Population and Sample Uncertainty

Thomas Plümper[a*] *and Eric Neumayer*[b]

[a] Department of Government, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK, tpluem@essex.ac.uk, Tel. +44.1206.873567.

[b] Department of Geography and Environment, London School of Economics, London WC2A 2AE, UK, e.neumayer@lse.ac.uk, Tel. +44.207.9557598.

Abstract

Causal inference seeks to identify a causal mechanism and the set of cases for which the causal claim makes sense. It is always based on a specific sample analyzed. However, whether results can be reliably generalized to the population depends on whether the sample represents a true random draw from the population. In principle, the population can be deductively derived as the set of cases for which a theory claims validity or inductively derived as the set of cases for which the studied sample could have represented a random draw. However, researchers are typically uncertain about what constitutes the relevant population for testing their hypotheses as well as uncertain about what population their studied sample might represent. Both uncertainties render the underlying assumption of econometrics – that one studies a random draw from the population – fictional. We explore the principal sources of population and sample uncertainty and discuss the inferential threats they pose. We suggest tests for probing the robustness of inferences toward population and sample uncertainty, illustrating our suggestions with an example from the study of civil war onset.

1. Introduction

Researchers necessarily study samples, but always wish to make inferences to a wider population of cases. Thus, empirical analyses intend to establish results that can be generalized to a larger set beyond the sample studied. Scholars rarely are exclusively interested in the set of observations they happen to study in their specific chosen sample, nor can they typically study the entire set of observations for which they would like to make inferences.

In this paper, we tackle threats to making inferences from the sample to the population arising from scholars not knowing with certainty what the true population is for their analysis. Similarly, we analyze inferential threats from scholars either not knowing with certainty whether the sample they actually study represents the population or, conversely, not knowing with certainty to which population their results can be generalized. Population and sample uncertainty are connected: if uncertainty exists about what constitutes the population then it becomes impossible to ensure that one studies a sample randomly drawn from the population. Hence, population uncertainty generates sample uncertainty. Conversely, if one is uncertain about whether one has managed to draw a random sample or whether one can sufficiently correct for any non-random sampling (selection), then it becomes difficult to know to which larger set beyond the sample one can generalize one's findings. Hence, sample uncertainty generates population uncertainty.

Population and sample uncertainty do not necessarily lead to biased estimates and wrong inferences, but bias and wrong inferences are clearly possible and indeed very likely. In order to validly generalize from the sample to the population, sample and population properties should be identical or vary only arbitrarily. Systematic differences between the sample and the population cause bias and may invalidate the inferences.

We argue that demonstrating robustness is the best strategy for tackling population and sample uncertainty. We therefore, firstly, identify the principal sources of population and sample uncertainty and analyze how they threaten inferences and, secondly, discuss techniques that allow researchers to test the robustness of inferences in the presence of such uncertainty. These tests do not eliminate population or sample uncertainty, but they test whether inferences are robust towards changes in the definition of a population, the sampling technique, or the methods used for controlling for selection or missing data.

Given the importance in empirical analysis of generalizing from the sample to the population, the scant attention this process receives in econometric theory is baffling. For example, 'population' is not listed in the index of such standard textbooks as Baltagi (2008), Davidson and MacKinnon (2004), Dougherty (2010), Greene (2012), Hsiao (2003), Kennedy (2008) and Maddala and Lahiri (2009).[1] Textbooks on survey data analysis fare somewhat better. Scheaffer et al. (2006: 8), for example, clearly state: "A population is a collection of elements about which we wish to make an inference". However, even survey analysts do not seem to sufficiently grasp the full extent of population and sample uncertainty they face. More importantly, we will show that population and sample uncertainty become much more pronounced outside the field of survey analysis.

We start with a discussion of population and sample uncertainty in survey analysis and other social science research. We then suggest there are two very different ways to define a population, namely deductively and inductively. A more detailed discussion of the sources of population and sample uncertainty is then followed by our suggested robustness tests for establishing whether inferences are robust in the presence of population and sample

[1] Wooldridge (2000: 699) is a notable exception of an econometric textbook offering a definition of population that is, however, problematic as will become clear further below.

uncertainty. We illustrate some of these tests by exploring population and sample uncertainty in Fearon and Laitin's (2003) analysis of civil war onset.

2. What is Population and Sample Uncertainty?

Population, sample, case and observation are important ontological elements of statistics. Ontology discusses whether entities exist or when they exist how such entities can be grouped, related within an order, and subdivided according to similarities and differences. A population is a set of subjects (such as individuals, groups, institutions, countries etc.). It exists if and only if its subjects can be distinguished from other subjects that do not belong to the very population. A case is a single subject that can be observed more than once. A sample is a strict subset of the population. Samples can be selected or random. To be a true random sample, all observations of the population need to have an identical a priori probability of being drawn into the sample.[2]

Survey analysts enjoy the advantage over other social scientists that in principle their population is well defined by the set of individuals whose opinions, vote intentions or other characteristics they wish to survey. For example, their population may be the set of individuals eligible for voting in the US Presidential elections or the set of households in an administrative unit in Malawi or the set of Londoners.

Yet, despite a well defined population in principle, even survey analysts are faced with population uncertainty. For example, according to the Office for National Statistics (ONS), London's population in 2010 was both 7,825,200 and 8,278,251.[3] This difference of roughly

[2] If every case or subject is observed only once, then this translates into an identical a priority of cases or subjects being drawn into the sample.

[3] See data.london.gov.uk/datastore/. Last visited June 04th 2013.

450,000 individuals results from a slight difference in the definitions of London employed: While the ONS used London's administrative area to calculate the first figure, it uses London's urban area to calculate the second figure. In other words, London's urban area has a larger population size than the number of citizens administered by London's mayor. London is a moderate case with respect to the effect using different definitions of populations has on reported population size. Other cities show much bigger discrepancies between the size and the population of the administrative and the urban area. In Tokyo, the difference reaches an astonishing 22.5 million people. While the administrative city of Tokyo is reported to be 13,189,000, Tokyo's urban area reportedly has a population of 37,270,000.[4] It is very likely that a random sample drawn from Tokyo's administrative population has properties that significantly differ from a random sample drawn from Tokyo's urban area.

Another problem is that there is uncertainty about who forms part of a defined population and what their characteristics are, which renders it impossible to draw a truly random sample. Consider for example the problem posed by illegally residing residents. According to the UK's Home Office, the number of illegal immigrants in the UK lies between 310,000 and 570,000.[5] Let's assume that 440,000 plus or minus 130,000 is a fairly good approximation, though Migrationwatch UK,[6] a non-governmental organization that campaigns against immigration, publishes considerably higher estimates. Many of these illegal immigrants will live in London. Not knowing how large the population really is and who belongs in it will render true random sampling impossible because a sample is random if and only if all

4 See www.metro.tokyo.jp/ENGLISH/PROFILE/overview03.htm and esa.un.org/unup/. Last visted June 04th, 2013.

5 See news.bbc.co.uk/1/hi/uk_politics/4989874.stm. Last visited June 4th, 2013.

6 See www.migrationwatchuk.co.uk. Last visited June 4, 2013.

subjects belonging to the population have the same ex ante probability of being drawn into the sample.[7] This cannot be guaranteed if there is dissent about the boundaries of a population and many subjects are illegal. Clearly, it is not likely that an illegal immigrant has the same ex ante probability of being drawn into the sample as, say, Mr Eric Smith living at 12 Roundhouse Avenue.

These two examples illustrate population uncertainty and sample uncertainty, respectively. The two threats to inferences often come together such that scholars simultaneously face population uncertainty and sample uncertainty: It is not possible to precisely define the boundaries of the population and it is not possible to draw a random sample from that population such that every subject has an identical probability of being drawn into the sample. Moreover, as already stated in the introduction, population uncertainty typically generates sample uncertainty and vice versa.

