Proceedings of

COMPSTAT 2014

21st International Conference on

Computational Statistics

hosting the **5th IASC World Conference**

Geneva, Switzerland

August 19–22, 2014

Manfred Gilli
Gil Gonzalez-Rodriguez
Alicia Nieto-Reyes (Eds.)

ISBN: 978-2-8399-1347-8

Proceedings of COMPSTAT 2014

Manfred Gilli
Geneva School of Economics and Management
University of Geneva
Switzerland
Manfred.Gilli@unige.ch

Gil González-Rodríguez
Department of Statistics
University of Oviedo
Spain
gil@uniovi.es

Alicia Nieto-Reyes
Department of Mathematics, Statistics and Computer Science
University of Cantabria
Spain
alicia.nieto@unican.es

ISBN 978-2-8399-1347-8
19th August 2014
Université de Genéve – 1211 Genéve, Switzerland

Preface

The 21st International Conference on Computational Statistics (COMPSTAT 2014) is held in Geneva. This year the Conference also hosts the 5th IASC World Congress. The Geneva edition coincides with the 40th anniversary of this biennial event which started in 1974 in Vienna and has been organized all over Europe. In the preface of the 1974 proceedings we can read: *'If we succeed in making statisticians aware of the great possibilities of modern computing facilities, which at any rate go beyond simple numerical computations, the Symposium serves its purpose.'* This goal has since been reached with certainty, as by now statisticians fully integrate computational tools in their work.

The Geneva edition seems to pursue 'the success story' with more than 400 participants and 370 presentations. The electronic Book of Proceedings includes a selection of 84 papers covering 700 pages, all peer reviewed.

Keynote lectures are addressed by Peter Bühlmann from the Swiss Federal Institute in Zurich, Anthony Davison from the Swiss Federal Institute in Lausanne and Xuming He from University of Michigan, USA. Two tutorials are offered, one by Dietmar Maringer, University of Basel, Switzerland and one by Stefan Van Aelst from KU Leuven, Belgium.

The editors thank the contributing authors, the referees and the members of the scientific program committee, and most importantly, all participants who are the soul of the conference.

The next edition of COMPSTAT will take place in Oviedo, Spain on August 23-26, 2016 and will be organized by Prof. Ana Colubi. We wish her the best success.

COMPSTAT 2014 Editors:

Manfred Gilli, University of Geneva, Switzerland.
Gil González-Rodríguez, University of Oviedo, Spain.
Alicia Nieto-Reyes, University of Cantabria, Spain.

Scientific Program Committee

Ex-officio:

COMPSTAT 2014 organiser and Chairperson of the SPC: Manfred Gilli, University of Geneva, Switzerland.

Past COMPSTAT organiser: Erricos John Kontoghiorghes, Cyprus University of Technology, Cyprus.

Next COMPSTAT organiser: Ana Colubi, University of Oviedo, Spain.

IASC-ERS Chairman: Vincenzo Esposito Vinzi, ESSEC Business School, France.

Members:

Alessandra Amendola, University of Salerno, Italy.

Ivette Gomes, University of Lisbon, Portugal.

Sandra Paterlini, European Business School, Wiesbaden, Germany.

Anne Philippe, University of Nantes, France.

Elvezio Ronchetti, University of Geneva, Switzerland.

Marieke Timmerman, University of Groningen, The Netherlands.

Consultative Members:

Representative of the IFCS: Anuska Ferligoj, University of Ljubljana, Slovenia.

Representative of the ARS of IASC: Jung Jin Lee, Soongsil University, Korea.

Representative of ERCIM WG CMS: Stefan Van Aelst, KU Leuven, Belgium.

COMPSTAT 2014 Proceedings Management Committee:

Manfred Gilli, University of Geneva, Switzerland.

Gil González-Rodríguez, University of Oviedo, Spain.

Alicia Nieto-Reyes, University of Cantabria, Spain.

