Counterfactual Temporal Trends for Program Evaluation

Andrea Piesse, David Judkins
Westat, 1650 Research Blvd., Rockville, MD 20850

Keywords: Counterfactual projection weights, Dose-response, Media campaigns, Propensity scoring

1. Introduction

This paper presents an interpretative tool for use with survey data in the evaluation of ongoing programs, such as media campaigns. Temporal trends in the main outcomes of interest are often cited as evidence for or against the success of a campaign. However it is well known that trends are subject to the influences of other societal forces. Alternative methods, such as dose-response analysis, with adjustments made to remove the unwanted effects of confounders can provide more direct estimates of the effects of a campaign. For several reasons, these two approaches may lead to different conclusions. We propose counterfactual temporal trends—that is, trends for different exposure groups adjusted for measured confounders—as a means for understanding the nature of campaign effects when the evidence from temporal trends and dose-response analyses is not in exact agreement. The approach involves propensity scoring and survey weights.

2. Temporal Trends

Media campaigns are often used to try to change attitudes and/or behaviors, such as AIDS awareness, drug use, or exercise habits. The estimation of temporal trends from sample surveys is a component of many evaluations of ongoing campaigns. One advantage of the use of trend data for this purpose is the ease with which the results can be communicated to funders, policy makers, and the public in general. However in the context of program evaluation, reliance on temporal trends alone can be misleading. Consider, for example, the percentage of the population who are non-smokers. A significant upward trend in this outcome during the period of implementation of an anti-smoking campaign might be interpreted by many as evidence of campaign success. However if the price of tobacco products rose steadily over the same time period, there would be competing explanations for the observed trend. In general, it is difficult to determine the cause of any trend in outcomes.

3. Dose-response Analyses

One method that seeks to detect cause-and-effect relationships is dose-response analysis. In its application to the evaluation of media campaigns, the underlying theory is that if advertising is effective, then a larger dose of advertising exposure should be at least as effective as a smaller dose. If such a relationship is established, then a strong case for the effectiveness of the campaign has been demonstrated.

In dose-response analysis, one must assume that variation in dose is random after controlling for known background variables. Media campaign exposure is not randomly assigned, being self-selected by choices in media consumption and filtered through subject recall. Hence one must assume that all sources of nonrandom joint variation in exposure and outcomes have been measured, and then employ some form of confounder control. This is a strong assumption and requires data collection on a wide range of variables that might affect campaign exposure and outcomes of interest.

The effects of measured confounders can be removed using a variety of statistical techniques (Imbens, 2004). A popular choice is propensity scoring (Rosenbaum and Rubin, 1983) which will be described briefly in our current context. With propensity scoring, a statistical model is built for the probability of exposure in terms of potential confounders, and the associations between exposure and outcomes are conditioned on the estimated exposure probabilities (known as propensity scores). An attractive feature of propensity scoring is the existence of formal tests (called balance tests) that assess whether covariates have been effectively controlled. Another advantage of the technique applies when there are many different outcome variables. In this case the propensity scores are developed only once and applied for each outcome.

With two exposure groups—exposed and unexposed—the propensity scoring methodology first develops a model (often logistic) to predict group membership based on the potential confounders. One approach is then to use inverses of the predicted propensity scores as weights in the analysis. An alternative approach which reduces variance while eliminating slightly less bias is to divide the sample into a small number of strata—say five—based on the propensity scores. The

groups are then re-weighted to make the weighted count in each stratum the same in the exposed and unexposed groups. The re-weighting thus approximately equates the groups in terms of the confounders. If both exposure groups are weighted to the overall population distribution across the strata, each group will generate weighted estimates of the outcomes under the counterfactual scenario that the whole population received that level of exposure. For this reason, we refer to the propensity-score-adjusted weights as counterfactual projection (CFP) weights (Judkins *et al*., 2006).

The construction of the CFP weights for two exposure groups can be described as follows. For individuals with exposure level j in stratum s, compute a weighting adjustment factor

$$f_{\mathrm{CFP},s,j} = \frac{\sum_{i \in s} w_i}{\sum_{i \in s} \delta_{ij} w_i},$$

where w_i is the sampling weight for the i^{th} individual, and

$$\delta_{ij} \begin{cases} 1 \text{ if the } i^{\text{th}} \text{ individual has exposure level } j, \text{ and} \\ 0 \text{ otherwise} \end{cases}$$