Outside survey analysis, the dual problems of population and sample uncertainty are much more severe. If we wish to study the wage returns to an additional year of schooling, what is the relevant population? The American working population, the American population whether working or not, the population in all Western developed countries or all school leavers worldwide? Is the relevant population for the study of the effects of social welfare policies the set of Western developed countries from 1980 to the present? Or should the same countries before and after this period also count to the population? If so, how far back in time and how far into the future? Should other countries also form part of the population? If so, where to draw the boundary?

[7] Presuming all subjects are only observed once. If they were observed multiple times, then a true random sample would call for the same ex ante probability of all observations being drawn in to the sample.

In the next section, we discuss how one can determine what constitutes the population, i.e., the set of observations one wishes to generalize to. We suggest there are two fundamentally different ways to do this, one working with deduction, the other with induction. As we will see, determining the population is a much more complex task for most social science research designs than deciding between the urban or administrative boundaries for a city population as in our illustrative example from survey analysis.

3. Deduction, Induction, and the Population

In this section, we develop a dual view on how to determine the population, drawing on the established distinction between deduction and induction. Deductively, the population is determined by theory as the set of observations for which the theory claims validity. Depending on the precision of the theory, the population may become more or less clearly defined. Inductively, the population is the set of observations to which the results from a selected or given 'sample' can be generalized. Population in this approach can only be defined for a particular result based on a particular sample and is determined by the set of observations to which the empirical finding can be generalized. Since this generalization can only be valid for cases that have sufficiently close properties to the analyzed sample, the set of observations to which researchers can generalize findings can be infinitely small. In fact, the population can be indistinguishable from the sample. We now discuss both approaches for determining the population in more detail.

3.1. Deduction and the Population

A deductively derived population is the set of observations for which a given theory claims validity, about which researchers want to make statements, and to which researchers intend to generalize their findings. A deductively derived population is thus determined by theory and the population includes all observations for which the theory claims validity.

This stands in stark contrast to the definition of population offered by Jeffrey Wooldridge in his textbook *Introductory Econometrics* as "any well-defined group of subjects" (Wooldridge 2000: 699). Following such a definition, all white men with a registered telephone landline that lived in the Borough of Westminster in 2007 could be a population. And so would the registered working population in the United States on the 19th of August 2010 or the 2nd year undergraduate students of the University of Oxford that happen to be on campus at 4pm on the 19th of October 2012. Wooldridge's definition is a-theoretical and assumes away any selection bias, the existence and consequences of which have played such a prominent role in econometrics (Heckman 1979, Heckman 1990, Vella 1998), by simply defining the population as a well-defined group of subjects however non-representative they may be for the set of observations for which a theory claims validity.

If researchers were free to define the population simply on the basis that the group of subjects can be distinguished from other subjects, then the group of Harvard alumnis would be a possible 'population' for the question whether an additional university year adds to income. However, it is an entirely different question to ask whether an additional year at Harvard adds to lifetime income than to ask whether an additional university year affects lifetime income. An estimate based on a sample of Harvard alumnis would represent a biased estimate for the true effect as Harvard students represent a self-selected sample. Similarly, the debate about whether internet surveys lead to biased results (Couper 2000, Birnbaum 2004) would be foolish at best if Wooldridge's definition of a population made sense. Scholars could just define the population as the group of internet users, rendering selection bias impossible. A sufficiently large random draw from this population will have features very similar to the group of internet users, which was just defined to be the population. Finally, the debate about selection effects caused by systematically missing values (Whitehead et al. 1993) would not

make sense if we could just define the population as the set of subjects that have available data.

A deductively derived population is thus not just any ‘well-defined groups of subjects’ but the one well-defined group of subjects or, more generally, the well-defined set of observations for which a theory makes a prediction and the researcher would like to learn something about. For deductively derived populations, it is therefore equally important to carefully think about the boundaries for which a theory claims validity as it is to draw a random sample from the population. Only if all observations of a population have the same probability to be drawn into the sample will the sample (if it is large enough) represent the population. In contrast, a random sample of a selected subset of the population – a convenience sample – is not a random sample of the population and will have divergent sample properties in expectation, unless the convenience sample is identical to the theoretically defined population. For this reason, inferences from a randomly drawn sample of a pre-selected set of subjects suffer from selection bias and can be misleading. Likewise, if a treatment is randomized for a pre-selected set of subjects, inferences to the population suffer from selection bias. Thus, selection bias can be avoided if and only if researchers know the true population for which a theory claims validity and study a sample randomly drawn from this population or, if random sampling cannot be guaranteed, researchers sufficiently correct for any sample selection. Population uncertainty stems from the fact that most theories do not clearly specify the limits of their validity, which in turn results in sample uncertainty. Additional sample uncertainty can stem from the difficulty in drawing a random sample from the population or sufficiently correcting for sample selection.

3.2. Induction and the Population

An inductively derived population is the set of all observations to which an empirical result (finding, regularity) can be generalized. Though induction does not rest on random sampling but on a selected sample, the inductively derived population can, at least in principle, have the same properties as the sample. If this assumption holds, then the selected sample could have been a random draw of the induced population and if that is true, then the generalization to the population is valid. In other words, the inductively derived population is the one for which the actually studied sample could have been a random sample.

A special case of an inductively derived population is research that uses all available observations to test theories. For example, in our research on disaster mortality we use samples that include all cases for all time periods for which information is available (Plümper and Neumayer 2009, Keefer et al. 2011). Even for such exhaustive samples, the question of generalization to the same set of cases or other cases in the past or in the future is relevant, however. If, for example, in the future a structural break occurs – say, because geologists can forecast volcanic eruptions or earthquakes or because countries become politically, economically or socially very different from the type of countries we know today – our findings would no longer be valid to make predictions.

As an inductively derived population, Wooldridge's "well-defined group of subjects" can make sense as long as one is aware of the limits of generalization. Wooldridge (2000: 699) offers an example of a well-defined group of subjects: "In the population of all working adults in the United States, labor economists are interested in learning about the return to education." Results from a randomly drawn sample of the population of working adults in the US can be generalized to this population. But it is unclear whether they can be generalized to the *entire* adult population of the US since education influences the decision to bear children,

the age of childbearing, the choice of partner, and so on, such that education exerts an influence on the probability that an individual offers his or her time to the labor market and thus becomes part of the population of all *working* adults in the US. Likewise, it is unlikely that the results can be generalized to other countries since the impact of education on income depends on political, economic and social institutions that regulate the openness of societies, the institutional structure that determines pay, and the tax and social system of the country. Hence, a random sample of the US adult working population would not allow us to estimate the return to education in general. Instead, it would allow us to obtain an estimate for the average conditional return on education for a selected subset of American citizens.

Inductively derived populations are thus defined by the empirical research design chosen. If scholars define the set of registered British telephone landline owners in 2007 as "the population", they can easily generate a random draw, analyze it, and then generalize back to the population of registered British telephone landline owners in 2007. Yet, researchers typically want to generalize beyond such convenience samples. In the above example, scholars would probably want to generalize to all British citizens or even to all human beings. Yet, the further away from the convenience sample one goes, the more problematic becomes the generalization. Generalizing from British telephone landline owners in 2007 to landline owners in 2006 and 2008 is less problematic than generalizing further away in time or to other individuals. Unless one can argue that the properties of British telephone landline owners in 2007 are representative of these other larger populations, then this inference is likely to be invalid. Yet, British telephone landline owners will differ in some ways from British citizens and in more important ways from citizens in other countries. To provide another example: researchers that study the US political system tend to generalize their findings to other countries, yet the US system differs from the political system of other

countries in non-negligible ways, putting a question mark over whether such generalization is valid.

