3 The Logic of Robustness Testing

The analysis of specification error relates to a rhetorical strategy in which we suggest a model as the "true" one for the sake of argument, determine how our working model differs from it and what the consequences of the differences are, and thereby get some sense of how important the mistakes we will inevitably make may be. Sometimes it is possible to secure genuine comfort by this route.
Duncan (1975: 101–102)

3.1 INTRODUCTION

We are not the first to argue that empirical models are misspecified. As George Box states, "all models are wrong, but some are useful" (Box 1976; Box and Draper 1987). Similar claims have been made over and over again. Martin Feldstein (1982: 829), former president of the National Bureau of Economic Research and former Chairman of the Council of Economic Advisers, warned that "in practice all econometric specifications are necessarily false models." Political scientist Luke Keele (2008: 1) states: "Statistical models are always simplifications, and even the most complicated model will be a pale imitation of reality." According to Peter Kennedy (2008: 71), author of one of the best-known introductory econometrics textbooks, "it is now generally acknowledged that econometric models are false and there is no hope, or pretense, that through them truth will be found." These authors do not argue that empirical models can be misspecified. Instead, they articulate a widespread consensus that all models are *necessarily* misspecified; they cannot and do not match the true data-generating process.

If all models are necessarily misspecified, authors and readers alike cannot trust any single estimation model to provide a valid estimate of the effect of a variable x on outcome y. This casts doubt on inferences derived from the estimate. Nearly all scholars are aware of the limits of model specification: if they did believe that their model specification was correct

(and that peers and reviewers shared this belief), they would present the results of a single estimation model. But usually they don't.

This chapter discusses the logic of robustness testing. We start by acknowledging intellectual heritage: Leamer's sensitivity analyses, Rosenbaum's bounds, Manski's non-parametric bounds, Frank's robustness limit tests, and others. Yet, despite this heritage rooted in econometric theory, robustness testing is a grassroots movement with no identifiable inventor. Consequently, no common standards and practices toward robustness testing have been developed, which results in deficient current practices and robustness testing failing to achieve its full potential. We propose a more systematic approach to robustness testing that proceeds in four steps: baseline model specification, identification of potentially arbitrary modelling assumptions, robustness test model specification based on alternative plausible assumptions, and comparison of estimated effects and the computation of the degree of robustness. We also discuss the multidimensionality of robustness and argue that robustness is best explored for each test separately, rather than averaged over all robustness test models. Lastly, we describe what we regard as the main aims and goals of robustness testing.

3.2 ROBUSTNESS TESTING IN THE SOCIAL SCIENCES

Edward Leamer was the first to systematically justify robustness testing as a means to tackle model uncertainty without the unrealistic aim of eliminating it. In Leamer (1978: v), he justifies his departure from what was then contemporary methodology: "Traditional statistical theory assumes that the statistical model is given. By definition, nonexperimental inference cannot make this assumption, and the usefulness of traditional theory is rendered doubtful." To deal with uncertainty about model specification, Leamer developed what he called *sensitivity tests*. Leamer understood sensitivity testing broadly: "One thing that is clear is that the dimension of the parameter space should be very large by traditional standards. Large numbers of variables should be included, as should different functional forms, different distributions, different serial correlation assumptions, different measurement error processes, etcetera, etcetera" (Leamer 1985: 311). Despite his ambitions, those who have taken their inspiration from Leamer have almost exclusively focused on analyzing permutations to the set of regressors.[1]

1 Much of this early literature (Levine and Renelt 1992; Feld and Savioz 1997; Temple 1998; Sala-i-Martin 1997) was motivated by uncertainty with respect to the correct set of explanatory variables in economic growth models. Sensitivity tests soon reached other social sciences, but in political science (Neumayer 2002;

Some intellectual heritage of robustness testing derives from scholars like Paul R. Rosenbaum and co-authors (see Rosenbaum 2002, though there are many earlier contributions with many co-authors), Charles F. Manski (1990, 1995), Ken Frank and co-authors (Frank 2000; Pan and Frank 2003; Frank and Min 2007) and others who have developed what in chapter 5 we will call robustness limit tests, which represent one of five types of robustness tests. These tests explore how much a specific model specification needs to be changed for a baseline model's estimate to become non-robust.

Despite its intellectual heritage, contemporary robustness testing has arisen as an independent grassroots movement. Robustness tests have always been around. The first publication that presented two regression estimates with the same dependent variable implicitly conducted a robustness test of some kind. The oldest robustness test must be to add an additional regressor to the existing list of explanatory variables. This constitutes a robustness test even though it took decades until somebody used the label for such a simple change in model specification. Unfortunately, we have not been able to identify the first-ever use of the term "robustness test" with the specific meaning social scientists attach to it now. Over time, robustness testing became a *best practice of empirical research*, with authors integrating robustness tests into their manuscripts as a strategy to deal with anticipated model specification issues raised by reviewers.

The number of articles reporting robustness tests increases exponentially in the social sciences, though the growth rate appears to be higher in some disciplines, most notably in economics and political science, than in others, e.g., sociology and business studies. Despite this uneven take-up, robustness tests today form an important element of the scientist's toolbox. Between 2008 and 2013 alone, the number of articles indexed in the Social Sciences Citation Index that explicitly reported robustness tests doubled.[2] Today, robustness tests are used across all the social sciences.

While the increase in the number of articles reporting robustness tests over the last decade is impressive, roughly half of the articles published in leading political science journals over a ten-year period that we surveyed do not present the estimation results of the robustness tests in the body of the

Scheve and Slaughter 2004; Gerber and Huber 2010) and sociology (Frank 2000) the meaning of these tests broadened, varying many aspects of the model specification to test the robustness of results.

2 Many more articles presumably report robustness tests without being explicit about it. If we cross-check this trend using Google Scholar, we find that the number of articles mentioning robustness tests increased from 1,280 in 2008 to 2,600 in 2013.

article, the appendix, or the online appendix.[3] Robustness is more often than not a mere claim, and it hardly ever becomes an explicit part of the research strategy.

To make matters worse, the vast majority of these articles do not justify their choice of robustness tests. Authors could, therefore, potentially have reported tests selected not because they really test the robustness of the estimates in the presence of model uncertainty, but simply because their baseline model proves to be robust to the carefully selected tests. Unless scholars provide a good justification for their chosen set of robustness tests or robustness tests are chosen not by the authors but instead by journal editors or reviewers, the value of robustness tests remains questionable.

According to our review of political science journals, the most widely reported robustness tests are the inclusion of additional control variables, alternative measures of the dependent or central explanatory variables, changes in the sample, and alternative measurement scales or functional forms. These tests are conducted in 20–30 percent of articles that report robustness tests published in leading political science journals we surveyed. Alternative estimators, alternative functional forms, and alternative dynamics are used in about 10 percent of those articles. All other robustness tests are even less frequent. They occur occasionally, but scholars do not use these tests systematically. Nevertheless, some researchers conduct tests that account for structural breaks, alternative lag structures, conditionality, spatial dependence, missing observations, crucial cases (jackknife), or endogeneity (instruments). Yet, a glaring gap remains to be bridged between the number of model uncertainties potentially relevant to a baseline model and the frequency with which they are explored in robustness tests.

While the arbitrary selection of robustness tests might be the most evident and important problem in the current practice of robustness testing, it is not the only one. In addition, few scholars justify the robustness tests they conduct. Not justifying robustness tests is as bad as not justifying the baseline model. A lack of robustness between a plausibly specified empirical model and an implausibly specified model is irrelevant. More importantly, a robustness test model that only makes a minuscule change to the baseline model specification or that has been carefully selected because it supports the baseline model estimate rather than because it represents a real test does not add much to the validity of inferences, if at all.

3 We identified more than 500 articles published in selected political science journals in which authors reported at least a single robustness test. Overall we found that explicit robustness tests are employed in approximately one out of four empirical papers published in these journals.

Exceptions of good practice exist. Consider Scheve and Slaughter's (2004) analysis of the influence of foreign direct investment on what they call "economic insecurity" defined as volatility in the demand for labor that causes volatility in wages and employment. They analyze an individual's perception of job security. In their baseline model, Scheve and Slaughter regress this perception of economic volatility on a dummy variable capturing the presence of foreign companies, an individual's education, age, income, union membership, manufacturing employment, and sector unemployment rate plus year dummies. Scheve and Slaughter report five well-justified robustness tests: first, they replace the FDI dummy with "alternative measures of FDI exposure" (2004: 670), namely FDI total share and FDI inward share. Second, they admit that their baseline model "does not allow (. . .) to differentiate between the idea that persistence in observations of insecurity is accounted for by the influence of past experiences of insecurity on present perceptions and the alternative idea that certain individuals just have unobserved characteristics that lead them to have certain types of perceptions" (2004: 670). Accordingly, they use Arellano-Bond's first-differenced estimator to account for dynamics. As a third robustness test, Scheve and Slaughter admit that perceptions may influence an individual's choice of industry (which would render their model partly endogenous). They deal with endogeneity by lagging their FDI variable, which does not, however, solve the endogeneity issue if FDI is announced one year before it actually occurs or if FDI is serially correlated. In a fourth robustness test, Scheve and Slaughter add six additional covariates to explore the extent to which estimates depend on the choice of regressors. In their fifth and final reported test, they repeat their analyses based on a broader sample. What makes their article a candidate for good practice in robustness testing is not so much the specific choice of robustness tests, but the discussion they devote to derive robustness tests from specific dimensions of model uncertainty.

For a second noteworthy example consider Gerber and Huber (2010), who study the association between partisanship and economic assessments. They find that large partisan differences between Republican and Democratic voters in the United States exist and conclude that the observed pattern of partisan response suggests partisan differences in perceptions of the economic competence of the parties. Naturally, the self-description of voters in a survey can be subject to measurement error. Not only does the existence of neutrals pose a specification issue, the degree to which a survey respondent supports Democrats or Republicans also varies largely in a way not appropriately reflected by the survey. Gerber and Huber's robustness tests seek to address these issues. In different model specifications, they exclude independents, they change the coding scale of party identification

toward fewer and more categories, they allow for a flexible effect of partisan affiliation by converting the categorical measure into separate exhaustive dummy variables, they employ matching to test whether the effect of partisanship is influenced by the linear functional form specification of the control variables, they control for spatial sorting of individuals by including a measure of partisanship at the aggregate state level, and they control for unobserved state heterogeneity by including state fixed effects. They conclude from conducting these tests (Gerber and Huber 2010: 167):

> [T]hese robustness checks suggest that the pattern of partisan response (...) is not driven by particular functional form assumptions or the behavior of independents. Rather, across a variety of measurement and model specifications, Democrats reacted to the 2006 election by becoming more optimistic in their economic forecasts for the national economy, while Republicans became more pessimistic.

3.3 ROBUSTNESS TESTING IN FOUR SYSTEMATIC STEPS

Since robustness testing developed as a grassroots enterprise, few if any common standards and practices have been developed. To provide a more systematic approach, we suggest that analyses of robustness require four steps:

1. Define a model that is, in the researcher's subjective expectation, the optimal specification for the data-generating process at hand, i.e. the model that optimally balances simplicity against generality, employing theory, econometric tests, and prior research in finding it. Call this model the baseline model.
2. Identify assumptions made in the specification of the baseline model which are potentially arbitrary and that could be replaced with alternative plausible assumptions.
3. Develop models that change one of the baseline model's assumptions at a time. These alternatives are called robustness test models.
4. Compare the estimated effects of each robustness test model to the baseline model and compute the estimated degree of robustness.[4]

The first step – the singling out of a model – appears to be the most controversial decision researchers have to take. Why do we think that the formulation of a baseline model and the resulting hierarchy between baseline model and robustness test models are good ideas? Conceptually similar approaches such as Leamer's sensitivity analysis and model averaging across

4 The next chapter deals with the fourth step. The entire second part of the book identifies potentially arbitrary modelling assumptions for different aspects of model specification and develops robustness tests for tackling them.

a large number of models refrain from suggesting the choice of a baseline model with all models in the model space having an ex-ante equal probability of being the best model.

The formulation of a baseline model and the testing of robustness against the baseline model require the provision of theoretical or other justifications for each model specification choice because the specification of the model has to be plausible. In fact, baseline models should be the researchers' *best bet* for an optimal specification of the data-generating process balancing simplicity versus generality. In addition, robustness test models need to represent plausible alternatives to specific baseline model specification choices.

Known or demonstrable misspecifications disqualify models from serving as a robustness test model. We distinguish here between three categories of empirical models. First, models known to be correctly specified – a category that is empty, at least in the study of observational data. Second, and on the other end of the spectrum, models that are known or at least strongly suspected to be misspecified. Structure in the residuals provides hints for model misspecification, but does not offer final proof. Likewise, if models are used for testing theories, they have to be consistent with the theory's assumptions and test its predictions. Models that do not are known to be misspecified. Third, models which are not obviously misspecified, but equally are not known to be correctly specified either. Even if the odds are diminishingly small, they could in principle or potentially at least be correctly specified. We call the specification of such models plausible. Both baseline and robustness test models should fall into this category.

Thus, the difference between Leamer's approach and robustness testing is the latter's focus on plausible model specifications. Leamer's definition of model space, by contrast, does not avoid the inclusion of models known to be misspecified. Rather than estimating and averaging across millions of models, many of which must be misspecified, robustness testing relies upon estimating a small number of plausibly specified models or a small number of sets of plausibly specified models (since randomized permutation tests can themselves employ hundreds or thousands of models varying one specific dimension of model specification).

3.4 THE MULTIDIMENSIONALITY OF ROBUSTNESS

Robustness is a multidimensional concept. The robustness of empirical estimates to changes in, say, the sample and assumed conditionality structure differ. Just like steel constructions are subjected to different robustness tests, so baseline models should be subjected to different robustness tests addressing different potential sources for potential lack of robustness.

Hence, the question is not whether the baseline model's estimates are robust in general, but whether they are robust to a specified change in a particular aspect of model specification. In other words: robustness is not an overall or general property of an estimate, but a property that differs from robustness test to robustness test.

While it seems possible to distinguish between important and irrelevant robustness tests, it is not possible to conduct "all relevant robustness tests." Every empirical model consists of multiple specification decisions. Scholars are usually uncertain about numerous aspects. In addition, few specification decisions are dichotomous, so that in each dimension a large number of alternatives may appear to be plausible. As a consequence, the possibility space for plausible model specifications is typically too large to try all permutations of all plausible assumptions of all aspects of model specification. There will always be possible alternative models which remain unknown or at least unchosen. In other words: the robustness tests that a researcher conducts are a selected subset of the entire model space of plausible models.

Given the multidimensionality of robustness, and the diversity of robustness tests that can be conducted within each dimension, averaging results over a large number of robustness tests is not useful either. While it is technically not difficult to average over different point estimates by taking weighted or unweighted means of point estimates or by adding the sampling distributions of these estimates, the result will inevitably depend on assumptions concerning the plausible model space and the model selection algorithm.[5] It would be convenient to compute a single parameter and single measure of uncertainty for the "overall averaged model" (baseline plus robustness test models) or for the "average robustness test model." Yet, the multidimensionality of model specifications should not beguile researchers to summarize over these dimensions in order to identify "the robustness" of an estimated effect. Strong robustness in one or more dimensions should not cover up the lack of robustness in other dimensions. At the very least, the multidimensionality of robustness testing requires that – regardless of the definition of model space and the averaging algorithm used – a single measure of overall robustness necessitates dimensionality reduction. This

5 Both Bayesian and frequentist methodologists have developed model averaging approaches (Hoeting et al. 1999; Claeskens and Hjort 2008). In model averaging, quantities of interest (point estimates, standard errors) are expressed as a weighted average of the same quantities from the models to be averaged. The weights used in these procedures differ. Where measures of model fit are used as weights results tend to be substantively similar since measures of model fit tend to be highly correlated.

will result in a loss of information unless all robustness tests were perfectly correlated, which they are not.

