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.

PART 1

Robustness – A Conceptual Framework

2 Causal Complexity and the Limits to Inferential Validity

Doing statistics is like doing crosswords except that one cannot know for sure whether one has found the solution.
John W. Tukey according to David Brillinger (2002: 1547)

Once the model is known, the inferential puzzles that remain are trivial in comparison with the puzzles that arise in the specification of a model.
Edward Leamer (1978: v)

2.1 INTRODUCTION

Causal inference is much more than the mere identification of a cause–effect relationship which dominates the current debate in the social sciences. Social scientists do not only wish to establish the existence of a causal effect, they also intend to estimate the size of the effect, they wish to find out whether the estimated effect represents all cases included in the sample, they aim at understanding the mechanisms by which causes take effect and they attempt to identify the population to which the estimated causal effect can be generalized.

Causal complexity of the social world renders the traditional science-derived concept of causality unsuitable for social science research. The traditional concept of causality assumes deterministic relations and homogeneity – assumptions that are problematic for the social sciences. Given causal complexity, all models – theoretical *and* empirical models – of social outcomes necessarily simplify and no empirical model can ever capture the true data-generating process.

The way forward, we suggest, begins with moving away from the concept of model misspecification toward model uncertainty. Rather than trying to specify models correctly (an impossible task given causal complexity), researchers should test whether the results obtained by their baseline model, which is their best attempt of optimizing the specification of their empirical model, hold when they systematically replace the baseline model

specification with plausible alternatives. This is the practice of robustness testing. By providing additional evidence from plausible alternative models, robustness tests potentially increase the validity of inferences compared to inferences based only on the baseline model.

In this chapter, we put forward our understanding of causal inference in social science research. We show that causal complexity of the social world renders the traditional science-derived concept of causality ill-suited and the quest to find the correct model specification futile. No model can be known to be correctly specified and we introduce robustness testing as the logical answer to the ensuing model uncertainty. Lastly, we make the case that robustness testing can improve the validity of causal inferences despite the fact that each single estimation model is likely to be misspecified.

2.2 SOCIAL SCIENCE RESEARCH AND CAUSAL INFERENCE

For Heckman (2005: 1), "causality is a very intuitive notion that is difficult to make precise without lapsing into tautology." He argues that two concepts are central for a scientific definition of causality: a set of possible outcomes and manipulation of one (or more) of the determinants. However, causality may exist where manipulation of causes is impossible and it may exist without change. For example, a perfectly stable equilibrium that resists change will have causes. A black hole does not emit light and will never do, but this state is *caused* by its gravitational force. In other words, causality exists beyond the realm of causes that can be manipulated.

In the currently dominant paradigm, social scientists refer to the idea of counterfactuals to express causality (Rubin 1974; Holland 1986; Pearl 2000; Morgan and Winship 2015). In this perspective, causality could be observed if at the same time a research design allowed to treat and not treat a single case and observe the consequences thereof. In our view, causal inference reaches well beyond the identification of the existence of causality, however. Causal inference consists of five distinct elements:

1. the identification of a causal relation between two variables (cause and effect);
2. the estimation or computation of the strength of the effect;
3. the identification and understanding of the causal mechanism;
4. the generalization of the estimated effect to all cases included in the sample;
5. the generalization from the observed cases to the set of cases defined as the population.

Importantly, causal inference not merely encompasses the identification of a causal effect and an estimate of its strength, but also the generalization

of causal findings to all cases included in the sample (internal validity)[1] and the larger set of cases defined as the population (external validity). It also includes the identification and understanding of a causal mechanism since the causality of an "identified" association is not understood unless the mechanism is understood.

For some methodologists, the provision of sufficient plausibility that an observed co-variation between two variables is actually causal is more important than the provision of evidence that the identified causal effect exists in an identifiable class of cases beyond the ones studied (Rosenbaum 2010: 56). This implied preference for internal over external validity is potentially dangerous for the social sciences (for a similar view, see Cronbach 1982): if theories developed by social scientists ought to be useful for guiding political, social, cultural, or economic decision-making, then stakeholders of social science research do not merely need to know that a causal effect has been established for a particular sample of cases; they also need to know under what circumstances the treatment effect – the effect of a cause – can be utilized for other cases.[2]

The identification of a causal effect and an unbiased estimate of its strength differs from understanding the causal mechanism – the chain of events that ultimately brings about the effect. Consider the causal effect of Aspirin on headache. The headache does not disappear because a patient swallows an Aspirin pill. It disappears because the pill has an ingredient, salicylic acid, stopping the transmission of the pain signal to the brain. Consequently, Aspirin does not necessarily eliminate the origin of the pain

