Building on solid ground:

Robust case selection in multi-method research

Ingo Rohlfing (University of Cologne, rohlfing@wiso.uni-koeln.de) and

Peter Starke (University of Bremen, peter.starke@sfb597.uni-bremen.de

)

Earlier versions of this paper were presented at the ECPR Joint Sessions of Workshops 2008, the Annual Meeting of the Working Group on Political Methodology of the German Political Science Association in 2008, and the Odense Workshop on Political Methodology in 2010. We thank the participants of these meetings for their helpful comments. Valuable research assistance was provided by Ross McCalden, Nancy Deyo and Tobias Schafföner. The usual disclaimer applies. This work is funded by the Deutsche Forschungsgemeinschaft (DFG).

After years of debate between qualitative and quantitative researchers in the social sciences, a growing number of scholars now take the view that multi-method research (MMR) is the way forward (Brady et al. 2006). While there is a great variety of ways in which MMR can be performed (Creswell and Plano Clark 2011), the recent methodological and empirical MMR literature focuses on the combination of some variant of regression analysis and process tracing (Bäck and Dumont 2007, Coppedge 1999, Kenworthy 2007, Lieberman 2005, Tarrow 1995). Even those who are skeptical of the utility of stand-alone regression analysis acknowledge that they have value if combined with process tracing based on an informed choice of cases (Shalev 2007).

We fully share this perspective on MMR, but also contend that the current perspective on the combination of regression analysis and process tracing fails to take some significant issues into account (Ahram 2011, Goerres and Prinzen 2012). In our paper, we address a serious problem related to *case selection on the basis of regression results*, which is the nexus between the two methods that has received considerable attention in the past (Bennett and Braumoeller 2005, Lieberman 2005, Rohlfing 2008, Seawright and Gerring 2008, Wolf 2010).

The rationale for case selection on the basis of a regression analysis is process tracing with the aim of generalizing these insights on mechanisms beyond the selected cases (Collier and Mahoney 1996, Lieberman 2005, Seawright and Gerring 2008, Shively 2006). In the literature on case selection in MMR, the general alternatives are random selection (Fearon and Laitin 2008) and the deliberate choice of cases (Bäck and Dumont 2007, Lieberman 2005, Seawright and Gerring 2008), both against the background of relevant regression results. As we explain in section two, the goal is usually to select either cases that are well-predicted by the regression model, called *typical cases*, or badly predicted cases, called *deviant cases*. Well-predicted cases are considered suitable for the analysis of causal mechanisms that underlie the estimated causal effects, whereas badly predicted cases may hint at ways to improve the theory and corresponding regression model. By reviewing published MMR research, we show that this design is already implemented in the empirical literature.

The recent debate about the proper case selection strategy made a significant contribution to the rigorous conduct of MMR. However, we argue that it disregards the very real possibility of being unable to settle on a single regression model in the first place (Ho et al. 2007). We present an empirical example and a review of empirical MMR research demonstrating that *modeling uncertainty* renders it necessary to estimate multiple models among which one finds it impossible to discriminate on grounds of theory, model diagnostics, and model selection criteria.

The question that naturally arises in such a situation is: if there is no justification for singling out one model, what model and corresponding results should I choose for case selection? We demonstrate that it is fallacious to believe that all of the available models serve equally well as the basis for case selection because they do not. The perils of modeling uncertainty in MMR are elaborated by highlighting the hitherto ignored distinction between the *classification* and the *selection* of cases in MMR. The classification of cases as typical or deviant is indispensable for determining the subsets of cases to which process tracing insights can and cannot be generalized. Different models and quantitative results can yield different classifications of the same case, which leads to *classification uncertainty* about the status of that case when generalizing causal inferences. Since generalization of process tracing insights is the goal, this is a serious problem that holds regardless of the actual case selection strategy used. Furthermore, we demonstrate for intentional and random case selection how modeling uncertainty creates the potential for *misselection* by choosing a case that is differently classified across models.

We aim to solve these problems by developing an easy-to-implement robust case selection procedure for process tracing in MMR. The protocol allows for case classification and selection with the maximum possible degree of confidence in the face of non-robust regression results. We therefore build on the extant state of the debate on the choice of cases in MMR by further developing the methodological toolkit and equipping empirical researchers with guidelines for dealing with the implications of modeling uncertainty.

For this purpose, section two introduces the idea behind the analysis of typical and deviant cases and three case selection strategies recommend by the MMR literature. The section also presents some existing applications from the empirical literature which, by and large, follow this state of the art, but ignore the problem we carve out in the following sections. In section three, we then present our empirical example of party competition. It only serves illustrative purposes for the elaboration of our procedure with real data and we do not intend to make any substantive contribution to the field of party research. Using this example, we review tools for addressing modeling uncertainty in section four and show that they are of limited use here. Section five highlights the implications of non-robust results for case classification. Furthermore, we present examples showing that modeling uncertainty is common in empirical MMR research. A discussion of the ramifications for case selection is the subject of section six. On the basis of these insights, we develop a straightforward robust case selection procedure in section seven. The last section concludes.

2. Why and how to select cases for process tracing after regression analyses

In principle, it is possible to combine case studies and regression analysis in both sequences, i.e. the large-n analysis is followed by process tracing or vice versa. If the case study comes first, however, there is no basis for choosing cases on the basis of regression estimates. Consequently, our focus rests on designs in which the regression part comes first, followed by the case study.

The choice of cases in regression-based MMR designs, thus understood, can have different aims. The analysis of a *typical case* can be either the generation or a test of hypotheses on the mechanisms that account for the causal effects discerned in the regression analysis (George and Bennett 2005, chap. 10, Hall 2008, Shively 2006, p. 346).[1] In the first scenario, one simply does not know why a certain cross-case pattern is in place and refrains from the formulation of hypotheses prior to the empirical analysis. Process tracing is exploratory and leads to the generation of propositions that can be tested in a follow-up study. The alternative to exploratory process tracing in typical cases is a test of a hypothesis on mechanisms. This is particularly useful when several rival theories predict the same causal effect, but stipulate distinct underlying causal mechanisms (Campbell 1975). A case in point is the democratic peace phenomenon, meaning that peace between two democracies is the rule (George and Bennett 2005, chap. 2). While the cross-case pattern is very robust, the mechanism accounting for democratic peace is not known. The democratic peace phenomenon could be due to the commitment of political leaders to democratic norms, or an institutional constraint deriving from the public, just to mention two possible mechanisms here (Rosato 2003).

The analysis of *deviant cases* is a classic in political science (Lijphart 1971).[2] In Lijphart's influential definition, deviant-case analyses are 'studies of single cases that are known to deviate from established generalizations' (1971:692).[3] The analysis of anomalies entails the search for previously omitted factors that account for the deviance of a case.[4] One can then distinguish between a narrow and wide variant of the deviant-case study. We may be 'narrowly' interested in explaining why a substantively important case is not well-predicted by the explanatory model (Mahoney and Goertz 2006, 242-3), without having the intention of generalizing the insights to other cases in the sample. It would be possible, for example, to study such a case in depth in order to understand its special status – a goal similar to studies of national 'exceptionalism', traditionally of the U.S.-American, Japanese, or French variety.

