EGE AKADEMİK BAKIŞ / *EGE ACADEMIC REVIEW*

Cilt 19 • Sayı 4 • Ekim 2019
SS. 389 - 399
Doi: 10.21121/eab.638571
Başvuru Tarihi: 25/5/2019 • Kabul Tarihi: 26/8/2019

Using Panel Data for Macroeconomic Policy Evaluation: A Survey[1]

Ron P. Smith

Abstract

In order to measure the macroeconomic effect of some policy or event, a "treatment", we need to construct a "counterfactual", a prediction of what would have happened in the absence of treatment, which is unobserved. Panel data for countries and regions, where the number of units and time periods are large, potentially provide untreated control groups which can be used to construct the counterfactual. A number of different procedures have been suggested for such policy evaluations, including the synthetic control method, SCM, and the panel data approach, PDA. We survey these and other methods.

Keywords: panel data, counterfactuals, policy evaluation, synthetic controls, macroeconomics.

JEL classification: C18, C54, E65,

Economists often want to measure the effect of some "treatment", a policy or event, on an outcome in some aggregate unit, such as a country or region. For example, the treatment might be joining the euro; the outcome is GDP for a particular country that adopted the euro. In another example, the treatment might be a US state introducing a "Stand Your Ground"[2], SYG, law; the outcome is the murder rate in the state. The objective is to measure the effect on GDP or the murder rate of these policy interventions. There are now many panels available for a large number of countries or regions over long time periods. A number of macro policy evaluation procedures have been developed to take advantage of the fact that such panels potentially provide untreated control groups which can be used to construct counterfactuals and allow the estimation of the effect of policy interventions, like joining the euro or adopting SYG laws, on treated groups.

This is different from the typical macro policy evaluation exercise, which considers, for instance, a monetary policy shock, calculated as a one standard error displacement of the structural disturbance of a policy equation, such as a Taylor rule. The impulse response function (IRF) is the time profile of the deterministic component of the effect of such a displacement. The IRF yields ex-ante information about the way the model responds to such a displacement, not an ex-post evaluation of the effectiveness of a real policy intervention. In addition IRFs ignore the cumulative uncertainty associated with the stochastic component, the post-intervention disturbances. Pesaran & Smith (2018) discuss these issues in more detail. This paper examines procedures that compare the actual realisations of the outcome after the intervention with a counterfactual that predicts what would have happened in the absence of the intervention.

In microeconometrics there is a long tradition of measuring such treatment effects on individuals. Abadie, & Cattaneo (2018) provide a recent review of this program evaluation literature. The macroeconomic interest in measuring the effect on aggregates is more recent. A number of the approaches to policy evaluation have their roots in the microeconometric literature. However, the micro and macro issues are rather different. This paper reviews the various approaches and compares their relative advantages and disadvantages.

Section 1 sets out the framework used in macro policy evaluation and contrasts it with the microeconometric treatment effect literature. Section 2 sets out and compares the two main approaches currently

Birbeck University of London, Department of Economics, Mathematics and Statistics, e-mail: r.smith@bbk.ac.uk

used in macro-policy evaluation: the synthetic control method, SCM, and, the panel data approach, PDA. The relative advantages of these procedures has generated some controversy. Section 3 examines some extensions to those approaches. Section 4 considers a different approach, which like SCM has its origins in the microeconometric literature and uses propensity scores. Section 5 has some concluding comments. It appears that none of the approaches is universally best, so it seems important to use more than one and compare, or average, their estimates.

1. The Framework

All the procedures considered will work with multiple treated units, but for clarity consider the effect on a single unit. Suppose the outcome in unit *1*, in period *t*, is y_{1t}. The effect of the policy is

$$\delta_{1t} = y^1_{1t} - y^0_{1t}$$

the difference between the two potential outcomes: the outcome with treatment, y^1_{1t}, (being in the euro) and the outcome without treatment, y^0_{1t}, (not being in the euro). However, it is impossible to observe both y^1_{1t}, and y^0_{1t}, in period *t*. Country 1 is either in the euro or out of the euro, it cannot be both in and out.

Since only one of the potential outcomes can be observed, one needs a "counterfactual", a prediction of what would have happened in the unobserved case. Suppose unit 1 was actually treated in period T_0, adopted the euro, and post-treatment values, $y^1_{1,T0+h}$, are observed[3]. The estimate of the effect of treatment, in the post-treatment period T_0+h, $h=1,2,\ldots,H$ is

$$\hat{\delta}_{1,T_0+h} = y^1_{1,T_0+h} - \hat{y}^0_{1,T_0+h} \qquad (1)$$

where $\hat{y}^0_{1,T_0}+h$ is the estimate of the counterfactual: a prediction of what would have happened to country *1* in period T_0+h had it not joined the euro. Thus we have(i) an estimation problem, finding a model to construct the counterfactual outcome in the absence of the policy intervention and(ii) an inference problem, determining whether the difference between the realized and counterfactual is larger than would have been expected by chance. To the extent that suitable control units are available to provide the counterfactual, there is no need to construct a structural model of how the outcome is determined or worry about identification. In which case policy evaluation can be data-driven and relatively atheoretical.