Knowing that a baseline model's estimate is robust on average is less useful than knowing that it is robust in, say, six dimensions of model specification and lacks robustness in a seventh dimension. Lack of robustness in a particular dimension is important information that gets lost by computing average robustness. It is therefore best to explore robustness in multiple dimensions of model specification separately without averaging across all robustness test models.

3.5 AIMS AND GOALS OF ROBUSTNESS TESTING

The purpose of robustness testing is *not* the demonstration that estimates, results, or findings are robust and all inferences are valid. Though practically all reported robustness tests conclude with a statement like the above, robustness tests do not demonstrate, let alone prove, the validity of inferences – especially not when the tests are selected by the authors. Instead, we suggest three main aims and goals of robustness testing:

- exploring the robustness of estimates,
- identifying limits of robustness and
- spurring further research via learning from variation in estimates across model specifications.

Exploring whether estimates are robust to specific plausible changes in the model specification is the principal aim of robustness testing. It is not the same as setting out to demonstrate that estimates are robust. The former task is driven by a sincere and serious attempt at exploring the robustness of estimates whereas the latter task seems driven by a desire to move the manuscript past reviewers and editors. Exploring robustness is part of a well-designed research strategy, whereas setting out to demonstrate robustness is merely part of a publication strategy.

The second goal of robustness testing is the identification of limits of robustness. Many of the most misleading inferential errors in the social sciences result from the concentration on and over-generalization from average estimated effects. Social scientists tend to infer internal and external validity from statistically significant point estimates. By doing so, they overgeneralize findings and ignore that inferences tend to be limited in space and time. Robustness tests may bring the relevance of cases and historic time back into focus. Robustness tests can explore whether the estimated point estimate of the baseline model represents the effects in all units of analysis. Likewise, robustness tests can investigate whether the mean effect represents the entire period under investigation, or whether effect strengths vary over

time. Robustness limit tests explicitly ask where the boundaries are. For example, scholars may ask how large measurement error needs to become to render the baseline model estimate non-robust. However, even other types of tests beyond robustness limit tests can nevertheless shed light on the limits of robustness.

Finally, though robustness tests are not primarily an instrument for model improvement, they can, by recognizing model specifications that lack robustness, identify areas where further research seems most promising or even necessary. In an ideal scientific world researchers abandon their preference for "robust" findings and employ robustness tests to identify important future avenues for research. An identified lack of robustness in a dimension of model specification poses questions that complementary research might be able to answer. Consider the simple example that two or more competing proxies for a latent variable exist, as is the case with ethnic diversity or democracy. Now assume that replacing one by the other proxy reveals a lack of robustness. In addition to stating that estimates are not robust to a change in the operationalization of the explanatory variable, researchers could thus investigate which cases drive the differences in results and discuss which operationalization appears to be more appropriate and how an optimal proxy variable for ethnic diversity or for democracy would be defined and measured.

Single robustness tests typically do not achieve all of the aims and goals of robustness testing. For the first and primary goal, the best robustness test provides the best insight into the dependence of estimated effects on model specifications. From a learning perspective, the best robustness test potentially offers the deepest insights into the causes for the observed variation in estimates of effect sizes. While robustness tests are often specialized for single purposes and only achieve one or perhaps two of these aims and goals, a shrewd combination of robustness tests can achieve all aims simultaneously.

3.6 CONCLUSION

By providing insights into the stability of estimates and into the factors that may inhibit this stability, by identifying the limits of robustness and by illuminating relevant areas of further research, robustness tests can contribute to the production of scientific knowledge. If properly undertaken, robustness tests can dramatically improve the perceived validity of causal inferences based on regression analysis of observational data.

Yet, despite the success and rise of robustness tests in social science practice, we argue that we are far away from this ideal and that a change in the practice of robustness testing is required. In this chapter, we have

provided the foundation for a systematic approach to robustness testing. We have argued against general, average or overall robustness. The multiple dimensions of model uncertainty need to be explored separately and strong robustness in one dimension cannot compensate for lack of robustness in other dimensions.

Robustness tests are necessarily selected from a "possibility space." It is not possible to conduct all possible or all relevant robustness tests – just as it is not possible to conduct all permutations of plausible models. Nevertheless, there can be robustness tests that are so crucial that the baseline model's failure to pass them will cast serious doubt on the baseline model's estimated effect. While it is not possible to predefine robustness tests that are crucial for all research projects, it is often possible to predefine (and possibly even pre-register) crucial robustness tests for specific research projects.

4 The Concept of Robustness

Humans desire certainty, and science infrequently provides it. As much as we might wish it to be otherwise, a single study almost never provides definitive resolution for or against an effect and its explanation (. . .) Scientific progress is a cumulative process of uncertainty reduction that can only succeed if science itself remains the greatest sceptic of its explanatory claims.
Nosek and 268 co-authors (2015: aac4716)

4.1 INTRODUCTION

Robustness relates to the behavior of an object under stress and strain. In technical language, robustness refers to the ability to tolerate perturbations that potentially affect the object's functions. In order to fall in line with this concept of robustness, we need to answer three questions:

1. What is the object?
2. What is the stress and strain to which we subject the object?
3. How can we compute robustness?

The object of robustness depends on the research question and the inferences researchers wish to make. In most cases, an analysis aims at testing the predictions from a theoretical model about the effect of one or more variables on an outcome. In this case, the object of robustness tests is the baseline model's estimated effect of x on y. Note that we write "effect" here rather than "coefficient" because similar effects are consistent with dissimilar coefficients in non-linear models, models that allow for non-linear functional forms, conditionalities, and so on. Conversely, similar coefficients may also imply very different effects. In a simple linear model without non-linear or conditional effects the robustness of an estimated effect is identical to the robustness of its estimated coefficients. In all other cases, this does not hold and analysts need to compute effects and state at what values of the explanatory variables they assess robustness or,

preferably, analyze partial robustness, which we define further below. Yet, researchers may instead be interested in forecasting, in which case the predicted effect of the entire model becomes the object of robustness tests. If, for example, an analysis forecasts population growth, the object of robustness tests is the forecast that the baseline model makes. In the remainder of this book, we will talk about the effect of a variable x on y as the object of robustness, but readers should keep in mind that the object may be different.

Researchers impose stress and strain on the above object by changing the specification of the baseline model in systematic and plausible ways. It follows that implausible model specifications are not valid robustness tests. Models known or strongly suspected to be misspecified do not qualify as robustness test models.

In this chapter, we define robustness and propose as a measure of robustness the extent to which a robustness test model estimate supports the baseline model estimate. We suggest that this measure, which varies from 0 to 1, offers several useful properties, including that it measures robustness continuously rather than declaring an estimate as robust or non-robust at an arbitrarily chosen threshold. Our definition of robustness is independent of the level of statistical significance of either the baseline or robustness test model and we contend that robustness is most usefully understood as stability in the estimated effect, which is inconsistent with a definition that relates to statistical significance, even if this departs from how many interpret robustness in current practice. We also introduce the concept of partial robustness, which is relevant in all non-linear models and even linear models that estimate a non-linear, conditional or heterogeneous effect. The concept of partial robustness allows the degree of robustness to differ across observations in all such models.

4.2 DEFINITIONS AND CONCEPTS OF ROBUSTNESS IN CURRENT PRACTICE

Robustness tests are common practice. An increasing number of researchers report the results of robustness tests and an even larger number claims that their baseline model proved robust to specific changes in the model specification without showing the results. However, though robustness tests are fashionable, neither a common practice of core tests nor a common understanding of the meaning of robustness has evolved. Indeed, social scientists disagree about what they mean by robustness, what ought to be robust, and where they see the threshold between robust and not robust.

No commonly accepted definition of robustness exists. Researchers conducting robustness tests rarely ever define robustness when they claim

their results are robust. Instead, robustness is typically regarded as given by a situation in which estimates from robustness tests do not "deviate much" from the estimates of the baseline model. Specifically, scholars see robustness as given when estimates "are quite similar" (Bailey and Maltzman 2008: 379) or "close" (Gehlbach, Sonin, and Zhuravskaya 2010: 732), "results uphold," coefficients remain "substantively similar" (Lipsmeyer and Zhu 2011: 652) or "do not change" (Braumoeller 2008: 86; Hafner-Burton, Helfer, and Fariss 2011: 701; Mukherjee and Singer 2010: 50) and thus "remain essentially the same." Yet, how similar estimates have to be to qualify as "fairly similar," "essentially the same," or "close" is almost never operationally defined. The vagueness in the conceptual definition implies that virtually all authors can interpret their results as "robust." To make matters worse, social scientists do not agree on what ought to be robust: is it effects, their level of statistical significance or inferences? In the next section, we offer an operational definition of robustness.

4.3 DEFINING ROBUSTNESS

Robustness tests explore the stability of the baseline model's estimated effect to systematic alternative plausible model specification changes. We define robustness as the degree to which the baseline model's estimated effect of interest is supported by another robustness test model that makes a plausible change in model specification. Higher levels of robustness imply a higher degree of support for the baseline model's estimated effect, lower levels of robustness suggest a lower degree of support.

The baseline model provides a point estimate for the effect of interest. Naturally, sampling variability means that the point estimate is unlikely to exactly capture the population parameter. If analysts draw another finite random sample from the population, they will get another point estimate because the distribution of random errors will be different. The estimated standard error of the point estimate allows the construction of 90- or, more typically, 95-percent confidence intervals. A confidence interval provides an estimate for a plausible range for the estimated parameter, given sampling variability (Cumming 2012: 79).

As we have argued before, the baseline model, like the robustness test models, falls into the category of plausible models – that is, models which are neither known to be misspecified nor known to be correctly specified either. Hence, researchers cannot claim that the baseline model captures the "truth" with any level of confidence. Likewise, robustness tests do not seek and cannot find the truth, but they analyze the extent to which estimates from different model specifications support the estimate from a baseline model specification. In this sense, the baseline model marks a researcher's

best effort at constructing an estimation model. It is therefore not just any model, it is the model against which other robustness test models that make other plausible specification assumptions should be compared to.

Given our above definition of robustness, and taking sampling variability into account, robustness becomes the extent to which social scientists can be confident that the plausible range for the estimated effect of the robustness test model supports the plausible range for the estimated effect from the baseline model. To be precise, we define robustness as the degree to which the probability density function of the robustness test model's estimate falls within the confidence interval of the baseline model.

Assume for simplicity a linear and unconditional model such that coefficients represent effects. Formally, let

$$f(a_b, \hat{\beta}_b, \hat{\sigma}_b) = \frac{1}{\hat{\sigma}_b\sqrt{2\pi}} e^{-(a_b-\hat{\beta}_b)^2/2\hat{\sigma}_b{}^2} \tag{4.1}$$

be the probability density function of parameter estimate β_b, which is the point estimate of the effect of variable x and σ_b its standard error. This density function is normally distributed by construction: since econometric theory assumes that errors are normally distributed, the probability density function of the parameter estimate is also normally distributed. If methodologists make alternative assumptions about the error process, a different probability density function for ρ or some transformation of the original equation is required. Fox (1991) argues that the assumption of normally distributed errors appears arbitrary. We disagree for two reasons. First, the assumption is not arbitrary but roots in theories of random processes and in experiments with stochastic processes. And second, the central-limit theorem proves that, in the limit, the sum of random distributions approaches a normal distribution. We therefore know no other general assumption about error processes which is as plausible as the normal one.

As equation 4.2 suggests, we define the degree of robustness ρ (rho) as the share or percentage of the probability density function of the robustness test model that falls within the 95-percent confidence interval of the probability density function of the baseline model,[1] which is

$$\rho(\hat{\beta}_r) \equiv \frac{1}{\hat{\sigma}_r\sqrt{2\pi}} \int_{\hat{\beta}_b - C\hat{\sigma}_b}^{\hat{\beta}_b + C\hat{\sigma}_b} e^{-(a_r-\hat{\beta}_r)^2/C\hat{\sigma}_r{}^2} da. \tag{4.2}$$

Again, the probability density function of the robustness test model is assumed to be normally distributed by econometric convention.

1 Note that C decreases from approximately 2.04 to approximately 1.96 as the sample size grows toward infinity.

This definition has some useful properties. Assume, for simplicity, that the coefficients of the baseline and the robustness test models are identical. Under this assumption, the estimated degree of robustness ρ depends entirely on the standard error of the robustness test model compared to the one from the baseline model. If the standard error were exactly the same in both models, then $\rho = 0.95$. This makes sense: with the robustness test producing the exact same result as the baseline model we are 95 percent confident that the robustness test estimate falls within the 95-percent confidence interval of the baseline model. If the standard error of the robustness test is smaller than the baseline model, ρ becomes larger than 0.95 and converges to 1.00 as the robustness test standard error becomes smaller and smaller. This, again, represents a useful property: the smaller standard error of the robustness test model suggests researchers can be more confident that the robustness test estimate falls within the 95-percent confidence interval of the baseline model. Conversely, if the robustness test standard error is larger than the one from the baseline model, ρ is necessarily smaller than 0.95 and declines as the robustness test error becomes larger, if we keep point estimates constant.[2] Figure 4.1 illustrates the logic for a baseline model with a coefficient of 1.0 for the variable of interest and a standard error of 0.3 and a robustness test model with a coefficient of 1.0 and a standard error of 0.5 (light grey shading).

The calculated value of ρ for figure 4.1 equals 0.760. Thus, 76 percent of the probability density function of the robustness test model falls within the 95-percent confidence band of the baseline model. The robustness measure ρ provides information on the stability of the baseline model's estimated effect: since the robustness test model has the same coefficient but a larger standard error, the confidence in the baseline model's estimate declines.

We now relax the unrealistic assumption that the point estimates of the robustness test and the baseline models are identical. In this case, ρ is determined by both the difference in point estimates and the standard errors. Figure 4.2 displays the same baseline model but the robustness test model has a different point estimate. As a comparison of the two figures shows, ρ declines as the difference between the point estimates of the baseline model and the robustness test model increases. In example 2 (figure 4.2), with the robustness test model giving a point estimate of 1.5 and standard

2 The 0.95 threshold at which one becomes either more or less certain is precise for all robustness tests that hold the sample constant. If, however, a robustness test varies the sample, the sampling variation between the baseline and the robustness test model will push the value of ρ downward, all other things equal. In principle, one would therefore want a lower threshold of ρ for robustness tests that do not hold the sample constant.

