Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

DISCUSSION PAPER

Applying a propensity score-based weighting model to interrupted time series data: improving causal inference in programme evaluation

Ariel Linden DrPH MS[1] and John L. Adams PhD[2]

[1]President, Linden Consulting Group, Hillsboro, OR, USA
[2]Senior Statistician, RAND Corporation, Santa Monica, CA, USA

Keywords
aggregate outcomes, interrupted time series analysis, propensity score, treatment weights

Correspondence
Dr Ariel Linden
Linden Consulting Group
Hillsboro, OR 97124
USA
E-mail: alinden@lindenconsulting.org

Accepted for publication: 19 April 2010

doi:10.1111/j.1365-2753.2010.01504.x

Abstract

Often, when conducting programme evaluations or studying the effects of policy changes, researchers may only have access to aggregated time series data, presented as observations spanning both the pre- and post-intervention periods. The most basic analytic model using these data requires only a single group and models the intervention effect using repeated measurements of the dependent variable. This model controls for regression to the mean and is likely to detect a treatment effect if it is sufficiently large. However, many potential sources of bias still remain. Adding one or more control groups to this model could strengthen causal inference if the groups are comparable on pre-intervention covariates and level and trend of the dependent variable. If this condition is not met, the validity of the study findings could be called into question. In this paper we describe a propensity score-based weighted regression model, which overcomes these limitations by weighting the control groups to represent the average outcome that the treatment group would have exhibited in the absence of the intervention. We illustrate this technique studying cigarette sales in California before and after the passage of Proposition 99 in California in 1989. While our results were similar to those of the Synthetic Control method, the weighting approach has the advantage of being technically less complicated, rooted in regression techniques familiar to most researchers, easy to implement using any basic statistical software, may accommodate any number of treatment units, and allows for greater flexibility in the choice of treatment effect estimators.

1. Introduction

Often, when conducting programme evaluations or studying the effects of policy changes, researchers may only have access to outcome measures reported at the aggregate level. In health care research, these metrics generally include utilization rates of various services (hospitalizations, emergency department, office visits, prescription fills, etc.), medical costs (usually reported as per-member-per-month), and mortality rates. The outcome variable is typically ordered as a time series, with a number of observations captured in both the pre- and post-intervention periods. The study design is generally referred to as an *interrupted time series* because the intervention is expected to 'interrupt' the level and/or trend subsequent to its introduction [1–3].

Time series analysis (TSA) is considered a relatively robust observational study design, even in the absence of a comparison group, due primarily to its control over the effects of *regression to the mean*. Stated differently, when only two measurements are taken (i.e. pre-post), high (or low) initial values will likely be followed by observations closer to the average value, but over the course of many repeated observations, this natural variability narrows around the true mean, allowing the researcher to more accurately estimate the treatment effect of an intervention [4,5]. Nevertheless, without a concurrent comparison group, the treatment effect of a single study group may still be biased because of selection issues or secular trends. Therefore, a TSA can be much strengthened with the addition of one or more control groups.

As is the case with any observational study, the researcher conducting the TSA will attempt to emulate the randomization process of a randomized controlled trial (RCT) by finding (or creating) a control group that is approximately equivalent to the treatment group on known pre-intervention characteristics and hope that the remaining unknown characteristics are inconsequential and will not bias the results [6]. When only one comparison group is available for the TSA, conventional regression modelling may be the only viable approach to account for pre-intervention differences between the groups, even though there is evidence that these methods may provide biased results, most notably in the presence of time-dependent confounders [7,8].

The TSA is therefore more robust when several control groups are available, and when additional covariates (other than just the outcome variable under study) can be used to further adjust for differences between groups. When such data are available and a robust evaluation is desired, one option is the *synthetic control* method [9,10]. This is a recently developed technique that estimates the treatment effect by comparing the trajectory of an aggregate outcome for a treated unit to the evolution of the same aggregate outcome for a synthetic control group in terms of the outcome predictors. This synthetic control group is constructed using a data-driven regression-based method to obtain weights for each variable contained in the V-matrix. A constrained quadratic programming routine is then used to find the best fitting weights conditional on the regression-based V-matrix. A more complex yet better-fitting algorithm can be used instead, relying on a fully nested optimization procedure that searches among all (diagonal) positive semi-definite V-matrices and sets of weights for the best fitting convex combination of the control units. Once the synthetic control group has been created, the researcher can conduct a variety of placebo and permutation tests that produce informative inference [11].

In this paper an alternative method for estimating the treatment effect of an intervention using time series data is proposed. Our analytic approach is based on a weighted modelling technique originally developed for unit-level longitudinal studies [8,12], and follows a three-step approach: First, the propensity score is estimated for the treatment group and all potential controls [13]. Second, weights are constructed based on the propensity score and treatment assignment, and third, these weights are then used within a regression framework to provide a treatment effect estimate. We posit that researchers may prefer the proposed analytic method over the synthetic controls method because this approach: (1) is technically less complicated and rooted in regression techniques familiar to most researchers and can be implemented using any basic statistical software without elaborate programming; (2) may accommodate any number of treatment units (as opposed to the synthetic control method which is limited to only one treatment group); and (3) allows for greater flexibility in the choice of treatment effect estimators[i.e. average treatment effects (ATEs)].

