UNIVERSITÀ
DI SIENA
1240

Dipartimento di Economia politica e statistica

Corso di Laurea Magistrale in
ECONOMICS

Estimating measurement error in India's GDP :

A synthetic ccontrol approach

Relatore:
Prof. Federico Crudu
Correlatore:
Prof. Giuliano Curatola

Candidato:
Advait Moharir

Anno Accademico 2021-2022

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Introduction

Chapter 1

Motivation

Economic growth and its determinants are a key area of macroeconomic research. The question of achieving sustained and inclusive growth is the cornerstone of economic policy for most developing countries Accurate statistical measurement of GDP and related macroeconomic aggregates is crucial to identify the sectors of the economy that are performing well, and those that are stagnating. Consequently, the measurement of GDP and related magnitudes are an object of scrutiny from academia, media and policymakers.

However, India's GDP numbers have been scrutinized even more than usual, especially over the last decade. Commentators have identified a number of issues in the data post-2011, the major ones listed as follows. One, the assumption that the informal and formal sectors grow at the same rates has led to overestimation bias in the output of the manufacturing sector. Recent research has found significant difference in manufacturing output as computed by formal and informal sector surveys (CITE). Two, inappropriate deflators have been applied to nominal GDP data, and the application of the right deflators leads to significant difference in output estimates. Finally, the focal point of debate has centered on the use of a new database, called the MCA21 database. Critics have pointed that along with other issues like firm mis-classification and lack of state-level data, the assumption that non-reporting firms contribute positively to growth has led to a "blowing up" of manufacturing output.

The third issue has caused many to question the veracity of GDP estimates in the recent years. This followed by the lack of good quality data alternatives raise an important question - what is the extent of measurement error in India's GDP? In this chapter, I provide an overview of the debate on Indian GDP, focussing on

the exchange between R. Nagaraj and the Central Statistical Organization on the veracity of the MCA21 database. I then briefly examine the concerns raised by Arvind Subramanian, the former Chief Economic Advisor on the recent growth numbers as well as his alternative estimates. I end with my broad research question.

1.1 The Indian GDP Debate

1.1.1 The Nagaraj - CSO Debate

India's Central Statistics Office (CSO) is the agency responsible for publishing estimates of GDP since 1950.In accordance with best practices, the CSO also undertakes revisions of the base year used to measure GDP with constant prices. Updating the base year is a routine practice and typically does not affect the sectoral as well as the aggregate growth trend. However, the change in base year from 2004 to 2011 caused a number of significant changes in the new series, prompting an exchange between R Nagaraj, an economist at the Indira Gandhi Institute of Development Research (IGIDIR) and the CSO, in the pages of the *Economic and Political Weekly*.

In his article (Nagaraj, 2015a), Nagraj points out that there is a significant difference in growth rates (6.2% versus 4.8% for 2013), and that such a drastic difference in growth rates is a red flag. The primary difference between the old and new series is driven by changes in measures of gross value added, savings and investment due to a change in the database used by the ministry. The new database called MCA21, has been argued to be better as it contains firm returns filed with the Ministry of Corporate Affairs and has a much larger sample size. The veracity of the new series is questionable however, due to differences in in estimates reported by the sub-committee in its 2014 provisional report and the 2015 final report. PCS savings, investment and GVA shot up by 257%, 34% and 108% respectively between the old and new reports. This combined with the fact that the MCA21 is not publicly available, and that the estimates were not vetted independently was a cause of great concern.

In its reply (CSO, 2015), the CSO gave six reasons for the differences of estimates between the two reports, with the key explanation being that differences in timing of filing returns restrict the sample being considered, requiring the body to use a 'scaling factor' to scale up the numbers. The other reasons pertained to changes in definitions, usage of individual compo-

nents of revenue instead of total revenue, and usage of better quality data.

In his rejoinder (Nagaraj, 2015b), questions the usage of dis-aggregated revenue data instead of aggregate data, and argues that the drastic difference in estimates is indicative of lack of consistency in the MCA21 database. The main critique however, centers on the usage of the 'scaling up' factor: as the number of companies sampled was lower in 2013-14, than previous year, the scaling up factor is larger, causing overestimation. He further argues that the problem of PCS estimation precedes this procedure, as very few firms with large output being legally registered, already causing the PCS to skew upwards. Nagaraj concludes that reasons for doubting the new series hence remains valid. In a recent summary(Nagaraj, 2021) reiterates these concerns, and argues that PCS output remains overestimated since the concerns raised before.

The Nagaraj- CSO debate shows that measurement of manufacturing area output is an area of concern. Other technical issues with the new series are detailed in Nagaraj & Srinivasan (n.d.).

1.1.2 Subramanian's Cross- Country Approach

Arvind Subramanian, India's former Chief Economic Advisor makes the case that India's GDP is overestimated using a different methodology (Subramanian, 2019). As a 'smell test', he compiles a list of 17 indicators strongly correlated with GDP growth (electricity consumption, vehicle sales,real credit etc.), and finds that pre-2011, 16 of these are positively correlated with GDP growth, but the correlation shifts to a negative one for 11 of these indicators post-2011. This shows that there is a prima-facie concern in the post-2011 estimates, given that average growth rates in the periods pre-and post-2011 periods were similar.