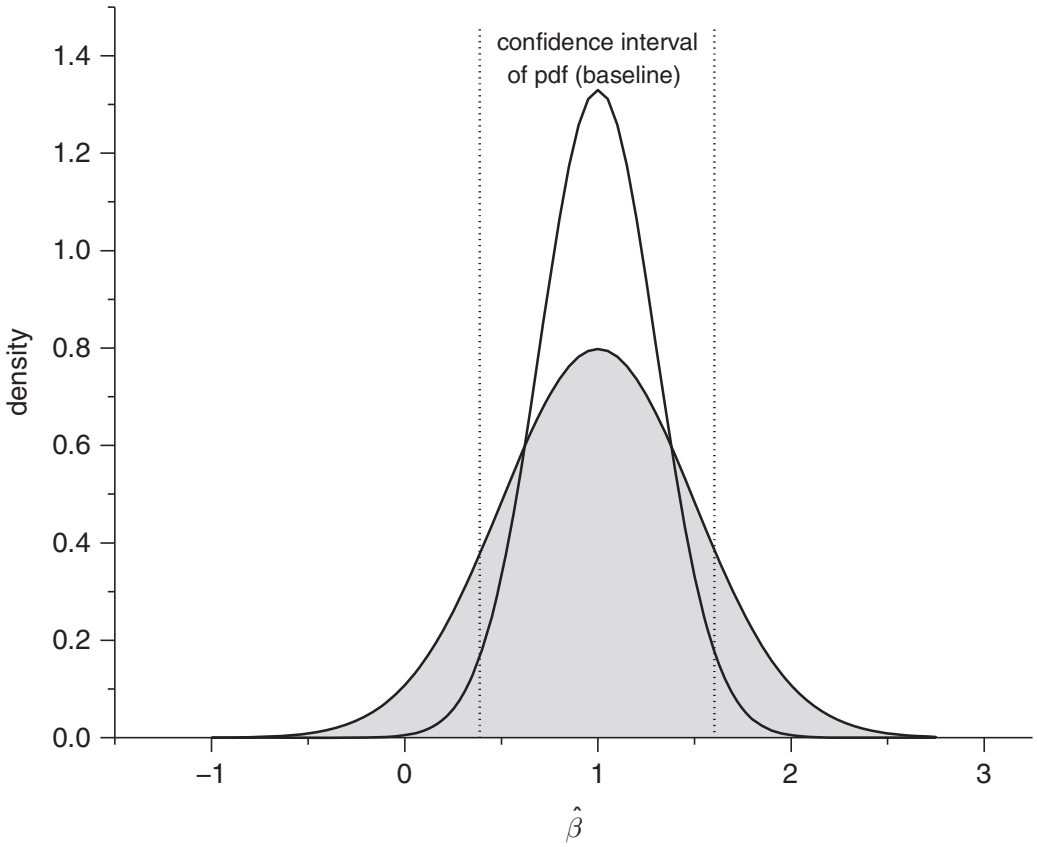

Figure 4.1: Example 1 of Degree of Robustness ρ

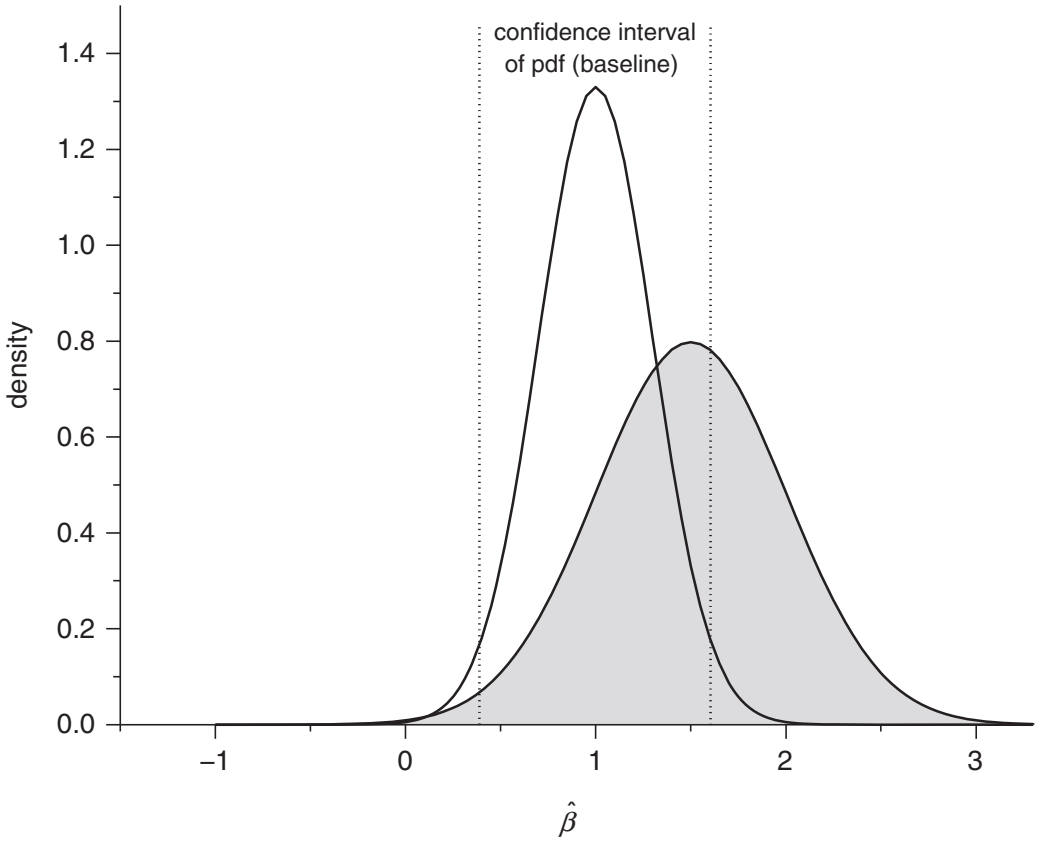

Figure 4.2: Example 2 of Degree of Robustness ρ

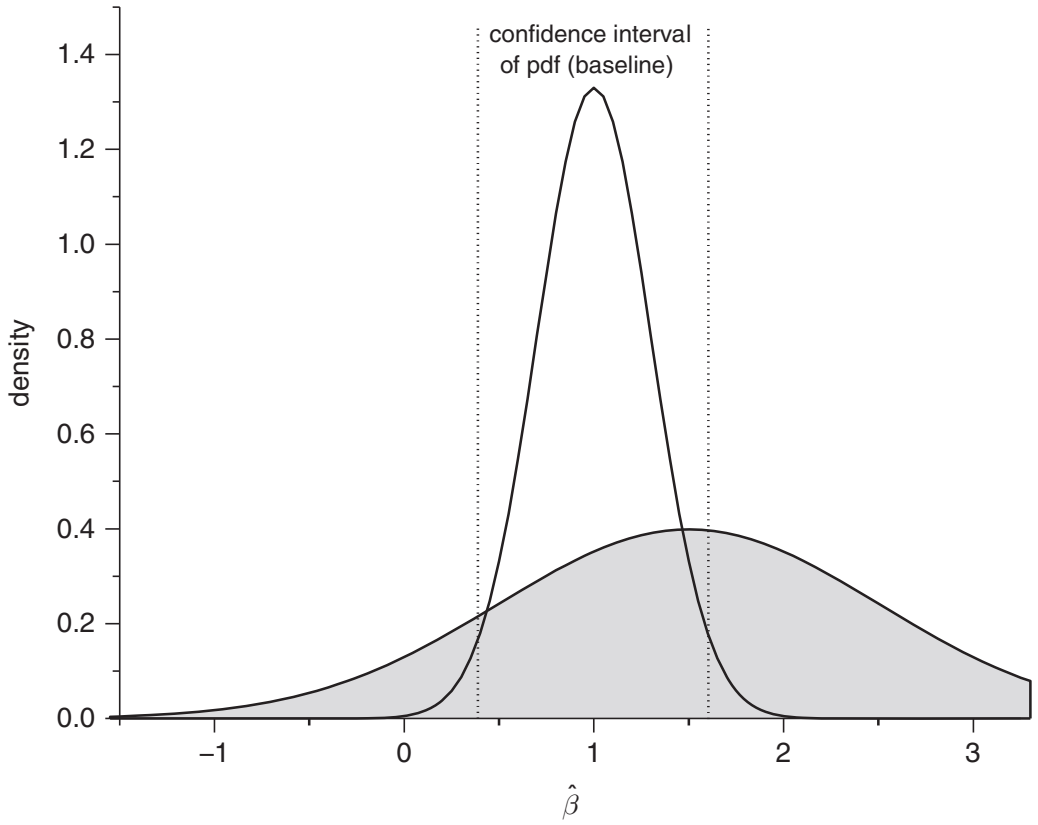

Figure 4.3: Example 3 of Degree of Robustness ρ

error 0.3, ρ is 0.615. Our third example (figure 4.3) doubles the standard error in the robustness test estimate to 1.0. By increasing the standard error, ρ declines to 0.397 in this example.

Table 4.1 shows the joint influence of the difference between the point estimates and the standard error of the robustness test model on ρ for a baseline model point estimate of 1 with standard error of 0.3.

Table 4.1 demonstrates several properties of ρ. First, robustness is left–right symmetric: identical positive and negative deviations of the robustness test compared to the baseline model give the same degree of robustness. It does not matter for ρ whether the estimate of the robustness test model is larger or smaller than the one of the baseline model. Only the difference matters. Second, if the standard error of the robustness test is smaller than the one from the baseline model, ρ converges to 1 as long as the difference in point estimates is small. If the robustness test coefficient is estimated with high precision, its probability density function can lie almost entirely within the baseline model's confidence interval even if the point estimates differ, as long as they do not differ too much. Third, for any given standard error of the robustness test, ρ is always and unambiguously smaller the larger the difference in point estimates. Not surprisingly, for any given level of uncertainty around a robustness test estimate, the larger the

Table 4.1: Degree of Robustness for Various Robustness Test Estimates

	s.e.=0.1	s.e.=0.3	s.e.=0.5	s.e.=0.7	s.e.=1.0	s.e.=2.0
β=–0.50	0.000	0.001	0.034	0.095	0.162	0.176
β=–0.25	0.000	0.014	0.093	0.168	0.221	0.191
β=0.00	0.000	0.085	0.204	0.266	0.284	0.204
β=0.25	0.053	0.295	0.369	0.381	0.345	0.216
β=0.50	0.811	0.615	0.555	0.490	0.397	0.224
β=0.75	1.000	0.867	0.704	0.570	0.431	0.229
β=1.00	1.000	0.950	0.760	0.599	0.443	0.231
β=1.25	1.000	0.867	0.704	0.570	0.431	0.229
β=1.50	0.811	0.615	0.555	0.490	0.397	0.224
β=1.75	0.053	0.295	0.369	0.381	0.345	0.216
β=2.00	0.000	0.085	0.204	0.266	0.284	0.204
β=2.25	0.000	0.014	0.093	0.168	0.221	0.191
β=2.50	0.000	0.001	0.034	0.095	0.162	0.176

Note: Baseline model β = 1.0; s.e. = 0.3, 95-percent confidence interval

difference in the point estimates the lower the support for the baseline model's estimate. Fourth, differences in point estimates have a strong influence on ρ if the standard error of the robustness test is small but a small influence if the standard errors are large. Robustness test models estimated with large sampling variability remain uninformative – they are not powerful enough to increase the certainty of the baseline model estimate but at the same time not powerful enough for signaling complete lack of robustness.

Perhaps surprising at first sight is the complex influence of the sampling distribution of the robustness test model estimates on ρ. The impact of increasing standard errors on ρ is ambiguous as it depends on the difference in point estimates between the robustness test and baseline model relative to the baseline model's confidence interval. If the difference in point estimates is such that the robustness test point estimate lies *within* the baseline model's confidence interval, i.e. if $|\beta_b - \beta_r| < C\sigma_b^2$, then increasing standard errors of the robustness test model's estimate unambiguously decrease ρ. The highest probability of the robustness test estimate lies within the baseline model confidence interval but increasing uncertainty around the robustness test estimate increases the uncertainty as to whether the robustness test supports the baseline model estimate. Conversely, if the difference in point estimates is such that the robustness test point estimate lies *outside* the baseline model's confidence interval, i.e. if $|\beta_b - \beta_r| > C\sigma_b^2$, increasing standard errors of the robustness test model first increase ρ and then decrease it as the standard error becomes larger and larger. This may seem counter-intuitive but is easily explained: the highest probability of the robustness

test estimate lies outside the baseline model's confidence interval. With small standard errors researchers can be fairly confident that the robustness test model does not support the baseline model estimate. In the extreme, almost the entire probability density function of the robustness test lies outside the baseline confidence interval and ρ converges to zero. As the standard error increases, one of the tails of the robustness test probability density function moves closer to or, if already inside, moves further into the baseline model's confidence interval. Researchers thus become less confident that the robustness test does *not* support the baseline model. Eventually, with larger and larger standard errors, the tail of the robustness test probability density function moves outside the other end of the baseline confidence interval and reduces the confidence that the robustness test supports the baseline model. Figure 4.4 displays the joint effect of changes in the difference between point estimates and changes in the standard errors of the robustness test model on ρ (based on the assumption of a baseline model point estimate of 1 with standard error of 0.3).

Figure 4.5 displays the same information in a different way, namely as a heat plot. It shows the nonlinear bivariate relation between the difference

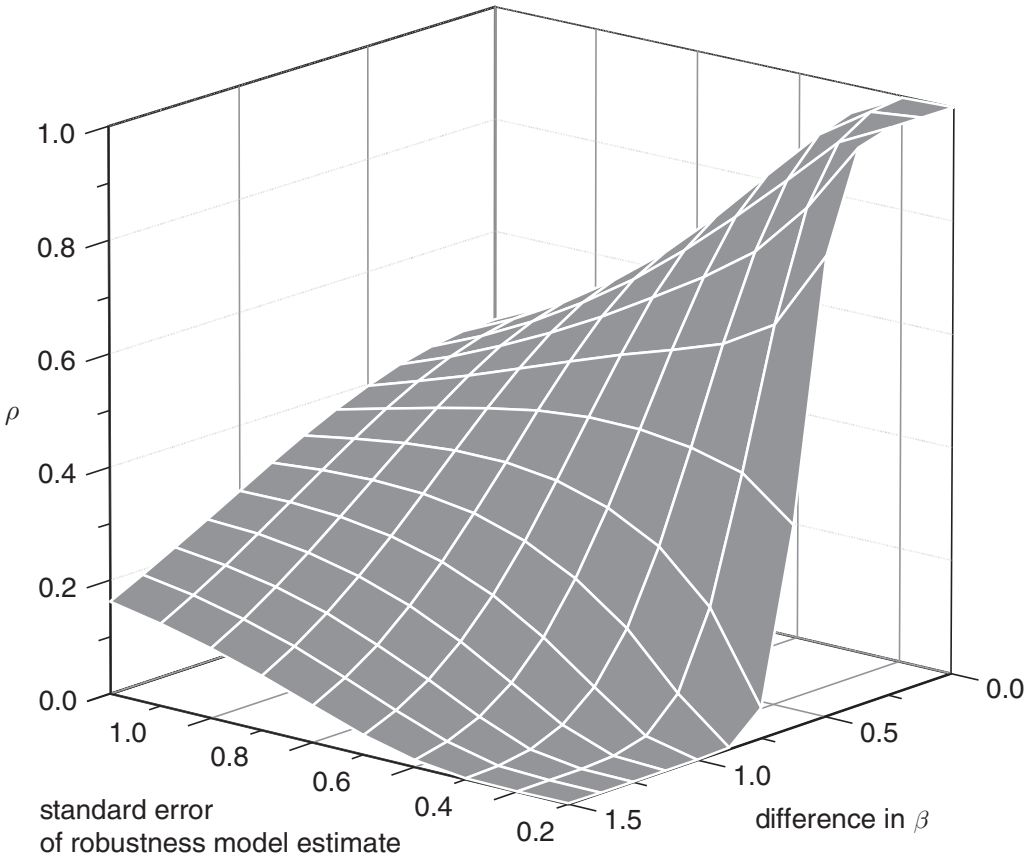

Figure 4.4: ρ as a Function of the Difference in Point Estimates and Standard Errors
Note: Baseline model β=1; s.e.=0.3.