This paper is organized as follows. In section 2 we briefly describe the dataset used for all analyses conducted here. Section 3 provides a basic tutorial on the most commonly used regression modelling techniques for single group and multiple group interrupted TSA and we then apply these models to the current data. In section 4, we describe the propensity score-based weighting framework applied to aggregated time series data and then apply this model to the current data. Section 5 discusses the results of our analyses using the models described here and compare them to those results generated using the synthetic control method, Section 6 provides a discussion and Section 7 concludes.

2. Data

In 1988, California passed the voter-initiative Proposition 99 which was a wide-spread effort to reduce smoking rates by raising the cigarette excise tax by 25 cents per pack and fund anti-smoking campaigns and other related activities throughout the state (for a comprehensive discussion of this initiative see Abadie *et al*. [10]). Per-capita cigarette sales (in packs) is the most widely used indicator of smoking prevalence found in the tobacco research literature [10], and serves here as the aggregate outcome variable under study, measured at the state level from 1970 until 2000 (with 1989 representing the first year of the intervention). The current data file was obtained from Abadie *et al*. [11], who originally obtained the cigarette sales data and average retail price of cigarettes from Orzechowski and Walker [14] and supplemented the file with the following covariates: per-capita state personal income (logged), the percentage of the population age 15–24, and per-capita beer consumption (for a complete listing of data sources, see appendix A of Abadie *et al*. [10]). Eleven states were discarded from the dataset because of their adoption of some other large-scale tobacco control programme at some point during California's intervention period under study between 1989 and 2000, leaving 38 states as potential controls [10].

Several of the covariates had data missing for certain years. Personal income was missing for years 1970, 1971, 1988–2000. Beer consumption was missing for years 1970–1983, 1998–2000, and the percentage of the population aged 15–24 was missing for years 1991–2000. For exposition purposes, we filled in all missing values using the 'impute' command in Stata, which runs regressions by best-subset regression, looking at the pattern of missing values in the predictor variables [15].

3. Basic regression models for interrupted time series analysis

Single group analysis

Regression (either ordinary or generalized least-squares methods) is the most commonly used modelling technique in interrupted time series analyses. When there is only one group under study (no comparison groups) the regression model assumes the following form: [16]

$$Y_t = \beta_0 + \beta_1 T + \beta_2 X_t + \beta_3 T X_t \quad (1)$$

Where Y_t is the aggregated outcome variable measured at each time-point t, T is the time since the start of the study, X_t is a dummy variable representing the intervention (pre-intervention periods = 0, otherwise 1), and TX_t is an interaction term. These terms are perhaps best explained using the lower half of Fig. 1. In the case of a single group study, β_0 represents the intercept, or starting level of the aggregated outcome variable. β_1 is the slope, or trajectory of the outcome variable until the introduction of the intervention. β_2 represents the intercept at the time of introduction of the intervention, and indicates whether there was a change in the level of the outcome immediately following the introduction of the intervention, and β_3 represents the change in slope or trajectory of the outcome after introduction of the intervention until the end of the study. Thus, we look for significant *P*-values in either β_2 or β_3 (or both) to indicate a treatment effect. Additionally, the magnitude of change in the outcome at any time-point after introduction of the intervention can be expressed in either absolute or relative terms [17,18].

Figure 2 visually displays the results of the single group TSA conducted on per-capita cigarette sales (in packs) in California

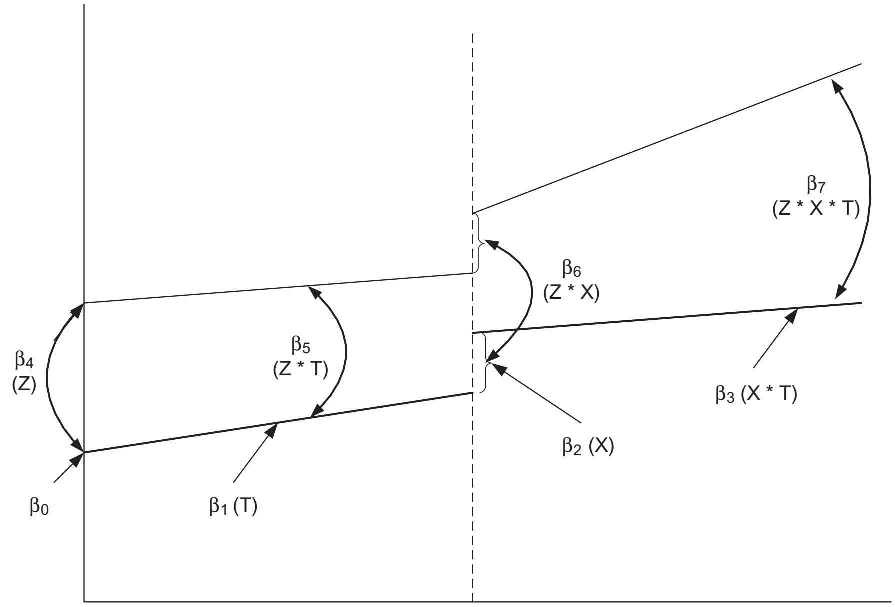

Figure 1 Regression intercepts and slopes for an interrupted time series analysis comparing a treatment group to control(s). T, time; X, intervention; Z, group.