Additional Referees:

Kohei Adachi, Ana Maria Aguilera, Marco Alfo, Andres M. Alonso, Tomas Aluja, Daniel Baier, Simona Balbi, Jose R. Berrendero, Sotiris Bersimis, Lucio Bertoli Barsotti, Patrice Bertrand, Concha Bielza, Angela Blanco-Fernandez, Xavier Bry, Carmen Cadarso, Daniela Calo, M. Angeles Carnero, Philippe Castagliola, Jose E. Chacon, Chun-Houh Chen, Victor Chepoi, Christophe Chesneau, Vartan Choulakian, Claudio Conversano, Mauro Costantini, Antonio Cuevas, Guglielmo D'Amico, Pierpaolo D'Urso, Michel Delecroix, Pedro Delicado, Marta Di Lascio, John Einmahl, Alesio Farcomeni, Arturo J. Fernandez, Silvia Ferrari, Peter Filzmoser, David Fletcher, Marta Garcia-Barzana, Luis Angel Garcia-Escudero, Laurent Gardes, Stelios D. Georgiou, Tomasz Gorecki, Sergei Grudsky, Bettina Grun, Serge Guillas, Armelle Guillou, Hwang Heungsun, Xianzheng Huang, Stephan Huckemann, Marie Huskova, Salvatore Ingrassia, Antonio Irpino, Yoshihide Kakizawa, George Karabatsos, Hyoungmoon Kim, Worapan Kusakunnirun, Agnes Lagnoux, Marc Lavielle, Michael Lechner, Anne Leucht, Christophe Ley, Gaorong Li, Xiang Liming, Chu-An Liu, Ann Maharaj, Dietmar Maringer, Marco Marozzi, Pablo Martinez-Camblor, Antonello Maruotti, Jorge Mateu, Agustin Mayo, Stefan Mittnik, Domingo Morales, Brenda Murphy, Fionn Murtagh, Matthew Nunes, Daniel Oberski, M. Carmen Pardo, Fulvia Pennoni, Carlos Perez-Gonzalez, Davide Pigoli, Ana Belen Ramos-Guajardo, Giovanna Ranalli, Philip Reiss, Holger Reulen, Havard Rue, Silvia Ruiz-Velasco Acosta Giorgio Russolillo, Luigi Salmaso, Antonio Salmeron, Theofanis Sapatinas, Gilbert Saporta, Johan Segers, Ana Sipols, Alwin Stegeman, Fabio Tardella, Andrew C. Titman, Valentin K. Todorov, Inmaculada Torres-Castro, Nickolay Trendafilov, M. Dolores Ugarte, Zidong Wang, Jinfang Wang, Keming Yu.

Contents

Propensity score matching with clustered data: an application to birth register data

Massimo Cannas, *University of Cagliari*, massimo.cannas@unica.it
Bruno Arpino, *Universitat Pompeu Fabra*, bruno.arpino@upf.edu

Abstract. In this paper we consider the implementation of propensity score matching for clustered data. Different approaches to reduce bias due to cluster level confounders are considered: matching within clusters and random or fixed effects models for the estimation of the propensity score. All the methods are illustrated with an application to the estimation of the effect of caesarean section on the Apgar score using birth register data from Sardinia hospitals.

Keywords. Causal inference, Propensity score, Matching, Multilevel data, Caesarean section, Apgar score.

1 Introduction

Methods based on the propensity score are widely used in many fields to estimate causal effects with observational data. When treatment assignment is not randomized but it is reasonable to assume that selection is on observables, matching (as well as weighting and stratification) methods are used to adjust for different distributions of the observed characteristics in the treated and the control groups [7]. Apart from few exceptions [2, 8, 11] these methods have been considered only for unstructured data. However, in many applications data show a hierarchical structure (e.g., students nested into schools, patients nested into hospitals, individuals nested into geographical areas). We consider situations where both individual and cluster-level (e.g., hospital) characteristics can influence both treatment intake and the outcome. In these contexts ignoring cluster-level confounding factors would introduce a bias.

In this paper, we consider different approaches to take into account the hierarchical structure of the data with the aim of reducing the bias due to group-level characteristics. These methods are particularly useful when it is not possible to measure all cluster-level confounders. To illustrate the methods, we consider estimating the effect of caesarian section on the Apgar score. In our application, the relevant structure is represented by a hierarchy of 2 levels (individuals nested into hospitals) and we will consider this type of data structure in the following. However, the approaches we consider can be easily adapted to more complex structures.

Propensity score matching with clustered data

Suppose we have a two-level data structure where N micro units at the first level, indexed by i ($i = 1, 2, ..., n_j$), are nested in J macro units at the second level (clusters), indexed by j ($j = 1, 2, ..., J$). We consider a binary treatment administered at the individual level, T, and an outcome variable, Y also measured at the individual level. Pre-treatment variables can be first (X) or second level (Z) variables.