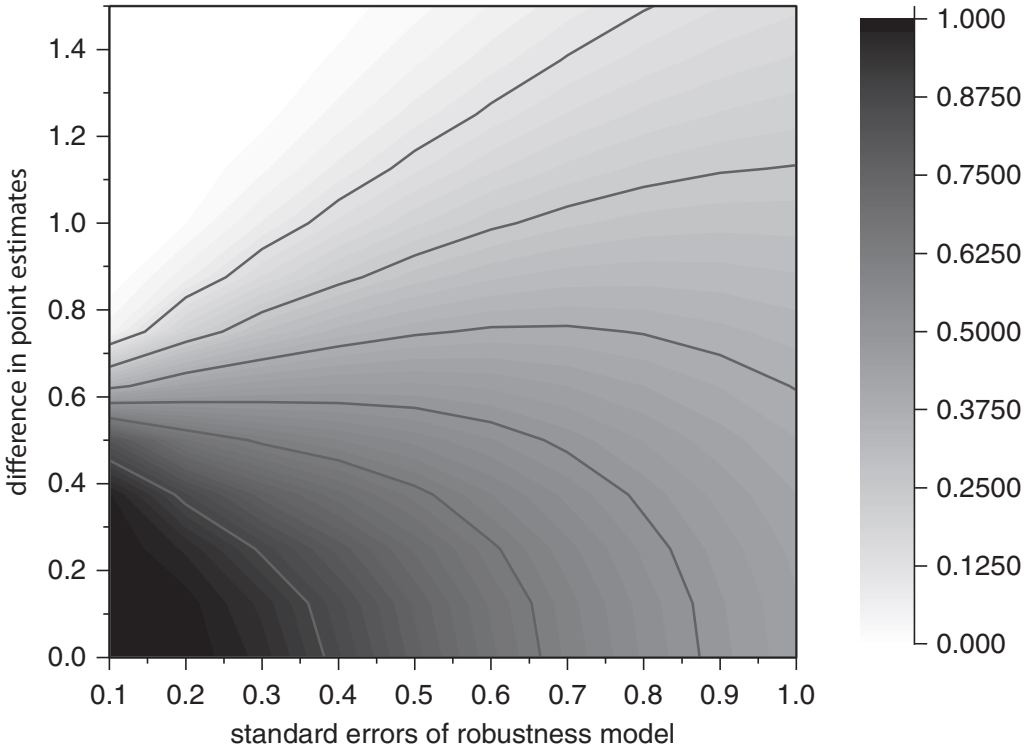

Figure 4.5: ρ as a Function of the Difference in Point Estimates and Standard Errors (Heat Plot)
Note: Baseline model β=1; s.e.=0.3.

in the point estimates of the baseline and the robustness test model and the standard error of the robustness test model (based on the assumption of baseline model point estimate of 1 with standard error of 0.3 as before). This distribution resembles a bivariate Weibull distribution. It asymptotically goes to 0 in three of the four corners and to 1 in the remaining fourth corner. Accordingly, if the difference in point estimates goes to infinity, ρ goes to 0 (top left corner). Similarly, if the standard error of the robustness test goes to infinity, ρ goes to 0 (bottom right corner). The same holds if both the difference in point estimates and the robustness test standard error go to infinity (top right corner). Conversely, ρ goes to 1 if either the difference in point estimates goes to 0 and the robustness test standard error is smaller than the one from the baseline model or if the robustness test standard error goes to 0 and the difference in point estimates remains sufficiently small.

Figure 4.5 also displays selected isolines, which represent equal degrees of robustness. The ones starting just above and just below 0.60 on the difference in point estimates scale illustrate that if the difference in point estimates is below the crucial threshold of approximately 1.96 times the baseline model estimate's standard error (here around 0.59), ρ monotonously decreases with larger standard errors of the robustness test model. If it stays above that threshold, ρ first increases and then decreases, i.e. it is

non-monotonous. Asymptotically, both converge to 0 as the robustness test standard error goes to infinity.

Note that for relatively small differences in point estimates, it takes large standard errors for ρ to converge to 0. For example, for a difference of 0.20 or 0.40 and a standard error of 0.90 (i.e. three times the size of the baseline model standard error), ρ equals 0.447 and 0.476, respectively. Even with a standard error as large as 5.00 for the same differences in point estimates ρ is 0.094 and 0.093, respectively. It takes a standard error of around 9.50 for ρ to drop below 0.05. When the robustness test estimate lies within the baseline model's confidence interval, the degree of robustness will not be low unless the standard error becomes sufficiently large. In other words, robustness test models with even fairly large uncertainty around their estimates do not render the baseline model estimate non-robust, unless the sampling uncertainty becomes extremely large. Conversely, if the difference in point estimates is relatively large, small standard errors signal lack of robustness, but even relatively large standard errors do not produce high degrees of robustness. If the point estimate of the robustness test lies outside the confidence interval, ρ can never be higher than 0.50 no matter what the standard error.

Critics might wonder whether our definition of robustness creates strategic incentives for authors to specify their baseline model sub-optimally in order to maximize the chances that their results will appear robust. After all, all other things equal, a baseline model estimate that has wider confidence intervals is more likely to be found robust than one with narrower confidence bands. This is only logical: the larger the uncertainty of the baseline model, the smaller the extent to which robustness tests can add further to the uncertainty. But all other things are not equal. If researchers intentionally specify their baseline model less well than they can, the degree of robustness of the baseline model's estimate will likely increase for some tests but decrease for others. Particularly if the choice of robustness tests is not left to researchers alone but partly determined by reviewers and editors, a strategic misuse of robustness may backfire, thus diminishing the incentive.

4.4 CONTINUOUS VERSUS DICHOTOMOUS ROBUSTNESS

According to our definition, the degree of robustness ρ is a continuous measure, ranging from 0 to 1. We think this has important advantages. A continuous concept of robustness reflects the fact that robustness comes in degrees and not as a dichotomy. Higher values of ρ represent a higher degree of robustness and lower values represent a lower degree of robustness. Robustness tests can increase the confidence in the baseline model's

estimated effect size if ρ exceeds 0.95. Yet, the majority of robustness tests will result in a ρ smaller than 0.95, which suggests a higher level of uncertainty than the baseline model implies. Robustness tests provide a more realistic picture of the uncertainty of the baseline model's point estimate. The true uncertainty stems not only from sampling variability expressed by the baseline model estimate's standard error but also from model uncertainty and its consequences.

Nonetheless, researchers are familiar with critical values of, for example, statistical tests and might crave a criterion for when to regard a robustness test estimate as suggesting non-robustness. Generally speaking, we do not believe that the arbitrary creation of critical values for robustness is useful, just as we do not believe that the arbitrary distinction between statistically significant and statistically insignificant – with its consequence of arbitrary rejection decisions (of null hypotheses and of manuscripts in the review process) – has served the social sciences well (Gill 1999). Arbitrary thresholds provide a major obstacle to the accumulation of scientific knowledge. With this caveat in mind, if scholars wanted to look for a critical value for ρ, it is likely to be 0.05 since in this case 95 percent of the probability density function of the robustness test estimate lies outside the baseline model's confidence interval. Accordingly, the robustness test model estimate does not support the baseline model estimate.

We nevertheless urge scholars to abstain from clinging to arbitrary bounds for "robust" versus "non-robust." The important element of robustness testing is not to define arbitrary thresholds in order to dismiss certain findings as irrelevant. Both high and low degrees of robustness provide important information. High degrees of robustness indicate that model specification does not exert much influence over the estimated effect. An apparent lack of robustness indicates large uncertainty about the estimated effect. We believe that all non-trivial estimation models will lack robustness to some degree in some dimension. A lack of robustness signals an important research question, it does not falsify a theory, a prediction, or a hypothesis.

4.5 ROBUSTNESS AND STATISTICAL SIGNIFICANCE

Despite all known shortcomings and flaws of Fisher significance (Gill 1999; Rainey 2014; Gross 2015), social scientists are used to basing their inferences on whether an estimated effect is statistically significant. The widespread recognition of model uncertainty and the rise of robustness testing put an end to the idea that a single model, a single parameter estimate and its sampling error can be used to make valid statistical inferences. It did not, however, put an end to statistical significance as the predominant

criterion for making inferences, which maintained its status through the back door by becoming the dominant way in which robustness is assessed.

Starting from Leamer's idea of sensitivity testing (Leamer 1978), most applied scholars even today define robustness through an extreme bounds analysis: a baseline model estimate is robust to plausible alternative model specifications if and only if all estimates have the same direction and are all statistically significant.[3] Let us call this "Leamer robustness" for short.

In stark contrast, our definition of the concept of robustness, and our measure of the degree of robustness ρ based on this concept, are independent of the level of statistical significance of the effects in either baseline or robustness test models. All that matters for computing ρ are the point estimates and the standard errors of the baseline and the robustness test model.[4] We contend that the logic of robustness testing is incompatible with Leamer robustness and that useful definitions of robustness must refer to stability in estimated effect sizes or effect strengths as in our definition.[5]

On a fundamental level, Leamer robustness ignores that the difference between a statistically significant baseline model result and an insignificant robustness test result need not be statistically significant. For the same reason a statistically insignificant result in a replication exercise does not necessarily demonstrate that a statistically significant prior result has proven non-replicable (Goodman 1992). Gelman and Stern (2006: 329) correctly point out that if one were to make statistical significance the criterion for inference from multiple estimations, then "one should look at the statistical

3 For example, in robustness tests for their analysis of the presence of multiple veto players on the credibility of monetary commitments, Keefer and Stasavage (2002: 772) find that the "test statistics are significant in most cases at the one percent level and in all but one case at the ten percent level of confidence." In a paper analyzing how the stock values of seven European defense companies respond to EU summit decisions on defense policy, Bechtel and Schneider (2010: 219) conclude their robustness test as follows: "The coefficient of the summit outcome variable (. . .) remains positive and statistically significant". Nordås and Davenport (2013: 934f.) find that "the results for youth bulges remain highly significant (at the 1% level)" in robustness tests for their analysis of the effect of large youth cohorts on state repression. We could cite many more examples, including from our own publications.

4 Thus, with a baseline model that has a point estimate of 1.0 with standard error of 0.3, ρ is 0.72 regardless of whether the robustness test point estimate is 0.5 with standard error 0.4 or is 1.5 with standard error 0.4: the difference in point estimates between the baseline and the two robustness test models, the standard errors of the robustness test models, and the 95-percent confidence interval of the baseline model are all identical.

5 We acknowledge that ours is only one way of defining robustness in terms of effect stability.

significance of the difference" in two results "rather than the difference between their significance levels."

Leamer robustness is at odds with an understanding of robustness as the extent to which the robustness test estimate is compatible with and supports the baseline model's estimate. This cannot be assessed without direct reference to the baseline model's estimated effect size and confidence interval. The robustness of the baseline model estimate is not tested by merely checking whether the robustness point estimate has the same sign and remains statistically significant when the actual point estimate and its associated confidence interval can be very different from the baseline model estimate. Would social scientists really call a baseline model estimate of 10 with small standard errors robust to a robustness test estimate of 2 with sufficiently small standard errors below 1 so that it too is statistically significant?

Equally importantly, due to the fact that multiple models can never all be assumed to represent the optimal trade-off between generality and simplicity, employing Leamer robustness to reject null hypotheses is based on a flawed inferential logic. At best, Leamer robustness provides a one-sided test: if all estimates have the same sign and remain significant, analysts can reject the null hypothesis with greater confidence. However, the opposite inference that the null hypothesis is correct – usually that there is no effect – cannot be derived from the fact that not all models generate estimates with the same sign and the minimum level of statistical significance since one of the models could be severely misspecified or inefficiently estimated. In other words, Leamer robustness has an extremely low probability for making false positives errors but an unreliably high probability for committing false negatives errors (Plümper and Traunmüller 2016). Rejecting hypotheses based on a lack of Leamer robustness, thus, potentially allows the worst specified model or the model estimated with lowest efficiency to determine the overall inference. Since both errors are equally problematic and can lead to costly faulty policy recommendations (Lemons et al. 1997), there is no "conservative research design strategy" excuse for adopting Leamer robustness.

This problem of one-sidedness is exacerbated by the fact that in a number of robustness test models standard errors increase *by design*. The estimated effect may well become statistically insignificant, but this does not necessarily cast doubt on the robustness of the baseline model estimate. For example, many authors employ robustness tests in which they restrict the sample in some way and thus discard some observations. Naturally, the reduced sample size lowers the efficiency of the estimates and renders finding a statistically non-significant estimate more likely.[6] This

6 In the study by Bechtel and Schneider (2010), for example, one robustness test restricts the sample to estimating immediate effects (abnormal returns on the day

similarly applies to other robustness tests that discard information – for example, unit fixed-effects robustness test models that drop all the between-variation in the data.

In contrast, in our definition of robustness, estimation models with large power exert a potentially large influence, namely when the estimated effects differ. In our definition of robustness, the consequences of efficiency for the degree of robustness depend on the location of the robustness test model's point estimate. If it is far from the baseline model's point estimate, small standard errors of the robustness test signal non-robustness, not robustness. Larger standard errors signal greater robustness but never high degrees of robustness. If the robustness test point estimate is close to the baseline model's point estimate, robustness test models that lack efficiency (that come with fairly large standard errors) are not informative: these estimates do not signal non-robustness unless the size of standard errors substantially exceeds the size of the baseline model estimate's standard error.

Finally, the hunt for statistical significance has always incentivized the selection of model specifications according to *p*-values. Adopting Leamer robustness as the inferential criterion with a small number of highly selected robustness tests – most social scientists report only few robustness tests – will fuel the undesirable tendency to find everything significant and hence robust in empirical analyses. Ever since Fisher's (1925) original proposal of null hypothesis significance testing, social scientists have learned how to "tweak" significance and to conceal the lack thereof. Coupled with the fact that "undisclosed flexibility in data collection and analysis allows presenting anything as significant" (Simmons et al. 2011: 1359), published empirical social science research seems to be robust to an astonishing degree.[7]

In sum, the logic of robustness testing conflicts with defining robustness as effects remaining statistically significant with the same sign. Defining robustness instead as stability in effect size embraces the logic of robustness testing. It perfectly fits with the call by a growing number of authors for

after the summit). The estimated effect becomes statistically insignificant at the 5-percent level. The increase in standard errors is obviously triggered by the sharp decline in the number of observations (from 1,554 to 222). Why would a result become not robust only because researchers artificially reduce the available information used to estimate the effects?

7 Negative findings are important and in need of robustness testing as well. We wholeheartedly agree with the editors of eight health economics journals who issued an editorial statement on negative findings that clearly states that results from well-designed studies "have potential scientific and publication merit regardless of whether such studies' empirical findings do or do not reject null hypotheses that may be specified" (Editors 2015: 505).

social scientists to focus on the substantive importance of their estimated effects (Ziliak and McCloskey 2008; Esarey and Danneman 2015). As Gill (1999: 657f.) has put it: "Finding population-effect sizes is actually the central purpose of political science research since any difference can be found to be statistically significant given enough data."

4.6 PARTIAL ROBUSTNESS IN NON-LINEAR AND LINEAR MODELS

Up to this point, we have defined robustness as stability of the estimated effect of a variable, implicitly assuming a single estimated effect. In all non-linear estimation models, however, coefficients do not represent effects and estimated effects are a function of the values of all explanatory variables in the model. Non-linear models, thus, do not have a single effect of variable x on outcome y. In practice, authors often report the marginal effect at mean values or at median values or at other specified variable values that for some reason are of particular interest, or at variable values as observed in the sample and averaged across all observations (called the average marginal effect).

Hanmer and Kalkan (2013) make the case for basing inferences to the population on the average marginal effect. Cameron and Trivedi (2010: 340) suggest that for policy analysis one might want to look at either the average marginal effect or at targeted specified values. We agree with the latter suggestion: if researchers know to what part of the population they intend to generalize findings, they should compute the effects for cases that are similar to the part of population to which they wish to generalize. In most cases, problems occur if no case represents the entire population and if researchers do not intend to make targeted generalizations. This is a problematic practice if effect strengths vary with covariates: in non-linear models or in linear models with non-linear effects, conditional effects, or causal or temporal heterogeneity. In all of these cases, predicted effects which are representative for the entire sample do not exist.