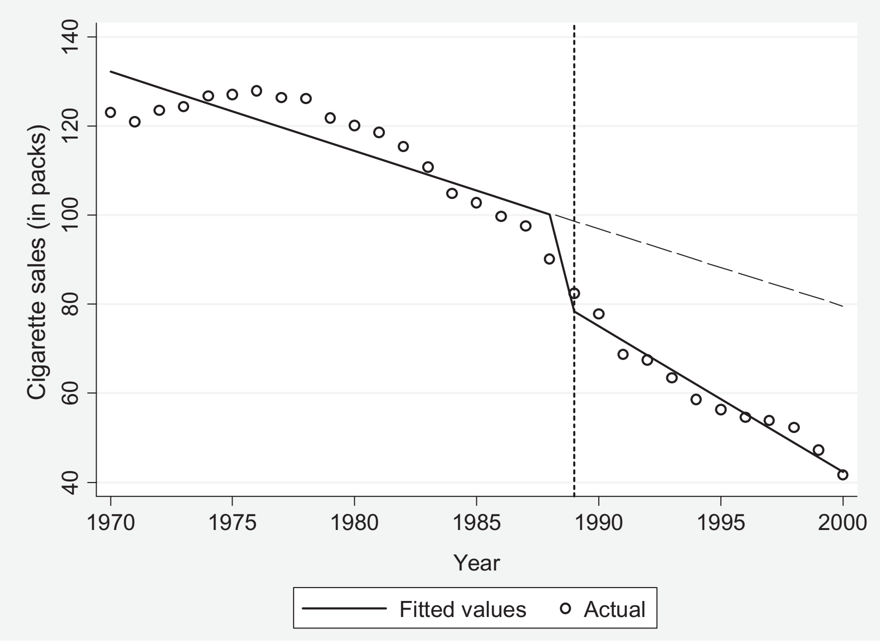

Figure 2 Actual and model-fitted per-capita cigarette sales (in packs) in California before and after introduction of Proposition 99. The broken line extending from the pre-intervention period trend-line represents the counterfactual of what cigarette sales would be in California absent Proposition 99.

before and after the introduction of Proposition 99. The starting level of the cigarette sales was estimated as 134 packs per capita, and sales appeared to be decreasing significantly every year prior to 1989 by −1.78 packs per capita [$P < 0.0001$, 95% confidence interval (CI) = −2.22, −1.33]. In the year immediately after the intervention (1989) there appeared to be an −18.56 packs per capita decrease in the level of cigarette sales ($P < 0.0001$, 95% CI, −26.62, −10.51) and a change in the slope (relative to the pre-intervention slope) of the annual sales of cigarettes of −1.49 packs per year ($P = 0.005$, 95% CI = −2.49, −0.50).

Also shown in Fig. 2 is a dashed-line that extends from the pre-intervention period until the end of the observation period. This line represents the counterfactual of what cigarette sales would have been in California had Proposition 99 not be initiated. We estimate from the gap between the actual cigarette sales and the counterfactual sales that in the period between 1989 and 2000 cigarette consumption in California was reduced by an average of 28.28 packs per capita.

An important feature of time series is that of serial dependence. Any outcome measured over time is potentially influenced by previous observations (referred to as autocorrelation or autoregression). When using linear regression models to fit time series data, it is important to test for autocorrelation since the error terms will likely be positively correlated, biasing the estimated standard errors downward, and thereby yielding F-tests with exaggerated significance [16]. We used Durban's test [19], which reported a significant ($P < 0.0001$, $\chi^2 = 25.15$) first-order autoregressive process (indicating autocorrelation with the most recent past value).

To correct for this, we then used the Prais–Winston estimator [20], which implements the generalized least-squares method to estimate the parameters, assuming the errors follow a first-order autoregressive process. After this adjustment, there appeared to be a 4.10 packs per capita decrease in the level of cigarette sales in the first year after introduction of Proposition 99 ($P < 0.0001$, 95% CI = −5.21, −2.98), and the change in the slope of the annual sales of cigarettes was now −2.0 packs per year ($P = 0.039$, 95% CI = −3.88, −0.11). One can conclude from a comparison of results from the adjusted versus unadjusted models that the change in annual cigarette sales after Proposition 99 was similar between the two models, but the autocorrelation-adjusted model reduced the magnitude of the change in the level of cigarette sales in the year immediately following Proposition 99 compared with the unadjusted regression model.