3.3. Inferences and Population Uncertainty with Induced and Deduced Populations

Given the importance of sampling, it is no surprise that deviations from random sampling caused heated debates between those who argue that non-random samples do not allow inferences and those that defend inferences from non-random samples (Smith 1983; Copas and Li 2002). If the former are correct, the majority of studies in the social sciences – those that fail to produce true random samples or that use convenience samples altogether – cannot make inferences and they would at best have a purely historical value. For others, however, inferences do not require random sampling. For example, Valliant, Dorfman and Royall (2000: 19) argue that the "claim that (…) probabilistic inferences are not valid when the randomization distribution is not available is simply wrong." For Copas and Li (2002), the generalizability of analyses of non-random samples is a continuum: for some data, sampling bias invalidates inferences, for other data it does not.

Yet, Valliant, Dorfman and Royall's statement is correct for inductively derived populations – that is for populations that by assumption or selection have the same properties as the sample. For every selected sample, there exists a population to which one can generalize the findings. In the worst case, this population is identical to the sample; in the best case the population is much larger than the sample. Thus, whereas Copas and Li (2002) argue that sampling bias is a property of the data, we would argue that sampling bias is a property of the generalization. The truth is, these are two sides of the same coin.

Uncertainty always remains about what constitutes the inductively derived population. A study of nationalization in Switzerland can possibly be generalized to a few other countries, but the result that committees are generally less inclined to discriminate against foreigners

than the general public (Hainmueller and Hangartner 2013) is unlikely to hold globally. If scholars analyze OECD countries, they might be able to generalize to other countries with similar properties as OECD countries, but probably not to developing countries; if they study a selection of British households they might be able to make inferences to other British households but probably not to households in other countries, and if they decide to conduct randomized controlled experiments with Oxford students they can make inferences to Cambridge students, we guess, but even that may already be overly optimistic.

With deductively derived populations, population uncertainty exists because the theory might not be conclusive enough to specify the limits of its validity and thus the limits of the population, or the boundaries of the population cannot be observed with certainty, or measurement error of the observations makes it difficult to judge whether observations are in or out of the true population. When scholars claim to study the entire population, it may be doubtful whether single observations belong to the population and thus to the study.

When the population is inductively derived, population uncertainty will be even larger. First, it is unclear whether all observations included in the sample belong to the population. And second generalizing beyond the convenience sample will always remain contested in the absence of a guiding theory. The gist of the problem encountered in inductively deriving the population is thus that scholars do not know 'how far they can go' in generalizing. Since the validity boundaries of their findings are not defined by theory, they must use analogies to transfer results from one set of subjects to another set of subjects. This 'other set' of subjects must be (close to) identical in all relevant dimensions – but since there is no theory to determine the relevant dimensions, research with induced populations cannot easily be generalized in a way that is known to be valid.

To conclude, researchers should carefully distinguish between deductively and inductively derived populations. Neither mode of determining the population is unproblematic, but the emergent problems are very different. With inductively derived populations, case selection is pragmatic but the identification of the population problematic. With deductively derived populations, the population is in principle determined by theory, but sampling is an issue since random sampling from the population is typically impossible, not least because theories are often unclear about the boundaries of the population for which they claim validity.

4. Sources of Population and Sample Uncertainty

As the previous section has shown, both deductively and inductively derived populations will suffer from population and sample uncertainty. In this section we discuss the main sources of population and sample uncertainty in more detail.

4.1 Under-specified Theories

Since theories often fail to specify the boundaries of their validity, the 'true population' remains as unknown as the 'true data-generating process'. Sometimes theories can sufficiently define the population, while the extent of population uncertainty can be very large for other theories. As an example for the former we refer to our own research on earthquake mortality (Keefer et al. 2011). Here the population for our theory of the political determinants of earthquake mortality is defined by all cases experiencing an observable event, an earthquake with potentially fatal consequences. Uncertainty can only occur as to the minimum strength of an earthquake that can kill human beings, which can be approximately known. As an example for a theory with a largely unknown population consider the effect of education and income. The set of cases for which an economic theory of the income returns to education claims validity is rather unclearly defined. Does it claim validity for the working population or the entire adult population? Does it claim validity for market economies or also

for countries and regions in which very little industry exists and in which service industries are confined to government agencies and other public service companies?

For deductively derived populations the problem posed by under-specified theories is clear. However, under-specified theories also pose a problem for inductively derived populations. Recall that the induced population is the one for which the actual sample studied could have been a randomly drawn sample. Theory can help in determining what would constitute the induced population for the sample studied and under-specified theories hamper the process of making judgments on 'how far' one can generalize the results from one's sample.

4.2 The Future, the Present and the Past

Theories rarely specify the temporal domain to which they apply. Many seem to be formulated as if they claimed eternal validity, others with the present day and age in mind. The future poses problems for both deductively and inductively derived populations. For deduced populations derived from theories that claim validity for the future as well, one obvious problem is that no randomly drawn sample can be established since future observations have not yet materialized. Yet, researchers need to be willing, to some extent at least, to generalize their findings into the future for their research to have any impact on policy-making. Otherwise a finding from a study of the past would provide no guidance on whether the future can be expected to be influenced in the way the finding suggests. For example, a finding that smaller class sizes improved the study performance of pupils in the past would only be of historical value if one did not believe that this result can be generalized into the future and thus smaller class sizes, if enacted, could be expected to similarly raise the study performance of present and future pupils. This is where the deductive and inductive modes converge: even for a deductively derived population, the sample studied can only reflect the past up to, at best, the present and scholars need to justify that the results can be

generalized into the future rather analogously to what one has to do with inductively derived populations. As with induced populations, the further into the future one ventures in generalizing the findings the more problematic the validity of such generalizing.

The past and present also cause problems. This is clearest for all those cases in the present or the past for which a theory claims validity, but that cannot be included in the sample because data are missing. This will cause uncertainty about the extent to which the sample is representative for the population. For inductively derived populations, the question arises how far the findings can be generalized back into the past. This question is particularly acute if one has no past data and cannot go back in time to collect data, but still wishes to generalize from one's findings to observations in the past.

4.3 Unobservable Boundaries

The third source of population uncertainty occurs when the population is clearly defined in theory, but the criterion that divides the population from similar cases that do not belong to the population cannot be directly observed. For example, assume that a theory claims validity for democracies only. Since democracy is a contested multidimensional concept that can neither be precisely measured nor observed, the boundaries of the population of democracies are blurred. While some countries are democracies with certainty, other countries only count as democracies with some probability smaller than one. Inductively derived populations similarly suffer from unobservable boundaries since it becomes unclear whether cases included in the sample should really be in the sample.

4.4 Sampling Issues

Fourthly, even if scholars know the population fairly well, it may not be possible to draw a true random sample. This problem occurs because for such a sample the ex ante probability

of being drawn into the sample has to be identical for all observations from the population. However, it may not be possible to guarantee such a perfect random draw because some units of analysis are difficult or even impossible to actually observe.

Sampling problems affect even survey analysts who have typically fewer problems determining the population than other social scientists. To give an example, assume the population that lives within the M25 ring around London defined the population. At any point in time, many of the individuals living within the confines of the ring are easy to identify: the registered population of London. Yet, a number of individuals living in the area will be unregistered, some of them legally residing, others illegally. The probability of sampling unregistered individuals is not zero (depending on the sampling technique), but certainly not identical to the probability of sampling registered individuals.