Whenever effect strengths differ across cases, the degree of robustness differs too. It may well be that robustness is high for some parts of the sample and low for other parts of the sample. For these situations, we have developed the concept of partial robustness that applies our definition of robustness to the predicted effect and its standard error for each observation. The predicted effect varies across observations in non-linear models and in all linear models that allow for non-linearity, conditionality or causal or temporal heterogeneity. Partial robustness means that the baseline model's estimated effect can be robust or more robust for some observations but non-robust or less robust for other observations.

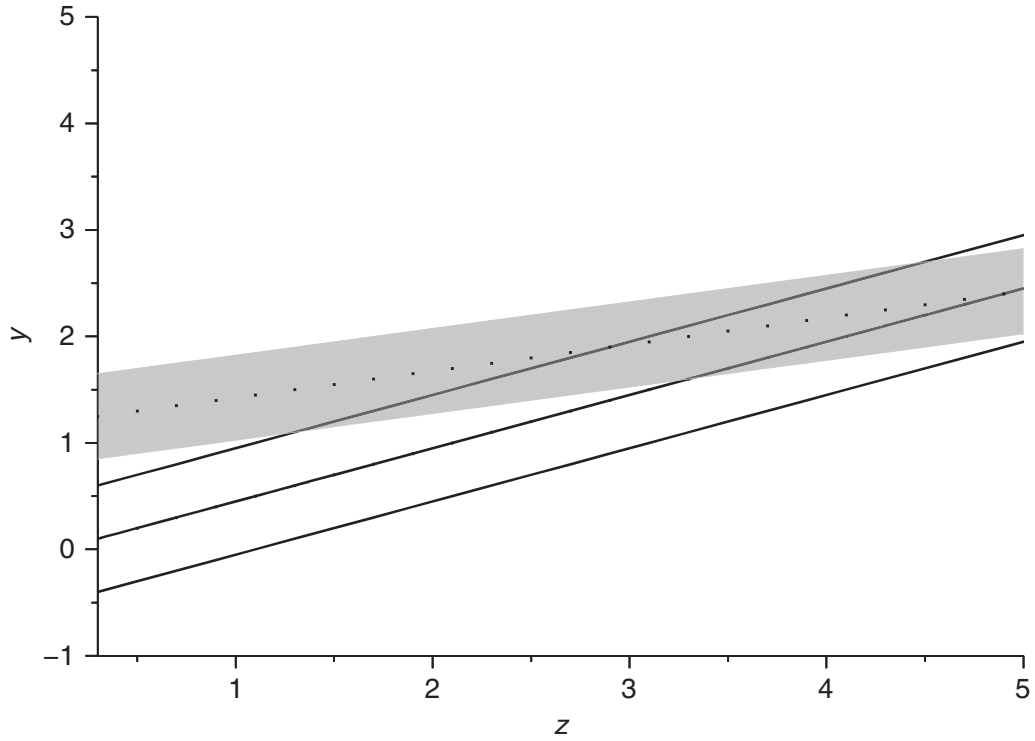

Figure 4.6: An Example of Partial Robustness
Note: Grey-shaded area represents confidence interval of baseline model.

To illustrate, assume a linear baseline and robustness test model that both estimate effects of x on y that are conditioned by z. The two models are specified differently; as a result, the estimated conditionality of x in the baseline model is weaker than in the robustness test model. Figure 4.6 shows point estimates with associated confidence intervals for the effect of x as a function of varying values of z for the baseline and robustness test model.

For values of z smaller than 1.6, the estimated degrees of robustness are below 0.05. At z=1.6, the degree of robustness is 0.05 and continues to increase to 0.98 as z increases to 4.9, the point where the point predictions are identical. Figure 4.6 thus demonstrates partial robustness: 0 or low degrees of robustness at low levels of z and high degrees of robustness at high values of z.

4.7 CONCLUSION

This chapter filled the concept of robustness with meaning. While the object of robustness depends on the research question and the intended inferences, the baseline model's estimate should be subjected to robustness testing in the form of systematic plausible changes to model specification. Robustness comes in degrees, we have argued, rather than as robust versus non-robust.

We have defined the estimated degree of robustness ρ as the degree to which the probability density function of the robustness test model's estimate falls within the 95-percent confidence interval of the baseline model. Put simply, ρ measures the extent to which the robustness test model supports the baseline model's estimated effect. Like all definitions, ours is neither right nor wrong. But it is useful, we have argued, since it has desirable properties. It both pays heed to how close the point estimates are and considers the sampling variability of both estimates.

There is another sense in which robustness comes in degrees. Whenever the estimation model is non-linear, the estimated degree of robustness will differ across observations, unless analysts restrict their analysis to effect sizes at specified variable values such as mean, median or targeted values or to the average of marginal effects across observations. Whenever they estimate non-linear, conditional or heterogeneous effects, the estimated degree of robustness will inevitably differ across observations even in a linear estimation model. We call this concept partial robustness: the effect can be more robust for some observations and non-robust or at least less robust for other observations.

Our definition of robustness conflicts with the implicit or explicit ad hoc definition of many who seem to equate robustness with an effect continuing to be statistically significant with the same sign in the robustness test model. According to our definition, an effect can be robust to a high degree or can be non-robust (be robust to a low degree) independently of whether the baseline or robustness test model's estimates are statistically significant. Robustness differs conceptually from statistical significance.

5 A Typology of Robustness Tests

We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous.
Ronald Fisher (1966)

5.1 INTRODUCTION

The number and variety of possible robustness tests is large and, if tiny details and small differences matter, potentially infinite. The research project and its design as well as the degree of uncertainty about specific modelling assumptions determine the choice of robustness tests. Not every possible robustness test is relevant for each research project. To the contrary: each project requires a distinct set of tests, as the relevance of each test depends on the specificities of model uncertainty, the intended inferences, and the data structure.

The great variety and large number of tests appears bewildering. To cut through this diversity, we suggest in this chapter a typology of robustness tests. Specifically, we distinguish between five types: model variation tests, randomized permutation tests, structured permutation tests, robustness limit tests, and placebo tests. Model variation tests vary one specific aspect of model specification in a discrete way. Randomized permutation tests randomly select robustness test models from a large space of plausible alternative models. Structured permutation tests exhaustively select all plausible alternative models from a small space or select a few models in a structured way with the aim of representing the entire distribution of models in the larger space of plausible alternative models. Robustness limit tests ask which model specification, which could represent a model misspecification, renders the baseline model estimate non-robust. Placebo tests either replace the dependent variable with a placebo variable to test that the variable of interest has no effect under conditions in which no effect is to be expected, or replace the treatment variable with a placebo variable to

test that this placebo variable has no effect. Alternatively, researchers assume that the baseline model is correctly specified in a certain dimension and placebo tests intentionally introduce a model misspecification given this assumption to test whether the baseline model estimate of the variable of interest remains robust despite the introduction of a model misspecification. Placebo tests of this kind represent the exception to the general rule that robustness test models should not be implausibly specified.

5.2 MODEL VARIATION TESTS

Model variation tests are as old as regression analysis. In model variation tests, researchers change their baseline model in discrete ways. The first scholar who added a control variable to his or her baseline model was probably the first person ever to conduct a robustness test.

Most researchers conduct this type of robustness test – usually without referring to it as robustness test at all. Yet, model variation tests go well beyond the addition or removal of control variables. They are flexible and can be applied to all dimensions of model uncertainty. In addition to adding and removing explanatory variables, it is possible to change the operationalization of the variable of interest and the controls, the sample, the functional form, to add or remove conditionalities, to change the specification of structural change, dynamics, spatial dependence, and so on.

Model variation robustness tests can be specified so that the baseline is nested in the robustness test model, so that the robustness test model is nested in the baseline model or so that the two models are non-nested. Nestedness requires that the baseline (robustness) test model is a special or constrained case of the robustness (baseline) model. For example, the baseline model might opt for greater simplicity by estimating a linear effect of variable x, whereas a robustness test model might add the square of the variable to allow for a non-linear (quadratic) effect and greater complexity. In this case, the baseline model is nested in the robustness test model, which contains the baseline model as a special case. Conversely, the baseline model might opt for greater complexity by estimating a conditional effect of variable x, whereas a robustness test model might estimate an unconditional effect, thus exploring the robustness of baseline model estimate to a simpler robustness test model that is nested within the baseline model. Lastly, the two models are non-nested if neither represents a special or constrained case of the other. For example, scholars might operationalize democracy in the baseline model with one measure and explore the robustness of the estimate with another measure of democracy in a robustness test model.

Examples of model variation tests in the literature abound. For example, in one of the most extensive early robustness tests, Tucker, Pacek, and

Berinsky (2002) use alternative survey questions, re-code the answer categories, and estimate models that exclude "don't know" answers to analyze attitudes towards EU membership in transition countries. Carey and Hix (2011: 389) are concerned about their arbitrary decision of a functional form: "We do not know whether some other functional form might describe the shape of the diminishing returns even better." They therefore replace their baseline specification by robustness test models adding first the squared, then the squared and cubic terms. Gibler and Tir (2010: 959) replace their baseline model autocracy–democracy threshold by both higher and lower thresholds.

A popular model variation robustness test seems to be to split the sample at a defined line. Muckherjee and Singer (2010) use this test in their analysis of the influence of the IMF on capital account liberalization. Likewise, Nielsen et al. (2011) split their sample into small and large conflicts. In an interesting study, Boix (2011: 819) analyzes the influence of democracy on per capita income using a structural equation model with instruments. As a robustness test, he varies the instruments, ranging from trade share to initial income ratio times the time trend. Though instrumental variables models estimate local effects and cannot be expected to give identical results, Boix's estimates at least all point in the same direction.

As these examples demonstrate, model variation tests are best suited for model uncertainty with a small set of discrete plausible alternatives. If, for example, a variable can be plausibly operationalized in two ways, the design of the model variation robustness test is fairly straightforward: use both operationalizations. In practice, these situations occur, but they are rare. For the majority of specification choices, a larger number of plausible alternatives exist and sometimes this number is very large or even infinite, in which case other types of robustness tests become attractive. Still, even in this case it is possible to test the robustness of the baseline model in comparison with the most common, the most plausible or the most drastic of alternative specifications. The argument for the most plausible test is simple, albeit a bit tautological: different potential robustness test model specifications differ in their degree of plausibility, and researchers should opt for the test that appears to be the most plausible. An alternative strategy employs the most common alternative specification. The advantage here is that the majority of peers will find the robustness test relevant. The most drastic specification is one that, in expectation, puts most strain on the baseline model, i.e. the test that offers the highest ex-ante likelihood to result in non-robust findings. Care must be taken not to choose a model that is not plausible since models known to be misspecified are not valid robustness tests. Despite these options for employing model variation tests even in situations where the number of plausible specifications in the

uncertain dimension is large, other robustness test types become more attractive. The next three types can all deal with a large number of alternatives.

5.3 RANDOMIZED PERMUTATION TESTS

Model uncertainties for which a large number of alternatives exist can be dealt with by either randomized or structured permutation tests. We refer to the former as randomized permutation tests because the uncertainty of a model specification is dealt with by randomly selecting a limited number of specifications from a larger set of potential specifications (the relevant model space) for the same specific dimension of model uncertainty. The number of random draws and of model iterations must be large enough to represent the relevant model space.

The challenge for randomized permutation tests lies in the definition of relevant model space. Care must be taken that only plausibly specified models are included by ex ante restricting the space to an exclusive model space or by minimizing the impact of implausibly specified models on the robustness analysis via ex-post evaluation of model specification (see Plümper and Traunmüller 2016). If the model space cannot easily be restricted either ex ante or ex post, the results from randomized permutation robustness tests become difficult to interpret. Findings that cast doubt on the robustness of estimates may be due to lack of robustness or due to the inclusion of implausibly specified models into the model space. The definition of the model space forms the Achilles heel of randomized permutation tests.

The problem of defining the model space becomes apparent in the best known randomized permutation test: Leamer's (1978) sensitivity analysis. Practically all of Leamer's followers have applied sensitivity analysis to the choice of explanatory variables, in which regressors are selected from a large set of possible variables via randomized permutation, though some variables are always included in order to limit the overall model space. In Leamer's original formulation, robustness requires that all estimates have the same sign and all estimates are statistically significantly different from zero. Since early sensitivity tests of economic growth theories (Levine and Renelt 1992) demonstrated that few variables pass this extreme bounds test, Xavier Sala-i-Martin (1997: 179) argues that extreme bounds analysis "is too strong for any variable to pass it: if the distribution of the estimators of βz has some positive and some negative support, then one is bound to find one regression for which the estimated coefficient changes sign if enough regressions are run. Thus, giving the label of non-robust to all variables is all but guaranteed." Instead, he suggests measuring robustness by the density

function of the weighted model averaged estimates. If 95 percent of this density function lies to either side of zero, the effect of a variable can be considered robust.[1] He suggests two variants of model weighting depending on whether the distribution of estimates is assumed to be normal or not normal, both based on the integrated likelihood of estimated models.[2] As Sala-i-Martin (1997: 180) himself admits, such goodness-of-fit weights need not be a good measure of the quality of model specification. In fact, clearly misspecified models can exhibit high goodness-of-fit, for example due to variables being endogenous. Consequently, we regard this kind of sensitivity analysis as failing the requirements for a valid robustness test.

Randomized permutation tests are not limited to sensitivity testing of the set of explanatory variables. For example, researchers may explore the robustness of a baseline model in the presence of measurement uncertainty. Researchers can randomize the extent of artificial measurement error injected into variable values across the plausible potential range of error. Instead of eliminating measurement error from the data, this test explores whether measurement error of a defined maximum magnitude affects the robustness of estimates and potentially invalidates inferences based upon them. The bounds of artificial measurement error should not be larger than the largest measurement error that likely occurs in reality. While this may sound cryptic, social scientists usually have information that allows them to justify the bounds of measurement error. For example, the January 2010 Haiti earthquake placed a plausible limit on measurement uncertainty in respect of mortality from large quakes in locations where reliable measurement appears difficult.

The split sample test provides another example of a randomized permutation test. The test aims at exploring the internal validity of causal homogeneity typically assumed in baseline models. The sample is randomly split in two halves and each observation in each half-sample is duplicated. If causal homogeneity holds, the baseline model estimate based on the causal homogeneity assumption will be robust to the estimates from these two split samples. While a single split sample estimate does not mean much, 1,000 split sample estimates can cover the relevant model space. This raises the question of how to assess robustness across the 1,000 estimates. Contrary to Leamer's sensitivity test, none of the models randomly selected by the split sample and the artificial measurement error tests is implausibly specified.

1 In a re-analysis of Levine and Renelt (1992), Sala-i-Martin (1997) found only one variable robustly related to growth based on extreme bounds. Using a model averaging approach, he identified 22 of the 59 tested variables as robust.

2 In Sala-i-Martin, Doppelhofer, and Miller (2004), the authors move to what they call a Bayesian averaging of classical estimates approach for model weighting.

We therefore suggest averaging the unweighted estimates of the permutations to assess robustness.

5.4 STRUCTURED PERMUTATION TESTS

In contrast to randomized permutation tests, structured permutation tests deal with uncertainty of a specific dimension of model specification by selecting either all alternative specifications or a limited number of specifications from a larger set of potential specifications according to some guiding principle. In other words, structured permutation tests are non-randomized and cover the model space either exhaustively or selectively but in a structured fashion.

The requirements for structured permutation tests are similar to the requirements of their randomized cousins: the model space and the robustness criterion have to be defined, but a rule to select models replaces the randomization algorithm. The same challenge to appropriately restrict the model space applies. In terms of assessing robustness, since with structured permutation tests the number of test models will typically be small to moderate, ρ can be computed and reported for each one.