Under the potential outcome framework, let $Y_{ij}(t)$ be the potential outcome if unit ij was assigned to treatment t, $t \in \{0, 1\}$. An individual causal effect is a comparison of $Y_{ij}(1)$ with $Y_{ij}(0)$, yet only one of the two potential outcomes is observed depending on the value of T_{ij}. Usually, the Average Treatment effect on the Treated (ATT) is considered as an interesting summary of individual causal effects: $ATT = E(Y_{ij}(1) - Y_{ij}(0) \,|\, T_{ij} = 1)$.

To identify the ATT with observational data, the following assumptions are often invoked:

- SUTVA: If $T = T'$ then $Y(T) = Y(T')$ for all T, T' in $\{0, 1\}^N$

- Unconfoundedness: $Y(1), Y(0) \perp T|(X, Z)$;

- Overlap: $0 < P(T = 1 \,|\, (X, Z)) < 1$.

The Stable Unit Treatment Value Assumption (SUTVA, [9]) requires that potential outcomes for a unit are not affected by the treatment received by other units, and there are no hidden versions of the treatment. Unconfoundedness asserts that the probability of assignment to a treatment does not depend on the potential outcomes conditional on observed covariates [9]. Unconfoundedness essentially assumes that within subpopulations defined by values of the covariates, we have random assignment of the treatment; it rules out the role of unobserved variables and therefore is often referred to also as selection on observables [7].

Rosenbaum and Rubin [9] showed that under the previous assumptions, adjustment on the propensity score eliminates bias due to observed confounders. The propensity score, e, is defined for each unit as the probability to receive the treatment conditional given its covariate values. In our setting, assuming that all covariates are observed we have $e_{ij} = Pr(T_{ij} = 1|(X_{ij}, Z_j))$. The propensity score is a one-dimensional summary of the multidimensional set of covariates, such that when the propensity score is balanced across the treatment and control groups, the distribution of all covariates are balanced in expectation across the two groups. In this way the problem of adjusting for a multivariate set of observed characteristics reduces to adjusting for the one-dimensional propensity score and this can be done using several Propensity Score Matching (PSM) algorithms that, for each given unit, determine a set of units in the opposite treatment condition with similar value for the propensity score.

In observational studies the propensity score is not known and must be estimated from the data, usually using logit or probit models. Obviously, an incorrectly estimated propensity score may lose its balancing property. More importantly, if one or more variables affecting the selection into treatment and potential outcomes are not observed, then unconfoundedness is violated and

ATT estimators based on PSM will be biased. In fact, PSM can only balance variables used in the propensity score model. In the following we shall assume that we have good measurement on all individual level confounders, X, but we may have no information on all or some of the second-level confounders, Z. We consider different approaches to implement PSM with a 2-level data structure. Two groups of strategies can be adopted in order to take into account the hierarchical structure of the data: implementing the matching within clusters; using a model for the estimation of the propensity score that takes the hierarchical structure explicitly into account. Therefore, the approaches we compare are as follows:

A Single-level propensity score; matching on the pooled dataset;

B Single-level propensity score; matching only within-clusters;

C Single-level propensity score; preferential within-cluster matching;

D Random-effect propensity score; matching on the pooled dataset;

E Fixed-effect propensity score; matching on the pooled dataset.

Approach A ignores completely the hierarchical structure. In this case, if we do not include all relevant confounders at the second level in the propensity score and obtain a good balance on all of them, our ATT estimator based on the PSM will be biased. Approach B deals with this problem by matching units within clusters only. This automatically guarantees that all cluster-level variables (measured and unmeasured) are perfectly balanced. This can come to a cost. Control units to be matched with treated units are only searched within the same cluster. In this way it could be that we lose some good match and so the balancing of individual level variables could be worse. Moreover, if we impose a caliper it could be that we do not find a control matched unit that we would find in other clusters. So, an additional problem could be losing some treated units.

To avoid these problems and combine the benefits of approaches A and B, approach C starts by searching control units within cluster. If none is found, control units are searched in other clusters. This approach improves the balancing of cluster level variables with respect to approach A and avoids the lost of units of approach B.