1 Our definition of "internal validity" differs from the currently dominating perspective, which focuses on the local average treatment effect. Assume a randomized controlled trial and assume there is causal heterogeneity across two groups of individuals. For treated individuals from the first group the treatment reduces mortality by 50 percent, whereas for the second group the treatment reduces mortality by 10 percent. If the treatment groups include an equal number of participants from each group, the local average treatment effect is 30 percent. According to the common definition of internal validity, this result is internally valid. For our definition of internal validity, the local average treatment effect is not internally valid, since it does not represent the true treatment effect in either group.

2 For others (Altmann 1974), internal validity serves as a logical prerequisite for external validity: "[T]o the extent that we have not eliminated alternative explanations for the results within our sample, we cannot rule them out of any generalization or interpretation derived from the sample" (Altmann 1974: 230). In our view, this conflates internal with external validity. External validity requires that the sample represents the population in all relevant characteristics, which is different from and independent of Altmann's concern whether what is generalized is internally valid.

but prevents the brain from noticing the pain. The molecules of salicylic acid attach themselves to COX-2 enzymes, which blocks these enzymes from creating those chemical reactions that will eventually be perceived as "pain." Clearly, identifying causation – the pain disappears after taking a pill – is distinct from understanding causal mechanisms.

Undoubtedly, causal mechanisms are an almost infinite regress (King et al. 1994: 86) slowly approaching the quest for the "first cause": knowing that Aspirin reduces the reception of pain is one thing, understanding why the treatment works is another. Accordingly, finding a causal mechanism gives rise to questioning the mechanism behind the mechanism. In our example, we may now ask how and why COX-2 enzymes produce the notion of pain and we may perhaps find another way of reducing the notion of pain.

The infinite regress of causal mechanisms ends with "acts of god" or the "big bang." Approaching these metaphysical questions runs into limits of knowledge or leads to metaphysical answers. This infinite regress may be one reason why many quantitatively oriented social scientists focus on the identification of a causal effect rather than on causal mechanisms. However, we need to identify and understand causal mechanisms for the purpose of policy and decision-making (Deaton 2010). The simple proof that medicines made from willow reduce pain did not suffice to develop Aspirin.

With the central focus on causal mechanisms, theory has a dominant role to play in making causal inferences. Theory provides the formulation and justification of a mechanism that links effects to causes. Claims of causal mechanisms remain shallow if no sound and logical theoretical basis for them is offered. Theory should guide every step of the research design for making causal inferences. And yet, with theory alone comes no knowledge, and empirical research is therefore needed to undertake causal inferences.

2.3 CAUSAL COMPLEXITY

The causal complexity that characterizes the social world hampers causal inference. In table 2.1, we compare traditional science-derived concepts of causality to the logic of causality in the social world. We stress that the traditional concept of causality used in the social sciences comes from traditional physics. It does not necessarily apply to other sciences and is increasingly questioned even by physicists as the rise of quantum physics testifies. We employ it as a backdrop to illustrate how ill it fits to the social world despite many modern social science methodologists putting their faith in inferential techniques that would be well suited only to the traditional science-derived concept of causality.

The first dimension of causal complexity is that practically all cause–effect relationships in the social world are probabilistic instead of

Table 2.1: Concepts of Causality and the Social World

	Traditional concept of causality	Data-generating process in social sciences
Causal effect	deterministic	probabilistic
Strength of causal effect	homogeneous and unconditional	heterogeneous and conditional
Dynamics of causality	determined by causal mechanism	influenced by agents' autonomous decisions
Sequence of causality	cause precedes effect	distorted by rational expectations
Effect on non-treated units	none (homogeneous)	possible, due to effect on expectations (placebo, nocebo) and to spill-overs

deterministic and that causes only contribute to effects instead of being sufficient. The probability of an effect is a continuum from 0 (a cause does not have an effect) to 1 (the cause is deterministic).[3]

The second dimension of causal complexity is the existence of conditional causal effects and heterogeneous causal effects. Some causes only have effects if certain conditions are satisfied (Franzese 2003). Unless these conditions are given, the causal factor has no effect. More common are conditional causal effects, in which other factors condition the strength of the effect of x on y. Causal heterogeneity exists when agents respond differently to the same treatment. Causal heterogeneity must be distinguished from stochastic error. That the mean sample effect size differs from the predicted effect for each case in the sample can be because of stochastic error in the estimation model and does not, as such, generate a problem for causal inference.