To further test this proposition, Subramanian runs two sets of cross-country regressions. The idea is to identify a sample of countries which have a robust correlation between growth and some select indicators, and check if India's growth numbers fall into this broad pattern or not. The first regresses GDP growth on credit, electricity, import and export growth for a sample of countries, with two samples: pre and post-2011. The second one includes an "India dummy" along with time fixed effects to examine the effect of change in methodology on India's growth and relationship with these indicators. He finds that the coefficient is

statistically insignificant pre-2011 and significant post-2011, indicating that India is an outlier compared to the rest of the sample. Furthermore, when he estimates a panel specification to exploit time variation in each of these indicators with a sample of 75 countries, adding a period dummy (0 for pre-2011 and 1 for post-2011). The India*period interaction coefficient, which indicates differential mis-estimation, is statistically significant at 1%. Using these results, Subramanian argues that there is a 2.5% overestimation in India's GDP annually from 2011-2016.

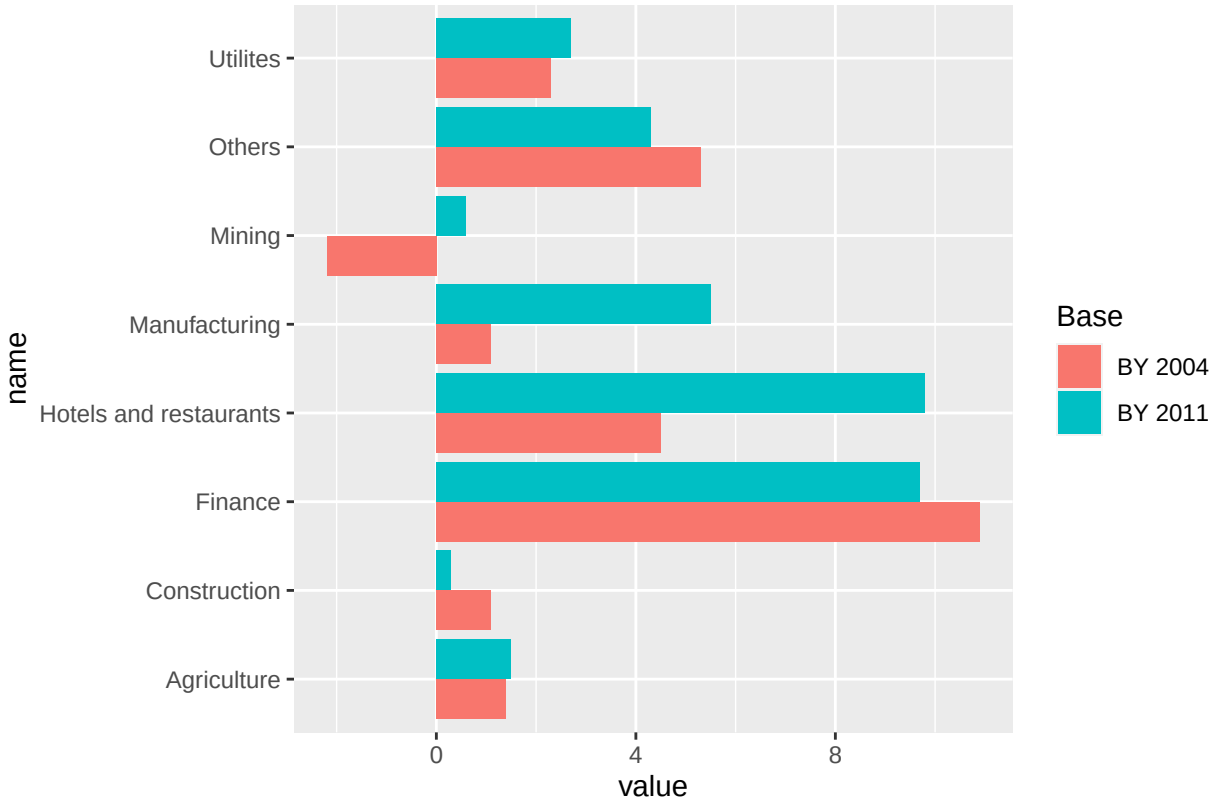

Figure 1.1a: Disaggregated GDP growth rates by sector for the year 2012−13.

