Model Uncertainty and Robustness Tests:
Towards a New Logic of Statistical Inference

Thomas Plümper\textsuperscript{a} and Eric Neumayer\textsuperscript{b}

\textsuperscript{a} Department of Government, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK, tpluem@essex.ac.uk

\textsuperscript{b} Department of Geography and Environment, London School of Economics and Political Science (LSE), London WC2A 2AE, UK, e.neumayer@lse.ac.uk
Abstract:

We augment the increasingly common practice of typically ad hoc robustness tests into a research methodology that allows reliable inferences when researchers do not know the true data-generating process. We identify three principal sources of model uncertainty. First, theories simplify and aim at isolating causal mechanisms under ceteris paribus conditions. Theories thus cannot sufficiently specify the data-generating process. Second, empirical researchers are uncertain about what constitutes the population – the set of cases for which a theory claims validity – from which to draw a representative sample as required for econometric estimation. Finally, model uncertainty also results from the fact that many concepts used in social science theories cannot be observed and therefore can only be measured by proxy variables, which deviate in unknown, but potentially systematic ways from the theoretical concepts. These three causes of model uncertainty result in at least 16 specific aspects of uncertainty about model specification that can be subjected to various robustness tests. Defining robustness as stability of causal inferences, we argue against the misplaced focus on single point estimates from any specific test. Instead, we suggest scholars employ model averaging techniques to assess robustness in terms of the entire set of empirical evidence from the baseline and all robustness test models. This method avoids the type-II error bias of current practice and encourages researchers to undertake a larger set of robustness tests instead of carefully hand-selecting only those tests that do not deviate from their baseline model results. We illustrate our proposal using an example of an existing study by the authors on the political determinants of famine mortality, which contained only two robustness tests and thus begs the question whether the causal inferences are fragile. Subjecting our baseline model to a battery of additional tests, we conclude that our causal inferences are in fact robust: democracies respond more elastically than autocracies to the simultaneous presence of international food aid and a large share of the population being affected, which results in lower famine mortality.
1. Introduction

Empirical researchers do not know the “true model”. Neither theory nor intuition, neither econometric tests nor data mining can give empirical researchers sufficient guidance to specify an econometric model which exactly mirrors the true data-generating process.\footnote{The recognition that model misspecifications are ubiquitous goes back at least to George Box, who alarmed his readers that “all models are wrong, but some are useful” (Kennedy 2008: 71). 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”, while Luke Keele put it this way: “statistical models are always simplifications, and even the most complicated model will be a pale imitation of reality” (Keele 2008). In the summary verdict of Peter Kennedy (2008: 71): “It is now generally acknowledged that econometric models are false and there is no hope, or pretense, that through them truth will be found.” Like us, these authors do not suggest that models can be misspecified. Instead, they agree that all models are necessarily misspecified.} Yet, if the econometric model differs from the true data-generating process, estimates are biased and standard errors do not reveal the true level of uncertainty. As a consequence, inferences remain unreliable.

Up to now, social scientists have offered three methodical ways of dealing with model uncertainty: testing for model misspecification, testing for the sensitivity of inferences to a very large number of model permutations (extreme bounds analysis), and removing model uncertainty via randomization of treatment effect status. All of these encounter major problems and therefore none of them provide a convincing answer to model uncertainty. Misspecification tests do not necessarily detect model misspecifications and positive test results can be caused by any number of misspecifications, providing insufficient guidance on how to move the model specification closer to the data-generating process. Only hopeless optimists nowadays believe that econometric testing leads to the true model. Testing the sensitivity of inferences toward a very large number of model permutations, in which the set of explanatory variables is partly randomized following Ed Leamer’s extreme bounds analysis, is bound to result in a large likelihood for committing type-II errors and does not even offer a guarantee that the correct model is included in the several million model permutations. Randomizing treatment status for a single variable across a controlled sample aims at holding all other factors constant, thus seemingly removing (most) aspects of model uncertainty. Yet, only in very large samples will randomized controlled trials successfully remove all correlation of the treatment effect with confounding factors and even then is the estimated effect likely to suffer from sample selection bias and thus from lack of external validity.
Most applied researchers have ignored these existing methodical ways of dealing with model uncertainty and have instead resorted to a rather ad hoc way in the form of a typically small, rather arbitrarily chosen and often poorly justified set of “robustness tests”. Our major contribution in this paper is to provide a methodical and systematic basis to the current practice of robustness tests. We discuss the three principal causes of model uncertainty, which give rise to at least sixteen different aspects of uncertainty about model specification that robustness tests can address. Our proposed approach abandons the illusory quest for obtaining one single unbiased effect estimate that also commands external validity. It thus stands in stark contrast to specification test approaches and is similar in spirit to Leamer’s extreme bounds analysis. Contrary to Leamer, however, we reject the idea of running millions of partly randomized regressions. Instead, we recommend applied researchers estimate their baseline model or models plus a small set of systematic and theoretically justified robustness tests which deal with each specific aspect of uncertainty about model specification that is pertinent to their research question at hand. Thus, which robustness tests are most appropriate depends on the research question studied. We define an estimation result as robust if the inferences a researcher makes with respect to the tested hypothesis or prediction do not change after the set of robustness tests have been conducted. In order to assess robustness, we argue that it is unhelpful to focus on any single point estimate of any single robustness test. Instead, one needs to take into account the entire set of evidence from all estimated models. We therefore draw on the model averaging literature (LITERATURE) and suggest different procedures to aggregate the results from the baseline model and the set of robustness tests.

We start this paper explaining why existing methodical ways of dealing with model uncertainty do not convincingly deal with the problem. We then review the current practice of ad hoc robustness tests in Political Science. With a view toward changing this practice and providing a methodical and systematic grounding for robustness tests we first discuss the principal causes of model uncertainty and the various aspects of uncertainty about model specification these causes give rise to, which robustness tests can address. Given space constraints, each section is necessarily brief and we cannot discuss the possible robustness tests at the disposal of researchers. Instead, we provide a clear definition of robustness and discuss how results from a range of robustness tests should be aggregated to assess robustness. We finish by illustrating our proposal with an example taken from one of our previously published studies,
in which we explore whether the reported results are robust to plausible changes in model specification.

2. Existing Approaches to Model Uncertainty and Their Shortcomings
We are, of course, neither the first to stumble over the problem of model uncertainty nor are we the first to suggest a solution. In fact, there are at least three competing approaches to deal with model uncertainty. This section discusses their shortcomings.

2.1 Can Misspecification Tests Identify the ‘True Model’?
For several decades, econometricians have developed econometric tests aimed at detecting model misspecification. The hope is that a battery of such tests will allow researchers to find the true model or at least get sufficiently close to it. Early and naïve tests related to model fit such as adjusted R² statistics, F-tests, Chi²-tests as well as more complex ‘goodness-of-fit’ measures in the form of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). It is easy to see why these tests fail to provide guidance. First, they are relative constructs, merely helping to identify models which fit the data better. If all compared models are misspecified and fit the data poorly, they will still identify the “best” model. Yet, the best fitting misspecified model is still misspecified. Second, a better model fit does not necessarily imply a better model specification. For example, the inclusion of vote intention in a model of voting spuriously improves model fit, but results in model misspecification. Unfortunately, it is easy to improve goodness-of-fit, but difficult to improve model specification.