Panels for countries and regions with data on y_{it}, $i=1,2,\ldots,N$ and $t=1,2,\ldots,T$, where *N* and *T* are large potentially provide untreated control groups which can be used to construct the counterfactual, $\hat{y}^0_{1,T_0+h}$, allowing the estimation of $\hat{\delta}^0_{1,T_0+h}$, the effect of a policy intervention on a treated group, and an evaluation of the policy.

A number of different procedures have been suggested, which are reviewed below, and there has been considerable controversy about their relative effectiveness. The two most prominent methods are the synthetic control method, SCM, and the panel data approach, PDA.

Abadie & Gardeazabal (2003) introduced the SCM to measure the costs of Basque terrorism, and it was subsequently applied by Abadie, Diamond, & Hainmueller (2010) to the California smoking program and Abadie, Diamond, & Hainmueller (2015) to German reunification. Since the package Synth became available on Matlab, R and Stata, SCM has been widely used. The PDA was introduced by Hsiao, Ching & Shui (2012) to measure the effects of Hong Kong's political and economic reunification with China. These studies evaluate the effects of events on single units, but both approaches can be used to evaluate effects on multiple units. Bove, Elias & Smith (2016) use both SCM and PDA to measure the effect on GDP of Civil Wars. Gobillon & Magnac (2016) compare the use of SCM and PDA in regional economics.

While we can learn from the microeconometric literature, the micro and macro issues are rather different. The classic micro method is difference in differences. If $\bar{y}_{C0}$ is the average over the units in the control group, C, in period *0* before treatment and $\bar{y}_{C1}$, in period 1, and similarly the treated group, A, averages are $\bar{y}_{A0}$ and $\bar{y}_{A1}$, then the difference in difference estimator is

$$\delta = \left(\bar{y}_{A1} - \bar{y}_{A0}\right) - \left(\bar{y}_{C1} - \bar{y}_{C0}\right) \qquad (2)$$

The first term measures the change in the averages for the treated group, the second term controls for any general trends, assuming that the trends in the control group are parallel to those for the treated group. Defining a dummy for group A, D_A, and a dummy for period 1, D_1, using the original observations it can be written as a two way fixed effect model plus a treatment effect:

$$y_{it} = \alpha + \alpha_A D_A + \alpha_1 D_1 + \delta D_A D_1 + \varepsilon_{it} \qquad (3)$$

where the four parameters in (3) are functions of the four means in (2). In more general cases, where one has, for instance, more time periods, covariates or endogenous treatment, (3) is a useful representation.

There are a number of points to note about this macro approach. Firstly, in micro cases *N* is large, *T* is small, often only *T=2*, as in (2) above, and there are major problems associated with endogenous selection into treatment. The endogeneity and sample selection biases that arise in the micro case from heterogeneity correlated with treatment, across the units, are not problems in the macro case. There the focus is on a single unit, and the "policy on/policy off" comparisons are done over time rather than across units. In micro terminology, the parameter of interest in the macro case is the effect of "treatment on the treated", not the average treatment effect over individuals. Because it is primarily a time series problem, the rules for assignment to treatment are not an issue. In terms of the examples given above, it makes no sense to consider either the effect of Hong Kong being integrated with West Germany or of East Germany being integrated with China. In addition, macro panels tend to exhibit cross-section dependence that results from strong factors driving all units. This means that other units can be used to construct controls that can be used to specify the counterfactual in the analysis of a treatment effect or the evaluation of a policy intervention. Unlike difference in differences, the parallel trends assumption is not required.

Secondly, a multi-horizon effect is estimated for each period for this unit.

There is no assumption that the effect is constant over time. One can average $\hat{\delta}_{1t}$ over time to get an average treatment effect, ATE:

$$\bar{\hat{\delta}}_{1H} = \sum_{h=1}^{H} \hat{\delta}_{1,T_0+h} / H \qquad (4)$$

but this ATE is quite different from the micro case which compares the average over treated units with the average over untreated units as in (2). The average in (4) is over time for a single unit not over units. Tests on the individual $\hat{\delta}_{1,T_0+h}$ are likely to be sensitive to the distributional assumption of the model, whereas tests on the average, $\bar{\hat{\delta}}_{1H}$, can rely on the central limit theorem if *H* is large to obtain a distribution and may be more robust.

Thirdly, the Lucas critique, which refers *to ex ante* policy evaluation is not a problem. Ex ante policy evaluations compare two predictions, one with the policy and one without and face the problem that the intervention may change the parameters. In estimating (1) the concern is with *ex post* evaluation of a policy intervention, where time series data are available before as well as after the policy change. The comparison is based on the difference between the realisations of the outcome variable of interest and counterfactuals obtained assuming no policy change. The counterfactuals, based on estimates using pre-intervention data, will embody pre-intervention parameters while the realized post-intervention outcomes will embody the effect of the change in the policy parameters and any consequent change in expectations.

