}{\eta}\left[\kappa - \text{Cov}\left(\frac{\bar{Y}_c}{\bar{N}_c}, \bar{N}_c\right)\right] \\ &= \Delta + \frac{1}{\eta}\left[\text{Cov}\left(\frac{\bar{Y}_c}{\bar{N}_c}, \bar{N}_c\right) - \text{Cov}\left(\frac{\bar{Y}_t}{\bar{N}_t}, \bar{N}_t\right)\right]. \end{aligned}$$

This proves Theorem 1. □

3.2. Consistency of the ILR

Consistency of a statistic under a finite population as such is defined given a sequence of k finite populations U_k where $M_k < M_{k+1}$, $m_k < m_{k+1}$ and $(M_k - m_k) < (M_{k+1} - m_{k+1})$ for $k = (1, 2, 3, \ldots)$. The estimator f_k is said to be a consistent estimator of Δ if $f_k \to \Delta$ (converges in probability) as $k \to \infty$.

Taking a cue from Brewer (1979), assume that as $k \to \infty$ the finite population U_k increases as follows: (1) the original population of M units is exactly copied $(k-1)$ times; (2) from each of the k copies, m clusters are allocated to treatment (such that $0 < m < M$) and the remaining $M - m$ are allocated to control; (3) the k subsets are collected in a single population of kM, with km clusters in treatment and $k(M - m)$ in the control; and (4) f_k is defined as the ILR estimator above, only now summation takes place across all kM units. A less restrictive set of assumptions is possible, but this setup is useful because U_k is easy to visualize and moment assumptions are built-in.

Theorem 2. *The ILR is consistent.*

Proof of Theorem 2. To show the consistency of the ILR note that it can be written,

$$\hat{B}_{ILR} = \left(\frac{\sum_{l=1}^{k} \bar{Y}_{tl}}{\sum_{l=1}^{k} \bar{N}_{tl}} \right) - \left(\frac{\sum_{l=1}^{k} \bar{Y}_{cl}}{\sum_{l=1}^{k} \bar{N}_{cl}} \right),$$

where l indexes over the k population copies. By the weak law of large numbers, $\frac{1}{k}\sum_{l=1}^{k} \bar{Y}_{tl} \to \tau$, $\frac{1}{k}\sum_{l=1}^{k} \bar{N}_{tl} \to \eta$, $\frac{1}{k}\sum_{l=1}^{k} \bar{Y}_{cl} \to \kappa$, and $\frac{1}{k}\sum_{l=1}^{k} \bar{N}_{cl} \to \eta$ as $k \to \infty$. Therefore, by Slutsky's theorem (e.g. Billingsley, 1968, p. 31, corollary 2)

$$\hat{B}_{ILR} \to \left(\frac{\tau}{\eta} \right) - \left(\frac{\kappa}{\eta} \right) = \Delta$$

as $k \to \infty$, demonstrating the consistency of the ILR estimator. □

Theorem 3. *The ILR estimator converges to Δ as $k \to \infty$ at a rate $O(1/k)$.*

Proof of Theorem 3. To examine the rate of convergence of the ILR, write the bias as

$$E\left(\hat{B}_{ILR} - \Delta\right) = E\left(\frac{\sum_{l=1}^{k} \bar{Y}_{tl}}{\sum_{l=1}^{k} \bar{N}_{tl}} \right) - E\left(\frac{\sum_{l=1}^{k} \bar{Y}_{cl}}{\sum_{l=1}^{k} \bar{N}_{cl}} \right) - \frac{\tau - \kappa}{\eta} \tag{5a}$$

$$= E\left(\frac{\sum_{l=1}^{k} \bar{Y}_{tl} - \sum_{l=1}^{k} \bar{N}_{tl}\tau/\eta}{\sum_{l=1}^{k} \bar{N}_{tl}} \right) - E\left(\frac{\sum_{l=1}^{k} \bar{Y}_{ctl} - \sum_{l=1}^{k} \bar{N}_{cl}\kappa/\eta}{\sum_{l=1}^{k} \bar{N}_{cl}} \right).^{4} \tag{5b}$$