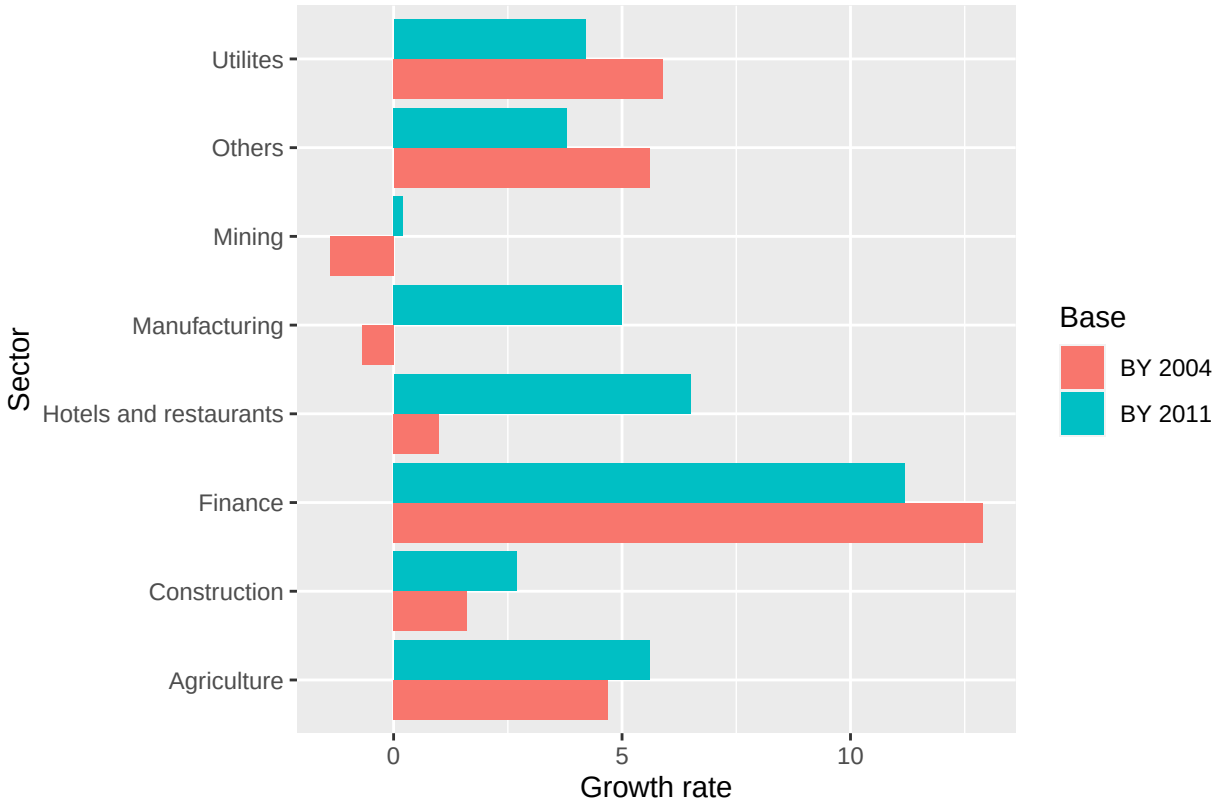

Figure 1.1b: Disaggregated GDP growth rates by sector for the year 2013−14.

Chapter 2

The Synthetic Control Method

The synthetic control method is a relatively recent innovation in causal inference methods. It was first introduced by Alberto Abadie and Javier Gardeabazal in their 2003 paper investingating estimating the impact of conflict on GDP in the Baque county(Abadie & Gardeazabal, 2003). Since then, the method has been used widely to study the impact of policy interventions involving large economic units like cities, states or countries(Abadie, Diamond, & Hainmueller, 2010, 2015). Nobel laureate Guido Imebns and Susan Athey call it "arguably the most important innovation in the policy evaluation literature in the last 15 years" in a recent review of applied econometrics(Athey & Imbens, 2017).

In this chapter, I present an overview of the synthetic control method (SCM), beginning with a non-technical introduction. In Section 2.2, I examine the theory behind the method, focussing on the setup estimation and inference. In Section 2.3, I present the relative advantages and limitations of the standard SCM, and briefly delve into the improvements provided by two new methods: the augmented synthetic control and generalized synthetic control respectively. Section 2.4 summarizes and concludes.

2.1 Introduction

Synthetic control methods emerged in the context of comparative case study. Case studies often study the effect of a policy or intervention on a particular outcome, by comparing it to other determinants of the outcome. For example, a medicine case study focusses on a ailment, while a political case study examines the impact of an electoral strategy. Comparative case studies compare one

or more units exposed to the event or intervention of interest to one or more unexposed units. For example, Abadie et al. (2010) examine the impact of Proposition 99, a tobacco control program implemented in California and compare its impact on cigarette sales relative to other states. It is essential that only some units are exposed to the intervention, while other comparable units are not.

Comparative case studies have been used in economics for a long time. However, studies typically relied on the comparison between one unit where a policy was implemented (also called 'treatment' unit) to a similar unit, where the policy did not take place (also called 'control' unit). A classic example is the study by Card (1990). Card examines the impact of the Mariel boatlift, which bought Cuban workers to Miami on the wages of low-skilled workers. He considers various single 'control' cities like Houston or Philadelphia, where no such event took place to estimate the difference between wages. The key assumption here is that wages in Miami and the control city would be the same, ahd the boatlift not taken place. This method is the classic difference-in-difference approach, which has been since used widely in economics(Card & Krueger, 1994; Dube, Lester, & Reich, 2010), with interesting updates to the originial methodology(**goodman-bacon_difference_differences_2021?**; **callaway_difference_differences_2021?**).

Synthetic control builds upon difference-in-differences, with the key difference being that instead of considering a single control, a weighted average of a group of controls in considered. The intuition behind this is that when the control group is small in number, their weighted average provides a better unit of comparison than any single entity(Abadie, 2021). Specifically, the estimator chooses the weights such that this combination of control units reflects the treated unit closely. This is the approach followed in Abadie et al. (2010). The authors use a weighted average of cigarette sales, composed of 38 states, to study the impact of Proposition 99 in California.

