Title:

Statistical Power in Studies that use the Synthetic Control Method

Authors
Chaplin, Duncan DChaplin@Mathematica-mpr.com (presenting)
Dotter, Dallas (DDotter@Mathematica-mpr.com)
Ingwersen, Nick (NIngwersen@Mathematica-mpr.com)
Mamun, Arif (AMamun@Mathematica-mpr.com)

In this paper we describe statistical power obtained when using the Synthetic Control Method (SCM) and investigate various factors affecting that power, namely, the number of comparison clusters, the number of units per cluster, the intraclass correlation (ICC), and the R-squared statistic.

Many policy-relevant interventions are implemented in a single location. This presents challenges for impact evaluations which SCM can help address. SCM compares outcomes from a treatment location to those from comparison locations weighted so that pre-treatment outcomes are similar on average. It provides researchers with more rigorous tools to select matching weights and describe the precision of the resulting estimates than were previously available for this situation. Unfortunately it is difficult to find clear descriptions of statistical power for this method. Many SCM studies lack standard errors of their estimates (Abadie et al. 2010, Barlow 2018, Billmeier and Nannicini 2013, Sills et al 2017, and Zhou 2018) or standard deviations of their outcomes (Bove et al 2014, Campos et al. 2014, Duodchenko and Imbens 2016, Linden 2017, Peri and Yasenov 2017, and Xu 2017) making it difficult to compare their results.

We present power calculations based on three completed SCM studies—the effects of California's tobacco control program (Abadie et al. 2010), German reunification (Abadie et al. 2014), and the Mariel boatlift (Peri and Yasenov 2017). We also present simulations based on plans for a fourth study on education reform in Washington, DC (Dotter et al. 2018) to see how much the number of comparison clusters, the numbers of individual units per cluster, the ICC, and the R-squared statistic affect the statistical power of this method.

Population/Participants/Subjects

Table 1 presents the topics of the four impact studies and the corresponding samples sizes. In each case the policy in question was mostly implemented in just one geographic area.

Table 1: SCM studies by topic areas, and their sample sizes at the cluster level

Study	Study topic	Treatment units	Comparison units
Tobacco sales (Abadie et al. 2010)	Impacts of taxes on cigarette consumption	1	29
German reunification (Abadie et al. 2014)	Impacts of German reunification on GDP	1	16
Mariel Boatlift (Peri and Yasenov 2017)	Impact of the Mariel boatlift (from Cuba to Miami, Florida) on wages	1	35
DC education reform (Dotter et al. 2018)	Impact of 2007 DC education reforms on student achievement	1	30

Research Design and Data

We present power calculations based on results from the three completed studies. We also use simulated power to see how much the number of comparison clusters, the numbers of individual units per cluster, the ICC, and the R-squared statistic affect the statistical power of this method for the fourth study. Data from the three completed studies is used to calculate the standard deviations of their outcomes. We use data for the fourth study to estimate the ICC and R-squared statistics that will be obtained in the context of statewide NAEP achievement. For each study, we assume that the same R-squared statistic holds at the cluster and individual levels.

Findings

Table 2 presents minimum detectable effect sizes (MDEs) in standard deviation units of the outcome for the three SCM studies. The cluster-level results are based on the standard deviation of the outcome at the cluster-level—the state for tobacco sales, the country for German reunification, and the city for the Mariel boatlift. The MDEs are large, ranging from 1.98 to 0.98 at the cluster level. The Mariel boatlift paper by Peri and Yasenov (2017) is an SCM version of a paper by Borjas (2017). In this case we were able to obtain an MDE based on the standard deviation at the individual level. This MDE is, as Table 2 shows, much smaller than when using the cluster-level standard deviation because the individual-level standard deviation is much larger than the cluster-level standard deviation.