Finally, whatever we use to predict $\hat{y}_{1t}^{0}$ must not be influenced by the policy intervention or treatment itself. For major changes, with spillover effects, this can be a very strong requirement: a change in one country, like the re-unification of West Germany, can affect many other countries. Similarly, there must be no large change in the control units that would not have affected the treated units.

The fact that the counterfactual is a prediction, means that we can learn from the forecasting literature about good ways to construct it. For instance, it is well known that averaging over forecasts improves performance and Hsiao & Zhou (2018) suggest averaging over counterfactuals produced by different procedures. Pesaran & Smith (2016) point out that simple parsimonious models tend to forecast better than ones with more parameters, so there are arguments for using simple models to generate the counterfactuals.

Counterfactuals differ from a conventional forecasts in that they are not about the future, they are conditional on observed data and there is no actual against which to compare them. This latter feature presents a major difficulty for the evaluation of the different methods. Because we never observe the truth, we cannot say which method gets us closest to the truth. To get around this problem, evaluation is based on simulations, where by construction we do know the truth. But the results are then dependent on the choice of data generating process, DGP, in the simulation. Gardeazabal & Vega-Bayo (2016) and Wan, Xie & Hsiao (2018) differ on the appropriate way to define the DGP in simulations used to compare the SCM and PDA.

The various approaches can be applied to multiple treated units very easily; for instance where a number of.countries join the euro or a number of US states adopt a particular law. But for ease of exposition consider a single treated unit. Suppose unit *1* is subject to treatment, some intervention at T_0. There is pre-intervention data $t=1,2,...,T_0$; post intervention data $t=T_{0+1},T_{0+2},...,T_0+T_1$; with $T=T_0+T_1$. We assume that there are enough pre-intervention observations to estimate models.

Defining $d_{it}=1$ if $i=1$ and $t>T_0$, zero otherwise we observe either the treated or the untreated, but not both:

$$y_{it} = d_{it}y_{it}^1 + (1-d_{it})y_{it}^0$$

In constructing the estimated counterfactuals, $\hat{y}_{it}^0$ we may use (i) observations on the outcome variables in other units: $y_{2t},y_{3t},..,y_{Nt}$ (ii) observations on a vector of covariates $\mathbf{x}_{it}$ in the treated or untreated units and (iii) lagged values in dynamic models. The models may be data driven (atheoretical or non-parametric), just based on correlations or similarity, or the models may be more theoretical parametric models. The synthetic control method, SCM, treats the other outcomes as providing controls and the panel data approach, PDA, treats them as providing predictors, both are data-driven.

Consider case (i) and suppose that there are *N-1* controls not subject to the intervention and not affected by the intervention in unit 1. The estimated counterfactual is a weighted average of the outcomes in these control or predictor units. The issue is how to choose controls and weights. In a static model the effect of the intervention is measured as

$$\delta_{1,T_0+h} = y_{1,T_0+h} - \sum_{i=2}^{N} w_i y_{i,T_0+h}; \quad h=1,2,...,T_1 \qquad (5)$$

It is important to examine the pre-treatment predictive power of the counterfactual, which is an important diagnostic. The method should predict well before treatment and the estimation errors $\hat{\delta}_{1,t}$ for $t<T_0$ should be small. The standard errors of the post-treatment estimate of $\hat{\delta}_{1,t+h}$ will reflect the pre-treatment fit. If the model does not fit well pre-treatment, the standard errors will be large and tests will have little power to detect the effect of a policy intervention.

It is also important to be precise about the nature of the policy intervention, what it is conditional on, and the plausibility of the counterfatual.

2. SCM and PDA

2.1 Synthetic Control

SCM uses the analogy with microeconometric treatment effect studies, where one chooses controls that are similar in characteristics to those that are treated. One would match patients treated with a drug to untreated controls with similar covariates such as age, sex, and health and compare the outcomes in the two groups. Similarity is usually measured by propensity score, the probability of being treated conditional on the covariates.

To determine the SCM weights w_i let $\mathbf{x}_{1kt}$ be a set of $k=1,2,...,K$ covariates or predictor variables for y_{1t}, with the corresponding variables in the other units given by $\mathbf{x}_{jkt}$, $j=2,3,...,N$. These variables are averaged over the pre-intervention period to get $\bar{\mathbf{x}}_{1k}^{T_0}$ and $\bar{\mathbf{X}}_k^{T_0}$ the *N-1* x *1* vector of predictor *k* in the control group. Then the *N-1* x *1* vector of weights $W=(w_2,w_3,\ldots,w_N)`$ are chosen to minimize

$$\sum_{k=1}^{K} v_k \,(\bar{\mathbf{x}}_{1k}^{T_0} - W'\,\bar{X}_k^{T_0})^2$$

subject to $\sum_{i=2}^{N} w_i = 1,\ w_i \geq 0$, where v_k is a weight that reflects the relative importance of variable *k*. Call the SCM weights $\tilde{w}_i$, many of them will be zero, for countries not included in the synthetic control.