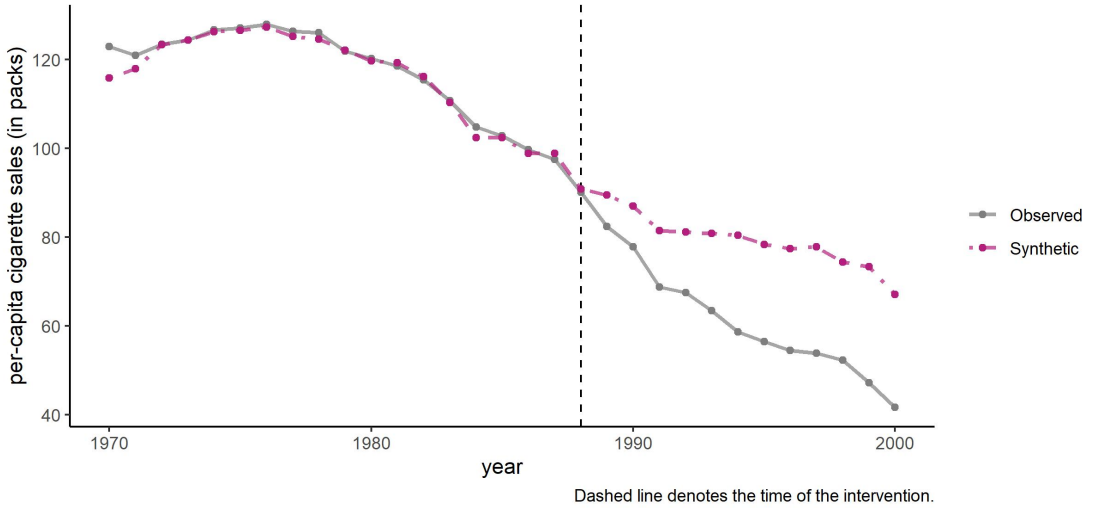

Figure 2.1: Synthetic Control of Californoa

Figure 2.1 shows the per capita cigarette sales in California and synthetic California from 1970-2000. According to the SCM model, synthetic California is an estimation of what cigarette sales in California would have been, had Proposition 99 not existed. The gap between the two trends post the treatment year (1988), shows that cigarette sales would have been significantly higher in the counterfactual, indicating that the law was successful in reducing tobacco consumption in California. Thus, the SCM offers a very clean and intuitive way to estimate the causal effect of a policy intervention with a small number of large units, like cities, counties or countries. In the next section, I introduce the formal model setup, and elaborate on how estimation and inference takes place.

2.2 Formal Aspects

To maintain consistency of notation, I stick to the model setup as seen in Abadie (2021).

2.2.1 Setup

We begin by observing data for $J+1$ units, where $j = 1, 2, ..., J+1$. Assume that $j = 1$, or the first unit is the unit where the intervention occured, or the treated unit. The rest of the units, from $j = 2, 3...J+1$, form the set of units unaffected by the treatment and are referred to as the donor pool. Let the data be observed for T periods, where the first $T0$ periods are those before the intervention. The outcome

of interest observed for unit j at time t is referred to as Y_{jt}. We also observe k predictors of the outcome j, $X_{1j}.....X_{kj}$. The vectors $\mathbf{X_1}, \ldots, \mathbf{X_{j+1}}$ contain the values of the predictors for units $j = 1, \ldots, J+1$, respectively. The $k \times J$ matrix, $\mathbf{X0} = [\mathbf{X2}, ..., \mathbf{XJ+1}]$, collects the values of the predictors for the J untreated units.

For unit j in time t, the potential response *without* the intervention is defined as Y_{jt}^N, while for the unit affected by the intervention $j = 1$ at time $t > T0$, the potential response is defined as Y_{1t}^I. Thus the effect of the intervention at $t > T0$ is given by

$$\alpha_{1t} = Y_{1t}^I - Y_{jt}^N$$

Since the unit of interest is observed after $T0$, we have $Y_{1t} = Y_{1t}^I$. The unknown response is the counterfactual, which is the evolution of the outcome of interest in absence of the intervention for $t > T0$. Now, let D_{it} take the value 1 if unit i is exposed to the intervention and 0 otherwise. Then the observed outcome Y_{it} is given by

$$Y_{it} = Y_{it}^N + \alpha_{it} D_{it} \tag{2.1}$$

As only the first unit is exposed to the treatment and only after $T0$, we have

$$D_{it} = \begin{cases} 1 & \text{if} i = 1, t > T0 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

The effect of intervention on unit 1 at time $t > T0$, given by $alpha_{1t}$, is

$$\alpha_{1t} = Y_{1t}^I - Y_{1t}^N = Y_{1t} - Y_{1t}^N \tag{2.3}$$

Since Y_{1t} is observed after the intervention, we simply need to estimate Y_{1t}^N to estimate α_{1t}.

2.2.2 Estimation

Assume that Y_{1t}^N is given by a factor model. [ADD A SHORT PARA ABOUT FACTOR MODELS]. The model used to estimate SC is given by

$$Y_{1t}^N = \delta_t + \theta_{\mathbf{t}} \mathbf{Z_i} + \lambda_{\mathbf{t}} \mu_{\mathbf{i}} + \epsilon_{it} \tag{2.4}$$

where δ_t is the time trend, $\mathbf{Z_i}$ is a vector of observed covariates not affected by the intervention and $\lambda_\mathbf{t}$ is a vector of unobserved common factors. $\theta_\mathbf{t}$ and $\mu_\mathbf{i}$ are vectors of unknown parameters and unknown factor loadings respectively. ϵ_{it} is the zero-mean error term. [WRITE ABOUT ADVANTAGES OF FACTOR MODEL]. Now, consider a $(j \times 1)$ vector of weights $\mathbf{W} = (w_2, .., w_{j+1})$, where each weight is positive and the sum of weights equals one. Thus, the weighted value of the outcome variable is

$$\sum_{j=2}^{J+1} w_j Y_{jt}^N = \delta_t + \theta_\mathbf{t} \sum_{j=2}^{J+1} w_j \mathbf{Z_j} + \lambda_\mathbf{t} \sum_{j=2}^{J+1} w_j \mu_\mathbf{j} + \sum_{j=2}^{J+1} w_j \epsilon_{jt} \tag{2.5}$$