In alternative to exploiting the hierarchical structure in the implementation of the matching, approaches D and E take it into account when modelling the propensity score. In particular, approach D and E use a random or fixed effect, respectively, to represent unmeasured cluster level variables. Arpino and Mealli [2] and Thoemmes and West [11] showed that PSM using random or fixed effects models are able to reduce the bias of ATT due to unmeasured cluster level variables. However, our simulation exercise is more realistic because it is inspired by a real case studies, it involves a larger number of individual level variables and strongly unbalanced dataset.

Estimating the effect of caesarian section on Apgar score

Apart from individual level variables, the literature suggested the relevance of hospital level factors both on the decision of taking a medical treatment and on the medical outcomes for several procedures. In other words, these cluster level variables may act as confounders and so the researcher should adjust the analysis accordingly. For example, Caceras et al. [4] and Bragg et al. [3] indirectly measured the impact of hospital variables on the likelihood of a caesarean delivery. Similarly, since the work of Hughes et al. [6] it is clear that these variables may also affect the quality of the outcome. When we refer to unobserved variables at the hospital level we are referring to variables whose role has been proved or conjectured by previous studies; for example variables which do not vary at the hospital level for a reasonably long period of time, like obstetrician practice, physicianÕs preferences and guidelines promoting or restricting the liberal use of caesarean sections. Clearly, it is not always possible to observe all hospital level factors that contribute to the decision of operating a caesarean section and may also impact on the infantÕs health as measured by the Apgar score. To this end we adopt the strategies detailed in the previous section.

2 Data

The data set we consider contains information on deliveries occurred in the 22 hospitals of the Italian region of Sardinia in 2010 and 2011. The source is the official form on the birth event (known as CedAP) filled by physicians after the birth and accounting for all hospitalized births in the specified period. The form is divided in three parts containing sociodemographic information on the mother, the pregnancy and the infant. From the initial population of 23,925 observations we extracted the subset of non-complicated pregnancies in order to better isolate the effect of the caesarian section on the target variable. In particular, we selected nulliparous women at 32 or more weeks of gestational age with a singleton and living infant in vertex (head-down) position, without birth anomalies. We further restrict the sample to mothers aged between 15 and 44. The subset of non-complicated pregnancies is widely used in observational studies related to cesarean section, for example [3, 4] make analogous variable selections, but the former study also limits the sample to hospitals with almost 500 deliveries per year. The selected subset contains 14,757 cases clustered in 20 hospitals (the observations of two hospitals were removed since after the selection they contained only treated or untreated women). Proportions of caesarean sections across hospitals vary from a minimum of 0.11 to a maximum of 0.64 with an average of 0.35 (see Table 1). We focus on the 5-minute Apgar score as the outcome variable. This score is a simple and widely established indicator of the infantÕs health. It is well known that low Apgar scores are strongly associated with high mortality rates [1]. In our sample the proportion of low (< 7) scores is 0.0064. The score distribution is highly skewed with an average score of 9.54.

We built the propensity score model for the probability of caesarean section relying on a set of clinical (X) and social (Z) variables that proved significant in previous studies. In the first group of predictor we have infant weight, motherÕs gestational age, induction of labour and pregnancy related pathologies. In the second group we have socio-demographic information like maternal age and maternal education

3 Empirical Results

We start by reporting in Table 2 the mean differences of covariates across treated and untreated women for each balancing strategies. The last row of the table averages the (absolute) differences over all covariates and it known as the standardized bias (ASAM), an overall measure of covariate balance. We report the balance before matching and compare it with the balance we obtain with approaches A, B, C, D and E. Several variables showed a standardized bias higher than

Hospital	N. births	N. caesarean sections	$^o/_o$ caesarean sections	$^o/_{oo}$ low apgar infants
1	2,532	1,166	46.0	16.5
2	1,788	623	34.8	2.7
3	1,687	540	32.0	5.3
4	1,473	632	42.9	14.2
5	1,253	410	32.7	0.7
6	1,197	428	35.7	3.3
7	980	240	24.4	2.0
8	875	238	27.2	5.7
9	529	190	35.9	3.7
10	434	135	31.1	6.9
11	403	164	40.6	0
12	396	117	29.5	7.5
13	351	134	38.1	8.5
14	266	74	27.8	7.5
15	208	99	47.5	9.6
16	191	122	63.8	10.4
17	103	40	38.8	9.7
18	50	9	18.0	20
19	32	13	40.6	0
20	9	1	11.1	0
Total	14,757	5,375		
Mean	737.8	268.7	35.0	6.75