Alternatively, we may look for an explanation with a wider scope that can be generalized beyond the single deviant case. In our view, this is the way in which this research design is usually understood and more widely applied (Eckstein 1975, 118, Levy 2008). In the context of MMR, there is a strong affinity with a generalizing deviant-case study because the analysis is centered on a sample or population of cases expected to be causally homogenous (Lieberman 2005). Given the assumption of causal homogeneity, it suggests itself as a means to generalize insights from one or few cases to other cases. Some of the more respected case-study designs – including Lijphart's (1968) own case study of the Netherlands – are based on the idea that process tracing of anomalous cases may help in the search for variables previously considered to be theoretically irrelevant (Rogowski 1995). Omitted-variable analysis thus proceeds through exploratory process tracing aimed at improving the overall model-fit in a theoretically intelligible way.

Typical and deviant cases thus both serve to generate insights that are generalized beyond the examined cases (see also below). If generalization were not the goal of MMR, systematic case selection based on regression results would be pointless because process tracing would be narrowly confined to explaining a single, substantively important case. In this instance, case selection would be simply guided by subjective assessments of a case's substantive importance (Rohlfing 2008). Moreover, the very notion of a typical case entails generalization because a typical case is representative of other cases. What can be inferred from process tracing in a typical case thus extends to other typical cases. Of course, there is a risk of overgeneralization (Collier and Mahoney 1996), but this is a risk one always faces when generalizing and is not limited to process tracing in MMR.

There is no full agreement in the MMR literature on how to choose typical cases and deviant cases for process tracing. The alternatives offered are varieties of *intentional case selection* and *random selection*. The intentional choice of cases is based on a comparison of a case's actual outcome with the score predicted by the regression model. In this perspective, a typical cases is well-predicted by the regression model (see below for a definition of "well-predicted"). Conversely, when attempting to choose a deviant case, one should look for those that are badly predicted by the regression model (Gerring 2007a, 105-108). According to one variant, it is recommended to choose the *best*-predicted case for the analysis of a typical case and the *worst*-predicted case for a deviant-case study (Seawright and Gerring 2008).[5] In contrast, Lieberman (2005, 444) proposes intentional selection of several typical cases with large variation on the independent variables in order to determine whether the same mechanism is operative in all the selected cases.[6] Fearon and Laitin (2008) raise concerns that researchers may engage in cherry-picking by (unconsciously or intentionally) selecting cases which are likely to confirm their preferred argument. In order to avoid the introduction of an investigator bias, they recommend random selection or, random selection stratified by relevant factors like the region to which a country belongs.[7]

[table 1 about here]

Table 1 presents a number of applications of MMR designs from different subfields of political science.[8] We find both the study of typical and deviant case, although typical-case analysis is more popular. Most contributions select more than one case for in-depth study or, at least, for qualitative illustration. What is more, most studies also present more than one regression result in the article. However, case selection is always based on a single regression model only, that is, even when more than one is presented in the same article in order to assess the robustness of the quantitative results. The method with which typical and deviant cases are classifies varies slightly across the applications. When it is specified in the article, it is always based on the size of case residuals (or average residuals of a substantive group of cases). Some authors follow Seawright and Gerring's (2008) advice and select the cases with the largest (or smallest) residuals (see below), others for example choose a benchmark of one standard deviation to separate typical and deviant cases. It is also interesting to note that nobody uses random selection. In total, this brief literature review shows that, based on Lieberman's systematic elaboration of nested analysis, MMR becomes popular and respected among political scientists (see also Collier, Brady and Seawright 2010). While those who employ 'nested design' following Lieberman's template follow different goals and use somewhat different classification methods, we emphasize that they invariably build their case selection upon a single regression model. In what follows, we show that the consequences of this widespread practice can be serious and suggest ways to avoid these problems.

3. An illustration of modeling uncertainty: the dynamics of party competition

To illustrate our arguments concerning hitherto ignored challenges in case classification and selection, we chose to use a substantive example from the literature on party competition. We chose this particular study simply because the underlying theoretical argument is easy to present and understand in this context which, in turn, allows us to focus on our methodological arguments. Equally important, Adams and Somer-Topcu (2009, 843) correctly assert that the cross-case results are indeterminate with respect to the underlying causal mechanisms because one can think of multiple reasons that parties respond to the average shift of their competitors. While we do not specifically seek to discern here which mechanism(s) explains the behavior of parties, as we discuss above and in the following section, cross-case results compatible with competing causal mechanisms are one of the main reasons for performing an MMR (Lieberman 2005).[9]

The data and the baseline model are taken from an article by Adams and Somer-Topcu (2009) on changes in party ideology in 25 democratic countries between 1945 and 1998. Adams and Somer-Topcu aim to determine the extent to which the ideological shift of a party (the focal party) is driven by the ideological moves of its competitors. The dependent variable is the ideological change of the focal party between two consecutive elections. The independent variable of interest is measured as the average shift of the competing parties at the previous election. It is lagged by one election because of the argument that the focal party needs some time to respond to the moves of its competitors. Additional independent variables that are included are changes in public opinion and the ideological change of the focal party at the previous election.[10] The lagged ideological change of a party is part of the model because of the policy alternation hypothesis (Budge 1994). According to this hypothesis, the current ideological shift of a party is in the opposite direction of its previous shift.

A second model additionally takes into account that the focal party may not treat all competing parties as equal. The focal party might be particularly sensitive to the shifts of parties belonging to the same party family. In order to test for a party family effect, the baseline model is expanded by adding a party family variable measuring the effect of the average left-right shift of parties from the same party family.[11] Our replication of Adams and Somer-Topcu's results, which confirms their findings, is presented in table 2. The left column includes the results for the baseline model. The first row shows that the hypothesis on the effects of rival parties' policy shifts has empirical resonance. The right column contains the results for the party family model. Both relevant variables, captured by the first two rows, are in line with the theoretical expectations. From a quantitative perspective, one can conclude that the result for the baseline model is fully robust to the inclusion of the party family variable because the sign of the effect of the all-parties variable, its level of significance, and the size of the effect are very similar in both models.[12]

[table 2 about here]

Before starting the discussion, we emphasize that the arguments we make in the following hold independently of particular estimation techniques in regression analysis and extend beyond regression analysis. The empirical example is built on time-series cross-section data, but case selection could also involve an ordinary cross-section OLS (Lange 2009) or a discrete choice model (Bäck and Dumont 2007). Moreover, the problems we identify for case selection in MMR, i.e. modeling uncertainty and non-robust cross-case results, pertain to other cross-case techniques that can also be combined with process tracing, most notably Qualitative Comparative Analysis (QCA) and matching.[13] For this reason, our concern with case selection after regression analysis should be seen as illustrative for a problem that is independent of the large-n technique applied in MMR.

4. Searching for the best model: diagnostics and selection criteria

The two models presented in the previous paragraph show that the estimates are fairly robust to the inclusion of a party family variable. Although both models are equally satisfactory from a quantitative standpoint, they represent partially different theoretical arguments. For this reason (and even if we were not to choose cases for process tracing), there is value in the application of regression diagnostics and model selection criteria so as to check whether we can settle on a single, best model.[14] Modeling uncertainty is a familiar topic in the social sciences (Bartels 1997, Ho et al. 2007), but is less so in the MMR literature (but see Lieberman 2005). The methods literature offers tools that might allow one to diminish, or even eliminate modeling uncertainty (Fox 1991), but not necessarily. As they are too numerous, we cannot discuss all the available tools here, but instead focus on two prominent tools from the literature on *regression diagnostics* and *model selection*.