Multiple group analysis

When one or more control groups are available for comparison, the regression model in Eq. 1 is expanded to include four additional terms (β_4–β_7) [16]:

$$Y_t = \beta_0 + \beta_1 T + \beta_2 X_t + \beta_3 T X_t + \beta_4 Z + \beta_5 Z T + \beta_6 Z X_t + \beta_7 Z X_t T \quad (2)$$

Where Z is a dummy variable to denote the cohort assignment (treatment or control) and ZT, ZX_t and ZX_tT are all interaction terms among previously described variables. Now when examining Fig. 1, the coefficients of the lower line (β_0–β_3) represent the control group and the coefficients of the upper line (β_4–β_7) represent values of the treatment group. More specifically, β_4 represents the difference in the level (intercept) of the dependent variable between treatment and controls prior to the intervention, β_5 represents the difference in the trajectory (slope) of the dependent variable between treatment and controls prior to the intervention, β_6 indicates the difference between treatment and control groups in the level of the dependent variable immediately following introduction of the intervention, and β_7 represents the difference between treatment and control groups in the trajectory of the outcome variable after initiation of the intervention.

The two parameters (β_4–β_5) play a particularly important role in establishing whether the treatment and control groups are balanced on both the level and trajectory of the dependent variable in the pre-intervention period. If these data were from an RCT, we would expect there to be similar levels and slopes prior to the intervention. However, in an observational study where equivalence between groups cannot be ensured, any observed differences will likely raise concerns about the ability to draw causal inferences about the outcomes.

Figure 3 illustrates the actual and predicted per-capita cigarette sales (in packs) of California and the 38 remaining states in the dataset outside of California (as controls), before and after introduction of Proposition 99, using Equation 2. The initial mean level difference between California and the remaining states (parameter β_4) was not significant ($P = 0.55$, 95% CI = −7.82, 14.58), but the difference in the mean baseline slope (parameter β_5) was significant ($P = 0.007$, 95% CI = −2.55, −0.414). This is verified upon visual inspection of Fig. 3, as the trajectory of mean cigarette sales for the 38 states appears to rise higher than in California and that level remains elevated throughout the duration of the observation period.

Given this differential pattern of change in the baseline, one could argue that the 38 other states were not comparable to California and thus any change in outcomes after Proposition 99 could be biased. Additionally, one can see that the linear model does not fit the baseline data well, calling into question the model's ability to accurately estimate the treatment effect parameters (β_6–β_7). We estimate from the gap between the actual cigarette sales and the

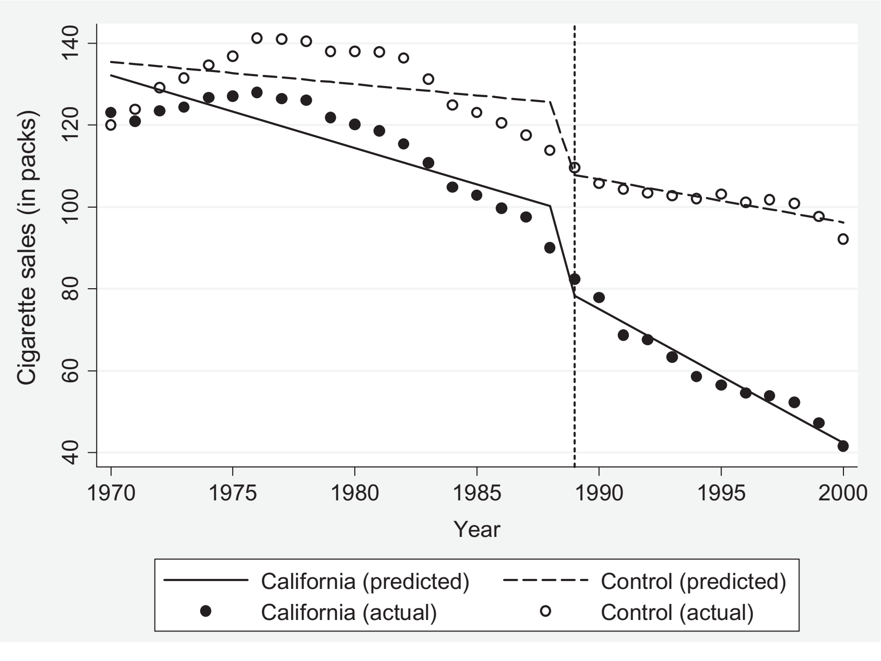

Figure 3 Actual and predicted per-capita cigarette sales (in packs) of California and 38 other states (as controls), before and after introduction of Proposition 99.

predicted values of the controls states that in the period between 1989 and 2000 cigarette consumption in California was reduced by an average of 41.71 packs per capita, much higher than what was estimated in the single group analysis. Given the differential pattern of change in the baseline between the two groups this gap is somewhat misleading. In fact, neither parameter β_6 nor β_7 were statistically significant (*P*-value for β_6 was 0.68, 95% CI = −10.55, 6.92, and *P*-value for β_7 was 0.06, 95% CI = −2.03, 0.05). Adjusting for autocorrelation using the Prais–Winston estimator did not improve these results (*P*-value for β_6 was 0.09, 95% CI = −3.06, 0.20, and *P*-value for β_7 was 0.53, 95% CI = −2.55, 1.31).