The more serious tests therefore did not attempt to improve model fit, but diagnostically test for specific model misspecifications. Some of the better known tests include the Durbin-Watson test for serial correlation, the Cook test and other test statistics for the detection of influential observations, the Breusch-Pagan, Cook-Weisberg and White tests for heteroscedasticity, the Chow test for structural breaks, the Ramsey regression specification error test (RESET), the Hausman test for unobserved heterogeneity in panel data analysis and other specification assumptions, the Durbin-Wu-Hausman test for variable exogeneity, likelihood-ratio tests for choosing among nested models or competing estimators, tests for choosing among non-nested models such as the Cox tests, tests for non-stationarity and co-integration tests, and so on.
Most tests search for structure in the residuals of an estimation model based on the assumption that non-spherical residuals imply model misspecification or compare the estimates of two or more models with each other under the assumption that one model is consistent such that systematic deviations from this model suggest that the alternative specification is inconsistent and to be avoided. For many reasons employing diagnostic tests will not result in the correct model specification (see Freedman 2010 for a similar assessment):

First, most tests make assumptions that are valid in asymptotia\(^2\), but not in the finite samples applied researchers use. Yet, residuals are only uncorrelated ‘white noise’ if the number of observations is infinite. With a finite number of observations, structure in the residuals does not prove model misspecification, but at best indicates the likelihood of model misspecification. Since it is impossible to generalize mathematical shortcuts from infinite sets to finite sets, strictly speaking, this renders all tests that rely on asymptotic properties invalid for the vast majority of empirical designs. Tests may still be useful, but only in an indicative sense.

Second, most available tests are individual misspecification tests, responding to a single specification error, with joint misspecification tests exceedingly rare and their properties still largely unknown (McGuirk, Driscoll and Alwang 1993). Yet, the odds are that empirical models suffer from numerous specification errors. Employing a battery of individual misspecification tests does not necessarily help either and raises questions about the appropriate overall significance level (Godfrey 2005).\(^3\)

Third, while diagnostic tests may signal misspecification, they offer little guidance on the causes of misspecification since a significant test result can be caused by a multitude of misspecifications. For example, if Ramsey’s “specification test” finds the residuals to be correlated with non-linear combinations of the regressors, the cause of the misspecification remains unknown. 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

---

\(^2\) Leamer (2010) calls it asymptopia to emphasize it is unreachable in this world.

\(^3\) In addition, for some misspecifications a multitude of tests exist which often produce different and conflicting test results since the null hypotheses differ or, if the null hypothesis is the same, the test procedure differs. For empirical researchers, it is exceedingly difficult to know which results to believe and which to discard. Tests for detecting unit roots in panel data represent a prime example (Baltagi 2008).
effect, wrong functional form of interaction effect, and so on. Moreover, the simple polynomial that Ramsey suggests may overlook the true structure in the residuals. Thus, even if a model passes the Ramsey test, there is no guarantee that the model is not misspecified.

Fifth, the concentration on and apparent solution to selected specification errors can render other specification problems worse. In other words, the apparent cure can do more damage than the disease. Take the use of fixed effects models in panels with trended time-series as an example. The fixed effects model eliminates all variation between units and only uses the within-variation (over-time variation within units) for estimating the coefficients. As a side effect, the influence of correlated trends on the estimation results becomes much larger. However, co-trended data do not need to be causally related. Fixed effects models can also exacerbate the bias stemming from omitted time-varying variables. The same holds for measurement error in explanatory variables (Angrist and Pischke 2009: 225). Thus, by making other problems worse, the fixed effects model may actually increase the probability of wrong inferences even though it ‘solves’ a specification problem.

2.2 Is there Truth in the Large Number of Models?

The major contribution to causal inference of Edward E. Leamer is also the reason why many econometricians regard him a heretic: he abandons the idea that econometric tests can identify the true data-generating process. Instead, Leamer’s central idea is to test the sensitivity of inferences to a very large number of specification changes: “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). Leamer’s suggestion for a “global sensitivity analysis” is thus a very radical one, calling for alternating all the parameters of model choice, which if undertaken simultaneously would result in a quasi-infinite number of sensitivity tests.

In actual reality, those who have taken their inspiration from Leamer have exclusively focused on alternating the set of explanatory variables in what has become known as extreme bounds analysis; despite his call for a global sensitivity analysis, this is also the specific focus of

---

4 The same is true for other diagnostic tests – see McAleer (1994: 330f.) for an overview of the large number of possibilities for rejecting the null hypothesis in each test.

5 We are not aware of a single application of extreme bounds analysis that did this.
Leamer (1983, 1985) and Leamer and Leonard (1983). Extreme bounds analysis comes in many varieties, but most studies permute the set of right-hand-side variables according to a simple algorithm, estimating a very large number of models – 2 million in the case of Sala-i-Martin (1997b) and it’s been 4 million before (Sala-i-Martin 1997a).

Since the vast majority of these millions of model will be misspecified, it is crucial to define ‘robustness’ in a way that minimizes the possibility that results from misspecified models do not invalidate the inference. Unfortunately, proponents of extreme bounds analysis have not solved the problem (yet). Levine and Renelt (1992) accept estimates as robust if all estimates from the extreme bounds analysis have the same sign and are statistically significantly different from zero. Sala-i-Martin (1997b) rejects the Levine and Renelt (1992) criterion as too demanding since one single insignificant estimate or one wrongly signed significant estimate would already lead to non-robustness, giving one single model “veto power” in the determination of robustness. Instead, he computes the cumulative distribution function of estimates, both under assumptions of normality and non-normality, and instead accepts estimates as robust if the average weighted estimate is significant, where the weights represent the integrated likelihoods of each model. In Sala-i-Martin et al. (2004), the authors move to a Bayesian-style analysis of aggregating results from the very large number of models.

Both inferential assumptions face the problem that in all but a few of its permutations, the randomized choice of explanatory variables causes a more or less severe misspecification of the estimation model. In other words, EBA randomizes model misspecifications that result from the inclusion of irrelevant variables and/or the exclusion of relevant variables. All but at best one of these randomized models are misspecified such that the estimates are biased and the standard errors wrong. It is thus not convincing to reject a hypothesis if a specific share of models rejects the hypothesis since every single model that rejects the hypothesis could be severely misspecified. As a consequence, EBA is best understood as a one-sided test: if some variables pass the extreme bounds analysis, then we can infer that these variables are robustly correlated with the dependent variable in the presence of randomized model misspecifications. If, however, some or even plenty of the model permutations suggest a non-significant or even oppositely

---

Proponents of extreme bounds analysis try “not to include variables in the conditioning set that, on a priori grounds, measure the same phenomenon as the variable of interest” (Levine and Renelt 1992: 943). But model misspecification can equally be caused by including variables which merely represent the causal process through which a variable of interest affects the dependent variable or simply by including irrelevant variables, which are sufficiently correlated with relevant variables.
signed coefficient, one cannot necessarily infer that the effect is not robust until one can justify the model specification of the deviant permutations – which of course is unlikely in the absence of knowledge of the true model. Thus, extreme bounds lends credibility to an estimated effect if it is found to be robust, but fails to provide convincing evidence for its fragility if the effect is found to be “non-robust”.

2.3 Can Model Uncertainty be Randomized Away?

Misspecification tests cannot identify the correct model specification and running millions of estimations for testing the fragility of inferences risks including a large number of clearly misspecified models. The third existing methodical way of dealing with model uncertainty is to render it irrelevant by way of randomized controlled trials (RCTs). The basic idea is to isolate one single causal factor and estimate its unbiased effect by perfectly randomizing selection into treatment status across a very large number of observations. The inspiration comes from the natural sciences, which can run true controlled experiments, and since this is impossible with human subjects, randomized controlled trials (RCTs) or experiments are the next best thing. Its proponents leave little doubt that they think RCTs are superior to all other research designs. Angrist and Pischke (2009: 12), for example, write: “The most credible and influential research designs use random assignment.”

In the ideal scenario, the treatment group and the control group are identical in all confounding factors (at the very least, confounding factors have to have similar moments in both groups) and the sample represents the population. If this is not the case, and randomization can only asymptotically achieve this ideal, then randomized trials do not identify the true causal effect. The expected ‘bias’ increases if confounding factors become rare and/or stronger (relative to the effect of interest). Clearly, the rarer a confounding factor in the sample, the lower the probability that it is evenly distributed between treatment and control group. In contrast, the bias declines as the number of independent participants in the randomized trials increases.