Suppose there exist a set of optimal weights $(w_2^*, ..., w_j^*)$, such that,

$$\sum_{j=2}^{J+1} w_j^* Y_{j1} = Y_{11}$$

$$\sum_{j=2}^{J+1} w_j^* Y_{j2} = Y_{12}$$

$$\sum_{j=2}^{J+1} w_j^* Y_{jT0} = Y_{1T0}$$

$$\sum_{j=2}^{J+1} w_j^* \mathbf{Z_j} = \mathbf{Z_1}$$

These set of weights $(w_2^*, ..., w_j^*)$ are such that the weighed sum of each observed outcome variable for the donor group, in the pre-intervention time period $T0$ $(Y_{j1}, ..., Y_{jT0})$ equals the respective outcome variable for the treated unit before the intervention $(Y_{11}, ..., Y_{1T0})$. Similarly, the weighted sum of donor group covariates equals the cvariates for the treated unit. In the Proposition 99 example, all the variables on the right hand side would be the observed per capita cigarette sales in California before 1988, while the terms on the left hand side denote the weighted sum of per capita cigarette sales in the donor states. From Equation (2.4), we have

$$\hat{Y_{1t}^N} = \sum_{j=2}^{J+1} w_j Y_{jt}^N \tag{2.6}$$

and

$$\hat{\alpha_{1t}} = Y_{1t} - \hat{Y_{1t}^N} \tag{2.7}$$

Equation (2.6) states that the estimated synthetic control for the unobserved

outcome is given by the weighted average of the observed outcome variables in the donor group. This is true as each observation of the estimated synthetic control (\hat{Yˆ{N}_{1t}) is the weighted average of each observation in the donor group (Equations 6-10). Once the synthetic control is estimated, the treatment effect for the affected unit is simply the difference between the observed outcome and the estimated unobserved outcome in $t > T0$, as shown in Equation (2.7).

How are the optimal weights determined? Abadie et al. (2010) argue that the weights should be chosen so that the synthetic control best resembles the pre-treatment values of the predictors of the outcome variable. As stted before, **X1** contains the pre-treatment values of the covariates which predict the outcome variable for the treated unit, while **X0** refers to the same for the donor units. Then, the vector of weights $\mathbf{W}^*$, minimizes

$$||\mathbf{X1} - \mathbf{X0W}|| = \sqrt{(\mathbf{X1} - \mathbf{X0W})'\mathbf{V}(\mathbf{X1} - \mathbf{X0W})}$$

The matrix **V** measures the discrepancy between the**X1** and **X0**. The value of **V** is an important one, and influences the quality of the pre-intervention fit. Abadie et al. (2010) chooses **V**, such that the synthetic control minimizes the mean squared prediction error (MSPE)

$$\sum_{t \in T0} (Y_{1t} - w_2(\mathbf{V})Y_2t - ... - w_{J+1}(\mathbf{V})Y_{J+1t})^2$$

The MSPE is a measure of the distance between the outcome variable of the treated unit (Y_{1t}) and the predicted outcome variable generated by the synthetic control. MSPE is crucial to inferring the outcome of the synthetic control, which is explained in the next section.

2.2.3 Inference

At a preliminary level, inference can be done using placebo tests. Placebo tests apply the synthetic control tests to each of the donor units i.e. the units where the intervention does not occur. If the gaps between the observed and synthetic values is of a magnitude similar to those seen in the unit of interest, then there is no significant evidence that the intervention had the desired effect. However, if the magnitudes are smaller, then it can be concluded that the intervention did have a significant effect on the unit of interest.

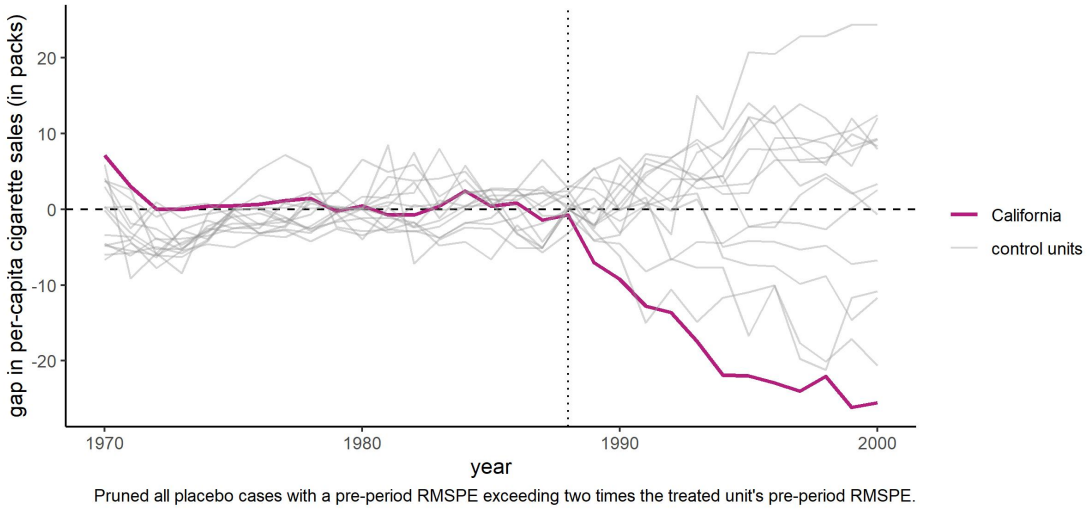

Figure 2.2: Per-capita cigarette sales gaps in California and placebo gaps in 19 control states

Figure 2.2 shows the gap in the per capita cigarette sales in the placebo states as well as California. The cigarette sale gap is highest for the state of California, showing that the effect of Proposition 99 was significant in size. The sample is restricted to 19 states, as units with a pre-intervention MSPE greater than 2 times that of California are discarded. This is done so as to exclude states which are too similar to California.