This problem is amplified in the territories of failed states, in war areas or areas affected by natural disasters. It is, all other things equal, also easier to draw a random sample in richer than in poorer countries. Consider survey research that requires a random sample of the rural population in, say, Burundi or Malawi. The absence of reliable administrative records, the scattered villages, poor infrastructure and so on render it impossible to guarantee a sample that has properties identical in expectation to those of the population. In sum, while the testing of theories depends on samples with properties similar to the population, it is often difficult to draw random samples, which creates sample uncertainty.

4.5 Model Complexity and Population

The population is not independent of the complexity of the estimation model. This should be obvious for inductively derived populations: to which subjects a particular result based on a particular sample and estimation model can be generalized will be a function of the complexity of the estimation model. However, if cases are heterogeneous, then the problem

also arises for deductively derived populations. If the theory claims validity for a set of cases that are heterogeneous then either one needs to model this heterogeneity with interaction effects and other complex modeling techniques or else – if heterogeneity cannot be adequately modeled – the reach of claimed validity of the theory needs to be curtailed to the set of cases that are homogeneous. In other words: population size and empirical model complexity are inversely related if case heterogeneity exists. The simpler the empirical model, the smaller the population for which it can produce valid estimates; conversely, the larger the population for which a valid estimate is sought the more complex the estimation model has to be.

As a simple example, an estimation model that does not control for democracy can either only include democracies or autocracies but not both if the democracy–autocracy distinction makes a difference for the dependent variable and democracy is not controlled for in the empirical model. However, once scholars control for democracy in the model, the population changes since now the variance in political regime type is accounted for. By contrast, if scholars only analyze democracies, they cannot generalize their findings to autocracies. As a perhaps more interesting example, randomized controlled trials (RCTs) try to identify a causal treatment effect for a selected group of individuals over which treatment status is randomized. If case heterogeneity in the effect of treatment exists, the population to which any identified effect can be generalized is confined to cases which if given the same treatment in a similar fashion would show the same effect. The findings of an RCT from a village in rural India may not be generalizable to urban dwellers in Brazil if rural Indian cases and urban Brazilian cases exhibit treatment heterogeneity. Randomizing treatment in a non-randomly selected sample thus does not produce valid inferences (Ho et al. 2007: 205). One can randomize treatment within a pre-selected group as much as one likes, the randomization will not solve the pre-selection bias, not even if the sample is very large. Unfortunately, in

social science experiments it is easy to randomize treatment, but very difficult (though not impossible) to do so for random samples of the population for which a theory claims validity (Shadish, Cook and Campbell 2002: 91f.).

5. Robustness Tests for Population and Sample Uncertainty

In this section, we discuss robustness tests for population and sample uncertainty. Most of these tests are well known and often used by applied researchers when they suspect the existence of heterogeneity across cases or when they face missing values or sample selection in their research design.

5.1 Sample Heterogeneity and Deviant Cases

In a well specified estimation model, a selected or randomly drawn case that does not 'belong' to the population should reveal its existence by a large residual. With normally distributed residuals, only approximately 1 in 22 observations will have residuals that are larger than twice the standard error and only 1 in 370 observations will be further away than three times the standard error. A more deviant residual than expected may thus indicate an erroneously included case. Unfortunately, non-normally distributed residuals may also be caused by any other type of model mis-specification, such that non-normality of residuals does not conclusively indicate outliers in the sense of not belonging to the population.

Besides standard outlier tests (Belsley, Kuh and Welsch 1980), another possible test for the existence of outliers is to estimate case-specific coefficients for the variables of interest. This is similar to a random-coefficients model of the following form, which requires panel or cross-sectional time-series data:

$$y_{it} = \alpha + \beta_i x_{it} + \gamma z_{it} + \varepsilon_{it}\,.$$

Outliers are presumed to be absent if the β_i show an almost normal distribution. Outliers are likely if, for example, the distribution of β_i has two peaks or if there are two or more distinct groups. Again, rejecting the assumption of normally distributed β_i does not necessarily indicate that the true population is smaller than the assumed population. It may also indicate the existence of other model mis-specifications such as an unobserved interaction effect or a mis-specified functional form of the interaction or measurement error.

Jackknife and bootstrap estimations of standard errors explicitly model the existence of heterogeneity as 'uncertainty', resulting in larger standard errors if observations are heterogeneous. They will also produce larger standard errors if observations are erroneously included in the sample. But jackknife and bootstrap techniques are not confined to alternative estimations of standard errors. They can also be useful in searching for overly influential observations. In a jackknife robustness test, one systematically excludes one or more observations from the estimation. With a 'group-wise jackknife' robustness test, researchers systematically exclude a number of cases that group together by satisfying a certain criterion – for example, countries within a certain per capita income range or all countries on a certain continent. Bootstrap is logically identical to jackknife, but holds the sample size constant by excluding some observations or cases but sampling some other observations or cases more than once so that the total number of observations remains constant. One can think of bootstrap as a resampling with replacement technique.

In sum, testing the robustness of the empirical results from one's baseline model toward the exclusion of detected outliers and toward single or group-wise jackknife or bootstrap techniques is our first suggestion for robustness tests in the face of population and sample uncertainty. Such tests will provide insights into the robustness of empirical results. Whether any apparent lack of robustness is caused by wrongly included observations in the sample or

other forms of model misspecification is difficult to say. As always, robustness tests do not obviate the need to specify one's estimation model as well as one can.

5.2 Missing Data: Interpolation, Out-of-Sample Prediction and Multiple Imputation

Samples can be problematic even if the population is clearly defined when relevant information is missing. Missing observations would not constitute a major problem if they were missing completely at random (they would still reduce the efficiency of the estimate then). Often, however, observations will not be missing completely at random (MCAR), but will be merely missing at random (MAR), in which case the probability of a value being missing is a function of observable values, or even missing not at random (MNAR),[8] in which case the probability of a value being missing is a function of both observable values and unobservable ones (Rubin 1976; King et al. 2001; Schafer and Graham 2002). This standard terminology is somewhat confusing or "unintuitive (for historical reasons)" (King et al. 2001: 50): values called missing at random many would call systematically missing values. Sticking to standard terminology, however, samples with observations that are not missing completely at random, i.e. that are either merely missing at random or are missing not at random, cannot constitute random draws from the population.

One could in principle try to correct for the effect of missing observations by sampling techniques (see section 5.5). For example, if observations for autocracies were more likely to be missing than observations for democracies then one could try to 'over-sample' more observations from autocracies. However, this is problematic because missing observations from autocracies may differ from non-missing observations from autocracies on other

[8] If the probability of an observation being missing also depends on unobserved values, then the process is resulting in MNAR observations is called 'nonignorable'.

dimensions relevant to the estimation model. The dominant technique for dealing with missing observations is therefore to recover these observations via imputing missing values. Imputation comes in three main variants: interpolation/extrapolation, theory-based out-of-sample predictions and multiple imputation techniques.

The oldest and least sophisticated variant is interpolation and extrapolation. Interpolation works best when earlier and later information is available, say when family income in 2004 and 2006 is known but is unknown for 2005. The odds are that the missing value is similar to the adjacent known values and probably falls between the adjacent values of family income. The most common form of interpolation is the linear one. It assumes that the missing family income of 2005 is the mean of the 2004 and the 2006 income – an assumption that is wrong (other than by coincidence), but probably close to the true figure. Linear interpolation is not the only possibility, however. Instead, one may wish to interpolate using the lower or the higher of the two adjacent values or some weighted average between the two. For example, in Plümper and Neumayer (2010) we interpolate missing values of the *polity* variable from the Polity IV project that are due to interregnum and affected transition periods once linearly, once by the minimum and once by the maximum of adjacent values. If extraneous information is available, one way of interpolating may be known or at least be suspected to be preferable to the other; otherwise multiple ones can be used in separate robustness tests. Of course, there is no guarantee that the true missing value will indeed lie between the adjacent known values. If the reason for missing values is that the case experienced a strong exogenous shock during the period of missing values then the true values may well lie outside the boundaries of adjacent known values.