These results highlight the importance of ensuring that treatment and control units are comparable on pre-intervention characteristics – in this case, pre-intervention level and trend of the outcome variable. Therefore, this model could be improved by limiting the choice of control groups to only those with similar values on these two metrics, or moving to a more comprehensive model as described in the next section.

4. Propensity score-based weighting approach for time series data

Estimating the propensity score

The propensity score, defined as the probability of assignment to the treatment group conditional on covariates [13], controls for pre-intervention differences between treated and non-treated groups. Propensity scores are typically derived from a logistic regression equation that reduces each participant's set of covariates to a single score. It has been demonstrated that, conditional on this score, all observed pretreatment covariates can be considered independent of group assignment, and in large samples, covariates will be distributed equally in both groups and will not confound estimated treatment effects [13].

In the propensity score model for the cigarette sales data, the outcome variable was the treatment assignment, coded either '1' for California or 0 for each of the remaining 38 states. Covariates included the pre-intervention (1970–1988) mean values for each of the four variables described in section 2, in addition to the mean cigarette sales during the pre-intervention years.

Construction of weights

Traditionally, the inverse probability of treatment has been used to construct these models [8,12]. The motivation for this originated in the survey sciences over 50 years ago to adjust for sampling probabilities [21] and are intended to provide an estimate of the ATE in the population. However, in the current study we are more interested in setting the distribution of covariates to be equal to that of the treated subjects and then estimating the average treatment effect on the treated (ATT). Thus, the treated group (California) is given a weight of 1 and the non-treated states are given a weight of the (propensity score)/(1 – propensity score) [22]. This ATT weighting mechanism makes the control group's outcomes represent the counterfactual outcomes of the treatment group by making the two groups similar with respect to observable pre-intervention characteristics (those variables included in the propensity score model) [22]. This ensures that balance is achieved between the treated and non-treated groups on pre-intervention characteristics and provides us with greater confidence that treatment effect estimates derived from observational data are unbiased (presuming that all sources of bias were accounted for in the estimated propensity score) [23].

Regression model estimation

Unbiased treatment effects can be estimated by fitting the appropriate regression model using the ATT weight (for example, in the Stata software package one would specify the ATT weight as either an analytic weight or sampling weight). Like any other outcome variable, the choice of regression model depends on the distribution of the outcome variable. This can be logistic regression for dichotomous variables, ordinary least squares (OLS) for continuous variables, Poisson for rates or rare events, and Cox regression for survival or censored cases. Some researchers prefer the use of generalized linear modelling (GLM) for its flexible distributional assumptions [24]. Regardless of which of these traditional regression models are used, standard errors must be adjusted to correct for within-subject correlation by either clustering at the individual level (the states) or using robust standard errors [25]. Alternatively, evaluators can choose from among more complex models specifically designed to account for within-subject (or group) correlation in longitudinal data).

After reviewing the distribution of cigarette sales using actual and various transformations (including separate analyses for pre- and post-intervention periods), we could not definitively conclude which model type was most suitable to fit the data. We therefore ran several GLM models specifying various distributional families paired with compatible link functions, followed by inspection of model output, Akaike's information criterion and graphic displays. Given that the relative treatment effect was similar in all models, we present the results using the standard OLS model. This allows us to keep the outcome variable on the original scale and directly compare the results of this analytic approach with that of the synthetic controls method.

Figure 4 illustrates the difference and 95% confidence intervals in annual cigarette sales for California over the other 38 states in the dataset using the ATT weights in the model estimation. As shown, the weighting mechanism provided good balance on the outcome variable in the pre-intervention years with all of the treatment effect confidence intervals crossing zero. Starting in 1990 California's cigarette sales dropped significantly below that of the other states and continued to drop until the end of the study in 2000. It should be noted that these confidence intervals may be inaccurate given that asymptotic limit theorems are being applied to a single treatment unit and a small number of controls. A perhaps more suitable approach would be to utilize a mixed model, which would estimate the random effects of the dependent variable for each state (over the years under study), and estimate the treatment effect as a fixed effect in the model. This is a more complicated approach and currently few software packages allow for the inclusion of weights in their mixed model procedures.