Exploring the entire relevant model space in a systematic fashion can be done in either one of two ways. If the number of plausible alternative models is small, all models of the model space should be selected. For example, a variable x can be conditioned by more than one potential factor or analysts can relax the functional form assumption by estimating polynomial models of increasingly higher order up to a certain degree. If the model space becomes large because minuscule variations are possible, researchers have to make a discrete choice of plausible models – a choice that represents the entire distribution of plausible models or, put differently, the relevant model space.

With a large model space, the question becomes whether the space can really be represented by structured selection. If this is questionable, randomized permutation might be preferable. Covering the entire model space can quickly become computationally infeasible. Consider the example of an exhaustive structured variant of Leamer's sensitivity analysis. Rather than randomizing models, one would either estimate all possible combinations of explanatory variables (inclusive model space) or estimate all combinations which are not considered to be misspecified (exclusive model space). The number of possible permutations equals 2^k–1, where k is the number of considered explanatory variables. Thus, if 20 variables are known to potentially influence an outcome, the number of possible models reaches a shade above 1 million. If the number of potential explanatory factors doubles to 40, the number of possible models increases to 1,099,511,627,775. Assuming that a single estimate takes

0.1 second on average, the sensitivity test would need almost 3,500 years to finish on a single computer and still 3.5 years on 1,000 efficiently clustered computers. While randomized permutation tests have their advantages when the model space becomes very large, structured permutation tests have their strengths with relatively small model spaces or if the selected structured permutations represent the entire distribution of models.

A frequently used structured permutation robustness test is based on the jackknife method, which drops one unit or one group of units of analysis at a time, thereby exploring the extent to which estimates depend on the inclusion of single units or groups. They thus indicate a lack of internal validity or – potentially – causal heterogeneity. Jackknife tests are popular. Egorov, Guriev, and Sonin (2009), Lipsmeyer and Zhu (2011: 654) and Martin and Swank (2004, 2008) exclude one country at a time in their analyses. As another example of a structured permutation robustness test, in an excellent robustness test section Scheve and Slaughter (2004: 672) gradually expand the sample size and move away from what they consider to be the sample "for which the theoretical framework most directly applies."

A common structured permutation test relates to the aggregation of a continuous or categorical variable into two or more sub-categories. In these cases, the "true" cut-off points are unknown. Accordingly, robustness tests can vary the chosen cut-off point to explore whether results are independent of the threshold. Take the *polity2* measure of democracy as an example. Besley and Reynal-Querol (2011) use a dichotomous distinction between autocracies and democracies setting the cut-off point at 0, i.e. democracies are defined as scoring 1 or higher on the scale that runs from –10 to 10. Cut-off points are arbitrary and thus controversial, and this example is no exception. Other authors prefer a higher threshold. Fearon and Laitin (2003), for example, use a cut-off point of 5, others use the even higher score of 6 (Bigsten 2013: 31) as threshold. A structured permutation test uses *all* plausible cut-off points.

As a final example of structured permutation tests, Michael Bailey (2005) employs a simple but appealing "varying control group approach" to analyze the migration response of poor single mothers who receive the treatment of a specific welfare benefit. Recognizing that his research design "requires that I include in the sample a 'control group' that is not eligible for welfare but otherwise resembles the 'treatment group' of poor single mothers" (Bailey 2005: 127), he follows two previous studies in using three different control groups of people who were not eligible for the particular welfare benefit. In his baseline model he uses poor women, but changes the control group to, separately, poor men without children and to married women with children. He chooses these groups to explore the robustness of findings (Bailey 2005: 127): "No group perfectly matches

the welfare population, but all match in some way the skill profiles and economic circumstances of poor single mothers. Using multiple specifications should increase confidence in the robustness of the results."

5.5 ROBUSTNESS LIMIT TESTS

The vast majority of robustness tests ask whether the baseline model estimates remain robust to plausible changes in model specification. However, not all robustness tests seek to check the degree of robustness given plausible alternative specifications. Robustness limit tests, which are inspired by "Rosenbaum bounds" (Rosenbaum 2002), though others like Frank (2000), Pan and Frank (2003), and Frank and Min (2007) have independently developed similar ideas, instead ask by how much the specification of a model needs to change to render the baseline model estimate non-robust.

Consider, as an example, the choice of functional form. Rather than analyzing whether the estimated effect is robust to a change in functional form of the variable of interest, researchers can ask to what degree the functional form needs to change to render the baseline model estimate not robust. Robustness limit tests work particularly well with model specifications that can be altered in a continuous fashion. For example, a set of estimates can explore what the correlation between a random placebo variable and the variable of interest needs to be to render the estimated effect non-robust.

Rosenbaum (1991, 2002) develops his idea of the bounds of hidden bias based on the example of the effect of smoking on lung cancer. His analysis draws on two previous works: Cornfield, Haenszel, Hammond, Lilienfeld, Shimkin, and Wynder (1959) were the first to use the logic of a bounds test. In an argument worth citing despite its convoluted English, they claim (p. 194):

> If an agent, A, with no causal effect upon the risk of a disease, nevertheless, because of a positive correlation with some other causal agent, B, shows an apparent risk, r, for those exposed to A, relative to those not so exposed, the prevalence of B, among those exposed to A, relative to the prevalence among those not so exposed, must be greater than r. Thus, if cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, but this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone X-producers among cigarette smokers must be at least 9 times greater than that of nonsmokers. If the relative prevalence of hormone X-producers is considerably less than ninefold, the hormone X cannot account for the magnitude of the apparent effect.

Based on this logic, the authors came to conclude that the evidence for smoking causing cancer is "beyond reasonable doubt," just as a Study

Group appointed by the National Cancer Institute, the National Heart Institute, the American Cancer Society, and the American Heart Association had proclaimed two years prior.[3]

The second study on which Rosenbaum relies is an analysis of matched pairs (Hammond 1964). This analysis identified 36,975 heavy smokers and nonsmokers who were (almost) identical in respect of age, race, time of birth, residence, occupational exposure to dust and fumes, religion, education, marital status, alcohol consumption, sleep duration, exercise, nervous tension, use of tranquilizers, current health, history of cancer and heart disease, stroke, and high blood pressure. Of these pairs 12 nonsmokers and 110 heavy smokers died of lung cancer. The lung cancer mortality rate among heavy smokers was thus below 0.3 percent (p=0.002975), but still more than 9 times higher than the lung cancer mortality among nonsmokers, which stood at 0.000325. The probability that the gap is random if we had a perfect random draw from a population was 0.0001.

Rosenbaum uses this information for what he calls a sensitivity test. He asks by how much an unobserved lung cancer propensity factor of heavy smokers has to exceed that of non-smoking individuals to render the causal effect of smoking statistically insignificant. Rosenbaum (2002: 114) concludes: "To attribute the higher rate of death from lung cancer to an unobserved covariate u rather than to an effect of smoking, that unobserved covariate would need to produce a sixfold increase in the odds of smoking, and it would need to be a near perfect predictor of lung cancer."

As Rosenbaum demonstrates, robustness bounds can be computed analytically. However, it is possible and in many cases easier to conduct robustness limit tests. These tests gradually increase the degree of "pressure" on the baseline model. It works best where researchers have a clear idea about a potential model misspecification that is difficult or impossible to correct, for example because of potential confounders that are unobservable or unobserved due to measurement problems. To stay in the example of the effect of smoking on lung cancer, the correlated artificial variable test proposed in chapter 9 on the choice of explanatory variables plays with two "moving elements" of an artificial variable that "by design" leads to cancer: the probability that a latent variable causes cancer and the correlation between this variable and smoking. As both

3 This logic depends on numerous untested assumptions. Most importantly, the authors assume that treatments are either present or absent and if absent either equally strong or, if strength matters, the strength is irrelevant for the causal effect. For example, the true causal effect may well be that smoking is correlated to an intensified production of hormone X, which has to exceed a certain threshold to stimulate the occurrence of lung cancer. Hence, it may well be that smokers only have 30 percent higher production of hormone X, but they may be nine times as likely to pass the threshold required for cancer.

factors go up, the predicted effect of smoking declines (while the uncertainty of an effect of smoking increases). In addition to the robustness limit scholars might be interested in the uncertainty of the estimate.

Robustness limit tests have drawbacks. Most importantly, the interpretation of results is not straightforward. Interpretation is easy if and only if researchers have sufficient information to conclude that the model that reaches the robustness limit is misspecified. Rosenbaum argues exactly this: the effect of a lung cancer phenotype correlated with smoking necessary to overturn the effect of smoking would have to be too large to plausibly exist. In other cases, it remains contested what to make of the robustness limit test: does the test suggest that the baseline model estimate fails the robustness test or does it instead suggest that the baseline model estimate is robust because the model that reaches the robustness limit is misspecified?

As a corollary, if the limit is known beyond which a model becomes clearly misspecified, analysts have two options: firstly, they can use a randomized or structured permutation test and assess the robustness of the baseline model estimate within the boundary. Alternatively, they can go beyond this boundary to study where the robustness limit lies, knowing that models which reach the robustness limit are misspecified. Take the example of measurement uncertainty. If the bounds of plausible measurement error can be established, a randomized or structured permutation test can explore robustness within the boundary. Alternatively, a robustness limit test can find the extent of measurement error that needs to be injected to render the estimate non-robust and can then assess whether this extent of measurement error falls within the boundary of plausible measurement error.

Robustness limit tests are rare in the social sciences. Imai, Keele, Tingley, and Yamamoto (2011) propose such tests as part of their methodological contribution on how to learn about causal mechanisms in observational and experimental studies:

> Given that the identification of causal mechanisms relies upon an untestable assumption, it is important to evaluate the robustness of results to potential violation of this assumption. Sensitivity analysis provides one way to do this. The goal of a sensitivity analysis is to quantify the exact degree to which the key identification assumption must be violated for a researcher's original conclusion to be reversed.
>
> (Imai et al. 2011: 774)

We agree with the authors that robustness limit tests always explore robustness in one specific dimension of model misspecification, not in other or all dimensions. Imai et al. (2011: 774) correctly warn readers about the limitations of their robustness limit test:

> Although sensitivity analysis can shed light on whether the estimates obtained under sequential ignorability are robust to possible hidden pretreatment confounders, it is

important to note the limitations of the proposed sensitivity analysis. First, the proposed method is designed to probe for sensitivity to the presence of an unobserved *pretreatment confounder*. In particular, it does not address the possible existence of confounders that are affected by the treatment and then confound the relationship between the mediator and the outcome.

Robustness limit tests are powerful in directing future research. Assume a researcher wishes to explore whether controlling for time-invariant "unobserved heterogeneity" renders the baseline model estimate non-robust. One option is a model variation test that includes unit fixed effects. As chapter 9 shows, this strategy may have severe drawbacks, including testing a hypothesis that differs from the theoretically derived hypothesis as well as potentially inappropriately throwing away variation that belongs to the estimated effect – thereby throwing out the baby with the bath water. As an alternative, we propose a between-variation test that can find the percentage of between-variation that needs to be dropped to render the baseline model estimate non-robust.[4]

5.6 PLACEBO TESTS

Up to this point, we have argued that, other than for robustness limit tests, models used in robustness tests must be plausibly specified. Placebo tests are different. To understand why, we first make a detour into medical trials.

Placebo analyses are most commonly used in experimental research with human participants. Placebo-controlled studies are a way of testing a medical therapy in which, in addition to a group of subjects that receives the treatment to be evaluated, a control group receives a placebo treatment specifically designed to have no real effect. Placebos have to be employed in blinded trials where subjects do not know whether they are receiving real or placebo treatment. Often, the experiment includes a third group that does not receive any treatment at all.

The placebo treatment aims at accounting for the placebo effect. This effect is caused by the treatment act – the psychological effect of receiving attention from health care professionals – rather than by the proper treatment, that is, a substance or procedure that supposedly has an effect. Typically, social scientists define the treatment effect as the net effect of the observed change in the treated group minus the observed change in the placebo group. If it were ethically possible and if an appropriate placebo

4 Additional research can analyze which factors usually assumed to be time-invariant such as history, institutions, culture, and geography can account for this between-variation and whether the estimated effect is robust with these additional time-invariant control variables included.

existed, researchers could study the effect of smoking on lung cancer by giving the treatment group cigarettes and the control group placebos which are identical to cigarettes in all dimensions except one: they do not contain carcinogenic substances.

Placebo robustness tests in regression analysis of observational data are similar, but not identical to placebo analysis in experiments. They come in two variants. In the first variant, researchers intentionally "misspecify" the model by either switching the dependent variable to one for which the variable of interest is expected to not have an effect (i.e., becomes a placebo variable) or by keeping the same dependent variable but switching the treatment variable to a placebo variable which is expected to not have an effect on the original dependent variable. Instead of testing the robustness of the baseline model's estimate, a placebo test asks whether the placebo variable that replaces the treatment variable does or does not have an effect on the original dependent variable or whether the variable of interest loses its explanatory power if we replace the original dependent variable with a placebo variable.

In the second variant of placebo robustness tests, researchers make a specification change that, under the assumption that the baseline model is correctly specified in a certain dimension, represents a misspecification. For example, scholars can add a placebo variable to the estimation model that in expectation does not affect the robustness of the baseline model estimate for the variable of interest. In this variant, analysts continue to estimate the degree of robustness similar to other types of robustness tests. Placebo robustness tests of the second variant have to be permutation tests. For example, it does not make sense to add a single randomly distributed placebo variable to the model since by pure chance it could affect robustness. Rather, a large number of permutations are needed – we recommend at least 1,000 permutations if computationally feasible – to render such chance impact unlikely.

For placebo tests of the first variant a single model run will not be conclusive either since pure chance can suggest relevant effects where none exist and suggest no effects where they do exist. However, it may not be feasible to undertake permutations since there may exist only one alternative dependent variable for which the treatment variable should have no effect or only one option to switch the treatment variable into a placebo variable. For the same reason, placebo robustness tests of this variant are somewhat limited since they require either the existence of an alternative dependent variable which is independent of the variable of interest or the possibility to transform the variable of interest, the treatment variable, into a placebo variable.

Folke, Hirano, and Snyder (2011) provide a clever example of the first variant of a placebo robustness test. While the baseline model demonstrates

that parties in power were able to use patronage to improve their chances of winning at later elections, their placebo test demonstrates that no such relation exists for prior elections. Similarly, Gerber and Huber (2009: 415) show that partisanship has no placebo effect "under conditions where our model predicts partisanship and consumption should be unrelated."

Occasionally, placebo tests can be a part of a structured permutation test in which a treatment variable becomes more and more a placebo variable. Consider the effect of democracy on an outcome where analysts use a dichotomous cut-off point to distinguish autocracies from democracies, which raises the question where to set the cut-off point. Of course, no "true" cut-off point exists. Democracy is a latent variable, proxy variables will be measured with error, and no consensus exists on where the true cut-off point is, not least because even in theory regime type falls along a spectrum instead of into a clean dichotomy. At the same time, certain cut-off points are not plausible and can thus function as placebo tests. For the example of the *polity2* measure of democracy which runs from –10 to 10, reducing the cut-off point to lower values represents a structured permutation test. At some point, further decreasing the cut-off point should result in the test finding the baseline model estimate to be less and less robust, just as it is designed to be.

A prime example of a placebo test of the second variant adds a random variable. It is of no interest whether the estimate for the placebo variable turns out to be statistically significant or not. By construction, random variables become statistically significant at the 95-percent level in roughly 5 percent of cases because the placebo variable is correlated to the random deviation of the errors from the assumed normal distribution of residuals. If the random placebo variable turns out to be significant in a substantially higher share of cases, the placebo variable is likely to be correlated to systematic structure in the residuals. This finding suggests some form of model misspecification though identifying the type of misspecification based solely on the structure in the residuals is not normally possible.