Arguably, interpolation is more reliable than extrapolation. Extrapolation can only work reliably if there is no structural break in the data, but uncertainty grows with every additional data point extrapolated into the future or into the past. Linear extrapolation is again just one

possibility. For series known to grow exponentially, for example, extrapolation based on the assumption of exponential growth may be preferable.

We prefer theory-based out-of-sample predictions for missing values to inter- and extrapolation. The classical version of out-of-sample predictions relies on a theoretical model that explains the variable that has missing values. For example, in Plümper and Neumayer (2010) we use a theoretical model of the determinants of democracy to make out-of-sample predictions for missing values of the *polity* variable from the Polity IV project that are due to interregnum and affected transition periods in countries. These predictions can then be used in lieu of the missing values.

A modern variant of the out-of-sample predictions technique relaxes the assumption that researchers necessarily know a useful theoretical model to generate these predictions. The multiple imputation approach (King et al. 2001) uses all available information in the dataset to impute missing values and Markov Chain Monte Carlo (MCMC) techniques generate a probability density distribution of imputed values. Since standard imputation models often work very poorly for time-series cross-section data since such data violate the assumptions of conditional independence and exchangeability of observations, the lags and leads of the variables that have missing values can also be included in the imputation model in order to impose smoother trends in the variables to be imputed than what is typically generated by a standard imputation model. Priors in the form of existing expert knowledge can also be included (Honaker and King 2010). Using the various imputed values for the missing observations in the estimation model gives a probability density distribution of estimated coefficients.

The multiple imputation approach thus implicitly relies on the logic of a kind of robustness test. It has the further advantage that it takes into account the added uncertainty that comes

with the imputation of missing values, whereas interpolation/extrapolation and theory-based out-of-sample predictions replace missing values with single values as if they were known such that the standard errors in the estimation model are under-estimated since they fail to take into account the additional sampling uncertainty that results from imputation. Of course, in principle MCMC techniques could be applied to generate a probability density distribution of imputed values rather than unique imputed values for different interpolation/extrapolation variants or different runs of theory-based out-of-sample predictions as well. In any case, the results from multiple imputation techniques can and should be compared to other techniques for dealing with missing values including theoretically-based out-of-sample predictions and interpolation and extrapolation – see Plümper and Neumayer (2010) as an example of the simultaneous application of all three variants for dealing with missing values.

One can also use the fit of out-of-sample predictions as a kind of validation exercise for imputation techniques: better specified models should have better out-of-sample predictive power than less well specified ones (Keane 2010). One can use this insight to test the reliability of the used imputation technique by artificially producing a set of missing values, imputing the missing values with the technique, and re-estimating the model. If the new results based on the artificially imputed values are not robust to the old results before missing values were artificially created and imputed, then the used imputation technique is not reliable and should not be used. Of course, this validation exercise cannot be employed when observations have missing values for a reason and this reason is absent in other observations. In the example of Plümper and Neumayer (2010) referred to above, the validation exercise for establishing the reliability of imputation techniques cannot be used because observations that do not actually have missing values are systematically different from observations that have missing values.

5.3 'Over-sampling' and Sample Selection Correction for Dealing with Population Boundary Uncertainty

When population boundaries are uncertain, scholars can 'over-sample' by employing a comprehensive sampling strategy from the start that includes observations for which it is questionable whether they belong to the population and thus should be eligible for inclusion in the sample. Such a comprehensive sampling strategy is preferable to a less comprehensive one simply because it is easier to eliminate observations from a sample than to add observations or to control for selection after one has already sampled.

Assume we believe that p1 is the population for our theory, but there is some positive probability that the theory is also valid for observations included in p2. By sampling once from p1 for the baseline estimation model and sampling from the set union of p1 and p2 for the robustness test model, one can test whether the inclusion of observations also drawn from p2 exerts a statistically significant influence on the inferences based on a sample drawn from p1 alone. If not, then the results from the theoretically plausible sample drawn from p1 are robust to the inclusion of observations from p2.

If, for whichever reason, researchers failed to sample observations from p2, the possibility of the union set being the true population leads to a possible selection problem, given one only sampled from p1. Selection problems are more widespread than often meets the eye. Hug (2003) provides three examples: the identification of ethnic minorities, necessary for the study of ethnic violence or discrimination against ethnic minorities, is typically based on a selected sample of minorities that have previously experienced discrimination or that have self-selected into mobilization; studies of new party formation suffer from self-selection of those who believe their endeavor may result in electoral success; research on social or protest

movements typically have to rely on media reports which select certain stories and events over others.

Without being able to re-sample observations from p2, one can test for the robustness of one's inferences toward potential sample selection by estimating a selection model to correct for possible sample selection. Assume researchers excluded the observations from the set p2 because p2 includes countries which are not fully democratic. The selection model would then be a model that predicts the level of democracy. Predicting the level of democracy in a first stage model allows researchers to calculate the inverse of the Mill's ratio – the non-selection hazard – and control for this in the main second stage (Heckman 1979).[9] Importantly, the robustness test must compare the estimated effect from the baseline model that is based on the sample of p1 without potential selection bias correction to the estimated effect from the sample selection model that takes into account the effect of variables onto both stages of the estimation – not just the second or outcome stage, but also the first or selection stage (Sigelman and Zeng 1999). For example, if one wishes to test whether one's estimate of the wage return of an additional year of schooling estimated based on a sample of working adults only is robust toward potential selection bias then one should compare this estimate to the wage effect of additional schooling that is the combined effect of an additional year of schooling increasing the likelihood of one entering the workforce and increasing one's wage once in employment.

Much discussion about selection models has focused on whether the inclusion of a variable that affects selection, but does not affect the outcome in the second stage (the so-called

[9] Many more models going beyond Heckman's (1979) seminal model have been developed for the purpose of correcting for selection in specific research contexts – see, for example, Sartori (2003), Boehmke (2003), Hug (2003) and Boehmke et al. (2006).

exclusion restriction) is strictly necessary for Heckman-style sample selection models (Leung and Yu 1996; Vella 1998; Puhani 2000). Without such a variable the selection model depends on the non-linearity of the inverse Mill's ratio for identification, which some find highly dubious (Breen 1996; Wooldridge 2010). The problem is that variables which exclusively affect the first (selection) stage, but not the main second stage are very hard to find. This is similar to the problem of finding valid instruments in instrumental variable (IV) techniques. In fact, IV techniques represent an alternative to selection models if one can find a source of independent variation that affects the selection stage, but not the outcome in the second stage. Whereas selection models try to model the selection process completely, IV techniques merely try to purge the second stage from the endogeneity that results from the self-selection of observations.[10] Selection models depend on stronger distributional assumptions and a well-specified selection stage model (Caliendo and Hujer 2006). Yet, IV techniques face many problems of their own. Most importantly, it is extremely difficult to find instruments that are valid, i.e. do not have a direct effect onto the dependent variable, and yet are strongly partially correlated with the variable that determines selection, where partially means conditional on the other explanatory variables in the estimation model. Also, importantly, IV techniques will recover a causal effect only for those cases which have experienced variation in treatment as a result of variation in the instrument. As Angrist (2004: C56) has put it: "IV estimates capture the effect of treatment on the treated for those whose treatment status can be changed by the instrument at hand". This may result in an estimate that is not representative for the population since in general this so-called local average treatment effect (LATE) will differ from the population average treatment effect (PATE).

[10] If the determinants of selection are time-invariant, then differences-in-differences methods provide yet another option for dealing with selection since they erase time-invariant selection effects.