Another threat to finding an unbiased treatment effect stems from the fact that experiments take place in the shadow of the real world. Participants often (not always) know that they participate in a randomized experiment and may adjust their behavior accordingly. In other RCT settings, certain individuals will attempt their best to either get out of or into the trial group in ways that are non-random, which also biases the estimated effect. This will particularly be a
problem where researchers have no direct control of the random assignment process and thus cannot enforce it, but use instrumental variables or regression discontinuity designs to mimic random assignments. For example, Deaton (2010: 432) points out that using draft lottery numbers as an instrument for the effect of compulsory military service (loss of years of schooling) on earnings is problematic since potential draftee’s efforts to circumvent their draft status partly depends on their individual rate of return to schooling. Even if individuals stay within their assigned group, RCTs will not result in unbiased effects if the treatment effect of some individuals depends on whether specific other individuals also received treatment. Such spillover effects violate the stable unit treatment value assumption (Sekhon 2010: 492; Sovey and Green 2011: 199).

The major drawback of RCTs is their limited external validity, however. The artificial situation in the laboratory may affect the behavior of participants, thereby generating deviations from real world behavior of interest. In addition, often the selected samples do not even aim at representing the population in randomized trials undertaken in the social sciences. Applying random assignment to non-random selections of a given population violates a critical requirement of RCTs for producing valid inferences (Ho et al. 2007: 205). RCTs trade internal validity for external validity and any increase in the former often comes at the expense of a sharp decline in the latter. RCT proponents like Imbens (2010: 403) give clear priority to internal validity over external validity, but for the reliability of inferences both matter. There is no reason for preferring an internally valid estimate for a small subgroup to a slightly biased (internally invalid) estimate for the population.

Finally, the social sciences are full of interesting research questions, which simply cannot be answered with randomized trials and it worries even its strictest proponents that too much emphasis on this research design “may lead researchers to avoid questions where randomization is difficult, or even conceptually impossible, and natural experiments are not available” (Imbens 2010: 401).

Thus, randomization is likely to reduce the influence of model misspecification on parameter estimates. Yet, unless parameter heterogeneity is perfectly absent, all estimates based on randomized trials suffer from potentially severe selection bias. Unless scholars still believe in naïve falsification, the results from randomized trials are not conclusive.
2.4 Summary

In this section, we have argued why neither diagnostic econometric testing, nor extreme bounds analysis nor randomized controlled trials convincingly deal with model uncertainty. Of course, we are not against the use of these approaches per se. Econometric tests can be helpful and there is a role for extreme bounds analysis and randomized controls in the toolkit of social scientists. But they are not the solution to reliable inferences in the face of fundamental model uncertainty. We next review the current practice of robustness tests, which represents a somewhat ad hoc alternative.

3. Robustness Tests in Political Science: The Current Practice

With none of the existing methodical ways able to solve the problem of model uncertainty convincingly, applied researchers developed an alternative practice that – we argue – points into the right direction. In what has become widely known as robustness tests, applied researchers vary the model specification and estimate a small number of additional models based on specific concerns about aspects of the baseline model specification. Applied researchers thus no longer claim (or need to claim) that their baseline model is the true model with certainty, but rather accept the possibility that alternative model specifications could be correct or at least closer to the true data-generating process than their baseline model. If the estimates of the baseline model and the robustness checks are sufficiently similar (a rather vague notion, as we will show below), applied researchers call their estimates “robust”. Before we provide a more methodical and systematic grounding for robustness tests, we briefly discuss current practice in Political Science and its weaknesses.

There is no evidence that robustness tests have been invented by a single author. Rather, they seem to have evolved in different social science disciplines roughly at the same time. We have surveyed the leading Political Science journals for articles that systematically conduct robustness tests (beyond varying the right-hand side variables) over a ten year period. We found more than 500 articles in which authors reported at least 1 robustness test. The number of articles reporting robustness tests is growing over time. Overall we found that explicit robustness tests are employed in roughly one quarter of empirical papers published in these leading Political Science journals. This does not mean that even neighboring sub-disciplines put the same emphasis on the relevance of robustness tests. While, for example, the probability that an
emirical article employs robustness tests is close to zero in behavioral politics and electoral studies, this probability rises to almost 50 percent in political economy. Naturally, model uncertainty is not necessarily higher in political economy than in behavioral politics. Rather, it seems that robustness tests are a contagious social phenomenon that spreads within sub-disciplines.

The most common robustness tests are the use of additional controls, alternative measures of the dependent or central explanatory variables, changes in the sample, and alternative scales. These tests are conducted in 20-30 percent of articles that report robustness tests. Alternative estimators, alternative functional forms, and alternative dynamics are used in about 10 percent of the articles that report robustness tests. All other robustness tests are less frequent. They occur, but scholars do not use these tests systematically. Nevertheless, some applied researchers conduct tests that account for structural breaks, alternative lag structures, conditionality, spatial dependence, missing observations (multiple imputation), crucial cases (jackknife), or endogeneity (instruments). Yet, there is still a glaring gap to bridge between the number of potential model uncertainties and the frequency with which they are taken care of either in the baseline model or in robustness tests. Where scholars conduct robustness tests, they typically report or discuss the results of two to three tests on average, with a minority of papers employing five tests.

A small minority of papers theoretically motivate the tests employed. A good and noteworthy example is Gerber and Huber (2010) who study whether the association between partisanship and economic assessments holds irrespective of partisanship. They find that large partisan differences between Republican and Democratic voters 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 is a Republican 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 driven 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.”

However, these occasional positive examples of authors undertaking a large set of theoretically motivated robustness tests notwithstanding, the overall state of current practice is very unsatisfactory. Political scientists typically report only a very small subset of possible robustness tests. In most cases, these tests are poorly motivated by a discussion of the source and nature of model uncertainty. 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 specific carefully selected tests. Whether or not the baseline model is robust to additional tests remains an open question. In addition, applied researchers hardly ever define what they mean by ‘robust’. Without a concise definition of robustness, however, the claim that the empirical model is robust to alternative model specifications remains rather empty.

In what follows, we take the methodology of robustness tests several steps further. Specifically, we, first, derive from the three principal sources of model uncertainty a comprehensive list of specification issues that robustness tests can address. Second, we clarify the meaning of robustness by providing a definition and arguing that what matters is not whether point estimates are robust, but whether causal inferences are. And finally, we discuss procedures for aggregating the estimation results from the baseline model and the set of robustness tests in order to check whether causal inferences are indeed robust.

4. The Principal Causes of Model Uncertainty

Our first contribution to the practice of robustness testing is the identification of three principal causes of model uncertainty: underspecified theories that do not allow applied researchers to derive a data-generating process, unknown populations, and systematic gaps between concept
and measurement. We discuss each cause in turn and the various aspects of uncertainty about model specification they give rise to.

4.1. Theories Do Not Specify the Data-Generating Process

Theory ought to inform model specification. However, for two reasons theory does not allow identifying the data-generating process. First, theories aim at simplifying complex real world phenomena. They are not derived from an integrated model of the social world. Accordingly, theorists do not intend to formulate a comprehensive account of the data-generating process, but rather intend to isolate a small number of causal mechanisms. Accordingly, theories in general and social science theories in particular are underspecified. Social science theories are “ceteris paribus” theories: they claim validity if one holds all other factors constant. These other factors are clearly part of the data-generating process; yet, they are explicitly excluded from the social science theory. Therefore, theory cannot reveal the “true” model.

Second, social behavior is not fully determined by the social, economic and political structure actors take as given. Rather, biochemical processes in the brain, genetic predispositions, individual and social learning and experience, and free will exert an additional influence on social behavior that cannot be captured by deterministic theory. As a result, there are no eternal “laws” governing social interactions akin to the physical laws generated during the big bang. Instead, “causality” in the social sciences is partly time-dependent and changes over time, with the “laws” governing social interactions at least partly determined by social interactions themselves. Thus, the idea that better theory allows applied researchers to formulate an appropriate model of the true data-generating process should not guide contemporary social science research. There exists no such theory in the social sciences that has a scientific status and a degree of precision similar to, say, relativity theory or Newton’s theory of gravity.