Table 2: Minimum detectable effect sizes by study (in standard deviation unit)

Study	Cluster level	Individual level
Tobacco sales (Abadie et al 2010)	1.98	
German reunification (Abadie et al 2014)	0.98	
Mariel Boatlift (Peri and Yasenov 2017)	1.32	0.23

Notes: The MDEs were estimated as being about 2.9 times the standard error of the impact estimate divided by the standard deviation of the outcome. For the first two papers the standard errors were obtained from estimates produced by Duodchenko and Imbens (2017). The standard error for the third paper was from the SCM results (column 5) in the 1981-1983 row in table 5 of that paper.

The remaining tables present estimated statistical power for the SCM study of DC education reform. The outcome for that study will be scores on the National Assessment of Educational Progress, so we use the student-level standard deviation in test scores to create our MDEs. Table

3 shows how the number of comparison group cities is expected to affect our statistical power. As can be seen, the MDE drops noticeably as the number of cities rises from 7 to 30 but additional cities after 30 do far less to improve the statistical power. At that point, the power is driven primarily by the fact that there is only a single treatment city. To see this, note that in the absence of control variables or matching, the formula for the variance of the estimator, in standard deviation units, is V(β)=(1/Nt)+(1/Nc), where β=difference in outcomes between the treated units and the comparison units, Nt is the number of treatment units and Nc is the number of comparison units. Since Nt=1 this formula boils down to V(β)=1 + (1/Nc).

Table 3: Minimum detectable effect sizes by number of comparison group cities for the DC education reform study

Number of comparison clusters (cities)	Minimum detectable effect size
7	1.070
15	0.929
30	0.880
60	0.858
120	0.847
240	0.842
480	0.839

Note: Effect sizes assume an R-squared of 0.50, an intraclass correlation of 0.15 (based on NAEP data), and 30 students per city. Minimum detectable effect sizes are calculated using the student-level standard deviation, two-tailed tests with a 5 percent significance level,and 80 percent power.

Table 4 shows the relationship between statistical power and the number of individual students per city (that is, units per cluster) in the sample. Power increases noticeably when the number of students is few—below around 32, but at a decreasing rate for larger numbers.

Table 4: Minimum detectable effects by students per comparison city for DC study

Units per cluster (students per city)	Minimum detectable effect size
1	2.084
2	1.580
4	1.255
8	1.055
16	0.939
32	0.876
64	0.842
128	0.825

Note: Effect sizes assume an R-squared of 0.50, an intraclass correlation of 0.15 (based on NAEP data), and 30 comparison cities. Minimum detectable effect sizes are calculated using the student-level standard deviation, two-tailed tests with a 5 percent significance level,and 80 percent power.

Table 5 shows how the MDEs vary with the intraclass correlation. As can be seen in this table the MDEs rise substantially at all levels of the ICC. Increased variation within clusters makes the ICC smaller and the standard deviation larger which in turn makes the MDE smaller.

Table 5: Minimum detectable effects by intraclass correlation for DC study

Intraclass Correlation	Minimum detectable effect size
0.01	0.432
0.06	0.630
0.11	0.779
0.16	0.904
0.21	1.013
0.26	1.112
0.31	1.203
0.36	1.287

Note: Effect sizes assume an R-squared of 0.50, 30 students per city, and 30 cities. Minimum detectable effect sizes are calculated using the student-level standard deviation, two-tailed tests with a 5 percent significance level,and 80 percent power.

Table 6 shows how the minimum detectable effects vary with the R-squared statistic. In this case the R-squared actually makes relatively little difference until it gets over 0.40. Indeed, the impact of the R-squared continues to grow. This is because the unexplained variation is what is driving the uncertainty. Moving the R-squared from 0.10 to 0.20 only reduces unexplained variation by around 11 percent—from 0.90 to 0.80. This reduces the MDE by around 6 percent. In contrast, moving the R-squared from 0.90 to 0.95 reduces unexplained variation from 0.10 to 0.05—a decrease of 50 percent and reduces the MDE by about 29 percent.

