INTRODUCTION

Michael Shalev has turned his attention, once again, to the bad methodological habits that social scientists – like myself – often adopt. As always, he presents us with thoughtful, rigorous, and penetrating criticism, but also with a generous dose of constructive prescription. His target is the widespread use of regression techniques in cross-national comparative research. The gist of the argument is that multiple regression (MR) is a far too blunt instrument if our aim is to arrive at a robust identification of crucial causal mechanisms. MR, as he puts it (p. 42), renders the cases invisible and, hence, precludes researchers from having any dialogue with them. The case becomes a set of scores; the causal mechanisms are reduced to correlation coefficients. As a result, analytical power is sacrificed rather than gained. Shalev advocates simpler ‘low-tech’ approaches such as tabular representations, tree diagrams, or clustering techniques either as substitutes for, or as companions to, regression analysis.

It is almost impossible not to agree with Shalev. As one of the three main targets of his paper, my Three Worlds of Welfare Capitalism analyses are, I am happy to see, not completely torn to shreds. The distinctiveness of welfare regimes more or less remains when subjected to alternative treatments, such as Shalev’s factor analytical approach. I am more than ready to
concede that my use of MR to explain welfare regime differences was rather inappropriate for the purpose at hand. I am now older and perhaps also wiser, and would certainly have done it all differently today. But would I now follow Shalev’s prescriptions? Yes and no. I am in full agreement that triangulation, i.e. combining MR with qualitative inspection, should be a favoured approach in small-N comparisons. I am less persuaded with his call for substituting regressions with more qualitative, lower-tech alternatives. Neither am I convinced that substituting MR with factor analysis (as Shalev does in his reanalysis of my data) will yield more analytical insight compared to scrutinizing residual plots from MR.

MR estimation on small-N country samples implies that we easily violate basic key assumptions, such as monotonic linear effects, statistical independence, the absence of selection bias, and conditional independence. Small-N regressions are therefore not very useful – and easily counterproductive – if used primarily to identify the strength of the statistical relationship. But all this does not mean that we should abandon MR.

Below I shall argue two major points. Firstly if MR is utilized as a diagnostic tool, explicitly aimed at detecting such violations, it provides, in my view, unrivalled potential for identification. Secondly, the ‘low-tech’ alternatives that Shalev espouses are not superior with regard to distinguishing wrong from correct causal mechanisms, in particular under conditions of selection and endogeneity.¹ My view is that we should use MR not to identify causal mechanisms via the βs, but rather as a ‘Popperian’ devise. The strength of a statistical association will not tell us much about the real causal mechanisms at work, but the diagnostics that we can obtain from MR residual plots are a minefield of information, truly powerful instruments for fine-tuning and possibly correcting our hypotheses, and subsequently for selecting appropriate alternative instruments. If our true aim is identification (following Manski, 1995), we should not throw MR out with the bathwater.

IDENTIFICATION WITH MULTIPLE REGRESSION IN SMALL-N COMPARISONS

We very often face important macro-level questions that cannot be answered. We usually have few cases but many rival explanations. We easily confound nation characteristics with the dimensions we measure and, hence, it is basically unclear what explains what. The explicit aim of the ‘politics
matter’ literature is to demonstrate that left power (x) matters for welfare state development (y), conditional on a vector (z) of other plausible factors (such as economic growth). The standard approach is to sample the 20-odd OECD democracies and then regress the \( y = f(x | z) \).

As Fearon (1991) insists, the smaller the sample size, the greater is the need to make counterfactuals explicit. In our sample we will observe Italy’s welfare state size, and that Italy was ruled by Christian Democrats throughout the post-war era. The obvious counterfactual is that its welfare state would have been ‘bigger’ or ‘better’ had it been ruled by social democrats. In other words, our sample needs to include another Italy, a country that matches Italy on all relevant z values but differs on x. No such country is likely to be found in the 20-odd OECD sample and we are therefore left with no cell-match. When we add to this that small-N studies make it impossible to condition on all relevant z variables, true causal identification is for all practical purposes stifled.

Furthermore, as Shalev also argues, the choice of OLS regressions implies that we assume monotonic linear effects. Sweden is always the top-scorer on left power cum welfare statism. If MR gives us an \( x\beta \) estimate of 0.3, we are then led to believe that a 5-point increase in left power (Denmark closes the gap with Sweden) will result in a 1.5-point increase in the Danish welfare state size. For Germany, the equivalent effect would be a 4.2-point increase. This is pretty much a non-sensical estimation and is, besides, not what the researcher should aim to identify. I readily admit my guilt in falling into this regression-trap on more than one occasion.