Theories thus isolate a single or a few causal mechanisms while abstracting from others and typically predict no more than the direction of a causal effect. Being necessarily underspecified, theory leaves applied researchers uncertain about:

a) the correct set of explanatory variables,

b) the functional form with which each explanatory variable exerts an effect on the dependent variable,
c) the specificities of conditional effects between explanatory variables,
d) the temporal delay between cause and effect,
e) effect dynamics,
f) the (in-) dependence of cases,
g) the existence and nature of reversed causal effects, and
h) the existence of structural breaks (i.e. the time-dependence of effects).

4.2. Population Uncertainty and Sampling

The population is the set of all cases on which a theory makes predictions. Econometric theory assumes that applied scholars draw a sufficiently large random sample from the population. Given a truly random draw, the properties of the sample mirror that of the population – plus some sampling error. In the practice of applied research, random sampling is a negligible event. Many scholars do not even attempt to draw a random sample but rather analyze convenience samples, defined as the set of cases for which information is (easily) available. For example, it is an old habit of comparative welfare state research to exclusively study the “old” OECD countries. This arbitrarily includes Japan but excludes Korea, it includes Portugal but excludes Poland, and it excludes Mexico but includes New Zealand. Whether or not that is theoretically justified cannot depend on the date of membership in the OECD, but rather on political, social, and economic criteria that need to be justified and discussed – and since they cannot be taken for granted need to become part of a robustness test.

If scholars do randomize the sample, they usually start with a convenience sample from which they randomly draw cases. However, a random draw from a selected subset of the population does not give a random sample with properties similar to that of the population. Rather, the properties of the sample will be similar to that of the pre-selected subset, and hence, subject to selection bias.\(^7\) Put differently, a random draw from a convenience sample is still a convenience sample, not a random draw from the population.

---

\(^7\) The consequences of analyzing either convenience samples or random draws from convenience samples are twofold: first, the external validity of findings remains doubtful. While it may be possible to generalize in case of the developing country selection to future generations of developing countries, it is impossible to generalize from Oxford students to a global set of individuals. Generalizing from cases that do not represent the population to the population is impossible. And second, since the variation within convenience samples tends to be smaller than the variation within the population, the estimated level of uncertainty is too small.
In many cases, applied researchers do not know the population, simply because theory does not sufficiently specify the conditions under which a case belongs to the population. It is often very difficult, sometimes impossible to know the set of all observational units for which a theory can claim validity. It is easy to establish, however, what does not constitute a population: all UK citizens that had a telephone landline in 2007, the graduate students of Oxford University in 2010, all developing countries between 1980 and 2005. All these are convenience samples. Researches choose UK citizens with a telephone landline because they intend to sample from a telephone book, or graduate students from Oxford University because the experimental laboratory is in Oxford and it is easier to get hold of graduate students than of non-students, and developing countries before 1980 are excluded from the sample because data is missing before the 1980s.

What makes the situation worse is that the complexity of the data-generating process can be influenced by selection. If one holds a potentially confounding factor constant in a sample, it cannot exert an omitted variable bias on the estimation of the variable of interest. It will, however, bias the estimates because of selection. Thus, simplifying the data-generating process of the sample – a strategy not only chosen by qualitative researchers but also by proponents of randomized controlled trials – does not lead to unbiased estimates of the population effect.

For example, it is an old habit of comparative welfare state research to exclusively study the “old” OECD countries. This arbitrarily includes Japan but excludes Korea, it includes Portugal but excludes Poland, and it excludes Mexico but includes New Zealand. Whether or not that is theoretically justified cannot depend on membership in the OECD, but rather on political, social, and economic criteria that need to be justified and discussed – and since they cannot be taken for granted need to become part of a robustness test.

In consequence, with uncertainty about the correct population and about the degree to which sample properties deviate from population properties the inclusion or exclusion of cases becomes a potential source of wrong inferences. With finite samples, the influence of single cases or groups of cases on inferences also becomes an issue. One should be reluctant to accept a theory if the statistical support for that theory depends on a relatively small number of cases that are either in- or excluded from the sample. Uncertainty about the population affects all elements of the sampling process, including
i) the influence of included cases (or groups of cases) on inferences,

j) the influence of excluded cases (or groups of cases) on inferences,

k) the influence of case- and group-wise causal heterogeneity on inferences.

4.3. The Concept – Measurement Gap

Everything can be quantified, but not every quantification rests on measurement. To measure a concept requires a defined scale and the independence of the scale from factors that influence measurement. Scales are usually defined arbitrarily. A meter is a meter and a minute has 60 seconds by convention, not because there is some overarching principle that allows only one scale. Yet, to satisfy the definition of a convention, a measurement requires that if two persons measure the same phenomenon, they get (roughly) the same result. Thus, while a meter is arbitrary, a meter has to be identical independent on potentially intervening factors such as location, culture, institutions, and so on.

In the social sciences, measurement serves different purposes, one of which is that by measuring concepts, scholars open the hypotheses and predictions derived from theories to quantitative tests. Thus, for the purpose of theory testing, concepts used in theories need to be operationalized and measured. This double process generates problems because social science theories rely on abstract theoretical concepts such as power, happiness, socialization, interest divergence, independence, development, and democracy, which cannot be directly observed let alone measured. When researchers test theories that use “unobservable” constructs, they operationalize these concepts in a way that allows measurement by proxies. These proxies will differ, sometimes systematically, from the theoretical constructs.

Measurement of theoretical concepts thus is a third source of model uncertainty. Social scientists do not know the optimal measure for their concepts and they do not know to what extent the use of proxies systematically biases the results. Take the theoretical concept of democracy as an example. Theories formulate predictions about the determinants of democracy or the effects of democracy on other social phenomena. But democracy is not directly observable or measurable, so existing real world measures of democracy such as the well-known and much-used polity2 score of the Polity IV project (Jaggers and Gurr 1995; Marshall and Jaggers 2002) or the political rights and civil liberties measures of Freedom House (2012) will differ from each other (Munck and Verkuilen 2002), which will in turn differ from alternative conceptualizations.
such as that of Bueno de Mesquita et al. (2003). For example, the Polity IV project’s polity2 score gives the highest democracy value equally to Switzerland, the United Kingdom and the United States. Yet, the core of what constitutes democracy in these countries in the eyes of their citizens as well as outside observers differs quite strongly. In the United States, the separation of powers seems to be the American democracy’s central attribute, but in the United Kingdom it would appear to be parliamentary sovereignty, while in Switzerland political participation by principles of direct democracy is a crucial element of this country’s understanding of democracy.

Missing measurement for certain cases or groups of cases exacerbates the problem caused by the concept-measurement gap. If causal heterogeneity exists and if cases are not missing at random then listwise deletion of missing observations generates bias. Unfortunately, missing at random will often be an untenable assumption. In comparative politics, for example, missing cases are more likely to occur in small and autocratic countries – let alone failed states. In survey research, non-response is often associated with more radical perceptions and preferences. In both cases, missing observations are likely to exert an undue influence on inferences. In sum, then, the concept-measurement gap adds four additional model uncertainties:

l) the influence of the definition and operationalization of concepts on inferences,

m) the influence of unsystematic measurement error on inferences,

n) the influence of systematic measurement error on inferences, and

o) the influence of missing observations on inferences.

4.4. Discussion

In this section, we have identified 16 aspects of uncertainty about model specification derived from three principal causes of model uncertainty. We do not claim that these aspects represent an exhaustive list, but we believe to have captured the most important ones. Any estimation model is likely to suffer from several of these aspects of uncertainty about model specification. In fact, applied researchers can never be entirely certain about any part of the model specification. However, there are of course different degrees of uncertainty. We recommend that applied researchers consider whether and to what extent each of these aspects represents a potentially significant risk to inferences given the hypotheses and predictions tested and given the research
design. Robustness tests should be employed for each aspect likely to endanger inferences. Thus, for any given research design, some robustness tests will be more important than others.