The third dimension of causal complexity is the timing of cause and effect. Scholars all too often implicitly assume that an effect occurs immediately after a cause and, given temporal aggregation, contemporaneously with the cause. Yet, effects can occur with a delayed onset, the duration of the effect can be long or short, and the temporal functional form of the causal effect can be complex if effect strengths evolve over time, all of which renders causal inference more difficult. Response delays can also be systematic and caused by actors' strategy (Fernandez and Rodrik 1991; Rodrik 1996; Alesina and Drazen 1991; Alesina et al. 2006) or by institutions (Tsebelis 1995, 1999). Causal mechanisms have heterogeneous, or even

3 The likelihood of an effect given a cause needs to be distinguished from the likelihood of a cause and from the strength of an effect: a supernova explosion is an unlikely event, but if it occurs it will destroy all planets within a given range with certainty. In other words, supernova explosions, though unlikely, have a deterministic and strong effect.

unknown dynamics. For example, the mechanisms that link investment to growth are likely to have different dynamics depending on the sector in which the investment takes place, the organization of the firm that makes the investment, and so on.

The fourth dimension of causal complexity is that in the living world effects can precede causes. Human beings have rational expectations about potential future treatments and may already act on their expectations rather than on the treatment itself. Consequently, if actors expect an exchange rate intervention, they may adjust their behavior before the intervention is announced or takes place and an exchange rate effect may occur without any actual intervention. Investors do not have to wait until inflation rises to shift investments from bonds into stocks, they can shift their portfolio composition based on their expectations of a rise in the inflation rate. Political leaders do not have to wait until another country attacks their country – they can launch preemptive strikes. Expectations blur the "causes precede effects" law that scientists believe in. The *cause-precedes-effect* assumption could be rescued only if social scientists traced back all behavioral changes to alterations of agents' expectations – an avenue of causation few are willing to take.

The fifth and final dimension of causal complexity is that treatments can affect non-treated cases. Because human actors act on expectations and even their internal biological processes react to expectations, positive or negative treatment effects are possible even if individuals have not actually been treated (placebo and nocebo effects). In the social world spill-over effects from the treated to the untreated are likely. Known as spatial dependence among cases, spillover effects can render it challenging to identify a causal effect of treatment. More generally, failure to adequately account for spatial dependence can result in biased estimates of causal effects. Conversely, confounding factors hamper the identification of spatial dependence effects if these confounders generate spatial patterns in the data – an inference problem known as Galton's (1889) problem.

2.4 FROM MODEL MISSPECIFICATION TO MODEL UNCERTAINTY AND ROBUSTNESS TESTING

To capture the true data-generating processes of a complex world, analysts would need to precisely know the set of regressors, include all relevant variables and exclude all irrelevant variables, operationalize and measure these variables correctly, precisely and without systematic measurement error, model the functional form of each variable correctly, get all conditionalities right, correctly account for temporal heterogeneity, dynamics, and spatial dependence among units, and so on. There is no way of knowing.

Theories cannot provide this knowledge, for two reasons. On the one hand, all empirical evidence is theoretically under-determined (Duhem 1954; Quine 1951), that is, empirical evidence is consistent with more than a single theoretical explanation. And on the other hand, theories have to simplify in order to identify causal mechanisms that explain some observable variation. Social science theories do not aim at explaining the true data-generating process (Freedman 1991). Rather, with few exceptions they focus on identifying and clarifying a single or at most a few causal mechanisms. Accordingly, theories in general and social science theories in particular tend to be underspecified.

It is a short logical step from underspecification to misspecification. Social scientists know at least intuitively that estimation models do not even try to model the data-generating process of reality and instead have to simplify – just like theories do. And just as with theories, so with empirical models: simplification is desirable. The only alternative to simplify reality is to copy it – in which case social scientists would not get any closer to an understanding of reality, but merely copy the lack of understanding of the original world over to the new world. Simplification is not a necessary evil, but an inevitable and important component of the process of understanding. Simplification requires knowledge and the skill to understand and to detect the important patterns of causal mechanisms in the complexity of reality.