5.4 Re-Sampling Techniques

Another robustness test would call for testing whether results change if the same model is estimated on another sample that is in part or in full different from the sample used in the main estimations. For example, if two entirely different samples could both be considered representative of the population, then estimation results should be robust toward using the same model specification for one sample and then for the other. Such a robustness test would appear particularly warranted where researchers deliberately forego assembling a true random sample of the population in order to save on data collection costs as with case-control research designs for rare events data (King and Zeng 2001a, 2001b). In the sciences, cross-validations with different samples are frequent and belong to what is considered normal science. In the social sciences, however, cross-validation is rare, probably because scholars find it difficult to publish these studies. This makes it even more important that authors conduct cross-validation tests with different samples themselves.

Recent progress in re-sampling analysis stems from identification approaches. Regression discontinuity or matching designs can be understood as re-sampling strategies in which certain observations are selected in order to allow, it is believed, the identification of an unbiased causal effect. Importantly, since both techniques select, researchers trade the low omitted variable bias or functional form bias of these research designs against the selection bias that comes from not analyzing samples with properties identical to the population properties. One can treat the results from these designs as a robustness test for observational studies that do not follow the identification approach and thus do not select certain observations over others. In other words, since the observational study based on a random non-selected sample potentially suffers from bias due to insufficiently controlling for confounding factors, one can test whether inferences based on such a model are robust toward an 'identified' effect that does not suffer from such confounding but is based on a pre-

selected sample. Alternatively, since regression discontinuity or matched samples are perfect subsets of the original sample, one can endogenize the selection of observations into these samples and construct a Heckman-type selection model that controls for the selection effect of regression discontinuity or matching designs along the lines discussed in section 5.3.

5.5 Selection Uncertainty

Survey companies invest large sums of money to produce samples with properties close to the desired population. And yet, at least in conversations, survey researchers are willing to admit that the ideal of sampling – that all subjects of the population have an identical ex ante probability of being drawn into the sample – cannot be achieved. If this is true, then even purportedly 'random samples' of the true population suffer from a moderate selection problem because the sample will include too many subjects that are easy to sample and too few subjects that are difficult to sample.

Researchers can econometrically reduce this problem if they have information on the true properties of the population, for example if reliable census data of the same population is available. It is possible then to correct the random sample so that the properties of the sample and the properties of the census data population become more similar (Scheaffer et al. 2006; Lohr 2010). If reliable census data do not exist, then it becomes difficult to correct the sample to render it akin to a true random sample. One can then use over- and under-sampling techniques as well as the estimation of separate effects for different groups as robustness tests.

Assume, for example, one suspects that ethnic minorities and illegally residing individuals are under-represented in the sample, but one cannot fully correct for under-representation since no reliable census data exist for such individuals. In robustness tests, one can employ several different assumptions of presumed under-representation and over-sample subjects

drawn from ethnic minorities and illegally residing individuals already contained in the sample. One employs a Monte Carlo study that mirrors a bootstrap to repeatedly draw additional subjects from the under-represented group into the sample. Each of these estimates generates standard errors that are slightly too small, but researchers can easily correct this by estimating these resampling models repeatedly and by using the 'resampling variation' as additional source of uncertainty and adding it to the mean standard error of the estimates. One can also over- and under-sample cases interchangeably. Rather than drawing subjects from under-represented groups more than once into a sample, researchers can also eliminate subjects from the over-represented group from the sample. Over- and under-sampling strategies are best regarded as alternative robustness tests and can also be employed to test robustness to each other.

A further robustness test consists in the estimation of separate effects for the under-represented group in the sample, which is particularly helpful if one is very uncertain about the extent of under-representation of certain groups in the sample. If the group-specific effects do not statistically significantly differ from the effects for the remaining sample, then one can conclude that the inferences drawn from the sample that deviates from a true random sample are robust toward the potential, but uncertain under-sampling of the tested group.

Things are more complicated if groups of subjects are not merely under-represented, but entirely missing. The best researchers can then do is to construct a Heckman-style selection model that explains the selection of all other groups into the sample and the exclusion of this group from the sample. If such a selection model cannot be constructed, inferences should be formulated in a way that makes it clear that they do not hold for the excluded group. To give an example: Assume a polling company uses the telephone book to draw a random sample. This sample will exclude all subjects that do not own a telephone landline and subjects that have opted out of the inclusion in the telephone book. Census data may allow one to identify

the individuals/households which do not appear in the phonebook, but may not be comprehensive enough to allow adjusting the drawn sample to make the properties of the sample and the properties of the census data population sufficiently similar. Explaining the non-appearance in the telephone book (by income, age, sex, profession and other factors), one can estimate a non-selection hazard, which will be non-zero for many of the subjects in the phonebook which will allow one to test for potential selection bias in the estimations based on the random sample drawn from the telephone book population. If no data exist allowing the identification of individuals/households not listed in the phonebook, then scholars should acknowledge that their findings can be generalized to the population of telephone landline owners registered in the phonebook, but not necessarily to the true population.

6. An Example Application

In most research projects it will not be necessary to conduct all the above robustness tests but a selection of the ones most appropriate for the specific research design at hand. The importance of a particular test clearly depends on the extent and nature of sample and population uncertainty. In this section, we illustrate some of the tests by exploring the robustness of inferences in Fearon and Laitin's (2003a) study "*Ethnicity, Insurgency, and Civil War*".

Fearon and Laitin (hereafter F&L) test a large number of hypotheses – eleven main ones, several of which branch out into sub-hypotheses – in their analysis of the determinants of civil and ethnic war onset over the period 1945 to 1999 covering 161 countries that had a population of at least half a million in 1990. Two main hypotheses are that, firstly, ethnic and religious fractionalization are not determinants of either civil or ethnic war onset, whereas state capacity, as proxied by a country's per capita income, is a strong predictor of the

absence of war. Complementing the latter hypothesis is that "anocracies" – regime types that are neither clearly autocracies nor democracies – are found to be more prone to civil war onset, which F&L interpret as another indication of "weak" states – too weak to turn themselves into either clear autocracies or democracies – falling prey to internal violent conflict. It does not matter for our purposes here that a country's per capita income as well as the political regime type of anocracy may measure or proxy other characteristics of a country than state capacity. We simply accept F&L's interpretation and test for the robustness of their findings toward population and sample uncertainty. F&L (2003a) mention a very large number of robustness tests, which are documented in detail in F&L (2003b). They find robust support for their main hypotheses.

One of F&L's robustness tests relates to population and sample uncertainty. Specifically, they impute missing data for 283 (out of 6610) country years that account for 5 (out of 111) onsets and find their results to be robust. This is not surprising, not least given the relatively small number of missing data. Since F&L include this robustness test already, we omit it here, concentrating on other tests instead.

Model 1 of table 3 replicates model 3 of table 1 of F&L (2003a). In models 2 to 7 we apply regional jackknives, testing whether their results are robust to slicing away observations of certain regions from the sample, as suggested in section 5.1. Results for per capita income are very robust in terms of coefficients maintaining negative signs at high levels of statistical significance across all models. In terms of substantive size, the effect becomes significantly stronger if Eastern Europe or North Africa and the Middle East are excluded and significantly weaker if Asian or Western countries are excluded, but the effect remains substantively strong throughout. Results for ethnic and religious fractionalization are very robust, continuing to be statistically insignificant throughout. The finding that anocracies face a higher risk of civil war onset turns out to be somewhat fragile to this robustness test. The

anocracy dummy variable becomes statistically insignificant if Eastern European, Sub-Saharan African or Asian countries are excluded from the sample. However, due to the large standard errors the estimated coefficients from these models are not statistically significantly different from the baseline model. The same goes for the other jackknife samples.

