1 Introduction

Regression analysis of observational data has always been and, we predict, will remain at the heart of the social sciences methodological toolkit. The major problem with regression analysis of observational data, broadly defined,[1] is that in order to produce unbiased and generalizable estimates, the estimation model must be correctly specified, the estimator must be unbiased given the data at hand, and the estimation sample must be randomly drawn from a well-specified population.

Social scientists know this ideal is unachievable. Empirical models of real world phenomena are hardly ever – we would say: never – correctly specified. Better theory, diagnostic econometric tests, other methodological advice, thoughtful sampling, experience, and even common sense can all help in the art of specifying an estimation model and creating a sample of observations for analysis. However, the world of interest to social scientists, human nature and the interaction of human beings at all levels, is too complex for social scientists ever to achieve the ideal of a correct model specification – a specification that closely matches the true data-generating process. We argue that given the limited information in data typically available to social scientists, social scientists should not even aspire to develop a model that closely matches the true data-generating process. Instead, based on the principle of parsimony, the optimal model specification trades off simplicity against generality, thereby ignoring many complexities. Empirical models cannot, at the same time, simplify and capture the true data-generating process. Rather, for each research question, there will be an optimal simplification of the true data-generating process and social scientists should use the entire theoretical and methodological toolkit to specify their baseline model as well as they can. Yet, there is no guarantee that the optimal baseline model is sufficiently similar to the "true" model to allow valid inferences with great certainty.

1 By regression analysis we mean all kinds of generalized linear and non-linear estimation techniques like logit, probit, Poisson, negative binomial regression, survival analysis, and so on, including semi-parametric techniques.

Robustness testing offers one and perhaps *the* answer to model uncertainty – the uncertainty researchers face as to which model specification provides the optimal trade-off between simplicity and generality. In multiple dimensions and in a quasi-infinite number of ways in each of these dimensions, a model requires choices to be made – specification choices that, even if well justified, could have plausibly been made differently.

Robustness testing allows researchers to explore the stability of their estimates to alternative plausible model specifications. In other words: robustness tests analyze the variation in estimates resulting from model uncertainty. To be sure, model uncertainty is but one source potentially leading to wrong inferences. Other important inferential threats result from sampling variation and from lack of perfect fit between the assumptions an estimator makes and the true data-generating process. In our view, model uncertainty has the highest potential to invalidate inferences, which makes robustness testing the most important way in which empirical researchers can improve the validity of their inferences.

Robustness testing reduces the effect of model uncertainty on inferences. Robustness testing does not miraculously transform uncertain and potentially invalid inferences into inferences that are valid with certainty. Rather, it reveals the true uncertainty of point estimates – the dependence of estimates on model specification. Importantly, robustness testing challenges the established logic of social science methodology: instead of trying to achieve the unachievable – to perfectly fit the model onto the data-generating process – the logic of robustness testing accepts the uncertainty of model specification and asks to what degree estimated estimates and ultimately inferences depend on model specifications.

Analyzing the influence of model specification on estimates is not the only way in which robustness testing can improve the validity of inferences, however. Even when estimates are not robust, researchers can analyze the causes for the lack of robustness. In this way, robustness testing can result in estimation models that have a higher chance of providing valid inferences. All tests can help in the individual and collective process of learning even if, and sometimes particularly if, estimates are found to be non-robust, as this opens up the challenge and opportunity of new research. Research agendas profit from identifying the robustness limits of empirical findings.

But not all is good. Unfortunately, the current practice of robustness testing does not live up to its full potential. Social scientists like to include robustness tests to improve their chances of getting their papers past reviewers and accepted by editors, not because they intend to explore the consequences of uncertainty about their model specification and learn about the robustness limits of their analysis. Practically all reported tests conclude that findings are indeed robust to changes in model specification even if few

authors communicate to their readers what they mean by robustness. Yet, if we do not know what robustness means we cannot know what it means that results are robust.

1.1 CONTRIBUTION

This book contributes to the emerging field of robustness test methodology in three important ways. Firstly, we show that causal complexity of the phenomena that social scientists study imposes severe limits on inferential validity. We explain why all models need to simplify and therefore cannot closely capture the extremely complex true data-generating process. This generates uncertainty as to which model specification provides the optimal simplification and consequently uncertainty about the validity of inferences based on a preferred model or baseline model, as we call it.