The empirical example centers on the question of whether or not one variable should be included in the model, rendering diagnostics for omitted variables particularly relevant here. A common visual means for underspecification is the residuals-vs.-fitted plot (Cox 2004, Fox 1991). A plot of the residuals against the fitted values allows the assessment of systematic patterns in the distribution of the residuals.

[figure 1 about here]

Figure 1 combines the plots for both models. While there are minor differences between the distributions, overall, they both look fine and do not point to serious problems. There is neither an indication of cases that clearly stand out or heteroskedasticity, nor of a skewed or otherwise apparent distribution of cases that would point to misspecification. Given that the two models differ with regard to one variable, we could expect discernible differences between the two plots, such as a skewed distribution in the left plot. The fact that the distributions are very similar in both plots means that the residual-vs.-fitted plots give us little guidance for model selection here.[15]

In the literature on model selection, information criteria such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) occupy a central role (Burnham and Anderson 2004).[16] On a general level, the information criteria capture how well the model describes the data, taking into account the complexity of the model in terms of the number of parameters. Table 3 compares the AIC and BIC scores for both models.[17]

[table 3 about here]

The two models perform equally well as regards the two information criteria, which is attributable to the strong robustness of the quantitative results of the models. We observe marginal differences between the AIC and BIC for the same model and across models, but not large enough to justify declaring one model unequivocally superior and discarding the other. The illustrative application of two common tools from the field of regression diagnostics and model selection therefore do not help us in discriminating between the two models. In a multi-methods study, the diagnostics and selection criteria thus still leave us with no clue about the proper basis of case selection for process tracing.[18]

The literature on case selection in MMR bypasses the issue of model uncertainty by presuming either that one can settle on a single model (Seawright and Gerring 2008), or that the results are robust (Lieberman 2005). While it might be that diagnostics and selection criteria reduce the number of models to one, our example shows that this is not necessarily so. As is widely known and demonstrated by our empirical example, it is rarely possible to focus on one model only because of uncertainty what the best model is (Ho et al. 2007, Young 2009).

In this context, it is interesting to note that prominent examples of MMR in the social sciences do not fulfill the strong expectations of the methods literature about a single model of reference, either. Iversen (1999) is, among other things, interested in the effects of macroeconomic institutions on unemployment in 15 OECD countries. He estimates three regression models that are all in line with his theoretical argument, perform equally well, and show that the effect of the relevant macroeconomic variables is robust across these models. Dunning (2008, 122) examines the effects of oil rents on a country's level of democracy. He presents the results of two regression models demonstrating that the marginal effect of oil rents is robust to the inclusion of control variables.[19] Ziblatt's (2009) multi-method analysis relates the incidence of electoral fraud in emerging democracies to the inequality among landowners. The regression analysis includes five models that differ as regards the included controls, and all support the expectation of a positive association between inequality and the occurrence of fraud. Howard and Roessler (2006) theorize a positive correlation between a united opposition coalition on the occurrence of liberalizing electoral outcomes in competitive authoritarian regimes. Due to modeling uncertainty, they estimate numerous models that all substantiate their theoretical expectation.

We could carry on with more empirical studies, but leave it with noting that all the studies of which we are aware confront modeling uncertainty and report multiple models supporting the respective theoretical claims. In the following sections, we demonstrate that modeling uncertainty should be taken seriously in MMR because they do not automatically imply robust case classification and selection. Our example shows that these problems also hold in the face of robust quantitative results producing similar effects, signs and levels of significance across models.

5. Modeling uncertainty and case classification

So far, we have replicated quantitative analyses and applied regression diagnostics and selection criteria. Given that there does not seem to be a way to adjudicate between the two models, the quantitative part of the MMR analysis stops here. It might be tempting to conclude that either model should serve equally well as the basis for process tracing. In the remainder of the paper, however, we demonstrate that this is a flawed conclusion. The pitfalls can be highlighted by the important distinction between the *classification* or *designation* of cases as typical and deviant, and the actual *selection* of cases for process tracing. The present MMR literature centers on the second aspect and largely ignores the issue of classification, meaning that empirical researchers are currently left with little guidance about how to delineate typical from deviant cases in the first place.[20]

Lieberman (2005, 445) suggests the standard deviation of the predicted value for the purpose of classification, two standard deviations being a potential benchmark distinguishing typical from deviant cases.[21] This proposal goes in the right direction because it would allow the separation of well-predicted from badly predicted scores.[22] However, the simple calculation of the standard deviation of the predicted values falls short of offering a valid criterion because it does not take into account that the uncertainty of the prediction is not equal across the entire distribution of the predicted values. For this reason, we propose the *prediction interval* instead of the standard deviation because the prediction interval reflects the varying uncertainty of the predicted values across the distribution (Wooldridge 2003, 216).[23] Figure 2 includes the 90-percent prediction interval for the predicted scores in the all-parties model.[24]

[figure 2 about here]

The prediction interval must be interpreted as follows.[25] Given the calculation of a 90-percent prediction interval, the predicted value of a given case lies within this interval in 90 samples out of 100 samples. When the observed value on the outcome is located inside the interval, we classify a case as typical because it is within the range of scores in which we expect cases to be observed. Correspondingly, a case counts as deviant when the observed score is outside of the prediction interval as this is where the case should not be located. The prediction interval therefore offers a straightforward and easy-to-understand way to designate cases as either typical or deviant.[26]

Two additional general notes on the prediction interval are in order. First, as holds true for the specification of levels of significance more generally, the width of the prediction interval has implications for the classification of cases. A 95-percent prediction interval – or a 99-percent prediction interval, for that matter – would be broader than the 90-percent interval that we chose for figure 2. In general, the broader the selected interval, the more cases will be classified as typical and the fewer cases qualify as deviant. The calculation of different intervals therefore might lead empirical researchers to classify the same case differently. Second, the calculation of the prediction interval presumes that the regression model is sound according to the current state of theory and econometric standards.[27] When the model suffers from an omitted variable bias or fails to accurately model the data structure and data-generating process more generally, the residuals should not be taken at face value. Of course, we are rarely, if ever, completely certain about how to specify our model (Ho et al. 2007), which is exactly at the heart of the problems behind case classification and selection in MMR.

Having elaborated the technique for the delineation of typical and deviant cases, we now turn to the implications of modeling uncertainty on the classification of cases. In figure 2, we used the all-parties model for the designation of cases as typical and deviant. However, we could have equally well taken the party family model because we cannot discriminate between the two. When we apply the 90-percent prediction interval to both models, we obtain the classifications that are summarized in table 4. The cross-tabulation represents the extent of non-robust classification, or *classification uncertainty*, across the two models and forms the basis for a three-fold typology of cases in MMR. Most of the cases, about 90 percent, are typical in both models and are what we call *robust typical cases*. This high percentage does not come as a surprise because most cases naturally are well-predicted by the model (according to the selected standard).[28] Furthermore, 131 of all cases are deviant in both models and qualify as *robust deviant cases*. Neither of the two types of cases is problematic because regardless of what model is estimated, we classify a case in the same way.

[table 4 about here]

More problematic are the cases in the upper-right and lower-left cells because these are *non-robust cases*. Ten cases that we take as deviant according to the party family model are typical in the all-parties model, while nine cases qualify as deviant according to the all-parties model and are assigned as typical on the basis of the party family model. Given that we have to classify cases on the basis of two models because of modeling uncertainty, we obtain 19 cases about which we do not know whether they are better taken as typical or deviant.