Table 1: Number of cesarean sections and low Apgar infants by hospital.

commonly accepted threshold (5% or 10%) representing substantive unbalance before matching. All considered approaches were effective in reducing imbalance even if approaches B and C show a slightly worse balance. However, these methods compared to method A take into account possible hospital level confounding effects and give anyway acceptable balance of all individual covariates. In particular, method B should be the preferred one given that it automatically balances all hospital level factors but still guarantees good balance of individual observed confounders compared to the other approaches. Finally, approaches D and E give slightly better ASAM than B and C for individual level covariates even if the balance of unobserved covariates at the hospital level is not guaranteed as is in within cluster matching.

In Table 3 the total number of treated units dropped due to the caliper option is shown. Here the caliper is 0.25 in standard deviation units so all treated units with a propensity score (e) outside the range ($e-2\sigma_e$, $e+2\sigma_e$), where σ_e indicates the standard deviation of the propensity score, will be discarded. When matching within hospitals we keep the same criterion by using the standard deviation of the clusters as the reference value. The matched dataset were obtained using macros based on the Matching package [10].It is interesting noting that the number of drops is not a constant proportion of the cluster size (not shown), as the covariate distribution may vary across clusters.

In Table 4 we show the ATT estimate for unmatched (i.e. the raw effect prior to any

Variable	Before	A	B	C	D	E
Maternal Age (years)						
< 20	-14.942	-0.632	-0.552	-0.551	-0.936	-1.685
20-24	-12.461	1.254	2.278	2.269	1.593	1.223
25-29	-15.048	0.151	1.778	1.780	0.915	1.257
30-35	-6.119	-0.708	-0.854	-0.818	-2.297	0.288
> 35	26.672	0.128	-1.435	-1.461	1.035	-1.383
Maternal Education						
Less than High School	-2.534	0.239	-4.264	-4.360	-2.495	-3.746
High School	0.575	-1.359	1.794	1.1784	0.452	2.172
Graduate or more	2.802	-0.997	3.063	2.998	0.581	-0.235
Missing	-0.056	0.828	0.418	0.688	2.849	2.910
Infant Weight (grams)						
< 2500	21.498	0.524	0.413	0.402	0.620	-0.291
2500-4000	-23.880	-1.700	-2.542	-2.544	-0.120	0.193
>4000	9.138	2.187	3.782	3.856	1.160	0.104
Labor Induction	-5.038	-1.547	0.393	0.437	-2.562	-2.813
Gestational Age						
Preterm (< 37 weeks)	23.273	-1.789	-1.584	-1.622	-1.937	0.193
Early norm ($37 - 38$ weeks)	26.950	0.400	-1.583	-1.486	-0.099	-1.933
Late norm (≥ 39 weeks)	-40.737	0.798	2.522	2.495	1.367	1.697
*Pathology during pregnancy**	20.756	0.353	4.225	4.088	2.616	1.447
ASAM	14.863	0.917	1.970	1.981	1.390	1.386

$*$ This is a dichotomous variable set to 1 if one (or more) of the following diseases occurred during pregnancy: Diabetes mellitus, Eclampsia, Hypertension, Placenta Previa.

Table 2: Mean differences of mothers characteristics before and after matching.

Hospital	N. births	N. caesarean sections	N. drops A	N. drops B	N. drops C	N.drops D	N.drops E
	14,757	5,375	0	38	0	0	0

Table 3: Number of dropped treated units.

adjustment) and matched datasets. The effect of caesarean is consistently estimated to be positive: it increases the risk of low Apgar score. It is worth noting that approaches B and C that control for hospital factors show considerably lower estimates than approach A. This may signal a possible overestimation of the effect of caesarean section when hospital confounding effects associated to higher prevalence of this section mode are not taken into account. Similarly, also multilevel and fixed effect propensity score models (approaches D and E) yield a pooled estimate lower than that of approach A. Clearly, approaches B-C and D-E have a higher mean ASAM than approach A (1.197-1.198 and 1.390-1.386 versus 0.94) and this should be considered the cost of balancing the potential confounders at the hospital level. is not surprising: indeed these two matching strategies are expected to diverge when there is strong imbalance at the hospital level but not globally.