We estimate from the difference between the actual cigarette sales in California and the predicted values of the weighted control that in the period between 1989 and 2000 cigarette consumption in California was reduced by an average of 24.54 packs per capita. In summary, the analytic approach proposed here to evaluate the effect of an intervention using time series data involves a weighted

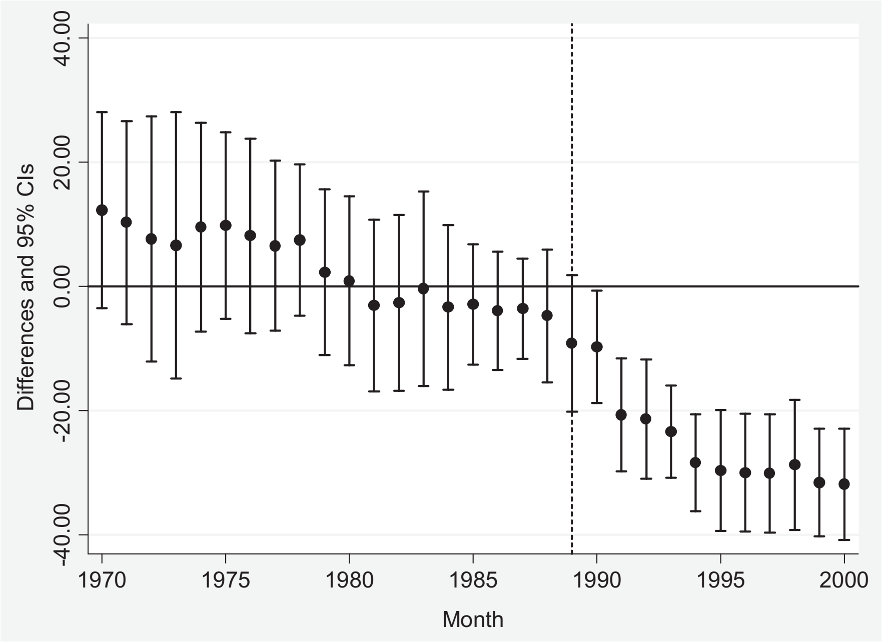

Figure 4 Weighted estimates of Proposition 99's effect on annual cigarette sales. Values and 95% confidence intervals (CIs, calculated using robust standard errors) represent California's annual cigarette sales relative to the other 38 states (solid horizontal line at zero).

regression of the aggregate outcome variable on treatment with the control group weighted to represent the average outcome that the treatment group would have exhibited in the absence of the intervention. When applying this model to the cigarette sales data, Proposition 99 appeared to have a significant effect commencing two years after the initiative was instated and continuing until the end of the study period.

5. Comparison of the various models

In reviewing the cigarette sales data results using the basic TSA models reported in section 3, an important limitation of these techniques became apparent. In the single-group model, a statistically significant treatment effect was found in the year immediately following the introduction of Proposition 99 as well as in the trajectory of sales until the end of the observation period. However, when California was compared with the other 38 states in the dataset, this treatment effect disappeared. Various arguments (both statistical and based on content knowledge) could be made in favour of one or the other model's findings. However, one could easily point out that in the multiple-group model, the groups were not comparable on the baseline slope, and thus the treatment effect estimates are suspect.

The propensity score-based weighting approach described in section 4 overcomes these limitations by making the two groups similar with respect to observable pre-intervention characteristics and thereby allowing for the estimation of unbiased treatment effects. The results from this model suggested that Proposition 99 indeed led to a significant reduction in cigarette sales in California.

We compared the results derived from our weighted regression approach to that of the synthetic control method [9,10] using identical predictor criteria. That is, we specified that the means of pre-intervention periods for all four covariates and the outcome variable be used to construct the synthetic control group. Figure 5 provides a visual comparison of the two models versus California over the course of the study period. The two techniques appear to track closely to each other along the entire continuum of observations, with the weighted model providing slightly lower estimates throughout. In relation to how the two models compared with California in the pre-intervention period, we see that between 1970 and 1979 the synthetic control group tracked closer to California than the weighted controls, and between 1980 and 1989 the weighted controls tracked closer to California than the synthetic controls. The two models provided very similar estimates for the intervention period. In fact, in comparing the gaps difference between actual sales in California to those of controls (from either model), we found that the difference between the two models in the estimated reduction of cigarette sales between 1989 and 2000 was only 1.55 packs (–26.09 for synthetic controls versus –24.54 for weighted controls).

We additionally compared the balance achieved on each of the predictor variables between California and both the synthetic and weighted control models. The synthetic controls achieved a mean percentage error (MPE) [4] close to zero and the weighted controls attained a very low MPE of 1.35% (data not shown).

6. Discussion

As illustrated in this paper, the ability to draw causal inferences about a treatment effect in observational time series data improves when a comparable control group is available (the groups similar with respect to observable pre-intervention characteristics). Moreover, it was shown that the propensity score-based weighting approach achieved similar results to that of the synthetic control method, inarguably, a very robust modelling approach to observational time series data.