Table 1. Regional jackknife robustness tests

Exclude:	None (baseline)	Eastern Europe	Latin America	Sub-Saharan Africa	Asia	North Africa & Middle East	Western countries
	1	2	3	4	5	6	7
Prior war	-0.916**	-0.914**	-0.892**	-0.920*	-1.105*	-0.939**	-0.957**
	(0.312)	(0.314)	(0.326)	(0.380)	(0.456)	(0.331)	(0.315)
Per capita income (lagged)	-0.318***	-0.368***	-0.316***	-0.307***	-0.241**	-0.399***	-0.258**
	(0.0714)	(0.0836)	(0.0765)	(0.0744)	(0.0782)	(0.0893)	(0.0801)
log(population) (lagged)	0.272***	0.260***	0.294***	0.274**	0.163	0.315***	0.266***
	(0.0737)	(0.0736)	(0.0804)	(0.0869)	(0.111)	(0.0786)	(0.0738)
log(% mountainous)	0.199*	0.184*	0.220*	0.201	0.272**	0.117	0.196*
	(0.0846)	(0.0866)	(0.0908)	(0.125)	(0.0992)	(0.0906)	(0.0856)
Noncontiguous state	0.426	0.380	0.209	0.355	0.651	0.504	0.586*
	(0.272)	(0.289)	(0.289)	(0.307)	(0.403)	(0.290)	(0.290)
Oil exporter	0.751**	0.739*	0.949**	0.841**	0.558	0.681	0.630*
	(0.278)	(0.295)	(0.292)	(0.325)	(0.324)	(0.348)	(0.285)
New state	1.658***	1.335***	1.586***	1.921***	1.629***	1.809***	1.633***
	(0.342)	(0.395)	(0.351)	(0.422)	(0.404)	(0.365)	(0.343)
Instability	0.513*	0.561*	0.270	0.581*	0.617*	0.562*	0.530*
	(0.242)	(0.248)	(0.277)	(0.290)	(0.291)	(0.262)	(0.243)
Ethnic fract.	0.164	0.0376	0.251	0.375	0.196	0.0671	0.114
	(0.368)	(0.374)	(0.398)	(0.506)	(0.471)	(0.409)	(0.372)
Religious fract.	0.326	0.426	0.330	0.251	0.680	0.0590	0.200
	(0.506)	(0.520)	(0.542)	(0.655)	(0.595)	(0.586)	(0.520)
Anocracy (lagged)	0.521*	0.317	0.797**	0.436	0.470	0.619*	0.484*
	(0.237)	(0.249)	(0.257)	(0.299)	(0.275)	(0.261)	(0.238)
Democracy (lagged)	0.127	0.137	0.321	0.357	-0.358	0.111	0.111
	(0.304)	(0.309)	(0.344)	(0.345)	(0.406)	(0.333)	(0.308)
Constant	-7.019***	-6.633***	-7.391***	-7.223***	-6.426***	-7.056***	-6.930***
	(0.751)	(0.764)	(0.837)	(0.951)	(0.968)	(0.807)	(0.764)
Observations	6,327	5,192	5,734	5,152	4,777	5,298	5,482

*** p<0.001, ** p<0.01, * p<0.05

In table 2, we report results from further robustness tests. Section 5.3 had suggested an over-sampling test for dealing with population boundary uncertainty. Some countries – more than half of those in the sample – never experience a civil war onset over the entire sample period. Another quarter experience one onset, 13 per cent of countries experience two, almost 4 per cent experience three and even fewer experience four or six onsets (none has five). One might wonder whether the process that generates cases (countries) that never experience conflict is fundamentally different from the process that generates cases (countries) that sometimes experience conflict and whether the theory underlying F&L's hypotheses only apply to those cases at risk of conflict. If so, then one might say that, given population boundary uncertainty, F&L 'over-sampled' by including also observations from p2 (those never experiencing civil war) into the relevant population of p1 (those at risk of civil war). Model 8 tests whether inferences are robust to basing estimates only on those countries that ever experience a civil war over the sample period. While the effect of per capita income is significantly smaller (less than half compared to the baseline model), it remains highly statistically significant. Ethnic and religious fractionalization remain insignificant, while the result for anocracy is fully robust and almost identical to the baseline model.

In section 5.4, we suggested cross-validation, i.e. the replication of estimation on another randomly drawn sample from the population, as robustness test. Since F&L employ a near exhaustive sample over the estimation period, their sample is not a random draw and we cannot replicate estimations on another random draw. However, a variant of cross-validation works as follows: We randomly split F&L's sample into two sub-samples of nearly identical sizes. If F&L's inferences are not driven by the particular set of observations they study, then their results should be robust if estimates are based on these randomly drawn sub-samples instead. This cross-validation variant faces the slight complication that with smaller sample size standard errors will increase due to smaller efficiency of estimations. To avoid this, we

simply sample each observation of the split sub-samples twice into the estimation, which results in identical coefficients, but a sample size and thus an expected efficiency of estimation and size of standard errors very similar to the baseline model. Models 9 and 10 present results from these cross-validation exercises. The effect of per capita income is significantly smaller in model 9 and significantly larger in model 10 compared to the baseline model, but the effect remains negative and highly significant. Fractionalization continues not to matter, while the effect of anocracy is practically identical in both models compared to the baseline model.

Table 2. Over-sampling and cross-validation robustness tests.

	baseline	exclude countries never experiencing civil war	randomly drawn split sample 1	randomly drawn split sample 2
	1	8	9	10
Prior war	-0.916**	-1.382***	-0.419	-1.452***
	(0.312)	(0.303)	(0.299)	(0.337)
Per capita income (lagged)	-0.318***	-0.150*	-0.242***	-0.390***
	(0.0714)	(0.0758)	(0.0677)	(0.0758)
log(population) (lagged)	0.272***	0.212**	0.198**	0.346***
	(0.0737)	(0.0745)	(0.0759)	(0.0734)
log(% mountainous)	0.199*	0.0104	0.356***	0.0807
	(0.0846)	(0.0910)	(0.0939)	(0.0797)
Noncontiguous state	0.426	0.185	0.342	0.409
	(0.272)	(0.279)	(0.285)	(0.266)
Oil exporter	0.751**	0.611*	0.838**	0.604*
	(0.278)	(0.294)	(0.310)	(0.262)
New state	1.658***	1.679***	2.546***	0.482
	(0.342)	(0.356)	(0.305)	(0.458)
Instability	0.513*	0.446	0.340	0.674**
	(0.242)	(0.242)	(0.271)	(0.224)
Ethnic fract.	0.164	-0.111	-0.378	0.685
	(0.368)	(0.369)	(0.387)	(0.363)
Religious fract.	0.326	0.417	0.862	-0.201
	(0.506)	(0.505)	(0.527)	(0.501)
Anocracy (lagged)	0.521*	0.472*	0.517*	0.548*
	(0.237)	(0.241)	(0.254)	(0.227)
Democracy (lagged)	0.127	0.159	0.120	0.188
	(0.304)	(0.303)	(0.328)	(0.289)
Constant	-7.019***	-5.325***	-7.099***	-7.134***
	(0.751)	(0.731)	(0.800)	(0.728)
Observations	6,327	2,775	6,444	6,210

*** p<0.001, ** p<0.01, * p<0.05

In sum, we find the main results from Fearon and Laitin's (2003a) analysis of civil war onset to be remarkably robust toward population and sample uncertainty – or at least to the specific set of tests we have undertaken.[11] Their inferences uphold when observations from specific regions are sliced off from the sample in regional jackknife estimates, when observations from countries never experiencing any internal conflict are excluded and when two randomly drawn sub-samples are employed for cross-validation.[12]

7. Conclusion

In the ideal econometric world researchers draw a random sample from a given population of cases, which justifies the nice optimality properties of estimators and allows valid generalizations from the sample results to the population. Alas, the real world looks very different from econometric wishful-thinking. Deductively derived, researchers are uncertain about the set of cases (the population) for which their theory claims validity, which renders drawing a random sample impossible. Inductively derived, researchers are uncertain about the set of cases (the population) to which a result based on a given sample of cases can be generalized. Such uncertainties pose clear, yet ill understood, threats to reliable causal inferences.