Strategy Metrics	Without match	A	B	C	D	E
Balance						
Drops	0	0	38	0	0	0
ASAM	14.8	0.91	1.97	1.98	1.39	1.38
# of outcomes (every 1000 individuals)						
in treated	10.9	10.9	11.0	10.9	10.9	10.9
in untreated	5.2	9.1	9.6	9.7	9.9	9.9
ATT	5.75	1.80	1.40	1.23	1.02	1.07

Table 4: Empirical results for unmatched and matched subsets (strategies A-E). For each strategy: Drops is the number of dropped treated units; ASAM is the average standardized mean difference in covariates values across treated and untreated units; ATT is the mean difference between the number of outcomes in treated and untreated groups.

Simulation study

Motivated by previous empirical analysis we made a simulation experiment which illustrates the implications of different matching strategies when there is unobserved confounding at the cluster level. We followed a semi-empiric simulation strategy (see for example Huber et al. [5]) in the sense that we kept the original set of covariates and introduced an additional hospital level variable (H) to analyze the confounding effect. The variable H is set up constant for all observations in the same hospital. We then simulated the effect of a null, mild and strong confounding effect of H on the balance and the ATT by increasing its coefficient (β_H) in the outcome and treatment equations.

Simulation results show that when there is no unobserved confounding ($\beta_H = 0$) approaches B-E yield a similar average balance, which is only slightly higher than the balance attained in approach A, which is the best approach in this situation. However, when the size of the confounding effect increases, approaches B-E yield considerably lower average balance and bias than approach A and so should be preferred when unobserved confounding at the cluster level is suspected.

4 Concluding remarks

In this paper we discuss the advantages and drawbacks of different techniques to implement propensity score matching with clustered data. We apply these techniques to a population dataset containing information on the birth event in a two year period, clustered in twenty hospitals. When clusters size are big as in our application and there is potential confounding due to unobserved hospital level variables, an effective approach consists in implementing the matching within clusters or starting with a within matching approach and then use the pooled sample for remaining unmatched cases.

Acknowledgement

We would like to thank the Autonomous Region of Sardinia for providing the anonymized data used in the empirical application.

Bibliography

[1] Annibale D.J., Hulsey T.C., Wagner C.L. et al. (1995) *Comparative neonatal morbidity of abdominal and vaginal deliveries after uncomplicated pregnancies,* Arch Pediatr Adolesc Med Aug, **149**(8),862-7.

[2] Arpino, B. and Mealli, F. (2011) *The specification of the propensity score in multilevel observational studies,* Computational Statistics and Data Analysis, **55**, 1770 -1780.

[3] Bragg, F., Cromwell, D.A., Edozien,L. et al. (2010) *Variation in rates of caesarean section among English NHS trusts after accounting for maternal and clinical risk: cross sectional study.* British Medical Journal; doi:10.1136/bmj.c506.

[4] Caceres, I.A., Arcaya, M., Declercq, E. et al. (2013) *Hospital Differences in Cesarean Deliveries in Massachusetts (US) 2004-2006: The Case against Case-Mix Artifact,* PLOS ONE **8**(3), doi:10.1371/journal.pone.0057817.

[5] Huber, M., Lechner, M. and Wunsch, C. (2013) *The performance of estimators based on the propensity score,* Journal of Econometrics **175**, 1-21.

[6] Hughes, R.G., Hunt., S.S. and Luft, H.S. (1987) *Effects of surgeon volume and hospital volume on quality of care in hospitals,* Med Care **25**, 489-503.

[7] Imbens, G.W. (2004) *Nonparametric estimation of average treatment effects under exogeneity: a review,* Review of Economics and Statistics, **86**, 4-30.

[8] Li, F., Zaslavsky, A. M., and Landrum, M. B.(2013) *Propensity score weighting with multilevel data,* Statistics in Medicine, **32**(19), 3373-3387.

[9] Rosenbaum, P.R., Rubin, D.B (1983) *The central role of the propensity score in observational studies for causal effects,* Biometrika, **70**, 41-55.

[10] Sekhon, J.S. (2011) *Multivariate and Propensity Score Matching Software with Automated Balance Optimization,* Journal of Statistical Software, **42**(7), 1-52.

[11] Thoemmes, F.J. and West, S.G. (2011) *The use of propensity scores for nonrandomized designs with clustered data,* Multivariate Behavioral Research, **46**(3), 514-543.