Building on our previous arguments about typical and deviant cases, classification uncertainty has implications for generalizing causal inferences on mechanisms. Classification is about the delineation of typical and deviant cases for the sake of knowing the group of typical cases to which causal inferences in typical-case analyses can be generalized. In light of our three-fold typology of cases, it now holds that causal inferences derived from robust typical cases should only be generalized to other robust typical cases. This is different for process tracing insights generated in robust deviant cases because they can be generalized to all cases in the population. We emphatically underscore that classification and generalization are independent of how cases were selected for process tracing in the first place. Classification uncertainty due to modeling uncertainty is a problem *regardless* of whether individual cases are randomly or intentionally chosen.

6. Modeling uncertainty and case selection

In addition to the implications for case classification, modeling uncertainty has important ramifications for the choice of cases. The precise implications hinge on the selection strategy that one applies in MMR. When robust typical and robust deviant cases are randomly chosen from the respective set of robust cases, the share of non-robust cases is equivalent to what we call the *misselection probability*.[29] For illustration, presume that one deviant case is selected on the basis of the all-parties model while the status of the case according to the party family model is ignored. Table 3 shows that 140 cases are deviant in light of the all-parties model, but that nine of these 140 cases are typical when relying on the party family model. If we chose a deviant case randomly from the set of 140 deviant cases, we have a chance of about six percent (9/140) of selecting a case that is deviant in the all-parties model and typical in the party family model.

Misselection is also an issue for the intentional choice, the consequences differing on the selection strategy. Lieberman's (2005) proposal to select typical cases representing the full range on the independent variables suffers from potential misselection because what is a typical case according to one model might be deviant for another model. In contrast to random selection, the actual probability of misselection cannot be determined in the abstract because it depends on the precise distribution of cases in the underlying sample. However, it holds that the upper probability bound for misselection is equal to the misselection probability under random selection. The reason simply is that one cannot choose more non-robust cases for process tracing than exist in a given MMR study.

The link between modeling uncertainty and case selection is somewhat different when cases are chosen intentionally by searching for the case with the highest and lowest residual, respectively (Seawright and Gerring 2008, 299-300). If the case with the smallest residual is considered the most typical case, non-robust case selection and a misselection probability of one is given when this case is not the most typical case in a second model. The same holds true for the choice of a deviant case for process tracing. The intentional search for the most typical and most deviant case is more susceptible to misselection than the other two strategies because there can be only one such case per model. Indeed, a look at our example shows that the most typical and most deviant cases differ in the all-parties model and party family model.[30] Drawing together the insights from this section and the previous one, it holds that modeling uncertainty has adverse consequences for the classification of cases and intentional and random case selection alike.

One could object to our example and diagnosis on the grounds that the number of non-robust cases is small and that these are negligible problems in MMR.[31] However, this position is incorrect for two reasons. First, there is no need to take unnecessary risks by inadvertently choosing a non-robust case for process tracing and generalizing causal inferences to cases in which they should not be generalized. Since the procedure that we propose below is easy to implement, it is always recommendable to address the problem of non-robust cases upfront. Second, the empirical example is an instance of perfect quantitative robustness because the marginal effects of all variables have the same sign, are of similar size, and reach the same level of significance in both models. If we confront classification uncertainty and selection uncertainty in this prime example of quantitative robustness, the former should be much higher when the regression analysis is characterized by regression results that vary more across models than they do in our example.

7. A procedure for robust case classification and selection

Our leading empirical example demonstrates that modeling uncertainty has far-reaching implications by creating a propensity for the misclassification and misselection of cases. Existing instruments for the diminishment of modeling uncertainty should be certainly applied in MMR, but section four – as well as our short review of MMR studies – showed that these tools do not necessarily reduce the number of viable models to one. Thus, there is an evident need for a procedure that places causal inference and generalization via process tracing on more solid ground. In the following, we elaborate this instrument in the context of a general exposition of a four-step procedure for the classification and selection of cases in regression-based MMR (table 5). The procedure partially recapitulates the preceding sections and is designed to maximize confidence in classification and selection by taking into account the perils of modeling uncertainty.

[table 5 about here]

The first two steps reflect nothing more than good practice of quantitative research. First, one should invoke theory in order to formulate regression models. When the theory is strong, one might be lucky and be able to specify only a single model. However, social science theory is weak and our cursory look at MMR studies shows that strong theories yielding one model are the exception to the rule. The need to estimate multiple models involving different variables or functional forms for the same variable can therefore be considered the norm (Freedman 1991). Similarly, the data structure rarely permits the application of a single estimation strategy (Kittel 1999, Kittel and Winner 2005). This means that given the same model in terms of included variables, one might have to estimate the model multiple times with different analytical strategies.[32] If we have to perform multiple estimations in step one, step two should aim at reducing the number of models by invoking regression diagnostics and model selection criteria.

If more than one more model remains after step two, we have shown that the current literature leaves MMR researchers with no guidance on how to maximize the validity of case classification and selection for process tracing. The third and fourth step of our procedure contribute to the advancement of MMR by providing researchers a novel and easy-to-implement instrument for the correct choice of cases in the face of non-robust case classification. Since these two steps follow the application of diagnostics and selection criteria, they do not replace but, instead, complement the established instruments in quantitative research.

In step three, one should engage in the *cross-model classification* of cases as was done in section five. Cross-model classification subsumes two elements. For each model that remains after the first two steps, cases are designated as typical and deviant on the basis of a specific prediction interval. Subsequently, cases should be denoted as robust typical, robust deviant, and non-robust, depending on whether their classification is identical or diverges across the models. We have shown above that the virtue of cross-model classification is two-fold. First, we maximize our confidence in the status of a case and the subsets of cases to which we should and should not generalize causal inferences. Second, regardless of the case selection rule, one obtains the best possible basis for the selection of individual cases for in-depth analysis. The extent of non-robustness can turn out to be quite large in this step of the procedure. However, since the models that we estimate are a reflection of the current state of theoretical uncertainty and specification uncertainty, we have to live with a great extent of non-robustness as it reflects the uncertainty that is inherent to our MMR study.

The choice of cases is subject to step four. When performing random case selection, cases should only be chosen from the set of robust typical and robust deviant cases, respectively. The group of non-robust cases should be set aside because we do not know whether the cases are suitable for the formation and test of hypotheses on mechanisms, or for the refinement of theory via exploratory process tracing. The classification of cases as robust and non-robust can be easily adapted to intentional case selection strategies. When one seeks to choose typical cases spanning a range of scores on the covariates, one should simply select these cases from the set of robust typical cases, for example by following Seawright and Gerring's (2008) recommendation and choosing the robust typical case with the smallest residual of all robust typical cases.

If the goal is to choose the case with the smallest residual or largest residual, all we must do is to discern whether the most typical and most deviant cases in one model also have the same quality in another model. Non-robust quantitative results are not an issue when cross-model classification shows that the same case is always the most typical or most the deviant. When most typical and most deviant cases differ across models, on the other hand, this case selection strategy actually lacks a basis because all cases that seem relevant at first turn out to be non-robust. In the presence of non-robustness, there thus is a salient difference between random and intentional case selection because random selection is likely to have a pool of robust cases at its disposal from which a case can be selected.[33]

Regardless of the case selection strategy that one follows, it is worth re-emphasizing that inferences generated from robust typical cases can only be generalized to robust typical cases. Due to their different nature, generalization follows a different logic for robust deviant cases. For the reasons explained above, insights from deviant cases are generalized to all other cases in the population, which includes non-robust cases. Still, it holds that one should only pick robust deviant cases in order to be as confident as possible that the selected case is appropriate for exploratory process tracing. It might seem counterintuitive to put non-robust cases aside during the case selection stage because of the belief that they are particularly insightful for diminishing modeling uncertainty. For our empirical example in particular, one could argue that non-robust cases could shed light on whether parties pay particular attention to the ideological shifts of parties from the same party family.