SCM chooses the comparison units to be as similar as possible to the target along the dimensions included in $\mathbf{x}_{ikt}$.The v_k are often chosen by cross-validation, which may be problematic for potentially non-stationary time-series samples. The pre-intervention outcome variable may be included in $\mathbf{x}_{ikt}$; it is argued that matching on the pre-intervention outcomes helps control for the unobserved factors affecting the outcome of interest.

Using the SCM weights the estimate of the counterfactual is

$$\tilde{y}_{1,T_0+h}^0 = \sum_{i=2}^{N} \tilde{w}_i y_{i,T_0+h} \qquad (6)$$

In the case of German Reunification, Abadie et al. (2015), use controls and weights w_i of Austria, 0.42, US, 0.22, Japan 0.16, Switzerland 0.11 and Netherlands, 0.09. The synthetic West Germany is similar to the real West Germany in pre 1990 per capita GDP, trade openness, schooling, investment rate and industry share. As they note there may be spillover effects. Since Austria, Switzerland and Netherlands share bor-

ders with Germany there is a distinct possibility that their post 1990 values may be influenced by German reunification. Those that are geographically the most similar are most likely to show spillover effects.

Given the way the SCM estimate is constructed inference, testing whether the effect of the intervention is significant, is not straightforward. Xu (2018) generalises the SCM in various respects to give generalised SCM, GSCM.

2.2 PDA

Hsiao, Ching & Shui (2012), HCS, measure the benefits of political and economic integration of Hong Kong with mainland China using PDA . They use (5), but choose the weights by regression of y_{1t}, growth in Hong Kong, on y_{jt}, $j=2,3,...,N$, growth in the control countries during the pre-intervention period. Then using the pre-intervention estimates they predict the post-intervention counterfactual as.

$$\hat{y}^{0}_{1,T_0+h} = \hat{\beta}_1^{T_0} + \sum_{i=2}^{N} \hat{\beta}_i^{T_0} y_{i,T_0+h} \tag{7}$$

where the $\hat{\beta}_i^{T_0}$ are the regression coefficients estimated on the pre-intervention period up to time T_0. Because the counterfactual is a forecast from a standard regression, inference is easier and HCS use robust standard errors to allow for serial correlation. The coefficients for most countries will be zero, only a subset of other countries are used to predict Hong-Kong. The subset is chosen by a model selection procedure, but other procedures have been used. HCS emphasize that Hong Kong is too small for the effects of integration with China to influence any of the control countries. The control group, weights chosen by AIC for the period 1993:Q1–1997:Q2 are Japan -0.69, Korea -0.3767, USA 0.8099, Philippines -0.1624, Taiwan 0.6189. They find that the political integration had little effect on the growth rate, but that the subsequent economic integration did; an example of the issues in choosing T_0.

Li & Bell (2017) show that the PDA works for a wider range of data generating processes than those HCS originally assumed; derive the asymptotic distribution of the average treatment effect estimator and propose using lasso to select control units. Lasso will work when the number of controls is greater than the sample size, when model selection methods will not. They argue that lasso is computationally more efficient than model selection methods and leads to better out-of-sample prediction.

The underlying interpretation of (7) rests on a factor model:

$$\begin{aligned} y_{it} &= \mu_i + \sum_{j=1}^{r} \lambda_{ij} f_{tj} + e_{it} \\ y_{it} &= \mu_i + \boldsymbol{\lambda}_i' \mathbf{f}_t + e_{it} \end{aligned} \tag{8}$$

The outcome in a unit is determined by a vector of r common factors, $\mathbf{f}_t$ which have different effects on different countries, reflected in their heterogeneous factor loadings, $\boldsymbol{\lambda}_i$, and an idiosyncratic factor e_{it}. In a macroeconometric context, it is natural to think of very different countries driven by the same common trends: the 2008 crisis hit most countries, though to different degrees. Hsiao et al. (2012) include the US in the controls, not because the US is like Hong Kong, the justification in the SCM procedure, but because US growth is a good predictor of Hong Kong growth. Factors are said to be strong if they influence almost every unit, weak if they only influence a sub-set of units.

2.3 Comparisons

Both equations can be interpreted as regressions, but (6) is a constrained regression, it has no intercept and the weights are non-negative and sum to one; whereas (7) is an unconstrained regression. SCM, just matching on covariates, can be estimated with fewer pre-treatment observations than PDA, which requires T_0 to be large enough to estimate a regression. SCM proponents criticize the fact that regression methods can give negative weights to controls. But this is to be expected if one interprets the procedure as involving prediction using common factors. Suppose Hong Kong before integration is largely driven by global factor A, the US by factors A and B, and Japan largely by factor B; then the US minus Japan provides an estimate of factor A, which drives Hong Kong.