In a forthcoming book we discuss in detail the various robustness tests that can be employed to tackle each of the identified 16 aspects of model uncertainty. Very often there are several options for tackling each aspect. For reasons of space, here we jump straight to the question of how robustness should be defined and how results from a set of robustness tests should be aggregated in order to assess robustness.

5. Robustness: Definition and Model Averaging

According to current practice, the vast majority of researchers – the present authors included on past occasions – conduct some small set of robustness tests and simply claim that their results are robust without saying what exactly they mean by robustness. Language is employed suggesting to readers that results are “similar”, results “uphold”, are “stable”, “do not change much”, “remain essentially the same” etc. This definitional ambiguity is clearly unsatisfactory. In this section, we provide definitional clarity to what robustness should mean and how to assess robustness from a range of models.

5.1 Defining Robustness

The purpose of an econometric model is the testing of hypotheses and predictions derived from theories. An estimation result is robust to changes in model specification if the inference a researcher makes with respect to the tested hypothesis or prediction does not change. What the inference is that a researcher makes will depend on the research question. Sometimes, theory allows one to derive a prediction with respect to the size of an effect of variable $x$ on variable $y$. In this case, robustness would unambiguously imply that the inference about the size of the effect does not change. More commonly, however, theory only allows one to derive a prediction with respect to the direction of an effect, in which case researchers may only be willing to make inferences about the sign of an effect. Even in this case though one might still wish to make inferences about the size of the actually estimated effect and robustness could therefore either imply that the inference about the direction of an effect does not change or, more demandingly, that the inference about the size of the effect does not change.
5.2 Assessing Robustness: The Misleading Focus on Single Point Estimates

If there were only one robustness test, it would be easy to check whether inferences are robust. However, with several tests conducted the question arises how to assess robustness. How does one know whether the inference is robust if there are several estimates? Predominant current practice seems to take an absolutist stance: if one robustness test leads to an estimation result in which the effect is no longer statistically significantly different from zero or differs statistically significantly in size from the main result, then the conclusion is that the inference about the direction of the effect or about the size of the effect is not robust. Conversely, a result is considered robust if it does not become statistically insignificant or does not statistically significantly differ in size in any of the robustness tests. This is similar to the criterion Levine and Renelt (1992) apply in their definition of robustness in extreme bounds analysis.

Definitions cannot be wrong; they are more or less useful. The absolutist definition employed by current practice is everything but useful. To start with, given model uncertainty any single one of the robustness test models could suffer from severe misspecification. To reject a hypothesis based on one single, possibly severely misspecified model is therefore not convincing. It would result in a large number of type-II errors. Second, this definition prompts applied researchers to only report a small number of carefully hand-selected robustness tests, no single one of which contradicts the inference from the main estimation model. In turn, all tests that change the effect direction or effect size are suppressed. The reason is simple: they fear that if they show a larger set of robustness tests then any single estimate may not be robust in this absolutist sense and their paper may be rejected by reviewers and editors. Third, and most importantly, what matters for causal inference is the entire empirical evidence that emerges from the entire set of models taken together, not any single point estimate of any single robustness test. How much a single point estimate that deviates in sign or size from the main estimate matters for the overall assessment of robustness depends on the results from all the other models in terms of their own point estimates and associated confidence intervals. What is therefore necessary for a more useful definition of how to assess robustness from several tests is an aggregation of the results from the various tests, to which we turn now.

---

8 The inference about the direction of the effect would also not be robust if one robustness test leads to an estimation result in which the effect remains statistically distinguishable from zero but switches the sign of the effect, i.e. switches direction.

9 For example, Linos (2011: 691) uses this definition when she stresses that “the coefficients on this measure are (...) similar in size in all specifications”.

20
5.3 An Alternative Assessment of Robustness: Aggregating Results via Model Averaging

We have argued that focusing on single point estimates is unhelpful and that, instead, results from a set of robustness tests together with the baseline model should be aggregated to come to a conclusion on whether the inference changes or not. The aggregation issue becomes even more important if applied researchers follow our suggestion to seriously consider each of the 16 aspects of uncertainty about model specification, which is likely to lead to a larger number of robustness tests undertaken than in predominant current practice. But how to aggregate? We start with two extreme options, which are both flawed, before we turn to model averaging techniques, which represent the superior way to aggregate the information that comes from each model.

To start with, one could look at the full range, from minimum to maximum, of estimated confidence intervals from all models. The lower bound of the confidence interval associated with the smallest point estimate provides the lower bound (minimum) of the overall aggregated confidence interval, while the upper bound of the confidence interval associated with the largest point estimate provides the upper bound (maximum) of the overall aggregated confidence interval. The full range from minimum to maximum represents the area in which one can be confident that the ‘true’ effect is located in, somewhere within this range. If this range does not include zero, then one could conclude that the inference with regard to the estimated effect direction is robust.

Looking at the full range is not helpful, however, since it says nothing about where within this overall aggregated confidence interval the effect lies. It is therefore also unsuitable for assessing robustness with respect to inferences about effect size rather than merely effect direction. Even in terms of assessing robustness with respect to inferences about effect direction, it essentially replicates the misplaced focus on single point estimates since both the minimum and maximum derive from single point estimates each. Any single, potentially severely misspecified model would thus have the power to determine whether the inference is robust.

At the opposite extreme of looking at the full range of estimated confidence intervals from all models, one could assess whether there is a common “core” of confidence intervals from all models, i.e., the area in which confidence intervals from all models overlap. If such a core exists, then one could conclude that inferences are robust. The reason is simple: by definition the core is the area of overlapping confidence intervals from all models and since the baseline model
is one of these models, then if the core is non-empty it must be true that the inference from the baseline model does not need to be changed after the robustness tests have been undertaken.

Again, however, this way of aggregating results is not helpful either. Most importantly, it resembles the misplaced focus on single point estimates since one single estimation model outside the core of the rest of models would have the power to determine whether inferences are robust. Clearly, what is needed is a way of aggregating results in a way that does not give undue veto power to any single model, but that assesses robustness on the basis of the entire evidence taking into account all models.

Both Bayesian and frequentist methodologists have proposed many different ways to aggregate results by weighted averages of individual models. In many applications, scholars use model averaging techniques merely for the choice of right-hand side variables while at the same time ignoring all other aspects of model uncertainty. For example, Hoeting et al. (1999: 386) describe Bayesian model averaging as a technique that “seeks to average over all possible sets of predictors.” Bayesian techniques of model averaging tend to start with an uninformed prior (a prior distribution that assigns the same probability to all models) and then use the updated prior distribution as the actual prior distribution in the Bayesian analysis (Ibrahim and Laud 1994), with the “predictive log score” – a close relative of the $R^2$ statistic as the criterion for measuring the “predictive performance” of a model (Hoeting et al. 1999). Unfortunately, when Bayesian model averaging techniques are applied on a large number of models, the priors often become inconsistent with each other.

Frequentist approaches to model averaging avoid inconsistent priors by not deviating from uninformed priors. In contrast to more traditional estimates of an ad hoc chosen ‘optimal model’, frequentist model averaging takes model uncertainty into account and computes confidence intervals which are larger than the confidence intervals of each of the averaged models but smaller than the maximum range of uncertainty (Buckland et al. 1997). In contrast to Bayesian model averaging that often employs a rather inconclusive approach to model selection and thus considers permutations of various original models found in the literature, frequentist approaches narrow down the number of models far more.

The frequentist approach to model averaging typically weights models by an algorithm based on some criterion of model fit. Early frequentist model averaging relied exclusively on the Akaike information criterion, which like the Bayesian information criterion, is an $R^2$ derivative.
Modern approaches increasingly use the focused information criterion (Hjort and Claeskens 2003, Claeskens and Hjort 2008), which explores the quality of a model from the perspective of the main variable of interest. However, difficulties emerge in empirical analysis since the concrete formulae and implementation for the FIC depends on the scientific and statistical context. In addition, the FIC approach is often inconsistent as it identifies one model as most appropriate for estimating a part of a distribution but another model as best for estimating the mean value.