We argue that this belief is fallacious for two reasons. First, process tracing is valuable, but cannot give more than a hint at what variables might be relevant, which is due to the need to focus on a very small share of cases. Evidence for the claim that parties from the same party family play a special role in party competition must be undertaken by regression analysis and the use of tools for model selection (which fail to discriminate between the models in our example).

Second, one might believe that non-robust cases are particularly insightful because the inclusion of the party family variable turns nine cases from badly predicted cases to well-predicted ones. It is common in the MMR literature to take the change in the status from deviant to typical as evidence for the need to include an additional variable into the model (Gerring 2007b, Lieberman 2005). However, this is a dubious strategy because more variables are likely to capture more variance on the outcome even if the new variable is not underpinned by a causal mechanism. Consequently, it should not come as a surprise that former deviant cases become typical in an expanded model (Rohlfing 2008). At the same time, the cross-tabulation of typical and deviant cases in our two models shows that ten cases take the reverse route and change from typical to deviant. According to the previous line of reasoning, this should not happen because, if at all, cases should have smaller rather than larger residuals in the expanded model. From a statistical point of view, however, it is apparently possible that typical cases turn deviant in an expanded model. One can hardly attach substantive meaning to this phenomenon because it is not obvious why a richer theory, represented by more variables in the regression model, should do a worse job in predicting the outcome of a case. But if we accept that we should not attach meaning to typical cases becoming deviant, neither should we attach substantive interpretations to deviant cases turning typical. Our procedure is immune to these problems because it ignores the direction in which a case changes its status and avoids making substantive sense of this as robust cases are the much better objects for process tracing.

8. Conclusion

The choice of cases for process tracing on the basis of a regression analysis is perhaps the most widely accepted way of combining small-n and large-n research. We have shown that the present prescriptions on case selection fail to give empirical researchers guidance on how to select cases when being confronted with modeling uncertainty and non-robust classifications of cases across multiple models. If problems of misclassification and misselection are not addressed in empirical research, the generation and generalizability of inferences on causal mechanisms is wholly uncertain.

While misclassification and misselection are salient problems in MMR, we have shown that there is no need to throw out the baby with the bathwater and abandon regression-based case selection altogether. In combination with regression diagnostics and model selection criteria, the classification and selection procedure we propose designates cases as typical and deviant on the basis of multiple regression models. Even if it is not possible to single out the best model, the procedure allows selecting individual cases for in-depth analysis in a systematic and robust manner. Our procedure therefore substantially improves the choice of case(s) in MMR and contributes to causal inference and generalization in designs combining regression analysis and case studies.

References

Achen, C.H. (2002): Toward a New Political Methodology: Microfoundations and Art. *Annual Review of Political Science* 5(1): 423-50.

Adams, J. and Z. Somer-Topcu (2009): Policy Adjustment by Parties in Response to Rival Parties' Policy Shifts: Spatial Theory and the Dynamics of Party Competition in Twenty-Five Post-War Democracies. *British Journal of Political Science* 39(4): 825-46.

Adcock, R. and D. Collier (2001): Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review* 95(3): 529-46.

Agrawal, A. and A. Chhatre (2011): Strengthening Causal Inference through Qualitative Analysis of Regression Residuals: Explaining Forest Governance in the Indian Himalaya. *Environment and Planning A* 43(2): 328-46.

Ahram, A. I. (2011): Concepts and Measurement in Multimethod Research. *Political Research Quarterly*: forthcoming.

Bäck, H. and P. Dumont (2007): Combining Large-*N* and Small-*N* Strategies: The Way Forward in Coalition Research. *West European Politics* 30(3): 467-501.

Bäck, H., H.E. Meier and T. Persson (2009): Party Size and Portfolio Payoffs: The Proportional Allocation of Ministerial Posts in Coalition Governments. *Journal of Legislative Studies* 15(1): 10-34.

Bartels, L. M. (1997): Specification Uncertainty and Model Averaging. *American Journal of Political Science* 41(2): 641-74.

Bennett, A. and B. F. Braumoeller (2005): Where the Model Frequently Meets the Road: Combining Statistical, Formal, and Case Study Methods. *working paper*:

Brady, H. E., D. Collier and J. Seawright (2006): Toward a Pluralistic Vision of Methodology. *Political Analysis* 14(3): 353-68.

Budge, I. (1994): A New Spatial Theory of Party Competition: Uncertainty, Ideology and Policy Equilibria Viewed Comparatively and Temporally. *British Journal of Political Science* 24: 443-67.

Budge, I., H.-D. Klingemann, A. Volkens, J. Bara, E. Tanenbaum, R. C. Fording, D. J. Hearl, H. M. Kim, M. D. Mcdonald and S. M. Mendes (2001), (ed.). *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments, 1945-1998*. Oxford: Oxford University Press.

Burnham, K. P. and D. R. Anderson (2004): Multimodel Inference: Understanding Aic and Bic in Model Selection. *Sociological Methods & Research* 33(2): 261-304.

Bush, S.S. (2011): International Politics and the Spread of Quotas for Women in Legislatures. *International Organization* 65(1): 103-37.

Campbell, D. T. (1975): Degrees of Freedom and the Case Study. *Comparative Political Studies* 8(2): 178-93.

Collier, D. and J. Mahoney (1996): Insights and Pitfalls: Selection Bias in Qualitative Research. *World Politics* 49(1): 56-91.

Coppedge, M. (1999): Thickening Thin Concepts and Theories - Combining Large N and Small in Comparative Politics. *Comparative Politics* 31(4): 465-76.

Cox, N. J. (2004): Speaking Stata: Graphing Model Diagnostics. *The Stata Journal* 4(4): 449-75.

Creswell, J. W. and V. L. Plano Clark (2011). *Designing and Conducting Mixed Methods Research*. Los Angeles: SAGE Publications.

Eckstein, H. (1975). Case Study and Theory in Political Science. In Greenstein, F. I. and N. W. Polsby (ed.), *Book title*. Reading, Mass.: Addison-Wesley (79-137).

Fearon, J. D. and D. Laitin (2008). Integrating Qualitative and Quantitative Methods. In Box-Steffensmeier, J. M., H. Brady and D. Collier (ed.), *Book title*. Oxford: Oxford University Press (756-76).

Fink, S. (2008): Politics as Usual or Bringing Religion Back In? The Influence of Parities, Institutions, Economic Interests, and Religion on Embryo Research Laws. *Comparative Political Studies* 41(12): 1631-56.

Fox, J. (1991). *Regression Diagnostics*. Newbury Park: Sage.

Freedman, D. A. (1991): Statistical Models and Shoe Leather. *Sociological Methodology* 21: 291-313.

George, A. L. and A. Bennett (2005). *Case Studies and Theory Development in the Social Sciences*. Cambridge, Mass.: MIT Press.

Gerring, J. (2007a). *The Case Study Method: Principles and Practices*. Cambridge: Cambridge University Press.