Table 6: Minimum detectable effects by R-squared for DC study

R-squared statistic	Minimum detectable effect size	Percent change in MDE from previous row
0.10	1.181	--
0.20	1.113	6
0.30	1.041	6
0.40	0.964	7
0.50	0.880	9
0.60	0.787	11
0.70	0.682	13
0.80	0.557	18
0.90	0.394	29
0.95	0.278	29

Note: Effect sizes assume an intraclass correlation of 0.15 (based on NAEP data), 30 students per city, and 30 comparison cities. Minimum detectable effect sizes are calculated using the student-level standard deviation, two-tailed tests with a 5 percent significance level,and 80 percent power.

Conclusions

The SCM produces very large MDEs, even with large numbers of comparison clusters or individual units within clusters. However, the MDEs can be smaller if matching with comparison clusters reduces unexplained variation in the outcome substantially (i.e. the R-squared is high) and if the standard deviation at the individual level is much larger than at the cluster level. Even then, the method may produce underpowered impact estimates if detecting true impacts smaller than the relatively large MDEs above is of interest.

References

Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American statistical Association*, *105*(490), 493-505.

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller (2014). Comparative politics and the synthetic control method, *American Journal of Political Science*, pp. 2011–25.

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2017). Matrix completion methods for causal panel data models. *arXiv preprint arXiv:1710.10251*.

Barlow, Pepita (2018). Does trade liberalization reduce child mortality in low- and middle-income countries? A synthetic control analysis of 36 policy experiments, 1963-2005, *Social Science and Medicine,* 205: 107-115.

Billmeier, A., Nannicini, T. (2013). Assessing economic liberalization episodes: a synthetic control approach. Review of Economics and Statistics. 95 (3), 983–1001.

Borjas, George (2017) ”The Wage Impact of the Marielitos: A Reapprisal” *Industrial Relation and Labor Review,* 70 (5), 1077-1100.

Bove, Vincenzo, Leandro Elia, and Ron P. Smith (2014). The Relationship between Panel and Synthetic Control Estimators on the Effect of Civil War. Working Paper, October.

Campos, Nauro F., Fabrizio Coricelli, Luigi Moretti (2014). Economic Growth and Political Integration: Estimating the Benefits from Membership in the European Union Using the Synthetic Counterfactuals Method. IZA Discussion Paper No. 8162, April.

Card, David (1990). The impact of the mariel boatlift on the miami labor market. *Industrial and Labor Relation,* 43(2):245–257.

Dotter, Dallas, Duncan Chaplin, and Steven Glazerman (2018) Proposed Analysis Plan for LJAF Grant, Study Title: Evaluation of Portfolio School Reforms in Washington, D.C.

Doudchenko, N., & Imbens, G. W. (2016). *Balancing, regression, difference-in-differences and synthetic control methods: A synthesis* (No. w22791). National Bureau of Economic Research.

Linden, Ariel (2017). Combining synthetic controls and interrupted time series analysis to improve causal inference in program evaluation, *Journal of Evaluation in Clinical Practice*, 24:447–453.

Peri, Giovanni and Vasil Yasenov (2017). The labor market effects of a refugee wave: Applying the synthetic control method to the mariel boatlift. Working paper 21801, National Bureau of Economic Research.

Sills, Erin O., Diego Herrera, A. Justin Kirkpatrick, Amintas Brandão, Jr., Rebecca Dickson, Simon Hall, Subhrendu Pattanayak, David Shoch, Mariana Vedoveto, Luisa Young, Alexander Pfaff (2017). Estimating the Impacts of Local Policy Innovation: The Synthetic Control Method Applied to Tropical Deforestation, *PLOS ONE*, July: 1-15.

Xu, Yiqing (2017). Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models, *Political Analysis*, 25:57–76

Zhou, Yang (2018). Do ideology movements and legal intervention matter: A synthetic control analysis of the Chongqing Model, *European Journal of Political Economy*, 51: 44-56.