The methodology of model averaging for dealing with model uncertainty is still in its early stages. What seems clear is that there is currently no single optimal algorithm for weighting different models and perhaps there never will be. Given uncertainty about the best way to undertake model averaging, we suggest turning the controversy about model averaging from its head onto its feet and employ the same robustness logic also for model averaging. In other words, we suggest that scholars average the results from their baseline model and the robustness tests using different model averaging techniques and assess whether their inference is robust across the range of techniques. We suggest scholars employ the following techniques:

1. Weight each model by its AIC.
2. Weight each model by its BIC.
3. Apply subjective weights for each model based on one’s subjective view on how well specified each of the models is.
4. Weight all models equally. In other words, do not weight at all, but take the unweighted model average.

The rationale for model averaging techniques 1 and 2 is to give models that fit the data better a higher weight. The problem with these techniques is that, as we have argued in section 2.1 on diagnostic tests, goodness-of-fit does not equal goodness of model specification. The two are correlated in the sense that well specified models will also fit the data well, but are nevertheless distinct and a better fitting model can still be a more misspecified model. In fact, models in which model fit increases spuriously by, for example, including endogenous variables are severely misspecified. The rationale for technique 3 is that researchers will have views on the relative merit of each of the robustness test models in terms of model specification. They also
must subjectively regard their baseline model as the best specified model – otherwise they would not have chosen it as the baseline. The disadvantage is that by allowing subjective weights one gives researchers leeway over the final robustness assessment as different weights can produce different overall assessments. Technique 4 can be justified on the basis that if there is strong model uncertainty, then researchers may not have a basis on which to prefer one model over the other and all should be equally weighted. The drawback of this technique is if all models are equally preferable, then it is somewhat unclear why one model was chosen as the baseline model.\textsuperscript{10} In the following section we illustrate some of these model averaging techniques with an actual example.

7. Robustness Tests and Inference: An Example
As an illustrative example, we take one of our own studies on the political determinants of famine mortality (Plümper and Neumayer 2009). Based on a theory of rational political activity, our argument claims that the number of fatalities ceteris paribus is a function of three interrelated variables: food aid provided by international donors, regime type, and the share of people affected by a food shortage to the total population of the country. Deviating from Amartya Sen’s famous claim that no democracy ever experienced a famine – a claim that does not hold up to empirical scrutiny – we argue that even democratically elected governments will at times find it political support maximizing not to prevent all fatalities in the event of famine. This occurs if the share of the affected population is small and the potential financial and political costs of redistributing food to affected areas high – a cost that is lowered by the availability of generous international food aid. However, we also argue that democracies still do more to prevent famine mortality than autocracies. Specifically, we argue that democracies respond more elastically to the simultaneous presence of international food aid and a large share of the population being affected. Thus, democratic governments have a higher propensity than autocratic governments to use international food aid effectively and the mortality gap between democracies and autocracies increases when both the share of the affected population becomes larger and when more international food aid is available.

\textsuperscript{10} One could apply the logic of robustness to the results from these different model averaging techniques as well. One could thus take the unweighted average over all averages from the various techniques to assess the overall robustness of one’s inference.
The empirical baseline model has the number of famine fatalities as the dependent variable and estimates the effect of food aid, the share of affected population and its interaction with separate effects for democracies and autocracies, where democracies have to score 5 or above on the combined (reversed) political rights and civil liberties measures of Freedom House and 6 or higher on the polity2 measure of Polity IV. The models are estimated with negative binomial regressions on a sample of developing countries over the period 1972 to 2000. In the main estimation model democracy is measured using Freedom House data in order to maximize sample size. In a robustness test data from the Polity IV project are used instead. Besides this one test and a Monte Carlo analysis in which random measurement error is injected into a small share of the estimates of famine mortality, no other robustness test is reported in the paper. Model 1 of table 1 replicates the baseline model from Plümper and Neumayer (2009). Note that in count data models such as Poisson and Negative Binomial regression, coefficients can be interpreted as semi-elasticities (Cameron and Trivedi 2009: 336), so we report coefficients rather than marginal effects at mean values or average marginal effects. The interaction effect of food aid with affected/population is negative and statistically significant, but larger in absolute terms in democracies than in autocracies as our theory predicts.

In translating the theoretical model into an empirical model we had to make a number of more or less arbitrary decisions with regard to model specification that can be subjected to robustness tests. These include:

1. **Uncertainty about the correct set of control variables and conditional effects among them.** In Plümper and Neumayer (2009), we added as control variables the incidence of civil wars, regime type, income per capita, population size and density, annual rainfall and net water availability. One major aspect of vulnerability to the mortal effect of food shortages neglected by these controls is the age composition of the population. Particularly, there is wide variation in the share of children among the population across developing countries. In model 2, we therefore add the share of the population under the age of 15 to the baseline model. As expected, we find that countries with a larger share of children and young adolescents exhibit greater famine mortality.

2. **Uncertainty about the independence of cases.** Civil wars are traditionally regarded as a major trigger of famine mortality. In our baseline model we control for the effect of civil wars within a country. However, civil wars in neighboring countries may have spill-over effects, either
triggering a famine or exacerbating it, for example because transport routes for channeling food aid into affected areas become interrupted. In model 3, we add a spatial component in the form of a spatial-x variable to our baseline model, by including the incidence of civil wars in geographically contiguous countries. We find that indeed a stronger incidence of civil wars in neighboring countries is associated with higher famine mortality.

3. Uncertainty about reverse causality. Civil wars might become exacerbated in the presence of a famine, hence the incidence of civil wars might be endogenous. In model 4 we therefore use a dummy for an existing conflict in the previous year, ethnic and religious fractionalization and the share of land area that is mountainous as instruments for civil war (data from Fearon and Laitin 2003), using a procedure described in Cameron and Trivedi (2009: 593f.). The insignificant coefficient of the residual from the first stage regression suggests there is no strong evidence for reverse causality.

4. Functional form uncertainty. In the baseline model, food aid and the share of affected population are interacted with each other and the effect of this interaction term is allowed to differ between democracies and autocracies by estimating two separate coefficients for each political regime type. This imposes a specific functional form restriction on our main variables of interest. In model 5 we explore whether inferences are robust toward a different, more flexible functional form. Specifically, we estimate a full 3-way interaction effect model between food aid, affected population share and the continuous measure of political regime type of Freedom House. With this alternative functional form specification, no easy check on our predictions are possible any longer, contrary to the baseline model and model 5 is therefore not included in our model averaging exercise. However, we can calculate average marginal effects in the form of semi-elasticities for international food aid at a specified level of affected population share (assumed to be one quarter) for various levels of the political regime variable. Consistent with our theory and baseline results, we find that the effect of food aid is indistinguishable from zero in the most autocratic regimes and becomes increasingly negative as the political regime becomes increasingly democratic.

5. Uncertainty about the measurement of our main concepts. Democracy is a latent concept. In Plümper and Neumayer (2009), we used the polity2 measure from the Polity dataset as an alternative measure of regime type to the Freedom House data used in the baseline model and we repeat the result here again as model 6. More importantly, famine mortality cannot be
counted, it can only be estimated. Accordingly, all measures of famine mortality are ridden with measurement error. In model 7, we employ an alternative data source using information from Munich Re’s NatCatSERVICE database. But even an alternative data source is fraught with measurement error of course. In Plümper and Neumayer (2009), we accounted for random measurement error by injecting an error of up to 50 per cent into approximately 15 per cent of observations. We showed that this makes little difference to our inferences, but one could reasonably argue that measurement error is likely to be systematic rather than random and that injecting measurement error into 15 per cent of observations is not radical enough.