While being a useful starting point, placebo tests are limited as the pre-treatment fit of all the placebo units may not closely match the trajectory of the outcome variable well. For this reason, Abadie et al. (2010) specify an exact test statistic, which measures the ratio of the post-intervention fit relatve to the pre-intervention fit. This is given by

$$r_j = \frac{R_j(T_0+1, T)}{R_j(1, T_0)} \tag{2.8}$$

where

$$R_j(t_1, t_2) = \left(\frac{1}{t_2 - t_1 + 1} \sum_{t=t1}^{t2} (Y_{jt} - Y_{jt}^N)^2\right)^{\frac{1}{2}} \tag{2.9}$$

Equation (2.9) defines the mean squared prediction error (MSPE) of the synthetic control estimator over periods t_1 to t_2. Hence, the numerator of Equation (2.8) is the MSPE of the post-intervention period, and the denominator is the MSPE of

the pre-intervention period. The p-value for this inferential procedure is given by:

$$p = (\frac{1}{J+1} \sum_{j=1}^{J+1} I_+(r_j - r_1) \tag{2.10}$$

where $I_+(.)$ is an indicator function taking value one for non-negative inputs and zero otherwise.

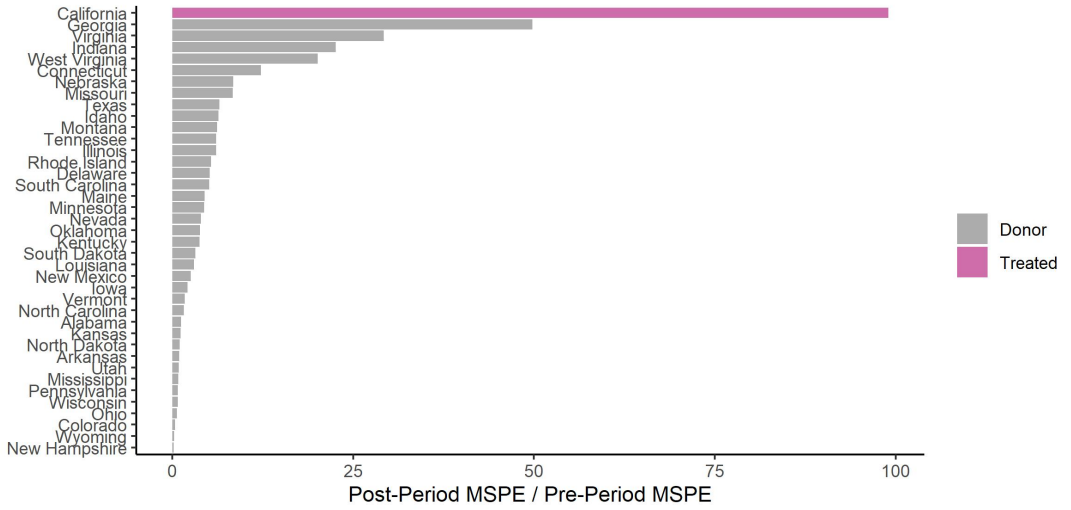

Figure 2.3: Ratio of post and pre-treatment MSPE in California and donor states

To better understand how to interpret this test statistic, consider again the Proposition 99 example. Figure 2.3 shows the value of the test statistic, i.e. the ratio of the post-treatment to the pre-treatment MSPE for the treated state (California) and each of the states belonging to the donor pool. The MSPE is the squared value of the absolute distance between the fitted cigarette sales and actual cigarette sales. If Proposition 99 had a significant effect on cigarette sales in California, then the pre-treatment trend of California and synthetic California overlap, and diverge in the post-treatment period. Hence, the ratio of the two MSPEs should be the largest for California, which is exactly what is seen in Figure 2.3. Furthermore, using equation (2.10) the corresponding p-values of this statistic can be computed, as shown below. The p-value for California is about 2.5%, making the test statistic significant at 5%.

Unit name	Type	p-values
California	Treated	0.0256410
Georgia	Donor	0.0512821
Virginia	Donor	0.0769231
Indiana	Donor	0.1025641
West Virginia	Donor	0.1282051
Connecticut	Donor	0.1538462
Nebraska	Donor	0.1794872
Missouri	Donor	0.2051282
Texas	Donor	0.2307692
Idaho	Donor	0.2564103

Table 2.1: p-values for the Post-Pre Intervention MSPE Ratio for select states

2.3 Advantages and Limitations of SCM

The standard synthetic control method, estimated as per Abadie et al. (2010) has several advantages. The first distinct advantage of the standard SCM is that it gives a transparent fit. Specifically, the discrepancy between the treated unit and the combination of the donor units can be seen clearly. This can be done by computing the difference for the pre-treatment averages of the variables of interest for the treated unit and the corresponding synthetic control. In continuation, the weights assigned to each of the donor units are also easily accessible, which allows the researcher to understand which units are more similar to the treated unit, and which are not. Secondly, synthetic control methods preclude extrapolation due to the constraint that the weights sum to one. A regression model, in contrast, allows for negative weights. Finally, SCM provides a safeguard against 'specification searches', which means cherry-picking models which yield a known result. This is avoided in SCM, as all the information used to compute the estimator is taken from the pre-intervention period, while the counterfactual for the post-intervention period remains unknown. Finally, SCM results are geometrically intuitive and can be easily visualized, and interpreted through placebo and exact test statistics.