At the same time, analyses based on misspecified models – models that simplify causal complexities – cause biased estimates[4] and may result in invalid inferences. The trick is to know how much an empirical model can simplify without losing the ability to recover the effect of interest in a sufficiently valid way. Researchers must find the optimal trade-off between simplicity and generality. To illustrate what we mean look at maps as a model of the real world.[5] All maps are wrong; they simplify and transfer three-dimensional spaces and contours onto a two-dimensional plane. Maps can also provide wrong information, though. For example, early maps like the Fra Mauro map of the fifteenth century represent the world as round, with oceans at the edge of the world. Mauro even failed to represent well the world known at his time. The British Isles exclude Scotland, Scandinavia is much larger than in reality, the Americas are – for obvious reasons – entirely missing, and so are the Bay of Bengal, Australia, New Zealand, and Japan. Accordingly, there exists an important

4 Under strong assumptions, it is possible to generate a data-generating process in which model simplification does not cause bias. For the analysis of observational data this possibility may also exist with a small probability, but one cannot be certain.

5 For a similar use of maps as analogy for theory, see Clarke and Primo (2012).

difference between simplification and misrepresentation – both of which open a gap between reality and model. Simplification is inevitable and desirable in the social sciences – misrepresentation is not. Scientific progress reduces misrepresentations, but not simplification.

An alternative to the impossible task of specifying a model that exactly matches the data-generating process is to accept model uncertainty. While the concept of model misspecification calls for all specification problems to be solved, the concept of model uncertainty and robustness testing takes the lack of knowledge about the correct model specification seriously. Model uncertainty is the uncertainty over which of all the necessarily misspecified models provides the best trade-off between simplicity on the one hand and generality on the other hand. Robustness testing explores the impact of taking alternative plausible model specification choices which, given model uncertainty, could have provided the best trade-off between simplicity and generality.

2.5 ROBUSTNESS TESTING AND CAUSAL INFERENCE

The probability that social scientists manage to specify an empirical model exactly right is fairly close to zero. The probability that social scientists derive valid inferences based on regression analysis of observational data despite estimating misspecified models is much larger. Yet, model misspecification and causal inference appears to be a contradiction in terms. At least, many social scientists seem to believe that unless an estimate is unbiased with certainty, causal inferences are invalid.

How can social scientists make valid causal inferences from misspecified estimation models and estimated effects that are likely to be biased? One of the first methodologists who argued that causal inference with observational data requires strong assumptions was Hubert Blalock (1964: 176):

> We shall assume that error terms are uncorrelated with each other and any of the independent variables in a given equation. (...) In nonexperimental studies (...) this kind of assumption is likely to be unrealistic. This means that disturbing influences must be explicitly brought into the model. But at some point one must stop and make the simplifying assumption that variables left out do not produce confounding influences.

Blalock correctly states that full validity of causal inferences depends on the absence of excluded confounders and he seems willing to assume that no such confounder exists. This is an assumption that fewer and fewer contemporary social scientists share.

Yet, causal inferences can be valid despite model misspecification. To start with, a trivial strategy to improve the validity of causal inferences

relies on formulating the hypothesis tested in a way that renders inferences trivially valid. The more "weakly" or "softly" causal inferences are formulated the more likely it is that they will be valid. Take the inference that higher education on average exerts a positive effect on income and social status (Griliches and Mason 1972). This effect can already be inferred from the fact that individuals voluntarily attend higher education – a behavior that would not occur as a mass phenomenon and as an evolutionary stable strategy if higher education had no positive effect for the average student.

More importantly, even non-trivial inferences can be rendered more valid despite model misspecification. The production of scientific knowledge is a social process. Social scientists do not make causal inferences based on any single study alone. Authors may do so but the community of scholars does not. Social scientists can learn about causality and make causal inferences from "misspecified models" because they

- do not use empirical analysis alone for making inferences but also rely on theory,
- do not rely on a single prediction (or hypothesis) when testing a theory, but instead derive and test multiple predictions of that theory,
- do not derive causal inferences from a single estimate, but rather validate findings by other estimates from other studies and interpret inferences by multiple analyses of various aspects of a theory using different methods, research designs, and samples.

Causal inference as a social or collaborative scientific process does not rely on empirics alone or on a single analysis. It relies on sophisticated theories that make multiple predictions and on a multitude of relevant analyses generating similar or consistent findings. In other words: causal inference is an outcome of academic debate and scientific progress that evolves over time.