As a second contribution, we develop the logic of robustness testing as the key way in which empirical researchers can tackle model uncertainty and thereby improve the validity of their inferences. We offer an operational definition of robustness and a typology of robustness tests. While a majority of social scientists seems to understand robustness in terms of statistical significance, we propose a definition of robustness that draws on effect size stability. As we discuss in chapter 4, our definition has a number of useful properties. It can be flexibly applied not just to frequentist analyses but also to Bayesian techniques. Having said this, all our examples use frequentist estimation methods. Still, robustness testing is all about model specification and not about a particular way of estimation. As we argue in chapter 6, no single methodology permits the formulation of perfectly valid inferences. Every design, procedure or estimation technique warrants subjecting its results to plausible alternative specifications to explore whether these generate sufficiently similar (robust) estimates. Exploring robustness tests for alternatives to regression analysis of observational data is beyond the scope of this book. We leave this important aspect of robustness testing to future research.

As a third contribution, for each dimension of model specification we show what the main uncertainties and therefore inferential threats are. We collect and systematize existing robustness tests that address these uncertainties but we also develop many new tests – or at least tests that we have not seen in the literature before. In this respect, this book seeks to demonstrate that the world of robustness tests is rich and diverse – much richer indeed than the limited number of tests that social scientists have used in the past suggests.

In sum, this book seeks to increase the take-up of robustness tests and improve the practice of robustness testing in the social sciences. It aspires to

overcome the narrow focus of most empirical researchers on model variation tests and open their eyes to the great potential that other types of robustness tests offer. If it fulfils these two objectives, it will significantly improve the validity of regression analyses of observational data.

1.2 OVERVIEW

We divide the book into two main parts. The first part discusses the theoretical and methodological foundations of robustness testing. In chapter 2, we clarify why causal complexity of the social world renders the quest to specify the correct model futile and requires all estimation models to simplify the complex data-generating process. Causal inferences will always remain uncertain and robustness tests explore the impact of model uncertainty on the validity of inferences, which can improve if it can be shown that results are robust independently of certain model specification choices taken.

Chapter 3 proposes a systematic approach to robustness testing in four steps – specify a baseline model that in the eye of the researcher optimally balances simplicity versus generality; identify potentially arbitrary model specification choices; specify robustness test models based on alternative plausible specification choices; and estimate the degree of robustness of the baseline model's estimate with respect to the robustness test model. With multiple dimensions of model uncertainty and multiple specification choices in each dimension, robustness is also multidimensional. We argue that robustness is best explored for each test separately instead of averaged over all robustness test models. We suggest three main goals and aims of robustness testing. Beyond its central focus of exploring the robustness of estimates, these tests allow identifying the limits of robustness and they spur learning and future research, particularly from specification choices that suggest a lack of robustness of the baseline model estimate.

Chapter 4 on the concept of robustness lies at the very heart of the book's first part. Here we define robustness as the degree to which an estimate using a plausible alternative model specification supports the baseline model's estimated effect of interest. We propose a quantifiable measure of robustness that varies from 0 to 1 and defend our continuous concept of robustness against a dichotomous arbitrary distinction into robust versus non-robust. We argue why our definition of robustness as stability in effect size is superior to conceptions of robustness as stability in the direction of an effect and its statistical significance. We introduce partial robustness, which becomes relevant in all non-linear models and even in linear models if analyses depart from linear, unconditional or homogeneous effects. In these cases, a baseline model estimate can be partially robust, that is, can be robust or more robust for some observations but less robust or non-robust for other observations.

Five types of robustness tests are distinguished in chapter 5: model variation, randomized permutation, structured permutation, robustness limit, and placebo tests. We discuss their relative strengths and weaknesses as well as the conditions in which they are appropriately used and refer to examples from leading political science journals in which they have been employed. Importantly, the different types of robustness are best seen as complementary, not substitutes for each other. In fact, the three main aims and goals of robustness testing – exploring the robustness of estimates, identifying the limits of robustness and learning from findings – positively require the use of multiple types of robustness tests.

Chapter 6 argues that there are no alternatives to robustness testing. Model specification tests and model selection algorithms cannot find the one "true" model specification. Model averaging across a huge number of specifications will include many models that are implausibly specified. Other research designs represent alternatives to regression analysis of observational data but, since their results are also based on a large number of specification choices that could have been undertaken differently, they too warrant robustness testing. While this book focuses on tests for regression analysis of observational data, we are confident that many proponents of case selection research designs, "identification techniques," and social science experiments will find the logic of robustness testing appealing and will want to adapt some of the tests we suggest for their own purposes.