However, the standard SCM has some limitations, two of which I highlight here. The first, elaborated by Ferman, Pinto, & Possebom (2020), relates to specification searching opportunites. The choice of contributors to the synthetic control is often a matter of discretion, and hence it is possible to try and test a number of combinations which yield a desired pre-treatment fit and post-treatment trend. There is

significant room for such searches when the number of pre-treatment periods is close to those used in common applications. While the distance-function is minimized algorithmically, which limits the bias, the function itself is endogenously chosen by the researcher, creating potential for bias and "p-hacking"(Cunningham, 2021).

Second,

2.4 Alternative Estimators

Since its introduction in Abadie & Gardeazabal (2003), the synthetic control method has become a popular tool of causal inference. However, it has also been criticised for reasons mentioned in the previous section; the critiques have also spawned a number of alternatives. Some of these are improvements to the standard SCM, while others use different estimation techniques altogether. Ferman & Pinto (2021) deals with the issue of imperfect pre - treatment fit. When the matching of pre - treatment predictors of the treated and the donor units is imperfect, the SCM estimator is biased, even for large pre - treatment period. However, using a de - meaned version of the SCM, reduces the bias of the estimator relative to difference - in - differences. Ben-Michael, Feller, & Rothstein (2021) also deals with the issue of bias, but instead propose a bias correction by modifying the regression specification. The predictor matching discrepancy is corrected for, in this case, by using a "penalty term", which is a synthetic control estimator applied to residuals. They refer to this as the Augmented Synthetic Control Estimator.

2.4.1 Generalized Synthetic Control

The generalized synthetic control was proposed by Yiqing Xu in 2017. This method combines the synthetic control method and the interactive fixed effects model (IFE). The IFE model, proposed by [CITE BAI], is used to model unobserved time-varying confounders. Time-varying confounders refer to unobserved variables that are correlated with both the dependent and independent variable, and which take different values over time. The IFE model interacts unit-specific intercepts, referred to as factor loadings and time-varying coefficients, referred to as latent factors. The generalized synthetic control method (GSCM), unifies the IFE model with the SCM in the following way: first, using only the control group (or donor pool) data, it estimates an IFE model to obtain a fixed number of latent factors. Then, the factor loadings for each treated unit are estimated by linearly projecting pre-treated

treated outcomes on the space spanned by this factor. Finally, it computes treated counterfactuals based on the latent factors and the factor loadings.

The key difference between the SCM and GSCM is that the latter does dimension reduction prior to re-weighting. Specifically, the model selects the number of latent factors to be used algorithmically. This is done using a cross-validation scheme (details in the inference section), which avoids specification searches. The GSCM is also more "general" in the sense that it can be extended to cases with multiple treated units and variable treatment periods.

Framework

In this section and the next one, I stick to the notation used in [CITE XU]. Let Y_{it} be the outcome of interest for unit i and time t, and let O and C denote the total number of units in the treatment and control groups, with N total units. All units are observed for T periods, with the pre-treatment periods denoted by $T_{0,i}$ for each unit. The post treatment period is given by $q_i = T - T_{0,i}$. It is assumed that Y_{it} is given by a linear factor model

$$Y_{it} = \delta_{it} D_{it} + x'_{it}\beta + \lambda'_i f_t + \epsilon_{it} \tag{2.11}$$

where D_{it} equals 1 if i is exposed to treatment, and zero otherwise. δ_{it} is the treatment effect, and x_{it} is a $(k \times 1)$ vector of observed covariates, while β is $(k \times 1)$ vector of unobserved parameters. f_t is a $(r \times 1)$ vector of unobserved common factors and λ_i is a vector of factor loadings. ϵ_{it} is the zero-mean idiosyncratic error term. The parameter of interest is δ_{it}; more specifically we are interested in estimating the average treatment effect on the treated unit, given by the sum of δ_{it} for each unit, divided by the treatment period.

Let $Y_{it}(1)$ and $Y_{it}(0)$ be the potential outcomes for unit i at time t, when variable D_{it} takes values 1 and 0 respectively. Therefore, we have

$$Y_{it}(1) = \delta_{it} + x'_{it}\beta + \lambda'_i f_t + \epsilon_{it}$$

$$Y_{it}(0) = x'_{it}\beta + \lambda'_i f_t + \epsilon_{it}$$

Hence, we have that $\delta_{it} = Y_{it}(1) - Y_{it}(0)$

Let the control and treated units be subscripted from 1 to N_{CO} and N_{CO+1} to

N respectively. Then, stacking the control units together yields

$$Y_{CO} = X_{CO}\beta + F\Lambda'_{CO} + \epsilon_{CO} \tag{2.12}$$

where Y_{CO} and ϵ_{CO} are matrices sized $(T \times N_{CO})$, X_{CO} is of dimension $(T \times N_{CO} \times p)$, while Λ_{CO} is a $(N_{CO} \times R)$ matrix.