What we truly aim to uncover are the precise causal mechanisms that link x to y. Deep historical scrutiny of all the countries will, no doubt, help the researcher identify how, exactly, left power in Denmark influenced welfare state growth. Doing this rigorously for 20-odd countries would amount to a lifelong project. The reason that I support Shalev’s call for triangulation is that MR can be employed very productively in the pursuit of more precise identification – in particular when aided by explicit and systematic use of counterfactuals.

The really valuable information in MR lies in the residual plots, not in the \( \beta \)s and \( R^2 \). Small N’s are frustrating, but they do have the advantage that scrutiny of the residuals is easy: you can quickly put a name on each point. In this sense, Shalev errs when he claims that MR impedes dialogue with the cases. There are three key issues related to identification where good diagnostic use of MR can become a major asset: dependence among the observations, selection, and endogeneity.
INDEPENDENCE

If our observations are not fully independent of one another, we will violate a basic MR assumption. This will show up as heteroskedasticity in the residual behaviour. Dependency is theoretically interesting because it suggests that some cases follow a similar logic by way of, for example, mutually influencing each other. In time series analysis the logic is similar. If last year’s values influence next year’s we have some evidence in favour of a path-dependency hypothesis. Frank Castles (1993) took dependency to its logical conclusion in his families of nations argument. Many MR practitioners simply do not bother about heteroskedasticity. My Three Worlds study was premised on the idea of clustered regime logics and I should, therefore, have actively tested for independency.

There are two options if the assumption is violated. One can correct for it via country-group dummies (say a Scandinavia dummy). If Castles is right and there are four families, the dummy solution results in paralysis because it will exhaust just about all degrees of freedom. More to the point, regressing with ‘family dummies’ will not get us much closer to identifying real causal mechanisms. The second and much more alluring option is to launch an in-depth study (as Castles did) of why or how diffusion came about – one example of why triangulation via MR can be scientifically fertile. Had I seized upon this opportunity I would probably have paid far more attention to whether regimes emerge from similar behaviour on x or from a policy diffusion process that is unrelated to x. My thesis would clearly have encountered problems were the latter true.

When we regress welfare state size on left power we will inevitably identify a cluster that ‘overshoots’. Putting names to the dots tells you immediately that they are countries with a strong Christian democratic tradition. Van Kersbergen (1995) noted this and subsequently conducted an in-depth examination of similarities and differences in the evolution of Christian Democratic welfare states. Here is another telling illustration of how dependency diagnosis plus qualitative analysis can yield good sociological research.

New countries are spawned from old ones almost on a yearly basis, and there are undoubtedly many MR practitioners that see this as a welcome addition of N’s. United Nations membership has leapt by 50 percent in the past decade with the birth of new nations. One should, of course, not assume that the new Slovakia and Slovenia are statistically independent from the old Czechoslovakia and Yugoslavia. We might also ask ourselves whether intensified EU integration has diminished the degree of independence that once existed between, say, Finland and France. Some clues may be found in the MR residual plot.
Our main challenge is to distinguish true from spurious relationships. Our
inferences will be seriously biased if \( y \) and \( x \) are both the outcome of some,
possibly unobservable, heterogeneity or if \( x \) is not truly exogenously deter-
mined. Selection bias and endogeneity are essentially two facets of a similar
problem, namely that if they are present we will make incorrect conclusions
regarding the causal mechanisms we care about. Using welfare state re-
search again as illustration, it is very possible that strong social democracy
and large welfare states are jointly determined by some unidentified factor
that, perhaps, lies deeply buried in history. Take Sweden: the seemingly
obvious connection between left power and welfare state growth may, in
reality, be incorrectly identified. It is theoretically equally possible that both
attributes of modern Sweden have their roots in any number of historical
peculiarities, be it patterns of landholding in pre-industrial ages, the nature
of absolutism, industrial structure, or the transition to democracy. We must,
likewise, assume that the welfare state – once in place – will have had
substantial influence on the social democrats’ electoral fortunes, both in the
short and long run. If so, the \( x \) for Sweden is influenced by \( y \) and the
assumption of conditional independence is violated.

Selection and endogeneity are often difficult to detect and manage. The
simpler methodologies that Shalev advocates are, as far as I can see, not
better equipped to handle either, at least when compared to MR.

Selection bias may be related to observables or unobservables (Heckman,
1988). The former occurs when the expected covariance \( E(u_j | u_\mu) \neq 0 \), but it
disappears once we control for the observed variables \( Z \), so that \( E(u_j | | u_\mu) = 0 \).
The latter is present when \( E(u_j | u_\mu) \neq 0 \) and \( E(u_j | | u_\mu) \neq 0 \). In this situation,
controlling for the factors observed by the investigator does not remove the
covariance between the errors in the outcome and the selection equations.
Now note that the regression coefficient \( \sigma_{ju} = \text{cov}(u_j u_\mu) / \text{var}(u_j) \). If selection
is on unobservables, controlling for some variable \( x \) in the outcome equation
may reduce the error variance \( u_j \) without equally reducing the covariance \( u_j u_\mu \).
Hence, the coefficient on the omitted variable will be larger and the bias will
be exacerbated.