We therefore let all observations be affected by measurement error and rather than injecting random noise we follow a similar strategy as in Keefer et al. (2011) who account for systematic measurement error in their estimates of earthquake fatality. Specifically, autocratic governments might have an incentive to downplay the actual number of deaths, but foreign observers might well over-estimate fatalities, knowing that domestic sources tend to underestimate them. This would result in systematically larger measurement error in autocracies, potentially going in either direction (under- or over-estimates). We therefore conducted a Monte Carlo study, in which we re-estimated the baseline model 100 times. In each re-estimation, we multiplied the value of the dependent variable of all observations in democratic countries by a random number drawn from the interval [0.5, 1.5], which mirrors measurement errors of up to 50%. For autocracies, it was drawn from the interval [0.25, 2] instead to reflect the larger degree of measurement error in these observations. Table 2 reports the full range of coefficients from the Monte Carlo study (minimum to maximum), thus taking in effect the impact of each single iteration onto inferences into consideration.
Table 1. Results from various robustness tests.

<table>
<thead>
<tr>
<th>Robustness test:</th>
<th>model 1</th>
<th>model 2</th>
<th>model 3</th>
<th>model 4</th>
<th>model 5</th>
<th>model 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>none</td>
<td>Additional control</td>
<td>Spatial-x of civil war</td>
<td>Civil war endogenous</td>
<td>Alternative regime type var (polity2)</td>
<td>Alternative source for DV</td>
</tr>
<tr>
<td>civil wars</td>
<td>1.490***</td>
<td>0.899***</td>
<td>1.377***</td>
<td>2.243***</td>
<td>1.276***</td>
<td>-0.240</td>
</tr>
<tr>
<td></td>
<td>(0.405)</td>
<td>(0.279)</td>
<td>(0.406)</td>
<td>(0.572)</td>
<td>(0.433)</td>
<td>(0.310)</td>
</tr>
<tr>
<td>democracy</td>
<td>-0.826***</td>
<td>-0.753***</td>
<td>-1.057***</td>
<td>-1.059***</td>
<td>-0.0889</td>
<td>-0.964***</td>
</tr>
<tr>
<td></td>
<td>(0.229)</td>
<td>(0.193)</td>
<td>(0.216)</td>
<td>(0.203)</td>
<td>-0.0652</td>
<td>(0.238)</td>
</tr>
<tr>
<td>GDP pc</td>
<td>-0.00365***</td>
<td>-0.00296***</td>
<td>-0.00373***</td>
<td>-0.00370***</td>
<td>-0.00420***</td>
<td>-0.000836**</td>
</tr>
<tr>
<td></td>
<td>(0.000410)</td>
<td>(0.000309)</td>
<td>(0.000321)</td>
<td>(0.000408)</td>
<td>(0.000512)</td>
<td>(0.000331)</td>
</tr>
<tr>
<td>Population (logged)</td>
<td>1.460***</td>
<td>1.821***</td>
<td>1.562***</td>
<td>1.315***</td>
<td>1.380***</td>
<td>1.393***</td>
</tr>
<tr>
<td></td>
<td>(0.281)</td>
<td>(0.262)</td>
<td>(0.305)</td>
<td>(0.399)</td>
<td>(0.304)</td>
<td>(0.236)</td>
</tr>
<tr>
<td>food aid</td>
<td>-0.000310</td>
<td>0.00161</td>
<td>0.000511</td>
<td>-0.000676</td>
<td>-0.000383</td>
<td>0.00510**</td>
</tr>
<tr>
<td></td>
<td>(0.00119)</td>
<td>(0.00108)</td>
<td>(0.00146)</td>
<td>(0.00124)</td>
<td>(0.000854)</td>
<td>(0.00256)</td>
</tr>
<tr>
<td>affected/population</td>
<td>72.58***</td>
<td>61.21***</td>
<td>77.67***</td>
<td>81.62***</td>
<td>68.19***</td>
<td>48.03***</td>
</tr>
<tr>
<td></td>
<td>(13.54)</td>
<td>(8.060)</td>
<td>(12.05)</td>
<td>(16.01)</td>
<td>(13.80)</td>
<td>(9.517)</td>
</tr>
<tr>
<td>food aid * affected/population in democracies</td>
<td>-0.118***</td>
<td>-0.0909***</td>
<td>-0.119***</td>
<td>-0.131***</td>
<td>-0.119***</td>
<td>-0.139***</td>
</tr>
<tr>
<td></td>
<td>(0.0285)</td>
<td>(0.0199)</td>
<td>(0.0236)</td>
<td>(0.0291)</td>
<td>-0.0332</td>
<td>(0.0410)</td>
</tr>
<tr>
<td>food aid * affected/population in autocracies</td>
<td>-0.0220***</td>
<td>-0.0155***</td>
<td>-0.0248***</td>
<td>-0.0255***</td>
<td>-0.0199***</td>
<td>-0.0370***</td>
</tr>
<tr>
<td></td>
<td>(0.00687)</td>
<td>(0.00422)</td>
<td>(0.00618)</td>
<td>(0.00789)</td>
<td>-0.00718</td>
<td>(0.0113)</td>
</tr>
<tr>
<td>population density</td>
<td>-0.176</td>
<td>-0.623**</td>
<td>-0.377</td>
<td>0.0520</td>
<td>-0.479</td>
<td>-3.066***</td>
</tr>
<tr>
<td></td>
<td>(0.395)</td>
<td>(0.316)</td>
<td>(0.379)</td>
<td>(0.413)</td>
<td>(0.436)</td>
<td>(0.333)</td>
</tr>
<tr>
<td>annual rainfall</td>
<td>-0.00160***</td>
<td>-0.00192***</td>
<td>-0.00168***</td>
<td>-0.00181***</td>
<td>-0.00153***</td>
<td>0.00438***</td>
</tr>
<tr>
<td></td>
<td>(0.000425)</td>
<td>(0.000415)</td>
<td>(0.000387)</td>
<td>(0.000455)</td>
<td>(0.000427)</td>
<td>(0.000756)</td>
</tr>
<tr>
<td>net water availability</td>
<td>-0.0156***</td>
<td>-0.0200***</td>
<td>-0.00556**</td>
<td>-0.0121*</td>
<td>-0.0162***</td>
<td>-0.00463***</td>
</tr>
<tr>
<td></td>
<td>(0.00572)</td>
<td>(0.00329)</td>
<td>(0.00228)</td>
<td>(0.00660)</td>
<td>(0.00456)</td>
<td>(0.000678)</td>
</tr>
<tr>
<td>share pop above 15</td>
<td>-0.468***</td>
<td>-0.468***</td>
<td>-0.468***</td>
<td>-0.468***</td>
<td>-0.468***</td>
<td>-0.468***</td>
</tr>
<tr>
<td></td>
<td>(0.153)</td>
<td>(0.153)</td>
<td>(0.153)</td>
<td>(0.153)</td>
<td>(0.153)</td>
<td>(0.153)</td>
</tr>
<tr>
<td>residual first stage regression</td>
<td>-0.993</td>
<td>-0.993</td>
<td>-0.993</td>
<td>-0.993</td>
<td>-0.993</td>
<td>-0.993</td>
</tr>
<tr>
<td></td>
<td>(0.621)</td>
<td>(0.621)</td>
<td>(0.621)</td>
<td>(0.621)</td>
<td>(0.621)</td>
<td>(0.621)</td>
</tr>
<tr>
<td>spatial-x (civil wars) W: contiguity</td>
<td>3.536***</td>
<td>3.536***</td>
<td>3.536***</td>
<td>3.536***</td>
<td>3.536***</td>
<td>3.536***</td>
</tr>
<tr>
<td></td>
<td>(0.933)</td>
<td>(0.933)</td>
<td>(0.933)</td>
<td>(0.933)</td>
<td>(0.933)</td>
<td>(0.933)</td>
</tr>
<tr>
<td>AIC</td>
<td>971.9</td>
<td>967.4</td>
<td>927.9</td>
<td>927.1</td>
<td>975.4</td>
<td>658.3</td>
</tr>
<tr>
<td>BIC</td>
<td>1047.1</td>
<td>1054.1</td>
<td>1007.8</td>
<td>1006.8</td>
<td>1050.1</td>
<td>732.5</td>
</tr>
<tr>
<td>---------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>Observations</td>
<td>2,399</td>
<td>2,399</td>
<td>2,224</td>
<td>2,193</td>
<td>2,304</td>
<td>2,240</td>
</tr>
</tbody>
</table>

Note: Negative binomial regressions. Robust standard errors in brackets. * statistically significant at 0.1, ** 0.05, or *** 0.01 level.
Turning to a discussion of the results, as mentioned already, table 1 reports the first set of our robustness tests. We find that the estimates of the effect of our main variables remain fairly stable. The coefficient for our central interaction effect, food aid multiplied by the share of the affected population, varies in democracies between -0.09 and -0.14 – our baseline model’s estimate being in the middle at -0.12. For autocracies, the interaction effect of interest varies between -0.015 and -0.037 – it was -0.022 in our baseline model.