We posit that researchers may prefer the weighted proposed analytic method over the synthetic controls method for several

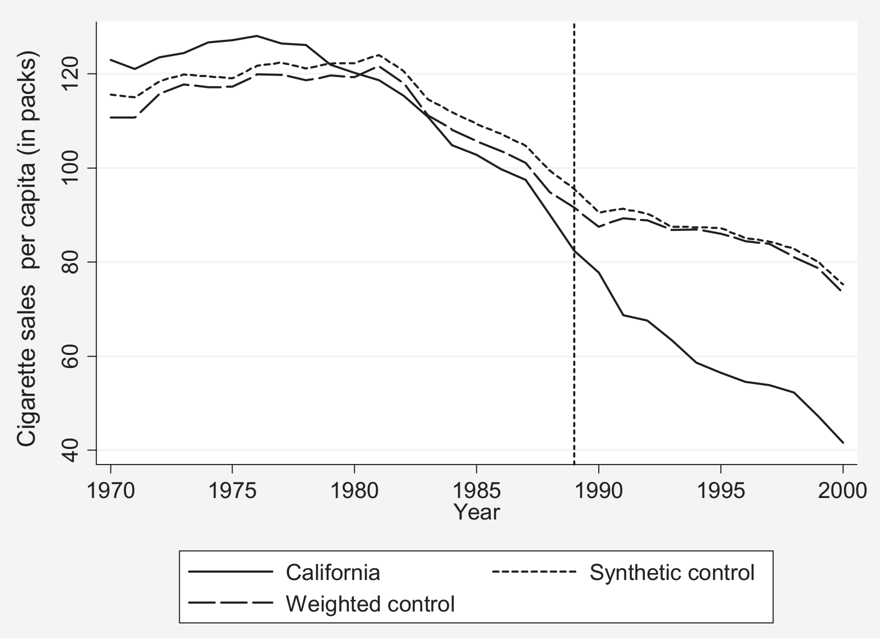

Figure 5 Comparison of actual cigarette sales in California to weighted controls and synthetic controls before and after Proposition 99.

reasons. First, the weighting approach is technically less complicated than the synthetic control method and rooted in the familiar regression framework. This means that the analytic process is easier to comprehend, perform, interpret and describe. Moreover, no special programming is required and the method can be implemented in any basic statistical software package.

Second, the weighting approach can easily accommodate any number of treatment units. Conversely, the synthetic controls algorithm only allows a single unit to be specified. If the intervention of interest affects several units, the researcher must first combine these units and then treat them as a single unit [11].

Third, the weighting approach allows for greater flexibility in the choice of treatment effect estimators. In the current study we used the ATT; however, the researcher can choose among other estimators as well, such as the ATE, which sets the distribution of covariates to be equal to that of the population, or the ATE on the controls, which sets the distribution of covariates to be equal to that of the control group [22].

While the current weighting approach has demonstrated its robustness, there are at least a couple ways in which the model could be strengthened. Boosted logistic regression [26] is worth considering as an alternative to the standard logistic model in estimating the propensity score. Regression boosting is a general, automated, data-adaptive modelling algorithm that can estimate the nonlinear relationship between the outcome variable (in this case, treatment assignment) and a large number of covariates including multiple level interaction terms resulting in greater accuracy over standard linear models [27].

Inferences about intervention effect

Abadie *et al*. [9,10] suggest that large sample inferential techniques are not well-suited to comparative case studies when the number of units in the comparison group and the number of periods in the sample are relatively small. They instead propose the use of placebo studies to draw inferences regarding treatment effect estimates. This method consists of iteratively casting non-treated States into the role of 'treated' and then applying the synthetic control method. The estimated gap between each of these 'treated' states and their synthetic controls provides an indication of whether the magnitude of the effect in California was meaningful. More specifically, one would not expect to see substantial gaps when comparing between non-treated states, but would expect a comparatively larger gap in California to its synthetic control, relative to the gaps in the placebo studies [9,10]. This inferential approach can be readily applied it to the weighting model as well, simply by recasting each non-treated unit iteratively as 'treated' within the propensity score estimation model and then rerunning the weighted regression on the outcome as usual.

The cigarette sales results from our weighted regression analysis (relying on the usual asymptotic limit theorems for inference implicit in this model) are in accord with the results generated using the synthetic control technique (see Fig. 4). The advantage here is that familiar statistical measures are provided and readily interpreted. As discussed previously a mixed modelling approach (which allows for the inclusion of weights) may offer more accurate confidence intervals.

Limitations of the propensity score-based weighting technique in time series data

As with any evaluation of observational data, the foremost limitation is that we presume that all biases and confounding have been adjusted for in the model, an assumption that cannot be tested outside of a randomized study. Also the control states available for the comparison must have substantial overlap with the treatment state. For example, if California had the highest use rate in the nation there would be no way to weight a combination of other states to make a comparable set. This problem can also occur for other covariates used in the model as well.