The CFP weight for the i^{th} individual is then

$$w_{CFP,i} = w_i \sum_{j=0}^{1} \delta_{ij} f_{CFP,s(i),j} \qquad (1)$$

where $f_{CFP,s(i),j}$ is the adjustment factor for the exposure-propensity stratum containing the individual. Thus, propensity scoring can be viewed as a natural extension of survey weighting. These CFP weights are calibrated in the sense that the total population estimated using the CFP weights for each exposure group is the same as that estimated from the entire sample with the regular sampling weights. That is,

$$\sum_i \delta_{ij} w_{\mathrm{CFP},i} = \sum_i w_i \quad \forall\, j\,.$$

The estimated effect of exposure to the campaign on the mean of some variable, y, is then given by

$$\frac{\sum_i w_{CFP,i} \delta_{i1} y_i}{\sum_i w_{CFP,i} \delta_{i1}} - \frac{\sum_i w_{CFP,i} \delta_{i0} y_i}{\sum_i w_{CFP,i} \delta_{i0}}.$$

The propensity scoring methodology is readily extended to cover several exposure groups, for example by using an ordinal logit model (Joffe and Rosenbaum, 1990). When there are more than two groups, the summation in Equation (1) is from 0 to $J-1$, where J is the number of groups.

The gamma statistic is a nonparametric measure of association for ordered data (Agresti, 1990). As explained above, the best evidence for a dose-response relationship is a monotonic rise in a favorable outcome of interest as exposure increases. While there is not a perfect test for distinguishing this type of relationship from flat or complex patterns, the gamma statistic is a reasonable choice. When the gamma statistic is estimated using CFP weights the effects of measured confounders are removed from the exposure-outcome association.

4. Counterfactual Temporal Trends

Results from trend analyses and dose-response associations are not always consistent. Some evaluators have argued that both types of effects should be statistically significant before declaring the campaign a success, but this requirement is highly conservative (Judkins and Zador, 2002). Moreover, it does not address the effects of societal factors on trends. How then can we reconcile the two types of analysis?

The idea arose of decomposing an overall temporal trend into separate trend lines for exposed and unexposed individuals. Trend estimates are usually computed using regular cross-sectional sampling weights. However we know there is a need to control for confounders when comparing estimates for different exposure groups. With the dose-response analysis this can be achieved by using CFP weights. Similarly, we can use CFP weights to estimate counterfactual temporal trends: separate trend lines for exposed and unexposed individuals that have been adjusted to remove the effects of differences between the two groups. These counterfactual trends are proposed as an interpretive tool for understanding campaign effects when the results from overall trend and dose-response analyses are not in exact agreement. They also may be useful when the two forms of evidence are consistent – particularly, as a communication tool.

The procedure consists of four main steps:

1. Fit and balance a separate exposure propensity model for each time period, using exposure at that time period;
2. Create separate CFP weights for each time period;
3. Compute three sets of cross-sectional estimates for an outcome of interest by applying regular sampling weights to the entire sample and CFP weights to the exposed and unexposed individuals separately; and
4. Plot the overall temporal trend, and the exposed and unexposed counterfactual trend lines, on the same graph.

Clearly, many trend-line patterns may result from the steps above. We consider some of these in the next section.

5. Illustrations

Imagine the following examples in the context of an ongoing anti-smoking media campaign. Assume that survey data are collected prior to the initiation of the campaign to provide a baseline measure for the outcomes. First, suppose that there is no significant overall trend in the percentage of non-smokers in the population, but at the most recent time period there is a significant and favorable dose-response association between campaign exposure and non-smoker status. Decomposing the overall trend into exposed and unexposed counterfactual trend lines, will result in a plot of the type shown in Figure 1.

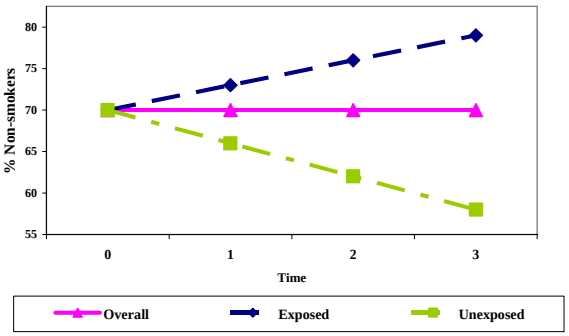

Figure 1. Counterfactual trends with flat overall trend

In general such situations may be described by an overall trend line that is flat, an exposed trend line that is upward, and an unexposed trend line that is downward. Reliance on the overall temporal trend would lead policy makers to the erroneous conclusion that the campaign is not working. However, the unexposed counterfactual trend shows that the outcome level would have declined if the campaign had not been implemented.

As another example, suppose that the significant favorable dose-response association is coincidental with a significant downward overall trend in the percentage of non-smokers, as illustrated in Figure 2.