In the case of microeconometric treatment effect studies, when the units are only subject to weak factors, SCM is sensible: choose controls that are similar in characteristics to those that are treated: match patients treated with a drug to untreated controls of similar age, sex, and health. Similarity is usually measured by propensity score, $p(x)$, the probability of treatment given the covariates; though since it is difficult to predict macroeconomic policies or civil wars, macro propensity scores may be less accurate than micro ones. We return to propensity scores below. In

addition the $\mathbf{x}_{jkt}$ are often poor predictors for y_{it}, which is why pre-intervention y_{it} are often used.

It is not clear that SCM is as sensible in macroeconometric time-series contexts, where there are strong common factors driving the y_{it}, so prediction from outcomes in other units y_{jt} may be more sensible than trying to identify units with similar x_{it} .

The SCM procedure requires that the matching variables of the treated units and the control units overlap, in microeconometric terms you have to find someone of the same age, sex and health as the treated person. In statistical terms this is referred to as the need for common support. The support of a variable is the range of values that it takes. Common support means that the range of the matching variables of the treated units is within the range of the matching variables of the control units. This may not hold in micro studies, if, for instance, there is a 100 year old man in the treated group and nobody over 90 in the control group. It is even more of a problem in macro studies. The common support assumption is unlikely to hold for Hong Kong, no other country is like Hong Kong, not even Singapore, the closest comparison. This is not a problem for the prediction method. SCM relies on interpolation, but the PDA can extrapolate. Of course one can question the relative accuracy of extrapolation relative to interpolation.

There is a growing literature on comparing SCM and PDA, which are likely to work well in different circumstances, depending for instance on the size of T_0 and whether one interpolates (the support of the controls covers the treated case) or extrapolates. Gobillon and Magnac (2016) investigate the use of interactive effect, Bai (2009), or linear factor models, difference in differences, (2), and SCM in regional policy evaluation. They show that (2) are generically biased, derive support conditions for SCM and use Monte Carlo to compare the methods, Wan et al. (2018) compare the panel time series approach and the synthetic control method. Hsiao & Zhou (2019) compare parametric, semi-parametric and non-parametric methods. SCM & PDA are non-parametric methods. Since they find that no method dominates in all circumstances they suggest model averaging.

Bove et al (2016) use both SCM and PDA to measure the effect on GDP of Civil Wars. They find that the results from the two methods are similar, perhaps because they both tend to weight the same countries. What makes a large difference is whether the outcome measure modelled is log GDP or the change in log GDP, the growth rate. These give very different results for the costs of civil war to a country. The sensitivity of results to specification is always an issue.

3. Extensions

There are many natural variants or extensions to these procedures, often allowing for exogenous variables and dynamics as well as unobserved factors. For notational simplicity, $\hat{y}^0_{1,T_0+h}$ is used for any estimated counterfactual and $\hat{\delta}^0_{1,T_0+h}$ for any estimated policy effect, not distinguishing between the methods used to generate them.

Suppose y_{it} is determined by an intercept, a vector of exogenous variables, $\mathbf{x}_{it}$, and a single factor f_t:

$$y_{it} = \alpha_i + \beta_i' \mathbf{x}_{it} + \lambda_i f_i + \varepsilon_{it} \tag{9}$$

Average across units to give

$$\bar{y}_t = \bar{\alpha}_t + \bar{\beta}_t' \bar{\mathbf{x}}_t + \bar{\lambda} f_t + \bar{\varepsilon}_t + N^{-1} \sum (\beta_i - \bar{\beta})' x_{it}$$

$$f_t = \bar{\lambda}^{-1} \{ \bar{y}_t - \bar{\alpha} + \bar{\beta}_t' \bar{\mathbf{x}}_t + [\bar{\varepsilon}_t + N^{-1} \sum (\beta_i - \bar{\beta})' x_{it}] \}$$

The term in [..] will average zero under fairly weak assumptions, so the $\bar{y}_t$ and $\bar{\mathrm{x}}_t$ provide a proxy for the unobserved factor, as long as $\bar{\lambda} \neq 0$. This is the basis of the correlated common effect, CCE, estimator of Pesaran (2006), which suggests filtering out the effect of the factors by adding the means of the dependent and independent variables to the regression for each unit

$$y_{it} = \alpha_i + \mathbf{x}_{it}' \beta_i + \gamma_{0i} \bar{y}_t + \gamma_i' \bar{\mathbf{x}}_t + u_{it} \tag{10}$$

Notice that the covariance between $\bar{y}_t$ and ε_{it} goes to zero with *N,* so for large *N* there is no endogeneity problem. The CCE generalises to many factors and lagged dependent variables, but requires *N* and *T* large.

Bai (2009) calls (9) with a homogeneous β and multiple factors an interactive fixed effect model

$$y_{it} = \alpha_i + \beta' \mathbf{x}_{it} + \boldsymbol{\lambda}_i' \mathbf{f}_t + \varepsilon_{it}$$

contrasting it with the usual additive fixed effect model which has $\alpha_i + \alpha_t$. Bai suggests an iterative principal component based method to determine the

r x *1* vector of factors $\mathbf{f}_t$, and to estimate the parameters. One advantage of CCE over alternative principal component based estimates of the factors is that one does not need to determine, *r*, the number of factors, which can be difficult.