Placebo robustness tests become more informative when researchers do not add a purely random variable, but give the placebo variable a certain structural property to account for a potential specification error that is non-existent in the baseline model. Assume that researchers believe that they have specified their baseline model correctly in a certain dimension. For example, they believe that they have included relevant control variables that sufficiently account for time-invariant unobserved heterogeneity or that they have adequately modelled dynamics and temporal heterogeneity. Introducing time-invariant placebo variables or strongly trended placebo variables should not affect the stability of the baseline model's estimate for the variable of interest. However, in cases like these, placebo variables may

become statistically significant in more than the expected 5 percent of permutations. In both cases, the impact of the inclusion of the placebo variable on the effect of the variable of interest matters.

In chapter 14 on spatial correlation and dependence we propose a structured spatial placebo test. Assume that the baseline model tests a theory of spatial dependence. As we have argued elsewhere (Neumayer and Plümper 2016a), the weighting matrix models the causal mechanism of a theoretical argument for spatial dependence. The connectivity variable employed in the weighting matrix and its specification must capture this mechanism. The placebo robustness test replaces the theoretically informed connectivity variable in the weighting matrix by a random variable. Since the weighting matrix is multiplied with the spatially lagged dependent variable, the spatial lag becomes statistically significant in more than 5 percent of permutations – not least because it will be correlated with the spatial-lag variable that employs the theoretically informed connectivity variable. Nevertheless, if the baseline model is not obviously misspecified, the effect of the spatial-lag variable based on the theoretically informed connectivity variable will remain robust to adding this spatial placebo variable.

5.7 CONCLUSION

The degree to which robustness tests contribute to the validity of inferences derived from regression analysis of observational data depends on the extent of uncertainty about model misspecification, on the theoretical justification and design of robustness tests as well as on the type of robustness tests chosen for dealing with this uncertainty. The vast majority of social scientists rely on simple model variation tests, which seem to have the advantage to scholars that they can be done easily and carefully selected to not render the baseline model estimates non-robust. These tests merely aim at providing additional arguments for journal editors and reviewers to accept a manuscript.

This standard practice stands in remarkable contrast to the best work in empirical social science. Indeed, an increasing number of authors conduct intelligently designed, increasingly complex robustness tests of multiple types, going well beyond simple model variation tests. To us, robustness tests are an essential element of causal inference based on regression analysis of observational data, where researchers cannot guarantee the correct specification of the baseline model. Yes, replacing one variable operationalization by another constitutes a robustness test as do other model variation tests, but randomized permutation tests, structured permutation tests, robustness limit tests, and placebo tests, if well specified, offer deeper insights into the validity of causal inferences. If used optimally, robustness

tests allow scholars to improve the validity of causal inferences and to identify their limits, for example, the limits of generalizability. However, to seize this great opportunity, social scientists need to take robustness testing seriously and stop misusing them as a means to increase their publication chances. In best practice, robustness tests are no longer part of a publication strategy, they become an essential part of the research strategy.

6 Alternatives to Robustness Testing?

No amount of experimentation can ever prove me right.
Attributed to Albert Einstein

6.1 INTRODUCTION

Robustness tests provide social scientists with the means to improve the validity of statistical inferences based on the analysis of observational data. Regressions analyses of observational data will remain at the heart of the methodological toolkit for quantitative social science research though they are fraught with model uncertainty.

Robustness tests are not the only methodological option on offer for solving the problem of model uncertainty. Methodologists have developed alternative methods for analyzing observational data. They have also suggested that specific research designs strongly improve the probability of valid inferences from observational data. In addition, there are many who believe that the analysis of observational data cannot be free from ambiguities and that social scientists should simply follow the experimental turn in the sciences.

In this chapter we argue that these alternative methodologies offer no alternative to robustness testing. In short, an alternative methodology that allows researchers to formulate inferences that are valid with certainty does not exist – at least not if researchers intend to go beyond mere description. This provides the most fundamental reason why all methodologies require robustness testing: if no single research design, estimation procedure or analytical technique allows the derivation of perfectly valid inferences then every design, procedure or technique warrants subjecting its results to plausible alternative specifications to explore whether these generate sufficiently similar (robust) estimates.

Our argument may surprise those who have fallen under the spell of identification techniques and, particularly, social science experiments, erroneously believing that these alternative designs are not only unambiguously

superior to regression analyses of observational data but also identify the true causal effect of a factor with certainty and therefore obliterate the need for robustness testing.

We concentrate here on what we regard as the most important alternatives to regression analyses. For each alternative methodology that we discuss, we identify the – in our view at least – most important specification uncertainties it suffers from. We make no pretense of being comprehensive in the choice of specification uncertainties we identify. Our objective is merely to persuade those readers who remain skeptical that indeed no methodology is free of specification uncertainty and consequently requires robustness testing for improving the validity of inferences based on these methodologies.

We start by staying within the realm of regression analyses and discuss why comprehensive model specification tests or model selection algorithms cannot result in identifying or at least sufficiently approximating the one "true" model. Acknowledging that it is not possible to find this model and that estimates based on any model are sensitive, some seek solace in averaging across a very large number of models. We then move to more recent methodological advances in the form of research designs based on the selection of cases (regression discontinuity, matching, and synthetic control), effect isolation via instrumental variable estimation and social science experiments.

6.2 MODEL SPECIFICATION TESTS

Econometricians have long since developed econometric tests aimed at detecting model misspecification. The hope is that a battery of tests will allow researchers to find the true model, or at least get sufficiently close to it. Econometric tests fall into three categories:

First, relative tests which say nothing about the absolute quality of a model, but compare two or more models. Yet, the best-fitting misspecified model remains misspecified. Often, as in the Hausman test, a comparison draws on whether the estimates from an estimator assumed to be consistent and an estimator assumed to be inconsistent are significantly different. Implicitly, these tests depend on the absence of other model misspecifications that render the "consistent" estimator inconsistent, e.g. time-varying omitted variables or misspecified dynamics in the classical Hausman test that compares a fixed-effects to a random-effects specification.

Second, model fit tests that assess the overall quality of a model with adjusted R-squared statistics, F-tests, and chi-squared tests as well

as more complex "goodness-of-fit" measures. Often, the indicators become inflated by econometric patches which do not belong into the true model but are included in the hope that they account for omitted variables the researcher cannot or does not want to control for, e.g. unit fixed effects. This invalidates the goodness-of-fit test. Moreover, strong and implausible assumptions are required for making the claim that better-fitting models are better specified. For example, proponents of this claim would have to argue that over-fitting is impossible and to claim that better-fitting models are always better specified. Unfortunately, goodness-of-fit can be improved much more easily than model specification.

Third, model fit tests that analyze the residuals for structure. From our perspective, these tests are better suited than other tests to evaluate the quality of a model specification. We advocate the careful use of such tests for finding a baseline model. However, these tests will not find the "true" model and whether they always improve a model specification is debatable.

The main problem with econometric tests, even those that analyze the residuals for structure, is that they may signal the existence of a problem, but they fail to identify the problem's cause. For example, Ramsey's specification test is usually employed for detecting functional form misspecification, but a rejection of the null hypothesis can indicate a large number of other specification errors. A correlation between the regressors and the residuals can be caused by too many regressors, too few regressors, wrong regressors, wrong functional form, wrong interaction effect, and wrong functional form of interaction effect. The same holds for other specification tests (McAleer 1994: 330f.).

But what about a test contest of plausible models? Researchers could develop a set of plausible model specifications and then subject each of these model specifications to those tests that appear to have sufficient power to identify misspecified models. In principle, there is nothing wrong with this idea. However, the problem is that econometric tests remain inconclusive: whatever the specification test, more than one model will pass the test. Even if we could interpret a set of tests as a single model specification test that empirical models need to pass, there would still be more than one model specification that passes all econometric tests simultaneously. As Peach and Webb (1983: 697) already demonstrated in the 1980s, "econometric testing as sole criteria for discriminating among competing (...) models is inconclusive."

Accordingly, we currently see no possibility of conclusively discriminating between plausible models, that is, models that are not obviously

misspecified. As a consequence, in order to improve the validity of their inferences researchers have to use robustness tests to analyze whether the estimated effects of these plausibly specified models are sufficiently similar.

6.3 MODEL SELECTION ALGORITHMS

The idea that model specification tests can find the "true" specification and that the process of finding "the truth" can be handed over to a computer program has been championed by David Hendry and his disciples. His "testing-down approach" (Krolzig and Hendry 2001; Hendry and Krolzig 2005) starts with a general statistical model that is "congruent" with the dataset: It "matches the data evidence on all the measured attributes" (Campos, Ericsson, and Hendry 2005: 7). The objective is to reduce the complexity of this model as much as possible by eliminating statistically insignificant variables. Here, "as much as possible" means that the more specific model must pass specification tests and must be congruent with the dataset. Specification tests are, in other words, used to establish the congruence of the general model and are repeatedly used to discard invalid reductions of the general model.

The general-to-specific approach has been criticized, because it "would require an enormously complex exercise, with a complete model of the joint distribution of all variables, allowing for non-linearities, heteroscedasticity, coefficient drift and non-Gaussian errors" (Hansen 1996: 1411). Magnus (1999: 61–62) similarly argues that the testing-down approach "does not work. If you try to estimate such a large model, which has everything in it that you can think of, you get nonsensical results."

Hendry and his followers are aware of this critique (Campos, Ericsson, and Hendry 2005: 6), but believe their approach can recover the so-called local data-generating process (Hendry 2002: 599). Yet, it remains unclear whether a "local data-generating process" exists at all and, if it does, whether the convenient error structure that researchers assume to exist is exactly matched in the real data-generating process.

In the specification search from the general to the specific model two principal errors can and indeed are likely to occur: variables are eliminated that should be retained in the model, while other variables are retained even though they should be eliminated. Its supporters contend that their highly sophisticated multiple-path search and testing algorithms minimize the risk of both errors and provide Monte Carlo evidence to this effect (Hoover and Perez 1999, 2004). Yet, there is no guarantee that relevant variables are retained and irrelevant ones eliminated, particularly not if variables are correlated with each other as they typically are and if models are misspecified.

In consequence, the results of testing-down approaches depend crucially on arbitrary decisions about model misspecification. For example, it is not possible to simultaneously include all possible functional forms, conditionalities, and dynamic specifications into the "general" model. Likewise, it is not possible to simultaneously include different samples into a testing-down experiment so that more than one plausible model and more than one effect estimate emerge as a result. In fact, testing-down approaches are hardly suited to deal with more than the selection of right-hand-side variables and perhaps functional form assumptions. In conclusion, model selection algorithms cannot address all dimensions of model uncertainty and cannot solve those dimensions they do address.

6.4 MODEL AVERAGING

Leamer (1978) pioneered the idea of basing inferences not on the results from a single model but on a potentially very large number of models. Methodologies which draw conclusions from multiple estimates require an aggregation rule. In Leamer's original analysis a single model exercised a veto right over inferences, as a particular result had to pass the test – of statistical significance in Leamer's definition of robustness – in every single model.

Followers of Leamer have adopted statistical significance as the criterion of robustness, which runs counter to our definition of robustness as stability in effect strength. More importantly, however, for the purpose of this chapter is that Leamer's disciples moved away from granting a single model veto right and instead adopted what is known as model averaging. Model averaging serves as a label for very different techniques; it can be combined with numerous inferential rules and with virtually infinite definitions of the model space included across which results are averaged. For example, our randomized and structured permutation tests also employ model averaging techniques within robustness testing. In this sense already, model averaging is an integral part of robustness testing and not an alternative. Many proponents of model averaging agree, including Bayesians. Montgomery and Nyhan (2010: 266), for example, suggest that Bayesian Model Averaging "is best used as a subsequent robustness check to show that our inferences are not overly sensitive to plausible variations in model specification."

This does not mean that model averaging is not regarded by some as an alternative technique to robustness testing as proposed in this book. Model averaging techniques require the following specification decisions:[1]

1 This discussion follows Plümper and Traunmüller (2016).

1. A definition of the parameter of interest and its computation;
2. A definition of the model space: the set of all models potentially included;
3. A selection rule that draws models from the model space;
4. A stop function that determines the number of estimated models;
5. An aggregation and weighting rule;
6. An inferential rule, which decides how to interpret the weighted mean of parameter estimates.

Consider model averaging for estimating the effect of educational attainment on economic growth as the parameter of interest. Even if researchers exclusively focus on one aspect of model uncertainty, namely the set of explanatory variables, the model space can become extremely large. The model space is usually constructed from the set of all combinations of explanatory variables that in the past have been used in growth regression studies, giving us at least 50 different variables. Including 50 variables into the set of explanatory variables from which to generate all potential permutations results in a model space of more than 1,000 trillion models. Yet, researchers who employ model averaging usually estimate only a few million models, that is, a very small fraction out of every possible model variant. A popular model selection rule has been suggested by Levine and Renelt (1992), who limit the model space by the 1+3+3 rule; the first variable is the variable of interest, then three variables are specified which are always included, and the final three variables are a random draw from the remaining 46 variables – thereby shrinking the model space to a little more than 15,000 models. Researchers have also used different weights, ranging from unweighted to the Bayesian or Akaike information criterion. Since model fit parameters tend to be correlated, the choice of a weight exerts considerably less influence than the decision to use weights rather than an unweighted mean of all estimates. Finally, at least in principle, scholars could use numerous different inferential rules. In practice, however, most researchers have followed Leamer and adopted statistical significance by looking at the share of the distribution around the (weighted) mean of all point estimates that crosses the threshold, usually of zero.

Model averaging lacks a clear intuitively plausible foundation. A model space of thousands, millions, let alone billions of models will include many utterly implausibly specified models. More importantly for our argumentation here, model averaging techniques rely on numerous model specification assumptions for which plausible alternatives exist – for example, on the functional form of an effect, the conditionality between variables, the definition of the population from which a sample is drawn, the definition of the model space, model selection rules, and the choice of

weights. As a consequence, model averaging techniques require subjecting their results to robustness tests for improving the validity of inferences based on this technique.

6.5 CASE SELECTION RESEARCH DESIGNS

The recent rise of identification approaches in the social sciences brought case selection techniques back into the social scientists' toolbox. When done well, a careful selection of cases has two consequences: on the one hand, the selection of cases increases the homogeneity of cases analyzed. The more homogeneous the cases, the fewer confounding factors need to be controlled for to make valid inferences. Ideally, cases become not just more homogeneous but identical in all but the relevant dimension (the treatment), in which case it becomes possible to directly compute the causal effect, distorted only by the difference in the average random errors between treatment and control group. On the other hand, however, selection reduces the number of cases and the types of cases included in the sample. Since real world cases are likely to be characterized by causal heterogeneity and context conditionality, as selection gets stricter and stricter the sample properties become increasingly different from the population properties.

Research designs that employ selection rules aimed at improving the internal validity of estimates include regression discontinuity, matching, and synthetic control. Regression discontinuity designs exclusively compare cases that only just met the criteria to receive treatment in the treatment group with cases that only just failed to receive treatment in the control group – by assumption both sets of cases ought to be very similar to each other provided subjects did not exercise control over whether they received treatment or not. Matching selects cases from a larger sample, matching treated cases to similar, and ideally otherwise identical, untreated cases on the values of observable variables, with all other unmatched cases being discarded. Synthetic control designs employ some weighting rule to artificially create a control case (or a set of control cases) that closely resembles a treated case (or a set of treated cases). Thus, rather than finding an identical or sufficiently close twin of a specific case, researchers synthetically produce an almost identical twin by taking shares of other cases.

Selection-based research designs can deal fairly well with two specification problems: observable confounders and their functional form. Yet, all other model uncertainties remain. For example, since it is not possible to match or create a synthetic case control based on unobserved confounders, these techniques provide no solution to unknown confounding factors, but only against potential misspecification of the functional form of known confounders (Sekhon 2007, 2009). Specifically, matching gives unbiased

results if and only if cases are matched based on the true model. If the matching algorithm excludes one or more variables which are included in the true model, matching estimates are biased.