--- (2007b): Is There a (Viable) Crucial-Case Method? *Comparative Political Studies* 40(3): 231-53.

--- (2007c): The Mechanismic Worldview: Thinking inside the Box. *British Journal of Political Science* 38(1): 161-79.

--- (2010): Causal Mechanisms: Yes, But... *Comparative Political Studies* 43(11): 1499-526.

Goerres, A. and K. Prinzen (2012): Using Mixed Methods for the Analysis of Individuals: A Review of Necessary and Sufficient Conditions and an Application to Welfare State Attitudes. *Quality & Quantity* 46(2): 415-50.

Hall, P. A. (2008): Systematic Process Analysis: When and How to Use It. *European Political Science* 7(3): 304-17.

Ho, D. E., K. Imai, G. King and E. A. Stuart (2007): Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* 15(3): 199-236.

Howard, M. M. and P. G. Roessler (2006): Liberalizing Electoral Outcomes in Competitive Authoritarian Regimes. *American Journal of Political Science* 50(2): 365-81.

Iversen, T. (1999). *Contested Economic Institutions: The Politics of Macroeconomics and Wage Bargaining in Advanced Democracies*. New York: Cambridge University Press.

Kennedy, P. (2008). *A Guide to Econometrics, 6th Ed*. Malden: Blackwell.

Kenworthy, L. (2007). Toward Improved Use of Regression in Macro-Comparative Analysis. In Mjøset, L. (ed.), *Book title*. Amsterdam: JAI Press (343-50).

Kim, H. M. and R. C. Fording (1998): Voter Ideology in Western Democracies, 1946–1989. *European Journal of Political Research* 33: 73-97.

Kittel, B. (1999): Sense and Sensitivity in Pooled Analysis of Political Data. *European Journal of Political Research* 35(4): 533-58.

Kittel, B. and H. Winner (2005): How Reliable Is Pooled Analysis in Political Economy? The Globalization-Welfare State Nexus Revisited. *European Journal of Political Research* 44(2): 269-93.

Kuha, J. (2004): Aic and Bic: Comparisons of Assumptions and Performance. *Sociological Methods & Research* 33(2): 188-229.

Lange, M. (2009). *Lineages of Despotism and Development: British Colonialism and State Power*. Chicago: The University of Chicago Press.

Levy, J. S. (2008): Case Studies: Types, Designs, and Logics of Inference. *Conflict Management and Peace Science* 25(1): 1-18.

Lieberman, E. S. (2005): Nested Analysis as a Mixed-Method Strategy for Comparative Research. *American Political Science Review* 99(3): 435-52.

Lijphart, A. (1968). *The Politics of Accommodation: Pluralism and Democracy in the Netherlands*. Berkeley: University of California Press.

--- (1971): Comparative Politics and the Comparative Method. *American Political Science Review* 65(3): 682-93.

Luetgert, B. and T. Dannwolf (2009): Mixing Methods A Nested Analysis of EU Member State Transposition Patterns. *European Union Politics* 10(3):307-34.

Mahoney, J. and G. Goertz (2006): A Tale of Two Cultures: Contrasting Quantitative and Qualitative Research. *Political Analysis* 14(3): 227-49.

Rogowski, R. (1995): The Role of Theory and Anomaly in Social-Scientific Inference. *American Political Science Review* 89(2): 467-70.

Rohlfing, I. (2008): What You See and What You Get: Pitfalls and Principles of Nested Analysis in Comparative Research. *Comparative Political Studies* 41(11): 1492-514.

Rosato, S. (2003): The Flawed Logic of Democratic Peace Theory. *American Political Science Review* 97(4): 585-602.

Seawright, J. and J. Gerring (2008): Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options. *Political Research Quarterly* 61(2): 294-308.

Sekhon, J. S. (2009): Opiates for the Matches: Matching Methods for Causal Inference. *Annual Review of Political Science* 12: 487-508.

Shalev, M. (2007). Limits and Alternatives to Multiple Regression in Comparative Research. In Mjøset, L. (ed.), *Book title*. Amsterdam: JAI Press (269-308).

Shively, W. P. (2006): Case Selection: Insights from *Rethinking Social Inquiry*. *Political Analysis* 14(3): 344-47.

Skaaning, S.-E. (2011): Assessing the Robustness of Crisp-Set and Fuzzy-Set Qca Results. *Sociological Methods & Research* 40(2): 391-408.

Tarrow, S. (1995): Bridging the Quantitative-Qualitative Divide in Political-Science. *American Political Science Review* 89(2): 471-74.

Teorell, Jan. 2010. *Determinants of Democratization: Explaining Regime Change in the World, 1972-2006*. Cambridge: Cambridge University Press.

Wilson, S. E. and D. M. Butler (2007): A Lot More to Do: The Sensitivity of Time-Series Cross-Section Analyses to Simple Alternative Specifications. *Political Analysis* 15(2): 101-23.

Wolf, F. (2010): Enlightened Eclecticism or Hazardous Hodgepodge? Mixed Methods and Triangulation Strategies in Comparative Public Policy Research. *Journal of Mixed Methods Research* 4(2): 144-67.

Wooldridge, J. M. (2003). *Introductory Econometrics: A Modern Approach*. Cincinnati, Ohio: South-Western College Pub.

Wu, A. and B. Zumbo (2008): Understanding and Using Mediators and Moderators. *Social Indicators Research* 87(3): 367-92.

Young, C. (2009): Model Uncertainty in Sociological Research: An Application to Religion and Economic Growth. *American Sociological Review* 74(3): 380-97.

Ziblatt, D. (2009): Shaping Democratic Practice and the Causes of Electoral Fraud: The Case of Nineteenth-Century Germany. *American Political Science Review* 103(1): 1-21.

Tables

Table 1: Review of formalized case selection in empirical research

Reference	Deviant case/ typical case analysis	Number of cases selected	Number of models tested/presented in text	Number of models used for case selection	Classification method	Notes
(Agrawal and Chhatre 2011)	both	10	1	1	5 smallest & 5 largest residuals	-
(Bäck and Dumont 2007)	both	2	3	1 (see notes)	Residuals of cabinets (predicted vs. actual values); no benchmark specified	Only consider countries that are, on average, not extremely well or badly predicted in order to ensure generalizability Information on status with regard to other model was considered for one case, p. 485 – in pathway-case fashion
(Bäck, Meier, and Persson 2009)	typical	2	1	1	Residuals (see notes); no benchmark specified	Bäck et al. select cases (bargains about government coalitions) based on average residuals per country, not on residuals per case (i.e. country is typical, not case)
(Bush 2011)	typical	1	5	1	Residuals; no benchmark specified	

(Fink 2008)	Typical	1 (see notes)	5	Unclear whether all 5 models were used to determine typical/deviant status	Not specified	3 other typical cases which confirm the mechanism are mentioned, but not studied in depth (p. 1640)
(Lange 2009)	both	4	3	1	Residuals; one std. dev. as benchmark	
(Luetgert and Dannwolf 2009)	typical	5	1	1	Deviance residuals; benchmark of 2	No actual case studies. Rather classification of existing case studies; very brief discussion of some typical cases based on secondary literature (pp. 328-9)