Using data up to T_0 the CCE estimates could directly predict the counterfactual as

$$\hat{y}^0_{1,T_0+h}=\hat{\alpha}_1^{T_0}+\mathbf{x}'_{1,T_0+h}\,\hat{\beta}^{T_0}_{1,T_0+h}+\hat{\gamma}^{T_0}_{01}\bar{y}_{T_0+h}+\hat{\gamma}_1^{T_0}\bar{\mathbf{x}}_{T_0+h} \quad (11)$$

For large changes like German or Hong Kong re-unification, or joining the euro, it is not clear that there would be any country specific exogenous variables unaffected by the change. Having the weights specified a priori, equal in this case, avoids the problems associated with model selection or lasso based choices. If there are strong factors, the choice of weights might not matter very much.

Hsiao & Zhou (2019) characterise the Xu (2017) GSCM as a parametric method using covariates. Xu suggests estimating the homogeneous β, interactive fixed effect model

$$y_{it}=\beta'\mathbf{x}_{it}+\boldsymbol{\lambda}'\mathbf{f}_t+\varepsilon_{it}$$

on the data for the control group using all the observations, *t=1,2,...,T*. These estimates of β and $\mathbf{f}_t$ can then be used to estimate the factor loadings $\boldsymbol{\lambda}_1$ for the treated group using the pre-treatment data, up to T_0. The intercept is treated as a factor. The Xu counterfactual is then

$$\hat{y}^0_{1,T_0+h}=\hat{\beta}'\mathbf{x}_{1,T_0+h}+\hat{\boldsymbol{\lambda}}'_1\hat{\mathbf{f}}_{T_0+h}$$

Geng & Zhou (2018) combine the CCE and PDA approaches, to suggest what they call a panel data with exogenous regressors, PDX estimator Geng & Zhou provide asymptotic distributions that allow inference: tests and confidence intervals for the ATE. The method involves getting the CCE residuals and applying the the PDA approach to them. The model adds an exogenous variable, not affected by the treatment, with a homogeneous coefficient to (8) to give

$$y_{it}=d_{it}\delta_{it}+\alpha_i+\mathbf{x}'_{it}\beta+\lambda'_i\mathbf{f}_t+\varepsilon_{it} \quad (12)$$

The first stage is to estimate the CCE regressions for each country on the pre-treatment period, obtain the CCE residuals $y_{it}-\mathbf{x}'_{it}\hat{\beta}^{CCE}$. One can then apply the PDA approach by regressing the CCE residuals from the treated group on those for the untreated groups, using lasso or model selection procedures to identify the relevant controls. These estimates can be used to construct the counterfactual for the post-treatment period, $\hat{y}^0_{1,T_0+h}$, from which $\hat{\delta}_{1t}$ can be calculated.

Compared to SCM or GSCM, Geng & Zhou argue that this approach has the advantages of: that there is no need to impose constraints on both observables and unobservables and the number of parameters to be estimated in the model is greatly reduced. They use their method to measure the effect of US Stand Your Ground (SYG) laws on state murder rates and find adopting SYG laws tends to increase the murder rate. They use a number of different methods to measure the effect and note that the PDX fits the data in the pre-treatment period well compared with the other methods and different methods give very different estimates for the effects of the SYG laws.

Chan & Kwok (2016) also extend the Pesaran (2006) procedure and extract principal components from the control group to form factor proxies.

None of these procedures are fully dynamic. Pesaran & Smith (2016, 2018) explicitly consider testing for the effect of a policy intervention in a dynamic context, either in the case of a parsimonious reduced form or final form equation, P&S (2016), or in the context of a complete DSGE P&S(2018).

They suggest estimating for *t=1,2,...,*T_0 the pre-treatment period, the equation

$$y_{1t}=\alpha_{10}+\alpha_{11}\,y_{1,t-1}+\gamma_1\,y^*_{1t}+u_{it} \quad (13)$$

Where y^*_{1t} is some country specific function of the outcome variables in the untreated units. This function could be freely estimated, as in PDA, the mean as in CCE, or some other weighted average, such as the trade weighted averages of the other countries as used in the Global VAR, GVAR, Chudik & Pesaran (2016). The counterfactual is then the dynamic forecast for y_{1,T_0+h} conditional on the observed $y^*_{T_0+h}$. They consider tests for policy ineffectiveness.