Research designs that are based on case selection can only be generalized to a population that has properties identical to the selected sample. Unless the population of cases is characterized by strict causal homogeneity, the selected cases do not represent a random sample of the true population and the population to which results can be inductively generalized from the selected cases is therefore not the true population for which researchers seek to make causal inferences. Matching almost inevitably produces deviations of the sample from the population particularly in those parts of the population where the number of observations remains small so that perfect "matches" become unlikely. As a consequence, the matched samples will no longer represent the population. While it may be possible to "match" a sample in which, in principle, all permutations of treatments and conditionalities are represented, we are not aware of any attempt to produce such a matched sample in any study. Similarly, in regression discontinuity designs, the deviation of cases at the discontinuity threshold from other treated and untreated cases is likely to be non-negligible. The return on investment in higher education for somebody who almost failed to be accepted by a university ought to be smaller than the returns to the average student or the superstar among students.

It is questionable, in our view, whether it pays to trade off a potential increase in internal validity against a certain loss in external validity. In any case, robustness tests with different selected cases, possibly with entirely different analyses, are required to establish the extent to which the estimated effect can be generalized.

Beyond uncertainty about causal heterogeneity, selection-based research designs are also subject to other model uncertainties regarding the population, concept validity and measurement, dynamics, and spatial dependence. Consider spatial dependence: to analyze learning effects and externalities, a full sample is required. Missings and selection usually bias the results from spatial analyses. Accordingly, spatial dependence and selection-based research designs do not go well together. Selection-based research designs have to match cases based on their spatial dependence, regression discontinuities have to demonstrate that networks are identical on both sides of the discontinuity, and the synthetic case has to have identical ties and links to real cases. We are not arguing here that it is not possible to achieve this, but there must be a reason that no selection-based research design we have seen incorporates spatial dependence into the selection. In sum then, selection does not comprehensively solve the problem of model uncertainty. Selection-based designs quite successfully solve some

dimensions of model uncertainty but maintain and even exacerbate others. As a consequence, selection-based research designs warrant robustness tests as much as regression analyses.

6.6 EFFECT ISOLATION VIA INSTRUMENTAL VARIABLE ESTIMATION

According to econometric wisdom, instruments can be used to isolate the causal effect of a variable from the effect of unobserved confounders. As Morgan and Winship (2015: 291) claim:

> If a perfect stratification of the data cannot be enacted with the available data, and thus neither matching nor regression nor any other type of basic conditioning technique can be used to effectively estimate a causal effect of D on Y, one solution is to find an exogenous source of variation that affects Y only by way of the causal variable D. The causal effect is then estimated by measuring how Y varies with the portion of the total variation in D that is attributable to the exogenous variation.

Instrumental variable (IV) estimation gives an unbiased estimate of the effect of D on Y if

1. the instrument is perfectly correlated with the exogenous part of the covariation of D and Y,
2. the instrument is perfectly orthogonal to the endogenous part of D,
3. the instrument is perfectly orthogonal to any other model misspecification.

If these conditions are not satisfied, then the estimated instrumented effect of D on Y is biased: it differs from the true effect of D on Y.

Since conditions 2 and 3 need to be assumed and cannot be tested for or taken for granted, IV estimation does not solve uncertainty about model specification. Robustness tests are required to establish the stability of IV estimates for different plausible model specifications. For example, IV estimation does not help against dynamic or spatial misspecification: the instrument may suffer from measurement error which can be correlated to the endogenous part of the covariation between D and Y or with other misspecifications, such as those resulting from population uncertainty or sampling uncertainty. If an endogenous variable is conditioned by other factors, there will be great uncertainty whether an instrument for the endogenous variable closely mirrors the conditionality structure. All of these uncertainties warrant robustness tests.

Condition 1 is never fulfilled. Instrumental variable designs therefore do not "identify" the average population treatment effect, but only the treatment effect for those cases which have experienced variation in

treatment as a result of variation in the instrument. Since different valid instruments are correlated with a different part of the exogenous variance, different but equally valid instruments often produce significantly different point estimates, sometimes even in different directions. This would not be possible if IV estimation identified the true average population treatment effect; it therefore gives researchers the possibility of fine-tuning the desired result. As a consequence, IV estimates are not conclusive but require robustness testing. This seems to be perfectly clear to Angrist (2004: C80), who admits that "the external validity of IV estimates is ultimately established (. . .) by replication in new data sets and, of course, by new instruments," in other words: by robustness tests.

6.7 SOCIAL SCIENCE EXPERIMENTS

The most severe problems in empirical social science result from the analysis of observational data. Observational data are messy and their data-generating process unknown. Researchers have no control over who receives treatment and cannot eliminate the influence of potential confounders. It therefore seems only logical to replace the analysis of observational data with the analysis of experimentally generated data.[2] At the very least, this research strategy has the advantage of bringing the social sciences closer to what is regarded by many as the gold standard for causal inference in the sciences (Banerjee 2007; Rubin 2008; Falk and Heckman 2009; Angrist and Pischke 2009; Imbens 2010).

Social science experiments come in three variants: *lab experiments* usually observe responses of selected participating individuals to experimentally provided and randomized stimuli (treatments) in an artificial (laboratory) setting. *Field experiments* randomize a real treatment in the real world. Participants usually have to consent to participating due to ethical concerns about experimenting with humans without their consent. *Natural* and *quasi-experiments* are real-world situations in which a treatment appears to be randomized by some naturally-occurring phenomenon or some policy intervention.

2 However, experiments cannot answer the vast majority of relevant research questions in the social sciences because experiments are simply not feasible, are too costly or would be unethical since experiments must not cause substantial harm to participants. As Winship and Morgan (1999: 659f.) correctly point out ". . . in most social science research done outside of psychology, experimental designs are infeasible. (. . .) For these reasons, sociologists, economists, and political scientists must rely on what is now known as observational data – data that have been generated by something other than a randomized experiment – typically surveys, censuses, or administrative records."

The case for experiments is simple: observational data does not allow eliminating the effect of confounders, which therefore have to be controlled for; experiments however can eliminate the effect of confounders by controlling the environment in which treatment is given, by blocking on all known confounders and by randomizing treatment status across a very large number of cases. As the number of participants approaches infinity, any differences between the treatment and control group vanish, leaving only random correlation between the confounders and the randomized treatment. If the experiment manages to hold all confounders constant, the causal effect of the treatment can be inferred from experiments simply by the difference in outcomes in the treated compared to the control group.

Real-world experiments are typically undertaken with small sample sizes, for which it becomes questionable whether uncertainty about potentially confounding variables has been solved. Whether the properties of confounding variables in the treatment and control group become sufficiently similar depends on how rare relevant properties are. The rarer these properties, the larger the number of participants has to become for any given level of bias. Robustness tests which either use different samples or which condition on potential confounders can answer these uncertainties.

Experiments therefore require robustness tests even for the dimension of model uncertainty for which they are most powerful. Other uncertainties also remain as does therefore the need to employ robustness tests. Critics often doubt the concept validity of treatments given in lab experiments, pointing to a potential gap between real-world treatment and experimental treatment. Knowledge about participating in an experiment can result in behavioral adjustment of participants to the experimental situation (Hawthorne bias).[3]

3 Experimenters might interact more with the treatment group than with the control group and participants might learn to "play" experimental situations (Bracht and Glass 1968). Natural and quasi-experiments are the only form of experiment in which the participants do not know that they are exposed to an experimental design. They are also the only "experiments" based on observational rather than generated data. They do not rely on a randomization of treatment, but on the assignment of treatment by a rule, which happens to be more or less orthogonal to structure in the covariates of behavior and in outcomes (e.g., assignment by alphabetical order or by lottery number). The hope is that the treatment assignment is haphazard and as if randomly distributed across cases. Natural and quasi-experiments are both rare and contested. Keane (2010: 12) and Shadish, Cook, and Campbell (2002) warn that many of the actual allegedly natural or quasi-experiments found in the social science literature are of low quality as often it remains questionable whether treatment was quasi-randomly assigned across groups and whether the two groups were sufficiently similar.

The conductors of lab experiments cannot necessarily know what real-world behavior their experimental findings represent, since the experimental treatment and setting differ from the real-world treatment and setting (Cobb-Clark and Crossley 2003), again casting uncertainty on the concept validity of the treatment. For example, many social science experiments transfer real-world stimuli and incentives into monetary lab incentives. The concept validity of the experimental treatment needs to be checked against other values of the monetary incentives and against other incentives.

In principle, field experiments perform better with regards to uncertainty about concept validity. However, even in field experiments participants typically know that they participate in a randomized experiment for a certain period of time and may adjust their behavior accordingly, whereas in the ideal experiment no participant knows in which group she is, nor do those handing out treatment know whether they are administering a genuine treatment or a placebo (Cartwright 2010: 63). Robustness tests therefore need to explore the impact of uncertainty about concept validity on estimated effects even in field experiments.

Other specification uncertainties also require more attention than they typically receive. Uncertainty about the functional form of treatment requires tests that go beyond dichotomized treatment status or simple, often linear, functional form assumptions. Uncertainty about dynamics requires tests that overcome the typical comparison of only two points in time: before the experiment and after the experiment. Uncertainty about spatial dependence is essentially assumed away in field experiments by ignoring spill-over and general equilibrium effects (Ravallion 2012: 105). In reality, however, many treatment effects depend on the degree of social interaction. Lab experiments may mirror this setting by placing participants in a situation of competitiveness or co-operation among participants. Yet, the choice of strategies played by real actors varies with the level of competitiveness or co-operation, and so robustness tests that vary these factors are required.

The largest uncertainty that experiments face is with regards to causal heterogeneity and context conditionality, however. Real-world lab experiments often draw participants from pre-selected convenience samples of, for example, students. Yet, randomizing treatment within a pre-selected group will not solve pre-selection bias, not even asymptotically, and will not produce an unbiased effect of the average population treatment effect if there is causal heterogeneity (Ho et al. 2007: 205; Heckman et al. 1997). Analyses of randomized treatments in convenience samples trade internal validity for external validity and any increase in the former often comes at the expense of a sharp decline in the latter. The limited external validity could in principle be overcome if researchers managed to randomize

treatment in a sample that is randomly drawn from the deductively derived population (Shadish, Cook, and Campbell 2002: 91f.). More realistically, robustness tests are needed to tackle uncertainty about causal heterogeneity in lab experiments based on convenience samples. The external validity needs to be checked by conducting experiments with participants drawn from alternative social groups and strata. Without such tests, the only inference from lab experiments based on convenience samples one can draw is that the treatment can make a difference in some samples.

Field experiments also take place in particular settings and conditions (Shadish, Cook, and Campbell 2002: 18; Cartwright 2010; Cartwright and Hardie 2012). It cannot be guaranteed that an estimate derived from a randomized controlled experiment among individuals from a certain village in a certain region in north India in 2010 can be generalized to different settings and to the population of interest. Uncertainty about causal heterogeneity and context conditionality implies that, in the absence of robustness tests, the experimental result cannot be known to be valid beyond the specific setting or beyond the values of the conditioning factors as found in the experiment. The results should not be transferred and generalized to alternative settings or alternative values of the conditioning factors without robustness tests demonstrating the stability of the estimated effect.

In sum, then, for every experimental design there exist numerous plausible alternative experimental designs. Rather than being assumption-free, experimental design requires making a very large number of modelling assumptions, just as regression analysis of observational data does. The experimenter cannot simply assume that results are robust to plausible alternative specification assumptions of her experiment. Accordingly, social science experiments represent no alternative to robustness tests but instead themselves require robustness tests.

6.8 CONCLUSION

The methodological toolbox available to researchers has never been better equipped. Recent advances enable social scientists to study the phenomenon of interest with a plethora of methodologies. As this chapter has made clear, no research design, estimation procedure or analytical technique solves the problem of specification uncertainty. They all depend on specific modelling choices taken for which plausible alternatives exist. Therefore, no research design, estimation procedure or analytical technique represents an alternative to robustness tests as such. At best, they can provide a partial substitute for some robustness tests. For example, moving to experimental design can reduce the necessity to conduct robustness tests dealing with omitted explanatory variables. However, at the same time it makes robustness tests

analyzing the impact of alternative concept definitions and operationalizations as well as uncertainty about causal heterogeneity and context conditionality much more important. Every tool in the methodological toolbox of social scientists requires robustness tests to improve the validity of inferences based upon it.

Modern experimental designs and identification techniques have their role to play in the development of the social sciences. However, we wish their proponents would develop a more realistic assessment of the strengths and weaknesses of these techniques and the validity of results generated. Social scientists should not equate *identification* with *valid inference* and the "methodological triumphalism" (Barrett and Carter 2010: 516) of experimentalists and proponents of identification techniques is an unjustified self-marketing exercise. Imai, King, and Stuart (2008: 493) similarly caution against any presumed inferential superiority of any technique: "Experimentalists may envy the large, randomly selected samples in observational studies, and observationalists may envy the ability of experimentalists to assign treatments randomly, but the good of each approach comes also with a different set of constraints that cause other difficulties."

Social scientists will learn with time that the findings from experiments, quasi-experiments, and identification techniques need to be subjected to robustness tests just like any other analytical technique. Robustness testing should play an important part in the social sciences regardless of the research design and recognition of this fact is starting to spread. To give a laudable example: Lassen and Serritzlew (2011), in their quasi-experimental analysis from a large-scale municipal reform in Denmark of the effect of jurisdiction size on political efficacy, have a "preferred specification" (we call it baseline model) but they undertake a large number of different estimates (we call them robustness test models) to explore the robustness of their results. They find that

> the result that population size has a causal, significant effect on IPE [internal political efficacy] is robust across samples and estimators, and we demonstrate that local variations in the amalgamation process, as well as changes in local public finances and municipal political control following reform, do not affect this relationship.
>
> (Lassen and Serritzlew 2011: 239)

When in one robustness test they employ matching additionally to their baseline differences-in-differences specification and find significantly stronger effects, they explore reasons for these differences. The differences could stem from the sample changing because matching drops observations or from the fact that matching makes no functional form assumption. To find the true reason, the authors re-estimate the baseline (parametric) model based on the sample of their matching (non-parametric) model and "observe

results essentially similar to those identified by the matching analysis" (p. 252). Accordingly, the observed difference is caused by the change in sample, which suggests the existence of causal heterogeneity or unobserved conditionality. What is remarkable about Lassen and Serritzlew's analysis is that firstly they judge the lack of the robustness of their findings based on the estimated effect size and not merely on the direction of the effect. This finding motivates them to dig deeper into the nature of administrative reform and municipality size. And second, the robustness analysis, while based on research design and empirical model specification, also has important theoretical implications, which open the avenue for additional research.

Like modern techniques, old-fashioned regression analyses continue to be a valuable tool for social scientists. Regression analysis retains an unrivalled strength: its modelling flexibility and its almost unlimited versatility. Social science is full of questions, which cannot be answered with randomized experiments (Ravallion 2012), and it worries even its most fervent proponents that the emphasis on this inferential technique "may lead researchers to avoid questions where randomization is difficult, or even conceptually impossible, and natural experiments are not available" (Imbens 2010: 401).

Regression analysis does not require a random treatment; it can be used with all sorts of observational data. Real-world causal complexity in the form of conditionalities, temporal dynamics, spatial dependence, and so on can in principle be modelled and their effects interpreted. More often than with alternatives, it is possible to draw a random sample from the population. As Deaton (2010: 445) argues: "... a biased nonexperimental analysis might do better than a randomized controlled trial if enrolment into the trial is nonrepresentative." But, without doubt, regression analyses of observational data are fraught with model uncertainty, and the remainder of this book suggests tests for exploring the robustness of results for important dimensions of model uncertainty.