Table 2: Estimates for two models on party competition

Variable	All-parties model	Party family model
Average shift of all other parties	0.186*** (0.040)	0.131*** (0.043)
Average shift of parties from same party family	-	0.142*** (0.039)
Past ideological shift	-0.374*** (0.029)	-0.398*** (0.030)
Public opinion shift	0.494*** (0.032)	0.497*** (0.032)
Constant	0.055 (0.376)	0.056 (0.376)
Adj. R^2	0.30	0.31
N	1463	1463
OLS with standard errors clustered by elections Two-tailed tests of significance: *** .01 Standard errors in parentheses		

Table 3: AIC and BIC for two models

Information criterion	All-parties model	Party family model
Akaike IC	12241	12228
Bayesian IC	12228	12254
N=1463		

Table 4: Classification of cases across the two models

		All-parties model	
		Typical case	Deviant case
Party family model	Typical case	1.313	9
	Deviant case	10	131

Table 5: Procedure for robust classification and choice of cases

Step 1	Quantitative analysis	Estimate models according to theoretical and econometric state-of-the-art
Step 2	Model selection	Apply model selection techniques (information criteria, diagnostics, etc.)
Step 3	Cross-model classification of cases	For all remaining models, classify cases as robust typical, robust deviant, and non-robust
Step 4	Case selection	Choose cases from the set of robust typical and robust deviant cases and ignore the non-robust cases for process tracing

Figures

Figure 1: Residual-vs.-fitted plot for two models

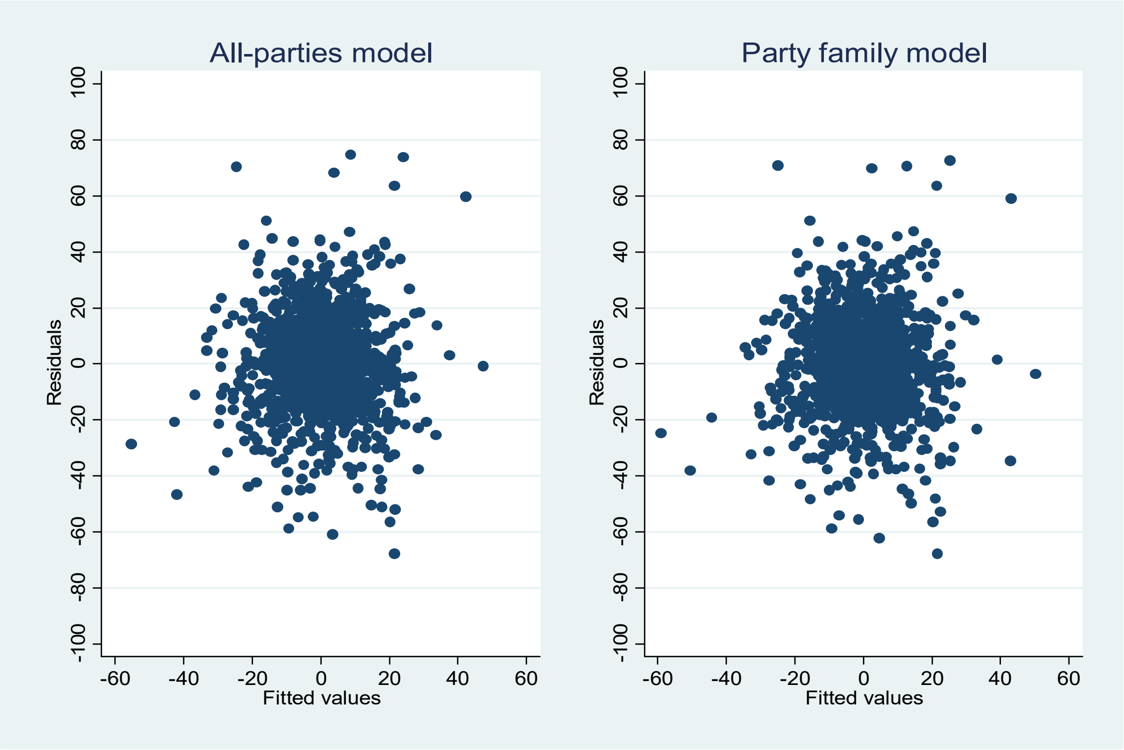
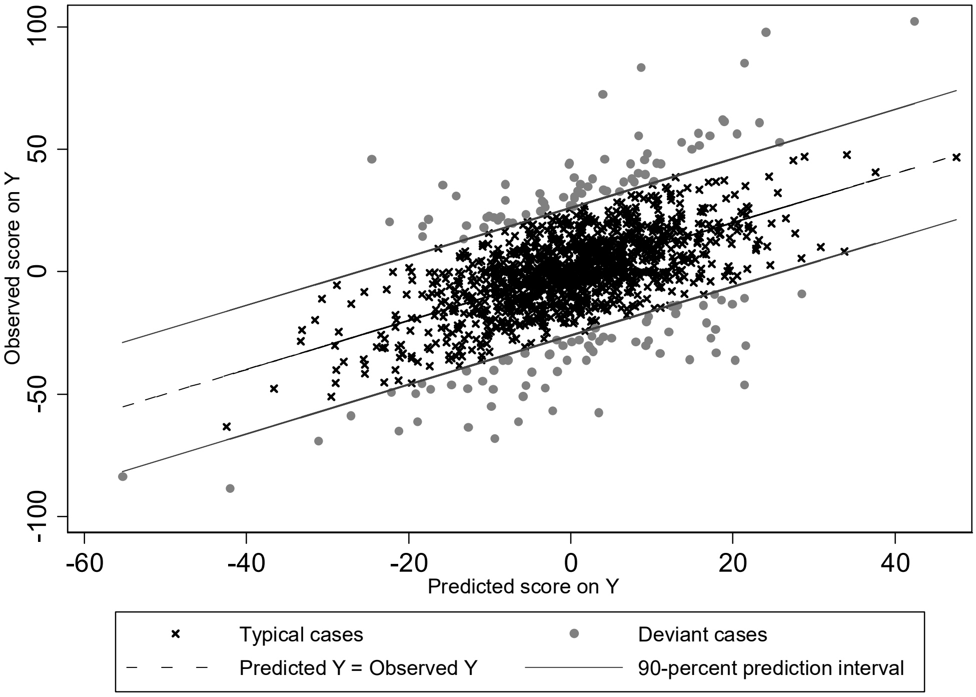


Figure 2: 90-percent prediction interval for all-parties model



Endnotes

[1] The meaning of the term ‘causal mechanism’ is ambiguous (Gerring 2010). For the sake of our argument, we do not become involved in this debate and simply use the term to signify the underpinnings of a causal effect (Gerring 2007c).

[2] A third type that could be mentioned is the ‘pathway case’ (2007b). The pathway case procedure presumes that one has identified the true model, which is an assumption that is usually untenable in practice. Leaving this issue aside, many of the problems that are covered by this paper extend to the pathway case.

[3] Throughout the paper, we refer to the choice of single cases, noting that all our arguments fully extend to comparative case studies.

[4] There are further reasons why some cases can be deviant, including a misspecified concept and measurement error (Adcock and Collier 2001). While one should consider these as potential reasons for a deviant case in practice (Seawright and Gerring 2008), this shall not concern us in our paper as it is specifically concerned with the classification of cases as typical and deviant and modeling uncertainty.

[5] Seawright and Gerring also introduce the influential case, which is discussed in note 26 below. The extreme case is not relevant here because it is defined by being extreme on the independent *or* the dependent variable, without referring to a relationship between the independent and the dependent variables. The pathway case (Gerring 2007b) is not considered further because the choice of a pathway case presumes that we have identified a true model. In contrast, our paper starts from the more plausible assumption that we do not know the true model. Finally, Seawright and Gerring introduce most-dissimilar and most-similar comparisons, but from a purely qualitative view. Since the implementation of these designs on the basis of regression results would be a separate topic, we do not consider it further in the following.