In some cases one wishes to forecast a number of outcomes and ensure that the conditional forecasts are consistent with each other, in the sense that they exhibit their historical association. To that end Akhmadieva, & Smith (2019) estimate, what is called a VARX* in the GVAR literature. This is a vector version of (13) explaining an *m* x *1* vector of variables for a country that joined the euro, say country 1, by a corresponding vector of constructed foreign variables

$$\boldsymbol{y}_{1t} = \hat{\alpha}_{10}^{T_0} + \widehat{\mathbf{A}}_{11}^{T_0}\mathbf{y}_{1,t-1} + \widehat{\mathbf{C}}_{1}^{T_0}\mathbf{y}_{1t}^{*} + \hat{\mathbf{u}}_{it}^{T_0} \quad (14)$$

$\widehat{\mathbf{A}}_{11}^{T_0}$ and $\widehat{\mathbf{C}}_{1}^{T_0}$ are *m* x *1* matrices, where *m=6*. This is estimated on pre-treatment data, up to T_0, experimenting with various dates. One can then construct dynamic forecasts conditional on the $\mathbf{y}_{1t}^{*}$, which will provide the counterfactual. Part of the change associated with joining the euro is that post-treatment different euro policy reaction functions determined interest rates and exchange rates than the pre-treatment national ones. One can distinguish the effect of changing policy rules from other changes by also constructing a counterfactual in which the exchange rate and short interest rate are treated as exogenous. Estimating large unrestricted VARs on relatively short periods can produce unreliable results and some of the counterfactuals did not seem reasonable.

Pesaran & Smith (2016, 2018) point out that there are various types of policy change including discretionary interventions where there is a deterministic change to the policy variable and rule based interventions where one or more parameters of a stochastic policy rule are changed. P&S (2018) consider both a standard case where all variables in the macroeconometric model, including policy variables, are endogenous and a general case where the DSGE model is augmented by exogenous variables. The latter case accommodates interventions that change exogenous policy parameters, such as a fixed money supply target, or when steady states of some of the variables are changed as occurs when the inflation target is altered. They make the point that in stable DSGEs estimated on deviations from steady states, any policy changes which do not change the steady states will only have transitory effects and thus be difficult to detect.

All the procedures discussed so far compare the actuals to a counterfactual. Pesaran, Smith and Smith (2007) in considering what would have happened if the UK, or Sweden, had joined the euro compare two counterfactuals. Since the UK is so large relative to the euro area, it would have changed the behaviour of the euro area. They simulate the GVAR, with and without the constraints that UK interest rates and exchange rates were equal to the euro area ones. Over the period they considered 1999-2005, UK interest rates were similar to euro area ones and the sterling euro rate was very stable, so the effects were not large.

4. Propensity Score Methods

Another recent approach to macro policy evaluation also borrows techniques from the micro literature to obtain an estimate of an average treatment effect. Angrist, Jorda and Kuersteiner (2018), AJK, develop flexible semiparametric time series methods for the estimation of the causal effect of US monetary policy on macroeconomic aggregates. While Jorda and Taylor (2015) use similar procedures to estimate the effect of fiscal policy in a panel of countries. AJK use local linear projection type estimators to measure the average effect of policy changes on future values of the outcome variables (inflation, industrial production, and unemployment), using inverse probability weightings, in a way similar to that used to adjust non-random samples, the probability weights obtained from policy propensity scores. They have a parametric model for the propensity score the probability of policy, changes in the federal funds rate target rate announcements, conditional on daily financial market data. Identification comes from the assumption that information revealed by an announcement, conditional on market rates the day before the announcement, is independent of future potential outcomes.

They say it captures the average causal response to discrete policy interventions in a macrodynamic setting, without the need for assumptions about the process generating macroeconomic outcomes. The proposed estimation strategy, based on propensity score weighting, easily accommodates asymmetric and nonlinear responses. Their main conclusion is about the asymmetric effects: interest rates rises can slow the economy down but cuts do not boost it much.

They rely on outcomes averaged across different (possibly heterogenous) policy episodes whilst the studies surveyed so far consider a single policy intervention and average the counterfactual outcomes over the post intervention sample corresponding to that intervention. AJK do not use a model for outcomes and their analysis is potentially subject to the Lucas Critique. Their approach requires that the underlying parameters are invariant to policy changes, since it is only policy changes within the same regime that are identified in their framework (see AJK, p.373). In addition, matching estimators of this sort require a lot of data whereas macroeconometric samples tend to be data-poor relative to microeconometric samples. This isrejected in the large confidence bands AJK report around the measures of their estimated effects of target rate changes on macro variables in figs 2 and 3.

Terzi (2019) uses a propensity score matching models to construct counterfactuals for the euro area crisis countries (Greece, Portugal,Ireland, Cyprus, Spain) based on over 200 past macroeconomic adjustment episodes between 1960-2010 worldwide. At its trough, between 2010 and 2015 per capita GDP had contracted on average 11 percentage points more in the Eurozone periphery than in the standard counterfactual scenario.

5. Conclusion

Although they have not yet become the standard approach to macroeconomic questions these procedures, which draw on microeconometrics, have been increasingly used in macro policy evaluation. However, many studies show that the estimates of the counterfactual and the treatment effect are sensitive to the method used and to the particular specification choices within methods. Of course, this is also true of more conventional macro methods. Hsiao & Zhou (2019) comment that no method appears capable of dominating all other methods under all different DGPs and different sample configurations of cross-sectional dimension N and pretreatment time dimension T_0. Since the true DGP is usually unknown and the statistical findings could be very different for different situations, they suggest model-averaging as a robust method for generating counterfactuals. They also note that the absolute magnitude of the model average estimates could be very different from the estimates based on a particular method to generate counterfactuals.