The second part of the book analyzes what we regard as the most important dimensions of model specification, identifies the causes of uncertainty for each dimension, and suggests robustness tests for tackling these model uncertainties. Examples illustrate many of these tests with real world data analyses. We start with the population and sample in chapter 7, which, because of the relentless focus on unbiased estimation (internal validity), has received little attention. Scholars are uncertain about the population for which a theory claims validity and uncertain which population the results from the analysis of any particular sample can be generalized to. We include the issue of missing observations as an aspect of sample uncertainty, which threatens both internal and external validity.

Hypothesis testing requires data and data need to be collected. Social scientists refer to the act of collecting data as measurement. Measuring the social world constitutes a more difficult task than measuring the natural world. In the social world, many or perhaps most concepts of interest cannot be directly observed. These unobservable factors need to be captured with proxy variables. Chapter 8 addresses uncertainty about the validity and measurement of social science concepts.

In contrast to both population and sample uncertainty and measurement uncertainty, if one dimension of model specification has attracted

much attention in the extant literature, it is the set of explanatory variables. Chapter 9 argues that including all variables of relevance to the data-generating process and excluding all irrelevant ones is impossible. In the vast majority of analyses, omitted variable bias is inevitable. Standard econometric fixes can do more harm than good. We thus suggest alternative and more flexible ways of dealing with uncertainty about potentially confounding unobservable and unobserved variables.

Linearity is the default functional form assumption and, if combined with robustness tests, not a bad choice given the need to simplify (chapter 10). Similarly, while the social world is marked by causal heterogeneity and context conditionality, the assumption of homogeneous and unconditional effects can be justified as a necessary simplification (chapter 11). Nevertheless, researchers are uncertain about when they need to deviate from these simplifying assumptions and robustness tests can explore if the baseline model's estimates and the inferences derived from them depend on these assumptions. Both dimensions of model uncertainty are closely linked since misspecified functional forms can erroneously suggest causal heterogeneity or context conditionality, and vice versa.

Chapter 12 discusses temporal heterogeneity, defined as variation in the effect strength of a variable over time. Temporal heterogeneity can be caused by structural change in the form of trends, shocks or structural breaks. Parameter homogeneity across time, the standard operating assumption of the vast majority of cross-sectional time-series analysis, seems a strong assumption to make in datasets covering several decades. Such samples cover a long enough period of time for disruptive events to have taken place or simply for actors to change how they respond to stimuli. Robustness tests set one or more of the estimated parameters free for all or a subset of cases, allowing the parameters to vary over time.

We turn to a problem related to temporal heterogeneity in chapter 13: dynamics. Researchers typically reduce dynamics to employing techniques that eliminate the serial correlation of errors and, almost haphazardly, impose simple and rigid dynamics on the effects of variables. However, the true data-generating process most likely contains more complex effect dynamics. If researchers strive to capture these dynamics, they need to model the onset and duration of effects and the functional form of effects over time and consider the possibility of dynamic heterogeneity across cases. Robustness tests either relax the constraints that the baseline model specification imposes on the dynamics of effects or model the dynamics differently from the baseline model.

Chapter 14 deals with a dimension of model specification that should in principle stand at the core of social science research: actors do not act

independently of each other. After all, social interaction and interdependence are constitutive elements of life. Actors not only learn from and exert pressure on each other, their actions (and non-actions) also generate externalities on others. As a consequence, we find it difficult to imagine a data-generating process that does not incorporate spatial dependence in one form or another. Even so, the vast majority of social science research treats spatial dependence as a nuisance to be ignored. Robustness tests for these baseline models give up the assumption of independence and model dependence in either the independent variables or the error term, typically assuming that geographically more proximate units exert a stronger spatial stimulus. Analyses that explicitly test theories of spatial dependence have recently surged, however. Robustness tests have to deal with the fact that true spatial dependence is difficult to identify since many causes are spatially correlated or units experience spatially correlated trends and shocks. Equally importantly, they have to explore the robustness of estimates toward modelling the spatial-effect variable differently.

Chapter 15 concludes with our thoughts on what needs to change for robustness testing to fulfil its great promise. We believe that robustness tests are too important to be left exclusively to authors. Instead, we advocate that reviewers and editors also take responsibility and identify relevant robustness tests and ask the authors to undertake them when they review and decide on manuscripts. Taken seriously, robustness testing requires significant additional investments in time and effort on the part of authors, reviewers, and editors but we know of no better way for improving the validity of causal inferences based on regression analysis of observational data.