[6] In principle, every typical case is equally suitable for process tracing. All typical cases are qualitatively identical (see below) and the assumption of causal homogeneity thus applies. (This contrasts with Teorell's (2010) claim that cases with high scores on the outcome are better choices than cases with lower scores.) Against this backdrop, Lieberman's strategy only serves to test this assumption by choosing cases creating a huge range of scores on the observed and predicted outcome. Moreover, this strategy refers to what Lieberman calls model-testing process tracing, which should be done when the regression results are deemed satisfactory. He recommends ignoring the covariates and comparing typical and deviant cases via model-building process tracing when the model does not perform well. We focus on model-testing case studies in the following because the large-n results are of questionable value for case selection when the underlying model is weak (Rohlfing 2008).

[7] The assumption that typical cases and deviant cases are useful for the purposes just described should not be taken in deterministic terms. After all, there can be many reasons – including pure chance and measurement error – that a particular case is well-predicted or badly predicted. But even if we follow a probabilistic logic, well-predicted and badly predicted cases are likely to offer relevant evidence (Fearon and Laitin 2008).

[8] These contributions were selected from all existing citations of Lieberman's (2005) seminal article on nested analysis in the Social Science Citation Index (SSCI). We emphasize that it is not our goal to discredit existing empirical studies. The total number of articles was 76 (in articles and conference proceedings). However, the vast majority consisted of general methodological contributions or review articles. Only about a dozen were empirical applications. Of these, a few either did not do a full-fledged MMR study or did not choose cases on the basis of regression models. The remaining five are listed in the table, together with one particularly relevant monograph (Lange 2009) which rigorously applies Lieberman's design.

[9] The problems and remedies that we discuss in the remainder of the paper are independent of the insights that one gathers from the selected cases. Since methodological principles should judged on methodological ground and not on the basis of the results that they might produce, we do not present any process tracing evidence.

[10] The ideological positions of parties are measured with the left-right positions provided by the Comparative Manifesto Project (Budge et al. 2001). The median voter position is estimated with the Kim-Fording formula (1998).

[11] Both models are estimated with OLS and standard errors clustered by elections in order to account for election-specific effects. All arguments we make extend to non-OLS regression estimation (see Bäck and Dumont 2007 for MMR estimating a discrete choice model).

[12] We recall that the example serves neither to make a substantive contribution to the literature on party competition, nor to criticize the analysis by Adams and Somer-Topcu.

[13] Besides that weak theory is a source of modeling uncertainty (see below and Achen 2002), QCA entails multiple elements (e.g. calibration of conditions) that require robustness checks potentially pointing to non-robust set-relational results (Skaaning 2011). Considering the plethora of matching algorithms and difficulties in settling on a single one (Sekhon 2009), it is equally likely to get non-robust results in matching leading to the same problems we discuss in the following.

[14] One should keep in mind that the best-performing models (whatever "best-performing" means) are not necessarily correct in the sense that they are substantiated by causal mechanisms and correctly capture the underlying data generating process.

[15] In light of the left plot in figure 1, one might argue that a deviant case study is futile because there does not seem to be a variable missing from the model. However, the results for the party family model suggest the opposite. Ultimately, the example shows that the plot (and other diagnostic tools) is not necessarily sensitive enough to detect omitted variables.

[16] See Kuha (2004) for a discussion and comparison of the AIC and BIC, which is not of further relevance here.

[17] Rather than the absolute figures, the relative performance of the two models matters here.

[18] While we cannot apply all diagnostic and model selection tools here, the overall picture does not change when applying further instruments.

[19] The extent of modeling uncertainty is larger because Dunning estimates more models (125-126).

[20] Fearon and Laitin (2008, 761-766) hint at the distinction between classification and choice without invoking these terms. Moreover, in other parts of their paper, they speak of "pure" random selection and their empirical example ignores this distinction altogether.

[21] In his analysis of post-colonial development of former British colonies, Lange (2009) uses one standard deviation as the benchmark distinguishing typical and deviant cases. Considering that 1.65 standard deviations represent a 90-percent confidence interval (two-sided test for significance), one standard deviation sets the threshold very low.

[22] The designation of typical cases automatically entails the classification of deviant cases because, given a specific model, any case belongs to one type only.

[23] The upper and lower bound of the prediction interval is obtained by calculating the predicted score, plus/minus the t-score for the critical value (e.g. 1.96 for the 95-percent interval), multiplied by the square root of the sum of the squared standard error of the estimate and the squared standard error of the mean prediction (Wooldridge 2003, 216).

[24] The lower and upper bounds of the prediction interval are concave, i.e., the interval is slightly narrower around the mean than at the extreme bounds.

[25] One might think that instead of choosing a full model, or at least a rather complex one, it should be more appropriate to run a bivariate model including the independent variable of key interest (see Lange 2009 for an application of this strategy). However, we argue against this minimal solution for two reasons. First, the mechanism is of theoretical interest because it underpins the causal effect related to the variable of interest. We therefore need to establish that there is a causal effect rendering it relevant to discern the mechanism. The correct estimation of the causal effect in turn requires a full-blown model instead of a bivariate correlation in order to control for potential confounders and correctly model the data generating process. The second reason is related to theory because the full model reflects the theoretical state of the art in terms of what variables to include, how to specify their marginal effect, and how to estimate the model (see section 7). When we turn to a bivariate model in the case selection stage, we are effectively putting theory aside.

[26] The prediction interval has a concave shape, meaning that the bounds are narrower at the center of the distribution of cases and wider at the margins. This implies that we tend to find more typical cases at the extreme bounds, though this is not necessarily so. Since all cases are causally homogeneous (see above), it does not matter whether we choose a case from the margin or the center of the distribution (pending empirical evidence suggesting that the causal homogeneity assumption is wrong).

[27] This implies that the estimates are not driven by *influential cases* that potentially introduce a bias and render residuals a fallacious criterion for the classification of cases (Kennedy 2008). Seawright and Gerring (2008, 303-304) propose to take the influential case as a third type of case in MMR. We do not follow this argument because it is not apparent what additional insight we gain from process tracing in an influential case. Seawright and Gerring contend that influential cases allow one to check for the soundness of the regression model. However, this is the same

reason for examining (influential and non-influential) deviant cases. On the other hand, why should we bother about an influential case that counts as typical when we control for its influence in the estimation stage (influential cases are not necessarily deviant)?

[28] This holds true even when the overall model fit is relatively low because the larger the standard error, the broader the interval and the more cases are inside the interval.

[29] This means we select cases randomly from the strata of typical and deviant cases, respectively.

[30] The Moledet from Israel in 1996 is the most typical case in the all-parties model, and the SDP from Finland in 1983, the most deviant case. For the party family model, the most typical case is the Portuguese PSP in 1999, while the most deviant case is Maki from Israel in 1969.

[31] We cannot estimate how severe this problem is in MMR studies. The number of non-robustness hinges on a variety of parameters such as the fit of the models and the number of models that is estimated.

[32] For example, Wilson and Butler (2007) discuss multiple ways of handling dynamics in time-series cross-section analysis.

[33] Technically, it is possible that all cases are non-robust. This is unlikely to hold in practice.