This sensitivity to choice of method and detail of implementation raises the possibility that researchers might search over specifications in order to obtain the counterfactual that produces an effect that confirms their prejudices. Again this problem is not confined to these methods but indicates the importance of replication studies to investigate the robustness of results.

REFERENCES

Abadie, A. & M. D Cattaneo (2018). Econometric methods for program evaluation. *Annual Review of Economics*, 10, 465-503.

Abadie, A., A. Diamond, & J. Hainmueller (2010) Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program, *Journal of the American Statistical Association*,105(490) 493--505.

Abadie, A., A. Diamond, & J. Hainmueller (2015) Comparative politics and the synthetic control method, *American Journal of Political Science*, 59(2), 495-510.

Abadie, A. & J. Gardeazabal, Javier (2003) The Economic Costs of Conflict: A Case Study of the Basque Country, *American Economic Review*, 93(1) 113-132.

Akhmadieva, V & R. P. Smith (2019) The macroeconomic impact of the euro, BCAM Working Paper 1903. Birkbeck, University of London, London, UK, forthcoming in *Scientific Annals of Economics and Business*.

Angrist, J. D., O. Jorda & G. Kuersteiner (2018): Semiparametric estimates of monetary policy effects: string theory revisited, *Journal of Business & Economic Statistics*, 36:3, 371-387,

Bai, J. (2009) Panel Data models with interactive fixed effects, *Econometrica*, 77(4) 1229-1279.

Bove, V, L. Elia and R.P. Smith (2017) On the heterogenous consequences of civil war *Oxford Economic Papers*, 69(3) 550-568.

Chan, M.K. & S. Kwok (2016) Policy Evaluation with Interactive Fixed Effects. University of Sydney Economics Working paper 2016-11.

Chudik, A & M.H. Pesaran (2016) Theory and Practice of GVAR Modeling, *Journal of Economic Surveys*, 30(1) 165-197.

Gardeazabal, J., Vega-Bayo, A., (2016). An empirical comparison between the synthetic control method and Hsiao et al.'s panel data approach to program evaluation. *Journal of Applied Econometrics,* 32 (5), 983--1002.

Geng, H & Q. Zhou (2018) Estimation and inference of treatment effects using a new panel data approach with application to the impact of US SYG Law on state level murder rate,

Gobillon, L. & T. Magnac (2016) Regional Policy Evaluation: Interactive Fixed Effects and Synthetic Controls, *Review of Economics and Statistics*, 98(3) 535-551.

Hsiao, C., H.S. Ching, and K. W. Shui (2012), A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong kong with mainland China, *Journal of Applied Econometrics*, 27(5) p705--740.

Hsiao, C & Q. Zhou (2019) Panel parameteric, semiparametric and nonparametric construction of counterfactuals, *Journal of Applied Econometrics*

Li, K., & Bell, D. (2017). Estimation of average treatment effects with panel data: Asymptotic theory and implementation. *Journal of Econometrics*, 197, 65--75.

Jorda O. & A.M.Taylor (2016) The time for austerity: estimating the average treatment effect of fiscal policy. *Economic Journal*, 126(1) 219-255.

Pesaran, M.H. (2006) Estimation and Inference in Large Heterogeneous Panels with a multifactor error structure, *Econometrica*, 74(4) 967-1012.

Pesaran, M.H., L.V Smith and R.P. Smith (2007) What if the UK or Sweden had joined the Euro in 1999? An empirical evaluation using a Global VAR. *International Journal of Finance and Economics*, 12, 55-87.

Pesaran, M.H., and R.P. Smith (2016) Counterfactual analysis in macroeconometrics: an empirical investigation into the effects of quantitative easing, *Research in Economics*, 70(2), 262-280.

Pesaran M.H. & Smith R.P. (2018) Tests of policy interventions in DSGE models, *Oxford Bulletin of Economics and Statistics*, 80(3), 457-484.

Terzi, A. (2019) The Euro Crisis and Economic Growth:A Novel Counterfactual Approach, CESifo working paper, 7746.

Wan, S-K, Y Xie, C. Hsiao (2018) Panel data approach vs synthetic control method, *Economics Letters* 164, 121-123..

Xu, Y. (2017). Generalized synthetic control method: Causal inference with interactive fixed effects model. *Political Analysis*, 25(1), 57--76.

(Endnotes)

[1]I am grateful to Veronika Akhmadieva for comments on an earlier version.

[2]These laws allow citizens more scope to use deadly force against others in situations like a burglary. Geng & Zhou (2018) analyse this case.

[3]The assumption that we know the date of treatment is not innocuous. One might argue that the relevant date for euro entry was in the early 1990s, after the ERM crisis, when countries began to adjust to meet the entry criteria; in 1999, the formal date; in 2002 when euro notes and coins were introduced; or in 2008 when the euro constraints began to bind.