Estimation and Inference

We are interested in estimating the treatment effect on the treated unit i, at time t, given by $\hat{\delta_{it}} = Y_{it}(1) - \hat{Y_{it}}(0)$. $\hat{Y_{it}}(0)$ is imputed using the following three steps:

1. Estimate an IFE model using only the control group data to obtain

$$(\hat{\beta}, \hat{F}, \hat{\lambda}) = \arg\min_{\tilde{\beta}, \tilde{F}, \tilde{\Lambda}} \sum_{i \in C} (Y_i - X_i\tilde{\beta} - \tilde{F}\tilde{\lambda}_i)'(Y_i - X_i\tilde{\beta} - \tilde{F}\tilde{\lambda}_i)$$

2. Estimate factor loadings for the treated unit by minimizng MSPE of the predicted treated outcome in the pre-treatment period:

$$\lambda'_i = \arg\min_{\tilde{\Lambda}_i} (Y_i^0 - X_i^0\hat{\beta} - \hat{F}^0\tilde{\Lambda}_i)'(Y_i^0 - X_i^0\hat{\beta} - \hat{F}^0\tilde{\Lambda}_i), i \in T$$

3. Using the estimated parameters for covariates, common factors and factor loadings, estimate the counterfactual outcome as:

$$\hat{Y_{it}}(0) = x'_{it}\hat{\beta} + \hat{\lambda}'_i\hat{f}_t, i \in T, t > T_0$$

How is the factor model used to obtain the GSC estimator chosen? Xu (2017) proposes a cross - validation scheme to choose the model before estimating the causal effect. Cross - validation is a resampling method i.e. it involves repeatedly drawing samples from a training set, which is a sub - sample and fitting a model of interest on each sample. In cross validation, the data is randomly divided into a training set, and a validation set. The model of interest is then fit on the training set as used to predict observations in the validation set. The mean squared error then allows us to check which model is the most precise. For the GSC estimator, a special case of cross - validation is used, namely the leave one out cross validation (LOOCV). In this case, the validation set consists of a single observation.

Broadly, the algorithm performing LOOCV performs the following steps. First, it estimates an IFE model with a given number of factors r, using only control group data. It then assigns one pre -treatment period of the treatment group to the validation set, and uses the rest of the pre - treatment data to estimate the held

back data. The corresponding MSPE is computed, and then this is repeated for different values of r. The number of factors that minimizes MSPE (r^*) is chosen. For technical details of this algorithm, see Appendix X.

How is inference carried out using the GSC model?

References

Abadie, A. (2021). Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Journal of Economic Literature*, *59*(2), 391–425. http://doi.org/10.1257/jel.20191450

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. http://doi.org/10.1198/jasa.2009.ap08746

Abadie, A., Diamond, A., & Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*, *59*(2), 495–510. http://doi.org/10.1111/ajps.12116

Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the basque country. *American Economic Review*, *93*(1), 113–132. http://doi.org/10.1257/000282803321455188

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32. http://doi.org/10.1257/jep.31.2.3

Ben-Michael, E., Feller, A., & Rothstein, J. (2021). The augmented synthetic control method. *Journal of the American Statistical Association*, *116*(536), 1789–1803. http://doi.org/10.1080/01621459.2021.1929245

Card, D. (1990). The impact of the mariel boatlift on the miami labor market. *ILR Review*, *43*(2), 245–257. http://doi.org/10.1177/001979399004300205

Card, D., & Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania. *American Economic Review*, *84*(4), 772–793. Retrieved from https://ideas.repec.org/a/aea/aecrev/v84y1994i4p772-93.html

CSO. (2015). No room for doubts on new GDP numbers. *Economic and Political Weekly*, *50*(16), 7–8. Retrieved from https://www.epw.in/journal/2015/16/discussion/no-room-doubts-new-gdp-numbers.html

Cunningham, S. (2021). *Causal inference; the mixtape* (1st ed.). Yale University Press. Retrieved from https://yalebooks.yale.edu/9780300251685/causal-inference

Dube, A., Lester, T. W., & Reich, M. (2010). Minimum wage effects acrss state borders: Estimates using contiguous counties. *The Review of Economics and Statistics*, *92*(4), 945–964. Retrieved from https://www.jstor.org/stable/40985804

Ferman, B., & Pinto, C. (2021). Synthetic controls with imperfect pretreatment fit. *Quantitative Economics*, *12*(4), 1197–1221. http://doi.org/10.3982/QE1596

Ferman, B., Pinto, C., & Possebom, V. (2020). Cherry picking with synthetic controls. *Journal of Policy Analysis and Management*, *39*(2), 510–532. http://doi.org/10.1002/pam.22206

Nagaraj, R. (2015a). Seeds of doubt on new GDP numbers: Private corporate sector overestimated?, *50*, 14–17.

Nagaraj, R. (2015b). Seeds of doubt remain: A reply to CSO's rejoinder. *Economic and Political Weekly*, *50*(18), 64–66. Retrieved from https://www.jstor.org/stable/24481913

Nagaraj, R. (2021). Revisiting the GDP estimation debate. *Economic and Political Weekly*, *56*(45), 10–13. Retrieved from https://www.epw.in/journal/2021/44/commentary/revisiting-gdp-estimation-debate.html

Nagaraj, R., & Srinivasan, T. (n.d.). Measuring india's GDP growth: Unpacking the analytics & data issues behind a controversy that refuses to go away.

Subramanian, A. (2019, June). *India's GDP mis-estimation: Likelihood, magnitudes, mechanisms, and implications*. {CID} Working Paper 354, CID Working Paper 354.