Using Taylor series, Raj and Chandhok (1999, pp. 136–138) show that the first term in Eq. (5b) can be approximated by,

$$\frac{V(N_j)\tau/\eta - \mathrm{Cov}\,(T_j, N_j)}{km\eta^2}.$$

Likewise, the second term in Eq. (5b) can be approximated by an expression of the same form. Now define

$$b_t = \frac{V(N_j)\tau/\eta - \mathrm{Cov}\,(T_j, N_j)}{m\eta^2}, \tag{6a}$$

and

$$b_c = \frac{V(N_j)\kappa/\eta - \mathrm{Cov}\,(C_j, N_j)}{(M - m)\,\eta^2}, \tag{6b}$$

[4] Note that by assuming fixed values of M and m, covariance between the X_i's is induced and consequently the terms $\sum_{l=1}^{k} \bar{Y}_{tl} / \sum_{l=1}^{k} \bar{N}_{tl}$ and $\sum_{l=1}^{k} \bar{Y}_{cl} / \sum_{l=1}^{k} \bar{N}_{cl}$ in (5a) will themselves covary. The bias of the estimator is not affected however since the *expected value* of a sum is equal to the sum of the expected values (e.g. Casella and Berger (1990, p. 56)).

and note that b_t and b_c are free of k. Returning to Eq. (5), the approximate expression for bias as a function of k can be rewritten as

$$E\left(\hat{B}_{ILR} - \Delta\right) \doteq \frac{1}{k}\left(b_t - b_c\right).$$

This proves Theorem 3. □

4. Collapsed data regression (CDR) for randomized experiments

The collapsed data regression (CDR) estimator is obtained by first averaging the individual outcomes within each cluster, then running the regression using the aggregated data.

The CDR can be written as,

$$\begin{aligned}\hat{B}_{CDR} &= \frac{1}{m}\sum_{j=1}^{M}\left(\frac{1}{N_j}\sum_{i=1}^{N_j} T_{ij}X_j\right) - \frac{1}{M-m}\sum_{j=1}^{M}\left(\frac{1}{N_j}\sum_{i=1}^{N_j} C_{ij}\left(1-X_j\right)\right)\\ &= \frac{1}{m}\sum_{j=1}^{M}\frac{T_j}{N_j}X_j - \frac{1}{M-m}\sum_{j=1}^{M}\frac{C_j}{N_j}\left(1-X_j\right). \end{aligned} \tag{7}$$

4.1. Bias of CDR

Theorem 4. *The bias of the CDR estimator is*

$$\text{Bias}\left(\hat{B}_{CDR}\right) = \frac{1}{\eta}\left[\text{Cov}\left(\frac{C_j}{N_j}, N_j\right) - \text{Cov}\left(\frac{T_j}{N_j}, N_j\right)\right]. \tag{8}$$

Proof of Theorem 4. Like the ILR estimator, the CDR estimator is a difference of two ratios with random denominators. It is therefore subject to ratio estimator bias. Taking the expected value of Eq. (7) we have

$$E\hat{B}_{CDR} = E\left(\frac{T_j}{N_j}\right) - E\left(\frac{C_j}{N_j}\right),$$

and applying Eq. (4) we arrive at

$$\begin{aligned}E\hat{B}_{CDR} &= \frac{1}{\eta}\left[\tau - \text{Cov}\left(\frac{T_j}{N_j}, N_j\right)\right] - \frac{1}{\eta}\left[\kappa - \text{Cov}\left(\frac{C_j}{N_j}, N_j\right)\right]\\ &= \Delta + \frac{1}{\eta}\left[\text{Cov}\left(\frac{C_j}{N_j}, N_j\right) - \text{Cov}\left(\frac{T_j}{N_j}, N_j\right)\right].\end{aligned}$$

This concludes the proof of Theorem 4. □

4.2. Consistency of CDR

Assumptions are the same as in 3.2.

Theorem 5. *The CDR estimator is not generally consistent.*

Proof of Theorem 5. As $k \to \infty$, $\text{Cov}\left(\frac{C_j}{N_j}, N_j\right)$ and $\text{Cov}\left(\frac{T_j}{N_j}, N_j\right)$ are unchanged. Therefore the bias remains constant as $k \to \infty$. This concludes the proof of Theorem 5. □

5. Conclusions

This paper shows that simple linear regression may be biased for cluster randomized experiments whether or not data are collapsed at the cluster level. Eq. (3) gives the bias for ILR. Theorem 3 shows that the ILR estimator is consistent for Δ, the average individual level effect, while Theorem 4 shows that the ILR estimator converges at a rate $O(1/k)$. Notable is that the magnitude of bias is a function of the treatment and control allocations, m and $M-m$, as shown by Eqs. (6a) and (6b).