Yet, not all variables in the model show the same stability. For example, the effect of civil wars remains stable unless we change the data source for the famine mortality rates. With Munich Re’s data the otherwise positive effect of civil wars on famine mortality disappears. Of course, in many cases it will be difficult to disentangle the causes for mortality in countries with famines and civil wars. Yet, if we use Munich Re’s estimates of famine mortality, other variables appear to have counterintuitive effects. For example, the effect of rainfall becomes reversed. Rather than reducing famine mortality, we find that more rainfall increases famine mortality. Given these somewhat implausible estimates, we would give model 6 a subjectively low weight in the model averaging exercise.

Table 2 displays the mean and the range of estimates from Monte Carlo simulations that add artificial systematic measurement error (larger in autocracies than in democracies) to the estimated famine mortality data. With very few cases of non-zero famine mortality, one might expect a large effect from adding measurement error into the estimates of fatalities. However, though we find that the measurement uncertainty is far larger than the sampling error (standard errors) of the baseline model, the addition of large measurement error leaves our main inferences basically intact: we obtain a range between -0.083 and -0.157 for our main interaction effect in democracies and a range of -0.013 and -0.033 for the same interaction in autocracies. Importantly, even at the minimum value of the Monte Carlo estimates, which represents the largest conditional effect of food aid in absolute terms, the effect is smaller in autocracies than at the maximum value in democracies, which represents the smallest effect in absolute terms on mortality. In other words, even with systematic measurement error and even looking at the two polar extremes, the inference that democracies react more elastically to food aid for any given share of affected population is robust. The results suggest that democratic governments use international food aid about 2.5 to 12 times more effectively than autocracies at every share of the affected population, holding all other factors constant.
Table 2. Summary statistics of Monte Carlo Analysis of Systematic Measurement Error in Famine Mortality Estimates.

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>civil wars</td>
<td>1.472</td>
<td>0.162</td>
<td>1.096</td>
<td>1.865</td>
</tr>
<tr>
<td>democracy</td>
<td>-0.835</td>
<td>0.124</td>
<td>-1.101</td>
<td>-0.537</td>
</tr>
<tr>
<td>GDP pc</td>
<td>0.233</td>
<td>0.009</td>
<td>0.213</td>
<td>0.254</td>
</tr>
<tr>
<td>Population (logged)</td>
<td>1.517</td>
<td>0.145</td>
<td>1.269</td>
<td>1.858</td>
</tr>
<tr>
<td>food aid</td>
<td>0.000</td>
<td>0.001</td>
<td>-0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>affected/population</td>
<td>72.166</td>
<td>7.069</td>
<td>56.336</td>
<td>90.309</td>
</tr>
<tr>
<td>food aid * affected/population</td>
<td>-0.120</td>
<td>0.014</td>
<td>-0.157</td>
<td>-0.083</td>
</tr>
<tr>
<td>in democracies</td>
<td>0.028</td>
<td>0.002</td>
<td>0.023</td>
<td>0.032</td>
</tr>
<tr>
<td>food aid * affected/population</td>
<td>-0.022</td>
<td>0.004</td>
<td>-0.033</td>
<td>-0.013</td>
</tr>
<tr>
<td>in autocracies</td>
<td>0.007</td>
<td>0.000</td>
<td>0.006</td>
<td>0.008</td>
</tr>
<tr>
<td>population density</td>
<td>-0.173</td>
<td>0.305</td>
<td>-0.821</td>
<td>0.380</td>
</tr>
<tr>
<td>annual rainfall</td>
<td>-0.002</td>
<td>0.000</td>
<td>-0.002</td>
<td>-0.001</td>
</tr>
<tr>
<td>net water availability</td>
<td>-0.015</td>
<td>0.004</td>
<td>-0.024</td>
<td>-0.006</td>
</tr>
</tbody>
</table>

Note: Based on 100 iterations. Larger measurement error injected into autocracies than democracies. First row shows coefficient, second row standard error.

To sum up, our main inference survives the extensive robustness tests. Importantly, however, the robustness tests reveal a far greater level of uncertainty than the size of the standard errors in our baseline model would suggest. The conclusion is simple and holds generally true: standard errors do not reveal the true size of uncertainty of causal inferences in quantitative analysis. We can take these results one step further by employing different data averaging techniques. Figure 1 displays the 95-percent confidence intervals for the models reported in table 1 plus results from several model averaging techniques for the two variables of main interest. One technique averages across models without weighting, the second and third employ, respectively, a model’s AIC and BIC as weights, whereas the fourth technique reproduces the sampling distributions of the point estimates, aggregates these over all models and computes the mean and the standard deviation. The result should indeed be very close to the analytical solution.
Figure 1a: Parameter Space and Model Averaging (M1-M6) for the Interaction Effect in Democracies

Figure 1b: Parameter Space and Model Averaging (M1-M6) for the Interaction Effect in Autocracies
Figures 1a and 1b demonstrate that a) our main results are robust, b) both variables of interest are significantly different from zero and c) the averaging technique does not make much difference for the assessment of robustness. We do not know whether the variation in results from different averaging techniques becomes relevant for other robustness tests or for other analyses, but the inference we make here are independent not only of the choice of models from the set of models reported in table 1 but also from the choice of model averaging technique.

8. Conclusion

Model uncertainty results from three sources: First, theory provides little guidance on the data generating process and thus on model specification. Second, theory also provides insufficient if any guidance on the true population for which a theory claims validity and – to make matters worse – data constraints often prompt researchers to base their estimates on convenience samples. Third, important concepts in the social sciences are not directly observable and therefore need to be measured by proxy variables. We argued that neither diagnostic econometric tests, nor millions of estimates based on partly randomized models in the form of extreme bounds analysis nor randomized controlled trials can solve the problem posed by model uncertainty.

Recently, applied scholars have begun to report the results of robustness tests. We argue that these tests provide a potential attractive alternative, but not as they are currently practiced. With no clear definition of robustness and an ad hoc selection of a small number of tests readers are left unsure about the reliability of causal inferences. Our main contribution has been to provide a more methodical and systematic grounding to robustness tests as a way of dealing with model uncertainty. We have defined robustness as requiring that causal inferences do not change with plausible permutations to the specification of one’s baseline model. Having identified at least 16 specific aspects of model uncertainty, we suggest that researchers need to conduct a much larger set of robustness tests than currently undertaken, even if it depends on the research question and design which aspects are more prevalent than others.

We have also proposed techniques to aggregate the information from each of the estimated models in order to assess robustness based on the entire available empirical evidence.
This implies that the focus on single point estimates, which seems to dominate current practice, is misplaced. Given model uncertainty, any one model could suffer from relatively severe misspecification such that no model should have veto power over causal inferences. Equally, none of the models will be the ‘correct’ one and none will therefore produce the ‘true effect’. It lies at the heart of our proposed new logic of statistical inference that we make no pretense to finding the true effect. However, robustness tests can tell us much more about the range in which the true effect is likely to lie than any single estimate ever could. Methodically and systematically undertaken, robustness tests will improve the reliability of inferences. Robustness tests are not the answer to all prayers. But they are an appropriate answer to model uncertainty.
References