---

[11] There are other, much more work-intensive tests one could undertake, such as for example extending the sample forward and backward in time.

[12] As we have shown elsewhere (Plümper and Neumayer 2010), the result on the effect of anocracy is not robust toward recoding the polity2 variable where we replace Polity IV's coding of interregnum and affected transition periods that is arbitrary and entirely void of face validity with imputed values derived from interpolation, theoretically derived out-of-sample predictions and multiple imputation techniques.

We have argued that under-specified theories, the temporal confines of any sample analyzed, unobservable population boundaries, sampling problems and model complexity are the principal sources of population and sample uncertainty. We have suggested outlier, jackknife and bootstrap tests for tackling potential sample heterogeneity and the existence of deviant cases, several imputation techniques for recovering excluded observations due to missing data, 'over-sampling' and sample selection correction tests for dealing with population boundary uncertainty, re-sampling and cross-validation tests for checking whether any studied sample truly represents a random draw from the population and several re-sampling tests for tackling under-representation of certain cases in the study sample. Researchers cannot eliminate population and sample uncertainty, but they can test whether their inferences are robust toward plausible changes in the sample and the definition of the relevant population. Robustness tests are the answer to modeling uncertainty.

References

Baltagi, Badi H. 2008. *Econometrics*. 4th edition. Berlin: Springer.

Belsley, D. A., E. Kuh, and R. E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Birnbaum, M. H. 2004. Human research and data collection via the Internet. *Annual Review of Psychology* **55**: 803-832.

Boehmke, Frederick J. 2003. Using Auxiliary Data to Estimate Selection Bias Models, with an Application to Interest Group Use of the Direct Initiative Process. *Political Analysis* 11: 234-254.

Boehmke, Frederick J., Daniel S. Morey and Megan Shannon. 2006. Selection Bias and Continuous-Time Duration Models: Consequences and a Proposed Solution. *American Journal of Political Science* 50 (1): 192-207.

Breen, R. (1996) *Regression Models: Censored, Sample-Selected, or Truncated Data*, London: SAGE.

Caliendo, M. and R. Hujer. 2006. The Microeconometric Estimation of Treatment Effects – an Overview. *Allgemeines Statistisches Archiv* 90: 199-215.

Copas, J.B. and H.G. Li, 1997. Inference for Non-Random Samples. *Journal of the Royal Statistical Society B* 59 (1): 55-95.

Couper, M. P. 2000. Web surveys - A review of issues and approaches. *Public Opinion Quarterly* 64: 464-494.

Davidson, Russell and James G. MacKinnon. 2004. *Econometric theory and methods*. New York: Oxford University Press.

Dougherty, Christopher. 2010. *Introduction to Econometrics*. 4th edition. Oxford: Oxford University Press.

Fearon, James D., and David D. Laitin. 2003a. Ethnicity, insurgency and civil war. *American Political Science Review* 97:75–90.

Fearon, James D., and David D. Laitin. 2003b. Additional Tables for "Ethnicity, insurgency and civil war". Working Paper. Stanford University: Department of Political Science.

Greene, William H. 2012. *Econometric Analysis*. 7th edition. Harlow : Pearson Education.

Hainmueller, Jens and Dominik Hangartner. 2013. Who Gets a Swiss Passport? A Natural Experiment in Immigrant Discrimination. *American Political Science Review* (forthcoming).

Heckman, James 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47: 153-161.

Heckman, James 1990. Varieties of Selection Bias. *American Economic Review* 80: 313-318.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis* 15: 199-236.

Honaker, James, and Gary King. 2010. What to do about missing values in time series cross-section data. *American Journal of Political Science* 54 (3): 561-581.

Hsiao, Cheng. 2003. *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Hug, Simon. 2003. Selection Bias in Comparative Research: The Case of Incomplete Datasets. *Political Analysis* 11: 255-274.

Keane, Michael P. 2010. Structural vs. A-theoretical Approaches to Econometrics, *Journal of Econometrics* 156: 1, 3-20.

Keefer, Phil, Eric Neumayer and Thomas Plümper. 2011. Earthquake Propensity and the Politics of Mortality Prevention, *World Development*, 39 (9), 1530-1541.

Kennedy, Peter. 2008. *A Guide to Econometrics*. 6th edition. Malden: Blackwell.

King, Gary, and Langche Zeng. 2001a. Logistic Regression in Rare Events Data. *Political Analysis* 9: 137-163.

King, Gary, and Langche Zeng. 2001b. Explaining Rare Events in International Relations. *International Organization* 55: 693-715.

King, Gary, James Honaker, Anne Joseph, and Kenneth Scheve. 2001. Analyzing incomplete political science data. An alternative algorithm for multiple imputation. *American Political Science Review* 95:49–69.

Leung, S.F. and Yu, S. 1996. On the Choice Between Sample Selection and Two-Part Models, *Journal of Econometrics*, 72: 197–229.

Lohr Sharon L. 2010. *Sampling: Design and Analysis*. Belmont: Thomson Brooks/Cole.

Maddala, G.S. and Kajal Lahiri. 2009. *Introduction to Econometrics*. 4th edition. Chichester: Wiley.

Plümper, T. and E. Neumayer. 2009. Famine Mortality, Rational Political Inactivity, and International Food Aid. *World Development*, 37 (1), 50-61.

Plümper, Thomas and Eric Neumayer. 2010. The Level of Democracy during Interregnum Periods: Recoding the polity2 Score. *Political Analysis*, 18 (2): 206-226.

Puhani, P. 2000. The Heckman correction for sample selection and its critique, *Journal of Economic Surveys*, 14: 53-68.

Rubin, Donald. 1976. Inference and Missing Data. *Biometrika* 63 (3): 581-592.

Sartori, Anne E. 2003. An Estimator for Some Binary-Outcome Selection Models Without Exclusion Restrictions. *Political Analysis* 11: 111-138.

Schafer, Joseph L. and John W. Graham. 2002. Missing Data: Our View of the State of the Art. *Psychological Methods* 7 (2): 147-177.

Scheaffer, Richard L., William Mendenhall III and R. Lyman Ott. 2006. *Elementary Survey Sampling*. Belmont: Thomson Brooks/Cole.

Shadish, W., Cook, T., Campbell, D., 2001. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Sigelman, Lee and Langche Zeng. 1999. Analyzing Censored and Sample-Selected Data with Tobit and Heckit Models. *Political Analysis* 8 (2): 167-182.

Smith, T.M.F. 1983. On the Validity of Inferences from Non-Random Samples. *Journal of the Royal Statistical Society A* 146(4): 394-403.

Valliant, Richard, Alan H. Dorfman and Richard R. Royall. 2000. *Finite Population Sampling and Inference: A Prediction Approach*, Wiley-Blackwell London.

Vella, F. 1998. Estimating models with sample selection bias: A survey. *Journal of Human Resources* 33: 127-169.

Whitehead, J. C., P. A. Groothuis, et al. 1993. Testing for Nonresponse and Sample Selection Bias in Contingent Valuation – Analysis of a Combination Phone Mail Survey. *Economics Letters* 41: 215-220.

Wooldridge, Jeffrey M. 2000. *Introductory Econometrics: A Modern Approach*. Cincinnati: South-Western College.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. Second Edition. Cambridge, Mass.: MIT Press.