Accordingly, the expected values of the observed cases will be biased
because they co-vary with the variable that determines which cases are ob-
served. This bias can be corrected by conventional controlling procedures.
But if bias stems from unobservables, such controls will only worsen the
bias. As Heckman (opp.cit: 7) argues, the dilemma is that different methods
of correcting for selection bias are robust if there is no bias to begin with; if
there is, there is no guarantee that the methods are robust.
The problem is similar whether we study large or small N’s (Fearon, 1991). De Toqueville provided a nice exemplification with his observation that revolutions do not seem to change anything. The reason might be that they occur only in countries where it is difficult to change society in the first place. Accordingly, even studies based on \( N = 1 \) may suffer from selection bias: the French revolution may have been caused by the same conditions that made social change so difficult. It is possible that a revolution in a country where social relations are easier to change would have provoked change. But then a revolution would not have been necessary.

This suggests that more qualitative case-specific methods that prioritize dialogue between researcher and the case hold no special advantage over MR as far as selection bias is concerned. In essence, the only genuine method to correct for selection bias is to construct counterfactuals, to fill in the unobserved values in the distribution of \( y \) for all \( x \)’s. Comparative analysis of the case-study variety cannot benefit from statistical distributions to generate the counterfactuals. In this respect there is accordingly something to be said for methods, like MR, because they provide such distributions and because they permit us to estimate covariance coefficients.

The problems related to endogeneity are virtually identical and require, therefore, less elaboration. There are, however, a few small points to add. Endogeneity is present when our \( x \)’s are conditionally dependent on \( y \). Using welfare state research again to exemplify, this can be because social policies directly influence the parliamentary fortunes of social democracy (the Swedes love their welfare state and vote Left to ensure its continuity). It can also be because Sweden’s welfare state and Sweden’s unique variety of social democracy are part and parcel of ‘everything that is Sweden’. In the latter case, the true \( x \) and \( y \) for Sweden is not left power, nor welfare statism, but a full list of all that is uniquely Swedish.

If this is so, the fixed-effects panel estimation approach will go wrong since it assumes that \( x \) will have an identical impact on \( y \) regardless of which country. But if the left power effect on welfare statism is ‘Sweden’ or, perhaps, ‘Germany’ specific we should assume non-identical effects. Similar to the identification of selection on observables, we might therefore introduce controls for everything that is Sweden or Germany specific. Small-N studies with strong endogeneity have little capacity to extend the number of potentially necessary controls. The solution is therefore, once again, to concentrate on the theoretical elaboration by means of counterfactuals.

One promising avenue is to redefine the dependent variable so that it is less likely to incorporate all that is Swedish, and/or so that it is less likely to directly pattern voters’ party preferences. Indeed, the welfare state literature
has to a degree moved in this direction by replacing aggregate measures (such as social expenditure) with narrower indicators that measure specific properties of welfare states. However, the underlying problem may still remain if such properties are, once again, the mirror image of ‘all that is Swedish’ rather than verifiably related to specific values of \( x|z \).

If MR is applied to a sufficiently large number of N’s and used for diagnostic purposes, it can be a powerful and efficient method for detecting endogeneity – certainly superior to the kinds of low-tech alternatives discussed by Shalev. We do have good testing procedures to detect non-identical \( x \)-effects in fixed-effects regressions or, alternatively, we can use an IV approach within two-stage least squares estimation. These options are typically precluded in small-N studies and we are, therefore, back again to the importance of counterfactuals as our only realistic alternative.

In brief, my response to Michael Shalev’s argumentation is that we should favour whichever method delivers superior information about the underlying statistical distributions. In some cases, MR may be the relevant choice; in others, possibly not. We should, above all, be careful not to throw the baby out with the bath water. MR has very powerful and easy-to-use diagnostic tools that can be mobilized for what statistical analysis really should pursue, namely to search for the true causal mechanisms. If, instead, we continue the past tradition of employing MR to show that our \( \beta \) and \( R^2 \) are bigger than others’, then I agree whole-heartedly with Shalev. His lower-tech alternatives are less likely to produce violations of basic estimation assumptions than is the uncritical MR-based search for superior \( R \)-squares.

NOTE

1. Many of the points to be covered in this paper were previously examined in Esping-Andersen and Przeworski (2000). For illustrative purposes, I will draw primarily on examples from my own work on comparative welfare states.

REFERENCES