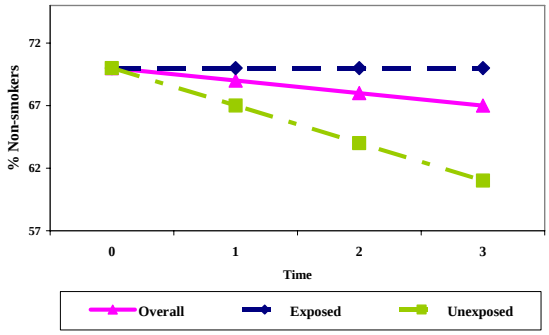

Figure 2. Counterfactual trends with downward overall trend

In such cases, the unexposed trend line has an even steeper downward slope than the overall trend, while the exposed trend line may be upward, flat, or downward at a lesser incline than the overall trend. Here, without the counterfactual trends, it may be even harder to convince a lay audience that the campaign is effective.

In both of the situations described above, the positive effects of a campaign are counteracted by external forces. The counterfactual trends provide convincing visual evidence of the success of a campaign that might well be abandoned on the basis of the overall trend alone. Of course, the reverse is also possible. An upward overall trend may mask the fact that exposed individuals are no better off than their unexposed counterparts.

6. Interpretation of Counterfactual Trends

While the counterfactual trends control for differences between exposed and unexposed individuals on measured covariates, they are each still subject to influences other than the campaign being evaluated. The exposed counterfactual trend estimates how other societal forces would have affected outcomes over time if the entire population had been exposed to the campaign. In contrast, the unexposed counterfactual trend estimates how other societal forces would have affected outcomes over time if none of the population had been exposed. The vertical difference between the two counterfactual trend lines at each time point can be

interpreted as a measure of maximum possible campaign effect at that time.

Similarly, the difference between the overall temporal trend and the unexposed counterfactual trend at each time point must be due to the campaign. These differences represent a measure of actual effect and take into account the estimated percentage of the population exposed to the campaign at each time period. Clearly, the greater this percentage, the closer the overall trend estimate will be to the exposed counterfactual trend estimate, and vice versa.

The estimates of both maximum and actual effects are likely to vary over time. This may be due to changes in the efficacy of the campaign (perhaps reflecting changes in advertising content), or due to differences in campaign reach which impact only the actual effects. In practice, there may be several ways to summarize the point-in-time differences between two trend lines. When there are true baseline measurements the overall temporal and counterfactual trends share a common estimate at the start of the series. One possibility would be to fit separate linear regressions to the two sets of counterfactual points and estimate the difference in the slope parameters of the two models. Another approach is to estimate the difference in gamma coefficients of association between time and outcomes for the two counterfactual trend lines. While the magnitudes of these gamma statistics represent crude summaries of the trend lines, the jackknifed gammas do provide useful tests for monotonically increasing or decreasing trends. Some users might also find it convenient to consider the same test statistic (i.e., gamma) for both the trend and dose-response analyses.

Before concluding, we note the following points. In the absence of true baseline outcome measures, the usefulness of comparing counterfactual trends may depend on the extent to which most of the campaign's effects preceded the evaluation's data collection. This is because the counterfactual trend estimates for the earliest available time period are likely to differ. Under certain conditions, the difference between the gamma coefficients (or slopes) of the exposed and unexposed counterfactual trends lines can be interpreted as a measure of change in maximum program efficacy over time. This assumes that the exposure-outcome association for time t^* does not depend on exposure at any time $t < t^*$.

7. Summary and Future Work

Counterfactual trends integrate two types of analysis into a coherent evaluation. They demonstrate the need for media campaign evaluations to look at dose-response or other confounder-controlled relationships. More fundamentally, they reinforce the need for a comparison and a treatment group. Significant differences between exposed and unexposed counterfactual trends (or between overall trends and unexposed counterfactual trends) can be directly attributed to the campaign. The approach can be applied to a wide range of surveys—cross-sectional, panel, in-person, RDD—and the graphical display communicates well to a wide audience.

Extensions to multi-level campaign exposure measures are currently underway, but the basic idea remains the same. As mentioned previously, the propensity scoring methodology is readily extended to cover several exposure groups. Counterfactual trend lines for different exposure levels can then be plotted on the same graph for visual inspection. The counterfactual trends may also be compared pairwise – for example, each to the unexposed group.

Acknowledgement

The authors thank Graham Kalton for valuable comments and suggestions for improvement on earlier drafts of the paper.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Imbens, G.W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, Vol. 86, No. 1, pp. 4-29.

Joffe, M.M., and Rosenbaum, P.R. (1990). Propensity scores. *American Journal of Epidemiology*, Vol. 150, pp. 327-333.

Judkins, D., Morganstein, D., Zador, P., Piesse, A., Barrett, B., and Mukhopadhyay, P. (2006). Variable selection and raking in propensity scoring. To appear in *Statistics in Medicine.*

Judkins, D. and Zador, P. (2002). Synthesis of alternate evaluation measures of public education campaigns. *Proceedings of the American Statistical Association*, pp. 1717-1722.

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