For special cases when the covariance terms in Eq. (3) are equal, the bias is zero. Unbiased cases include those where: (1) the number of individuals in every cluster is the same; (2) there is no relationship between cluster size and average outcome of individuals in the cluster; and (3) the number of clusters in the treatment is equal to the number of clusters in the control group ($m = M - m$) *and* $C_j = T_j$ for all j.

Eq. (8) gives the bias of the CDR estimator. Analysis shows that the CDR estimator is not consistent when cluster size is related to outcomes. This is an important caveat to Donner and Klar's (2000) assessment that CDR is appropriate for estimating intention-to-treat effects. Unbiased cases include those where: (1) there is no effect (i.e. $T_j = C_j$ for all j) and (2) there is no relationship between cluster size and average outcome of individuals in the cluster.

The above proofs underscore Freedman's (2008) insight that randomization does not justify regression assumptions. Compelling arguments for estimation without bias can be made. While parametric fixes are feasible (Panageas et al., 2007; Williamson et al., 2003, 2007) researchers might consider robust randomization-based methods. Contrary to widely held opinion, nonparametric methods can achieve robust point estimation, covariate adjustment, confidence intervals, and precision on par with regression (Middleton, 2007).

Acknowledgements

A special thanks goes to David A. Freedman for his candid feedback. I would also like to thank Donald P. Green for the helpful comments and conversations. An anonymous reviewer also gave an exceedingly informative critique, alerting me to key citations in the public health and statistics literature. Of course, any errata are my own responsibility.

References

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91 (434), 444–455.
Arceneaux, K., 2005. Using cluster randomized field experiments to study voting behavior. The Annals of the American Academy of Political and Social Science 601, 169–179.
Billingsley, P., 1968. Convergence of Probability Measures. John Wiley and Sons, New York.
Brewer, K.R.W., 1979. A class of robust sampling designs for large-scale surveys. Journal of the American Statistical Association 74 (368), 911–915.
Casella, G., Berger, R.L., 1990. Statistical Inference. Wadsworth Inc, Belmont.
Dabrowska, D.M., Speed, T.P., 1990. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science 5 (4), 465–480. Translation of Polish original by Jerzy Splawa-Neyman 1923.
Donner, A., Klar, N., 2000. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold, London, UK.
Freedman, D.A., 2008. On regression adjustments to experimental data. Advances in Applied Mathematics 40, 180–193.
Hartley, H.O., Ross, A., 1954. Unbiased ratio estimators. Nature 174, 270.
Little, R.J., Rubin, D.B., 2000. Causal effects in clinical and epidemiological studies via potential outcomes: Concepts and analytical approaches. Annual Review of Public Health 21, 121–145.
Maldanado, G., Greenland, S., 2002. Estimating causal effects. International Journal of Epidemiology 31, 422–429.
Middleton, J.A., 2007. Are the canonical studies in the voter mobilization literature externally valid? An Experimental Field Study of the 2004 Presidential Election. Yale University: Institute for Social and Policy Studies (unpublished manuscript).
Panageas, K.S., Schrag, D., Localio, R.A., Venkatraman, E.S., Begg, C.B., 2007. Properties of Analysis methods that account for clustering in volume outcome studies when the primary predictor is cluster size. Statistics in Medicine 26 (9), 2017–2035.
Raj, D., Chandhok, P., 1999. Sample Survey Theory. Narosa Publishing, London.
Rubin, D.B., 1990. [On the application of probability theory to agricultural experiments. Essay on principles. Section 9.] Comment: Neyman (1923) and Causal Inference in Experiments and Observational Studies. Statistical Science 5 (4), 472–480.
Williamson, J.M., Datta, S., Satten, G.A., 2003. Marginal analyses of clustered data when cluster size is informative. Biometrics 59 (1), 36–42.
Williamson, J.M., Kim, H.Y., Warner, L., 2007. Weighting condom use data to account for non-ignorable cluster size. Annals of Epidemiology 17 (8), 603–607.
Zucker, D.M., Lakatos, E., Webber, L.S., Murray, D.M., McKinlay, S.M., Feldman, H.A., Kelder, S.H., Nader, P.R., 1995. Statistical design of the child and adolescent trial for cardiovascular health (CATCH): Implications for cluster randomization. Control Clinical Trials 16 (2), 96–118.