7. Conclusion

In this paper we have described several approaches to conducting interrupted TSA. The most basic model requires only a single group and models the intervention effect using repeated measurements of the dependent variable. This model controls for regression to the mean and is likely to detect a treatment if effect if it is sufficiently large. However, many potential sources of bias still remain. Adding a control group to this model could strengthen causal inference if the groups are comparable on the baseline level and trajectory of the dependent variable. If this condition is not met, the validity of the study findings could still be called into question, as occurred here with the cigarette sales data.

The propensity score-based weighted regression model described here overcomes these limitations by weighting the control group to represent the average outcome that the treatment group would have exhibited in the absence of the intervention. This weighted model approach was comparable to the synthetic control method in studying the effect of Proposition 99 on cigarette sales in California. However, this approach has the advantage of being technically less complicated, rooted in regression techniques familiar to most researchers, easy to implement using any basic statistical software without additional programming, may accommodate any number of treatment units, and allows for greater flexibility in the choice of treatment effect estimators.

References

1. Campbell, D. T. & Stanley, J. C. (1966) Experimental and Quasi-experimental Designs for Research. Chicago, IL: Rand McNally.
2. Cook, T. D. & Campbell, D. T. (1979) Quasi-experimentation: Design and Analysis Issues for Field Settings. Chicago, IL: Rand McNally College Publishing.
3. Shadish, S. R., Cook, T. D. & Campbell, D. T. (2002) Experimental and Quasi-experimental Designs for Generalized Causal Inference. Boston, MA: Houghton Mifflin.
4. Linden, A., Adams, J. & Roberts, N. (2003) Evaluating disease management program effectiveness: an introduction to time series analysis. *Disease Management: MD*, 6 (4), 243–255.
5. Linden, A. (2007) Estimating the effect of regression to the mean in health management programs. *Disease Management & Health Outcomes*, 15 (1), 7–12.
6. Rubin, D. (2007) The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26, 20–30.
7. Freedman, D. (1999) From association to causation: some remarks on the history of statistics. *Statistic Science*, 14, 243–258.
8. Robins, J. M., Hernán, M. A. & Brumback, B. (2000) Marginal structural models and causal inference in epidemiology. *Epidemiology (Cambridge, Mass.)*, 11, 550–560.
9. Abadie, A. & Gardeazabal, J. (2003) Economic costs of conflict: a case study of the Basque Country. *the American Economic Review*, 93 (1), 113–132.
10. Abadie, A., Diamond, A. & Hainmueller, J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. NBER Technical Working Paper No. 335. January 2007.
11. Abadie, A., Diamond, A. & Hainmueller, J. (2007) *Synth Software Package for Stata*. Available at: http://www.mit.edu/~jhainm/software.htm (last accessed 27 January 2010).
12. Robins, J. M. (1998) Marginal structural models. In 1997 Proceedings of the Section on Bayesian Statistical Science, pp. 1–10. Alexandria, VA: American Statistical Association.
13. Rosenbaum, P. R. & Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
14. Orzechowski and Walker (2005) The Tax Burden on Tobacco. Historical Compilation, Vol 40, 2005. Arlington, VA: Orzechowski & Walker.
15. Stata Corporation (2007) Impute – fill in missing values. In Stata Statistical Software: Release 10.0 User's Guide, pp. 268–272. College Station, TX: Stata Corporation.
16. Simonton, D. K. (1977) Cross sectional time-series experiments: some suggested statistical analyses. *Psychological Bulletin*, 84 (3), 489–502.
17. Wagner, A. K., Soumerai, S. B., Zhang, F. & Ross-Degnan, D. (2002) Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics*, 27, 299–309.
18. Zhang, F., Wagner, A. K., Soumerai, S. B. & Ross-Degnan, D. (2009) Methods for estimating confidence intervals in interrupted time series analyses of health interventions. *Journal of Clinical Epidemiology*, 62, 143–148.
19. Durbin, J. (1970) Testing for serial correlation in least-squares regressions when some of the regressors are lagged dependent variables. *Econometrica*, 38, 410–421.
20. Prais, S. J. & Winsten, C. B. Trend estimators and serial correlation. *Cowles Commission Discussion Paper No. 383*. Chicago; 1954.
21. Horvitz, D. G. & Thompson, D. J. (1952) A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663–685.
22. Nichols, A. (2008) Erratum and discussion of propensity-score reweighting. *Stata Journal*, 8 (4), 532–539.
23. Linden, A. & Adams, J. L. (2010) Using propensity score-based weighting in the evaluation of health management program effectiveness. *Journal of Evaluation in Clinical Practice*, 16, 175–179.
24. McCullagh, P. & Nelder, A. (1989) Generalized Linear Models, 2nd edn. London: Chapman and Hall.
25. White, H. A. (1980) A heteroscedasticity-consistent covariance matrix estimator and a direct test of heteroscedasticity. *Econometrica*, 48, 817–838.
26. Ridgeway, G. (1999) The state of boosting. *Computing Science and Statistics*, 31, 172–181.
27. McCaffrey, D., Ridgeway, G. & Morral, A. (2004) Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9 (4), 403–425.