Here robustness testing comes in. It provides multiple analyses already in a single study, namely estimates based on multiple plausible model specifications. By exploring the robustness of the baseline model's estimated effects it provides additional evidence. The uncertainty about the baseline model's estimated effect size shrinks if the robustness test model finds the same or similar point estimate with smaller standard errors, though with multiple robustness tests the uncertainty likely increases. Either way, robustness tests can increase the validity of inferences. This follows from the fact that robustness tests provide information on the influence of alternative model specifications on results. This information is likely to reduce the certainty of the baseline model's specific point estimate, but not necessarily the validity of other inferences based on the estimates. For example, robustness tests may easily increase the range of an estimated effect size – thereby

reducing the validity about the specific point estimate. At the same time, the robustness tests may show a remarkable insensitivity to changes in the model specification within this range of estimated effect size – thereby increasing the validity of the inference that the effect estimated by the baseline model exists, has the estimated direction, and lies within the range suggested by the baseline and robustness test model estimates.

Not all types of robustness tests estimate effects based on plausible changes in model specification. Robustness limit tests ask which change in a specific model specification choice would render an estimate non-robust. Here, robustness tests improve the validity of inferences by testing whether a particular potential model misspecification can plausibly be so large as to invalidate the inference. Consider the effect of smoking on lung cancer. In responding to the evidence provided by analyses of observational data which demonstrated a significant effect of smoking on lung cancer (Hammond 1964), the tobacco industry argued that the models used to establish this cause–effect relationship were misspecified because an unknown gene could cause both a higher propensity for smoking and a higher propensity for lung cancer (Fisher 1958). In modern language: the effect of smoking on lung cancer is not identified. Rosenbaum (2010: 111ff.) nevertheless manages to demonstrate that smoking very likely causes cancer. He shows that the effect of smoking on lung cancer is robust to assuming the existence of a "smoking gene." Correcting for the plausible effect size of such a gene does not eliminate the effect of smoking.

This research shows that model misspecifications do not need to be eliminated to make valid inferences and that potentially misspecified models employed in the analysis of observational data can be used to derive causal inferences that are not valid with certainty but certain enough to be almost consensually believed to be valid. In other words: causal inference based on regression analysis of observational data is possible despite the potential for model misspecification if researchers analyze the relevance of model misspecification for estimated effects and causal inferences.

2.6 CONCLUSION

This chapter has made two important arguments, which both run counter to contemporary wisdom in the social sciences: first, we have argued that, due to causal complexity in the social world, not only is it impossible to specify an estimation model that perfectly captures the data-generating process;[6] it

6 Causal complexity is not the only reason though. There is the additional problem that variables are typically based on theoretical constructs that are not directly observable in reality. This results in measurement uncertainty and measurement

is not even desirable to try to do so since all estimation models must and should simplify. As a consequence, all estimation models obtain biased estimates. Second, valid causal inferences can and in fact have been made based on biased estimates from misspecified estimation models.

Our second argument draws on the notion of science as a social or collective process: while social scientists do not learn much from a single biased estimate, they may under certain circumstances learn a lot from different analyses of the same causal effect. Robustness tests mimic the social process that leads to better and potentially more valid causal inferences. Robustness tests are an important strategy to learn from potentially biased estimates. They should form an integral part of research designs in the social sciences.

Estimated effects and inferences from a single model should never be perceived as either valid or not valid. Rather, their validity should be perceived as uncertain and the individual and collective research strategy should focus on trying to increase their validity. It is possible for social science as a social enterprise to collect enough evidence over time to call inferences about a causal effect valid with near "certainty." Candidate examples include the inferences of a positive effect of schooling on income, a positive effect of proportional electoral systems on the number of parties in parliament, and the effect of living in cities on the experience of stress. If these effects – all understood as average effects, not as holding for all cases let alone holding for all cases equally – are not known with certainty, they are known with close to certainty.

A rather different question is whether social scientists know or even can know the strengths of these effects with certainty. The idea that social scientists can identify the one "true" effect size of schooling on income, for example, reveals a severe lack of understanding of social systems. The one true parameter does not exist and the search for its identification is necessarily futile. These effects are conditioned by many factors, including the wealth of the country, its political system, the technologies companies use, individuals' characteristics including intelligence and network externalities, and so on. Causal effects vary across space and time, potentially due to conditioning factors, but potentially also for endogenous reasons. Thus, there is not *one* effect size of schooling on income, nor even *one* average effect size, but many.

The baseline model ought to represent the researchers' best attempt at trading off simplicity versus generality to account for causal

error. As chapter 8 explains in detail, measuring the natural world is very different from measuring the social world.

complexity. The resulting model is almost necessarily misspecified and estimates based on the chosen model cannot fully generate either internal or external validity. Robustness testing does not necessarily prompt the development of better estimation models – models that result in higher internal or external validity. But robustness testing can increase internal and external validity by generating further evidence, namely that other plausible model specifications suggest varying degrees of robustness. The next chapter develops further this logic of robustness testing.